A FAIR APPROACH TO GENOMICS

Jasper Jan Koehorst

Thesis committee

Promotor

Prof. Dr Vitor A. P. Martins dos Santos

Professor of Systems and Synthetic Biology

Wageningen University & Research

Co-promotors

Dr Peter J. Schaap

Associate professor, Laboratory of Systems and Synthetic Biology

Wageningen University & Research

Dr Edoardo Saccenti

Post-Doc, Laboratory of Systems and Synthetic Biology

Wageningen University & Research

Other members

Prof. Dr Carole Goble, University of Manchester, UK

Prof. Dr Eugene V. Koonin, National Center for Biotechnology Information, Bethesda, USA

Dr Nikos Kyrpides, DOE Joint Genome Institute, Walnut Creek, USA

Prof. Dr Dick de Ridder, Wageningen University & Research

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

A FAIR APPROACH TO GENOMICS

Jasper Jan Koehorst

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 25 January 2019
at 4 p.m. in the Aula.

Jasper Jan Koehorst
A FAIR APPROACH TO GENOMICS,
247 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2019)
With references, with summary in English

ISBN: 978-94-6343-369-3
DOI: 10.18174/463364

Contents

1 General introduction and thesis outline

Bacteria represent the most abundant and species-rich group of organisms, demonstrating an enormous range of genetic variation, amidst the domains of life (Whitman, Coleman, and Wiebe, 1998). Due to the development of very affordable high throughput sequencing technologies, sequencing and analysis of this vast repertoire are in a period of exponential growth. Collecting bacterial sequences, however, is only the essential first step in gaining a systems level understanding of societal relevant microorganisms and microbial ecosystems. The next critical step is to give biochemical, physiological, and ecological meaning to the genome information obtained and to transform newly obtained actionable knowledge into applications of biotechnological, medical and environmental interest. The objective of this thesis is to increase our understanding on how microbial genome information leads to function. To achieve this, I will use a FAIR by design Semantic Systems Biology approach to genomics to study *i*) the relationship between the microbial genome and its functionome and *ii*) the impact of species diversity on the functional landscape.

In the following paragraphs I will discuss a number of essential elements of my research; the bacterial species concept, FAIR data and the requirement for data interoperability, and the role of Semantic Web technology in the development of the top-down workflows used to reach the goals.

The bacterial species concept

Comparative genomics is a branch of genomics in which the genomic features of different species are compared to study basic biological similarities and differences as well as evolutionary relationships. Species of animals and plants have come to be understood as cohesive groups because there are evolutionary mechanisms to constrain diversity within a species. In sexual species, such as most animals and plants, the force constraining diversity within species is understood to be genetic exchange, yielding offspring that remains reproducible as well. This evolutionary force ensures a flow of genetic information that remains within a species. Since bacterial species do not have a sexual reproduction cycle one would assume that also within bacteria the genetic pool remains fairly consistent within a microbial line. However, evolution in microbes happens at a much faster rate and through various means. One is a much faster growth rate which allows for a fast accumulation of mutations in their

genome, thereby acquiring or losing particular functionalities. The acquisition of new functions from the environment through Horizontal Gene Transfer (HGT) can, in a single step, also dramatically change the cell's repertoire of metabolic capabilities (Dutta and Pan, 2002). It has also shown that gene loss is a source of genetic variation that can cause adaptation to phenotypic diversity (Paul, Sokurenko, and Chattopadhyay, 2016; Albalat and Cañestro, 2016; Ochman and Moran, 2001). Additionally, in bacterial species gene fusion/fission and domain duplications are frequent events that lead to multi-domain proteins of different composition with new or altered functionalities (Doolittle, 1995).

While bacterial species are believed to exist, the nature of the cohesive evolutionary forces required to constrain the genetic diversity within a bacterial species is unclear. Since the current methodologies to define species groups such as 16S-rRNA sequence similarity (Kim et al., 2014) or Average Nucleotide Identity (ANI) (Konstantinidis and Tiedje, 2005) use a practical metric for species definition, these methods can lead to artificial species assignments because they are methodically unlinked to the biological mechanisms that lead to cohesion. Currently used methodologies in comparative functional genomics usually follow a sequence based (bottom-up) approach and assume that protein encoding genes within a clade have a shared common ancestor. Since bacterial genes have the potential of being horizontally acquired this mechanism can have a great impact on the overall genetic landscape and on the correct identification of members of a given species. Incorrect inclusion of new strains in a clade can have a large influence on the core genome, defined as the genetic core shared among the members of a given clade under study, while the pan-genome defined as the overall genetic diversity observed in a clade, can be overestimated.

FAIR data

Performing meta-analysis on large species groups and linking a broader range of genotypic diversity by computational means, requires that these results are Interoperable and Reusable for further analysis. FAIR data is a world-wide initiative, initiated by a group of academia, industry, funding agencies, and scholarly publishers, to make data Findable, Accessible, Interoperable and Reusable (Wilkinson et

al., 2016). These FAIR principles are in place with extra emphasis on enabling and enhancing the reusability of data by machines in addition to individuals. To make data interoperable by machines, Semantic Web technologies can be applied. Already several platforms have been developed that align with the FAIR principles. FAIR-DOM is a combination of SEEK (Wolstencroft et al., 2015), a web-based resource for sharing heterogeneous scientific research datasets, models or simulations, processes and research outcomes, and openBIS, a system for managing biological data (Barillari et al., 2016). Both systems have implemented a semantic representation of the data using the Resource Description Framework (RDF) data model. Other resources, such as UniProt, have converted the entire database into this representation allowing individuals as well as computers to mine these resources efficiently (UniProt Consortium, 2017). In addition various efforts have been performed on the fairification of existing datasets making them more FAIR (*Find FAIR Data tools* n.d.). FAIRification means that the resource data and the metadata are made machine-readable, in which the metadata clearly describes how the data can be accessed and reproduced, and that the metadata can be found by machines. The data FAIRifcation process includes:

1. Original data retrieval
2. Dataset identification and analysis
3. Definition of the semantic model
4. Data transformation
5. License assignment
6. Metadata definition
7. FAIR Data resource deployment (data, metadata, license)

When working according to FAIR by design principles, datasets are generated directly according to FAIR principles in which data and metadata are in a linked format. This can be achieved through the development and usage of computational applications and workflows in which data can be consistently appended with metadata.

Interoperability, formats and minimal information models

As Findability and Accessibility mostly concern organisational issues and data reusability grossly depends on the format, the key component in determining the FAIRness of a data set is the level of interoperability. Interoperability and reusability of data requires a larger effort at the data level. Translating computational predictions in interoperable data sets, requires sophisticated minimal information standards. These standards, often defined as ontologies describe the definitions of the properties and relations between concepts or data which can correspond to a broad range of domains.

As comparative genomics strongly depends on computer-based analysis and uses multiple heterogenous data sources to extract genomic features, data interoperability is essential. For bottom up approaches that work from sequence to function, a high level of interoperability already existed early on. This is due to standardisation of the FASTA sequence file format which originated from the development of FASTP, one of the first protein similarity search programs (Lipman and Pearson, 1985), Dictated by being the first, the FASTA format is accepted by most if not all sequence analysis tools. The format is also well understood by users and provides a simple representation of any given sequence. The format starts with a ">" symbol followed by an identifier and optionally a description. On the next line(s), the corresponding DNA or amino acid sequence is found using either a standard representation of DNA bases by single characters that specify either a single base or a set of bases as defined by the International Union of Pure and Applied Chemistry (IUPAC) (Comm, 1970) or the single characters amino acid code in case of a protein sequence (JCBN, 1983). A more line by line format is GFF, which is a tabular format consisting of 9 fields making it easier to generate, parse and index at the cost of only storing basic genetic information (*GFF and GVF specification documents* n.d.). Other formats such as the GenBank or EMBL format are capable to capture more complex data structures of which some are related to project meta-data such as project identifiers, articles with corresponding authors, releases/last update dates and sample related information such as taxonomy, location of isolation, strain and organism identifiers (*The DDBJ/EMBL/GenBank Feature Table: Definition* 2014).

```
1     gene            1430..1552
2                     /locus_tag="MS6671_01090"
3     CDS             1430..1552
4                     /locus_tag="MS6671_01090"
5                     /inference="ab initio prediction:Prodigal:2.60"
6                     /codon_start=1
7                     /transl_table=11
8                     /product="hypothetical protein"
```

Figure 1.1: Gene entry in Genbank format

Nevertheless, it remains extremely cumbersome to parse and integrate computationally derived genetic data from the current formats. Electronically inferred predictions such as the location of a gene on a sequence, mostly lack provenance related information. Therefore, when parsing this data it remains unclear what algorithm and which version was used to predict, for instance, the genes on a given genome. Although, hardly used, it is possible to provide such information using the inference tag (Figure 1).

In the example shown in Figure 1, a gene and its corresponding CDS have been predicted using Prodigal 2.60. Adding this information greatly improves the reproducibility for genome annotation although, this essential information is absent in nearly all annotated genomes. Additional information provided by a gene prediction program, for each given gene the element wise provenance, such as the putative ribosomal binding site, the GC content and more importantly, a confidence score are also absent as no standard is available for storing this type of information. As a consequence, with the current standards in genome annotation, the resulting data is far from being interoperable and therefore reusable.

Bottom up methods in comparative genomics

As previously discussed, the FASTA format displays a high level of Interoperability and if it were not for Darwinian evolution, a direct comparison between species scoring the presence or absence of gene sequences would be very well possible. Due to the occurrence of accepted mutations, insertions and deletions, between species the sequences of genetic elements will differ but still can show an acceptable level

of similarity. For protein encoding genes, mutations can lead to missense mutations resulting in an amino acid change but not necessarily a change of function. Important concepts here are homology, orthology and paralogy meaning that orthologs present in different species are derived from some common ancestor and therefore probably have the same or very similar function in these species (Wolf and Koonin, 2012). A number of methods have been developed to automatically cluster orthologous sequences e.g InParanoid (Sonnhammer and Östlund, 2014), PanOct (Fouts et al., 2012), OrthoMCL (Li, Stoeckert, and Roos, 2003) or ProteinOrtho (Lechner et al., 2011). The method that is applied in these orthology detection and clustering tools is an exhaustive all-vs-all BLAST (Basic Local Alignment Search Tool) comparison in which all protein sequences between two or more species are compared. When multiple genomes are studied, an all-vs-all pair-wise comparison is performed to identify cliques of proteins that are highly similar at sequence level (Figure 1.2). Due to protein architectures dynamics, in which protein domains are exchanged, shuffled, duplicated or deleted, a phenomenon called domain chaining can create articulation points resulting in a merger of different cliques. Moreover, each protein family has an intrinsic molecular clock and protein sequences change faster than the corresponding protein structures. Proteins with the same function might therefore end up in different cliques due to a low sequence similarity. Articulation points can of course be reduced using a more stringent sequence similarity threshold. However, this will also lead to a lower sensitivity.

In general, pair-wise sequence comparison methods have a number of shortcomings; the computational cost scales quadratically (Sonnhammer and Östlund, 2014) and comparisons are often limited to hundreds of genomes while currently we have more than 100.000 bacterial genomes in the public repositories. For certain species, especially the pathogens, such as *Pseudomonas aeruginosa* and *Streptococcus pneumoniae* we already have more than thousands of genome sequences at ENA (Leinonen et al., 2010) and PATRIC (Wattam et al., 2016), illustrating that a within species comparison of these organisms using a bottom-up approach to identify cliques of functionally similar proteins does not compute anymore, even when more advanced parallel set ups are used.

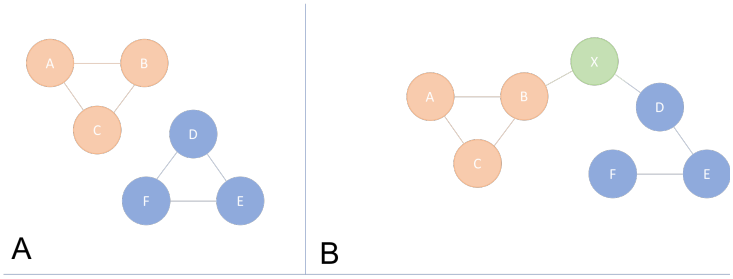

Figure 1.2: ***Orthology detection using an all-vs-all bi-directional BLAST method.*** *A) For each genetic element cliques are formed representing individual cluster of orthologous genes. B) Due to recombination events and domain chaining, subnetworks are formed in which individual genetic elements can function as articulation points merging different gene clusters. Separation of the subnetwork in individual clusters is highly dependent on the predefined threshold.*

Top down functional comparative genomics

It could be argued that since functional annotations in principle can be directly compared and analysed, thereby omitting the needs of comparing the underlying sequences, a top-down approach, working from function to sequence, would be more straight-forward and better scalable. However, the data and element wise provenance of a functional assignment to a protein sequence is rarely available and published functional assignments are currently not interoperable as different vocabularies are used.

To obtain a consistent functional annotation, different methodologies have been developed. Most commonly used are InterProScan (Jones et al., 2014), which is a collection of annotation modules and Eggnog-mapper (Huerta-Cepas, Forslund, et al., 2017), which uses precomputed clusters and phylogenies from the eggNOG database (Huerta-Cepas, Szklarczyk, et al., 2016). Using protein sequences as an input both methods allow to functionally annotate large sets of sequences according to highly curated reference sets. However, to warrant a high level of interoperability a direct link of the functional annotations obtained with the provenance is essential. For instance, the transition of PFAM 30.0 to 31.0 resulted in 415 new families and the removal of 9 families. If a set of genomes annotated with version 30.0 or 31.0 were compared this could result in a technical difference of 424 PFAM domains. Storage of a functional prediction should therefore be performed in a highly

standardised fashion that includes all element-wise and dataset-wise provenance enabling to completely retrace the annotation approach taken.

In summary, current large scale functional comparisons based on bottom-up sequence similarity approaches is challenged by methodological problems, such as the need of defining arbitrarily generalised minimal alignment length and similarity cut-off for all sequences analysed, and is hampered by the high computational cost, as it scales quadratically with the number of genome sequences to be compared. As the aim of this thesis is to study *i*) the relationship between the microbial genome and its functionome and *ii*) the impact of species diversity on the functional landscape, the main focus of this thesis lies mainly not in the evolutionary origin of a protein sequence but in similarity and diversity of bacterial functional landscapes, which is best studied in a Top-Down approach working from function to sequence. To mitigate problems associated with such an approach, Semantic Web and Linked Data technologies are used. In essence, Semantic technologies allow for a direct linkage of appropriate historical, data-wise, and element-wise provenance to a computational prediction.

Semantic Web

Semantic Web, an extension for the World Wide Web or the internet, is part of the World Wide Web Consortium and is in place to facilitate a common framework for data exchange across boundaries (Berners-Lee, Hendler, and Lassila, 2001). The two most common standards of the Semantic Web are the data format standard, which is the Resource Description Framework (RDF), and the corresponding query language called SPARQL (SPARQL Protocol and RDF Query Language).

The data infrastructure of the Semantic Web provides an efficient and flexible environment allowing to easily absorb heterogeneous data and enables to build database structures on the fly in a distributed and decentralised environment. This in contrast to currently more commonly used database systems such as MySQL, which is a relational database management system allowing to store data according to a fixed schema and is intended to be used as a central point of entry.

As shown in Figure 1.3 in the RDF data model, data is represented as triples, a subject, predicate and an object. Triples can be linked using the object node as a

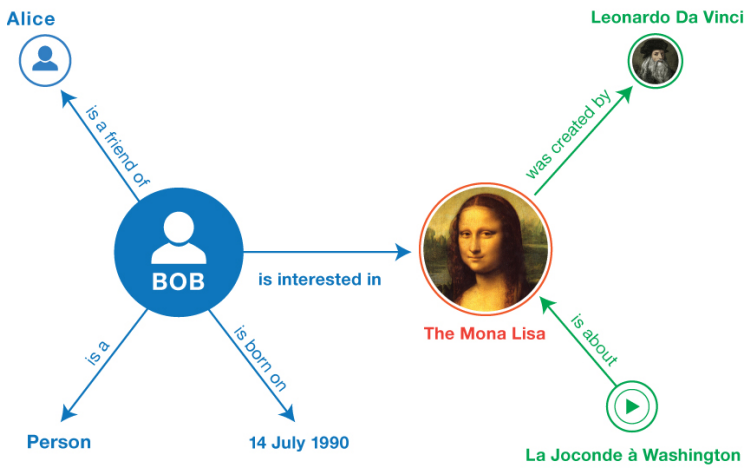

Figure 1.3: A visual representation of an RDF graph (Source Consortium et al., 2014)

subject node for another triple statement. Through this approach a graph database is created, and all relations are defined with predicates. By using a graph structure, any defined relation between essentially different "things" can be established. For example, although they are completely different "things" there is a direct link between the painter Leonardo Da Vinci and the "Mona Lisa". Using the RDF data model, all kinds of heterogenous datasets can be converted in a graph structure.

One of the largest biological datasets available in RDF from a single SPARQL endpoint is the UniProt database. This database consists of 41,016,681,408 triples (as of Augsust 2018) and 17 sub databases or so-called named graphs. It contains a wide set of information with regards to proteins and their functional annotation, pathway related information, article citations and taxonomic information. As it is a public endpoint, private and/or public databases, that use UniProt identifiers, can remotely connect and remain up to date when new information becomes available.

Data Provenance

When obtaining data from a public endpoint, it is important that the data is described including the provenance. For example, when a statement is made that Leonardo Da Vinci painted the Mona Lisa it is important to also mention where this information originated from and how this information was obtained. Prove-

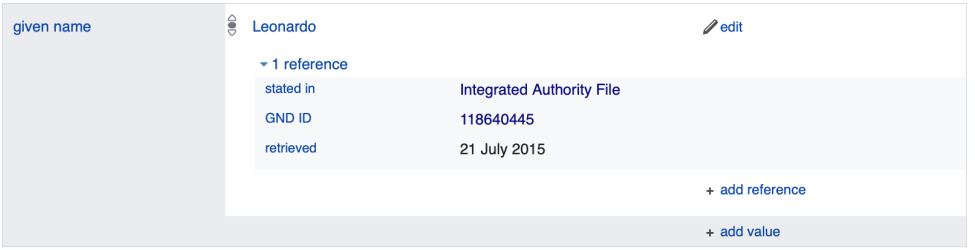

Figure 1.4: ***A WikiData entry****. Contains the given name of Leonardo da Vinci with asosciated provenance which shows that his name was stated in an Integrated Authority File and that this information was retrieved on the 21th of July in 2015.*

nance is additional information on top of the link that is established and is as important as the actual information. Provenance can be subdivided into dataset-wise and element-wise provenance. In the realm of comparative genomics, data-set wise provenance is additional information defining the programs used, versions thereof and selected parameters for the complete annotation of the (set of) sequences under study. The element-wise provenance is the statistical evidence of a computational prediction, such as the confidence score of a given gene prediction or the E-value of a blast similarity score.

When this important information is lacking, the reliability of the statement could be questioned. However, in practice this information is often still lacking in many public repositories. Currently WikiData, a free and open knowledge base is the only and largest resource that adds data and element-wise provenance as a rule (See Figure 1.4).

Ontologies

Due to the dynamic structure of RDF, everybody can create their own graph/ontology which could make it difficult to understand the complexity of such a database, resulting in a low level of Interoperability. As many topics are shared among different databases, using the same ontology would allow for distributed and decentralised environment. Ontologies such as FOAF (Friend Of A Friend), an ontology to describe persons and their social network, Prov-O for the storage of provenance and BIBO, the Bibliographic Ontology for the storage of documents and au-

thor related information are widely used for a large variety of databases (Brickley and Miller, 2007; Lebo et al., 2013; Giasson et al., 2008). When a public endpoint contains data aligning with an ontology, it is common that such an ontology is described in detail and made available for others to query its data. Websites such as `https://bioportal.bioontology.org` allow ontology developers to store their ontology with regards to biological information, which in turn simplifies exchange of already existing ontologies or finding new ones that suit your needs when transforming data into RDF.

To make sure that your dataset is according to the ontology that has been described, we developed RDF2Graph (Dam et al., 2015), an application to recover, understand and most importantly to validate or generate the ontology of an RDF resource.

In an effort to standardise the integration of biological information, several bio-ontologies already have been developed. The Gene Ontology, describes concepts/-classes used to describe gene function, and relationships between these concepts. (The Gene Ontology Consortium, 2015). FALDO is an ontology that can be used to describe the location of nucleotide and protein features in genome annotations (Bolleman et al., 2014). Similarly the Sequence Ontology (SO) (Eilbeck et al., 2005) was designed to categorise sequence features used in biological sequence annotation and BioPax, an ontology to facilitate integration, exchange, visualisation and analysis of biological pathway data (Community, 2010). These ontologies are all based on the semantic framework, allowing them to be easily integrated into a single database and to communicate with external resources containing additional information.

Unlocking genome information using a FAIR by design Semantic Systems Biology approach

To unlock the full potential of genomic information, for comparative genomes or for function-based mining, an ontology that describes and combines already existing ontologies into an entity allowing to transform already existing annotation and complement with new annotation methods is essential. The extendable Genome Biology Ontology Language (GBOL) and corresponding Stack of supporting tools has been

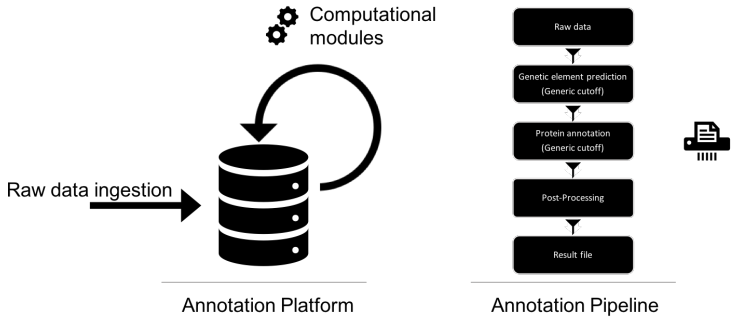

Figure 1.5: ***Comparison between a classical annotation pipeline and a semantic annotation platform.*** *In the semantic annotation platform (left) the tool meets the data and all computational results including data and element-wise provenance are stored in a linked format in a persistent data store. In the annotation pipeline (right) in each step, an (arbitrary) threshold is used to filter the data. Data provenance is usually not stored as this requires a more complex data management system.*

developed to cope with genomic data and is described in **chapter 3**. GBOL allows for a complete description of genetic information, localisation as well as functional annotation from a variety of applications linked to the full chain of provenance. As it remains difficult during the development of parsers or conversion code to adhere to the predefined ontology, the GBOL Stack supports the code developer through an Application Programming Interface (API) thereby safeguarding a match between the converted data and the ontology.

In the course of this thesis, SAPP was developed as an extension to the ontology and corresponding API's. SAPP is a tool independent platform and stands for a Semantic Annotation Platform with Provenance (Figure 1.5). The software integrated with the platform is extendable and currently consists of various modules for the conversion of existing annotation files in FASTA, GFF, GenBank and EMBL format in RDF. Additionally, SAPP can perform a FAIR by design *de novo* structural and functional prediction of genetic elements using a vast array of (complementary) prediction tools. SAPP will be discussed in **chapter 4**.

Development of dynamic top-down Semantic Systems Biology workflows

Through the development of SAPP and the incorporation of GBOL and corresponding Stack, biological data and derived computational predictions can be accessed

through a single SPARQL endpoint. By using this endpoint, it possible to compare various gene prediction approaches using a single query and to analyse and compare already existing functional annotation and descriptions with *de novo* predictions. For functional comparisons protein architectures are used, using (Pfam) protein domains as building blocks. By treating each protein domain as an independent functional unit, domain order can be used as a proxy for function. The full set of domain architectures thus represents the functional landscape of a given strain.

Thesis outline

To enable Semantic Web technologies and to ensure consistency within and between the described ontologies, Empusa was developed. The exact implementation is described in **chapter 2** and was used in the development of various ontologies. To transform biological data into a semantic structure, to enable a higher level of interoperability and reusability of these datasets, the here developed GBOL ontology is described in **chapter 3**. This chapter describes in detail the reasoning behind the ontology and the benefit of the corresponding Stack. **Chapter 4** describes the platform that is essential for the conversion and annotation of biological data formats through an RDF infrastructure. It enables biological data to become more interoperable and reusable through the incorporation of provenance from a variety of modules for the prediction and annotation of sequences. This platform has been widely used in a number of papers and throughout this thesis in **chapter 5-8**. In **chapter 5**, protein domain architectures were identified to be a fast, efficient and scalable alternative to sequence-based methods which in turn can have large applications in the field of comparative functional (meta-)genomics. By identifying the usability of protein domains for comparative genomics we applied the principles to *Pseudomonas* in **chapter 6**, where genome data was integrated with its functional landscape, combined with metabolic and expression data. Through this integration, essential genes were detected to be highly persistent according to the protein domain profile, showing that non-essential genes tend to more frequently exchange functional information. Since protein domains have proven to be useful for the characterisation of the functional landscape of an organism it was investigated to what extend 16S distances differ from the functional diversity. In **chapter 7**, a large study was performed on 5713 high quality genome assemblies obtained from public resources and it was shown that the functional landscape can provide a more detailed view on closely related species in contrast to a single gene such as 16S. The increased granularity in this approach allows to perform large scale comparisons which in turn paves the way to new applications to identify phenotypic properties from its functional landscape. In **chapter 8**, we translated this approach into a Knowledge Discovery in Databases workflow and applied it to identify strains capable of producing 1,3-Propanediol

using glycerol as a precursor. The initial analysis was performed on 84,329 bacterial genomes harbouring 2,661 species in total and resulted in the identification of 178 new candidate species that are able to degrade glycerol and produce 1,3-propanediol. **Chapter 9**, presents a new undergraduate course that is aimed to close the gap between computational and experimental studies. In this course students use moist-lab techniques to identify phenotypic properties of an unknown organism and learn to unravel the genotypic properties corresponding to these phenotypes. **Chapter 10** describes how Semantic Systems Biology has shaped the research field and how this is applied to functional and comparative genomics and how this can be incorporated in future research.

Bibliography

Albalat, Ricard and Cristian Cañestro (2016). "Evolution by gene loss". In: *Nature Reviews Genetics* 17.7, p. 379.

Barillari, Caterina et al. (2016). "openBIS ELN-LIMS: an open-source database for academic laboratories". In: *Bioinformatics (Oxford, England)* 32.4. DOI: 10.1093/bioinformatics/btv606.

Berners-Lee, Tim, James Hendler, and Ora Lassila (2001). "The Semantic Web". In: *Scientific American* 284.

Bolleman, J. et al. (2014). "FALDO: A semantic standard for describing the location of nucleotide and protein feature annotation." In: *Journal of Biomedical Semantics*. DOI: DOI10.1186/s13326-016-0067-z.

Brickley, Dan and Libby Miller (2007). *FOAF vocabulary specification 0.91*.

Comm, IUPAC-IUB (1970). "Abbreviations and symbols for nucleic acids, polynucleotides, and their constituents". In: *Biochemistry* 9.20, pp. 4022–4027.

Community, BioPAX (2010). "The BioPAX community standard for pathway data sharing". In: *Nature biotechnology*. DOI: 10.1038/nbt.1666.

Consortium, World Wide Web et al. (2014). "RDF 1.1 Primer". In:

Dam, Jesse CJ van, Jasper J Koehorst, Peter J Schaap, Vitor Ap Martins Dos Santos, and Maria Suarez-Diez (2015). "RDF2Graph a tool to recover, understand and validate the ontology of an RDF resource". In: *Journal of biomedical semantics* 6.

Doolittle, Russell F (1995). "The multiplicity of domains in proteins". In: *Annual review of biochemistry* 64.1, pp. 287–314.

Dutta, Chitra and Archana Pan (2002). "Horizontal gene transfer and bacterial diversity". In: *Journal of biosciences* 27.

Eilbeck, Karen et al. (2005). "The Sequence Ontology: a tool for the unification of genome annotations." In: *Genome biology* 6. DOI: 10.1186/gb-2005-6-5-r44.

Find FAIR Data tools (n.d.). https://www.dtls.nl/fair-data/find-fair-data-tools/.

Fouts, Derrick E., Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton (2012). "PanOCT: automated clustering of orthologs using conserved gene

neighborhood for pan-genomic analysis of bacterial strains and closely related species." In: *Nucleic Acids Res* 40. DOI: 10.1093/nar/gks757.

GFF and GVF specification documents (n.d.). https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md. Accessed: 2018-07-13.

Giasson, Frederick, Bruce D'Arcus, Bruce D'Arcus, Bruce D'Arcus, and Bruce D'Arcus (2008). *Bibliographic ontology*. Tech. rep. Technical report.

Huerta-Cepas, Jaime, Kristoffer Forslund, et al. (2017). "Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper". In: *Molecular biology and evolution* 34.8, pp. 2115–2122.

Huerta-Cepas, Jaime, Damian Szklarczyk, et al. (2016). "EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences". In: *Nucleic Acids Research* 44. DOI: 10.1093/nar/gkv1248.

JCBN, IUPAC-IUB (1983). "Nomenclature and symbolism for amino acids and peptides. Recommendations.(1984)". In: *Biochem. J* 219, pp. 345–373.

Jones, Philip et al. (2014). "InterProScan 5: Genome-scale protein function classification". In: *Bioinformatics* 30. DOI: 10.1093/bioinformatics/btu031.

Kim, Mincheol, Hyun Seok Oh, Sang Cheol Park, and Jongsik Chun (2014). "Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes". In: *International Journal of Systematic and Evolutionary Microbiology*. DOI: 10.1099/ijs.0.059774-0.

Konstantinidis, Konstantinos T. and James M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes". In: *Proceedings of the National Academy of Sciences* 102. DOI: 10.1073/pnas.0409727102.

Lebo, Timothy et al. (2013). "Prov-o: The prov ontology". In: *W3C recommendation* 30. URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/.

Lechner, Marcus et al. (2011). "Proteinortho: detection of (co-) orthologs in large-scale analysis". In: *BMC bioinformatics* 12.1, p. 124.

Leinonen, Rasko et al. (2010). "The European nucleotide archive". In: *Nucleic acids research* 39.

Li, Li, Christian J Stoeckert, and David S Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." In: *Genome research* 13. DOI: 10.1101/gr.1224503.

Lipman, D. and W. Pearson (1985). "Rapid and sensitive protein similarity searches". In: *Science* 227. DOI: 10.1126/science.2983426.

Ochman, Howard and Nancy A Moran (2001). "Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis". In: *Science* 292.5519, pp. 1096–1099.

Paul, Sandip, Evgeni V Sokurenko, and Sujay Chattopadhyay (2016). "Corrected Genome Annotations Reveal Gene Loss and Antibiotic Resistance as Drivers in the Fitness Evolution of Salmonella Typhimurium." In: *Journal of bacteriology*, JB–00545.

Sonnhammer, Erik LL and Gabriel Östlund (2014). "InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic". In: *Nucleic acids research* 43.D1, pp. D234–D239.

The DDBJ/EMBL/GenBank Feature Table: Definition (2014). URL: http://www.insdc.org/files/feature%7B%5C_%7Dtable.html (visited on 08/01/2015).

The Gene Ontology Consortium (2015). "Gene Ontology Consortium: going forward". In: *Nucleic Acids Research* 43. DOI: 10.1093/nar/gku1179.

UniProt Consortium, The (2017). "UniProt: the universal protein knowledgebase". In: *Nucleic acids research* 45.D1. DOI: 10.1093/nar/gkw1099.

Wattam, Alice R et al. (2016). "Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center". In: *Nucleic acids research* 45.D1, pp. D535–D542.

Whitman, William B, David C Coleman, and William J Wiebe (1998). "Prokaryotes: The unseen majority". In: *Proceedings of the National Academy of Sciences* 95. DOI: 10.1073/pnas.95.12.6578.

Wilkinson, Mark D et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.

Wolf, Yuri I and Eugene V Koonin (2012). "A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes". In: *Genome biology and evolution* 4.12, pp. 1286–1294.

Wolstencroft, Katherine et al. (2015). “SEEK: A systems biology data and model management platform”. In: *BMC Systems Biology*. DOI: 10.1186/s12918-015-0174-y.

2

The Empusa code generator: bridging the gap between the intended and the actual content of RDF resources

Jesse C.J. van Dam, Jasper J. Koehorst,
Peter J. Schaap, Maria Suarez-Diez

arXiv 1812.04386; url: https://arxiv.org/abs/1812.04386

Abstract

The RDF data model facilitates integration of diverse data available in structured and semi-structured formats. This flexibility makes RDF an efficient alternative to develop resources integrating heterogeneous data sets. To obtain an RDF graph with a low amount of errors and internal redundancy, the chosen ontology must be consistently applied. However, with each addition of new diverse data the ontology must evolve thereby increasing its complexity, which could lead to accumulation of unintended erroneous composites. Thus, for the development of graph databases that are continuously enriched with new, heterogeneous diverse data there is a need for a gatekeeping system that compares the intended content described in the ontology with the actual content of the resource.

Here we present Empusa, a tool that has been developed to facilitate the creation of composite RDF resources from disparate sources. Empusa can be used to convert a schema into an associated application programming interface (API) that can be used to perform data consistency checks and generates Markdown documentation to make persistent URLs resolvable. In this way, the use of Empusa ensures consistency within and between the ontology (OWL), the Shape Expressions (ShEx) describing the graph structure, and the content of the resource.

Background & Summary

Semantic Web technologies provide information retrieval and management systems to integrate heterogeneous data from disparate sources (Berners-Lee, Hendler, and Lassila, 2001). The RDF data model is a W3C standard for storage of information in the form of self-descriptive Subject, Predicate and Object triples that can be linked in an RDF-graph (Brickley and Guha, 2004; organisation, 2014). The use of retrievable controlled vocabularies enables integration of heterogeneous diverse data from different sources in a single repository and SPARQL can be used to query the so generated resources (Prud'hommeaux and Seaborne, 2008; Aranda et al., 2013).

By themselves, RDF graphs have no predefined structure nor a schema, and the structure of an RDF resource can vary as new triples are added. Therefore, a formal definition of the relations among the terms, called an ontology, is required to efficiently retrieve linked information from these resources. Structural information can be encoded using Web Ontology Language (OWL) files (Bao et al., 2012). RDFS is another, related, standard to define the structure of an RDF resource (Brickley, Guha, and McBride, 2014). In this standard, each object can be defined as an instance of a class and each link as the realisation of a property. Shape Expressions (ShEx) is a standard to describe, validate and transform RDF data. One of the goals of this standard is to create an easy to read language for the validation of instance data (Solbrig and Prud'hommeaux, 2014; Boneva et al., 2014; Prud'hommeaux, Labra Gayo, and Solbrig, 2014).

In previous work, we developed RDF2Graph, a tool to automatically recover the structure of an RDF resource and to generate a visualisation, ShEx file and/or an OWL ontology thereof (J. C. v. Dam et al., 2015). Application of RDF2Graph to resources providing data in the RDF data model in the life sciences domain such as Reactome, ChEBI, UniProt, or those transformed by the Bio2RDF project (Belleau et al., 2008; Croft et al., 2014; Hastings et al., 2013; Jupp et al., 2014; Bateman et al., 2017) showed mismatches between the retrieved data structure and the one described in the OWL definition of the particular resource. The main reason for this lack of consistency is the flexibility provided by RDF: the data graph is a free format, the ontology defines the structure but does not enforce it.

In the development of RDF resources, transformation of existing data into the RDF data model is often a source of errors such as typing errors in the predicates, instances with missing attributes, instances that did have a non-unique IRI, and instances that had no type defined, among others. Development of tools that directly use the RDF data model as means to store their output may therefore be essential to unlock the potential of these technologies in the life sciences. An example of a such tool is the Semantic Annotation Platform with Provenance (SAPP) (Koehorst et al., 2018), that can automatically annotate genome sequences using standard tools and directly store the annotation results and their provenance in the RDF data model using the Genome Biology Ontology Language (GBOL) (J. C. J. v. Dam et al., 2017). Development of such tools would be greatly facilitated by supporting tools able to read an ontology definition and generate code that can be used for data generation, export and validation.

Here we present *Empusa*, that has been developed to facilitate the creation of RDF resources, which are validated upon creation (figure 2.1). Empusa uses an OWL and a simplified version of ShEx, defining an ontology, and generates an associated application programming interface (API) that can be used to perform data consistency checks. The use of Empusa ensures consistency within and between the ontology (OWL), the Shape Expressions (ShEx) describing the graph structure and the content of the resource. In addition, Markdown documentation is generated, making URLs related to the ontology resolvable (Gruber, 2004).

Methods

The input definition of Empusa is a combination between OWL and a simplified version of ShEx, which can be edited within Protégé (Musen, 2015). The classes are defined in OWL, whereas the properties are defined in each class under the annotation property *propertyDefinitions* encoded within a simplified format of the ShEx standard. Additionally predefined value sets can be defined by adding a subclass to the *EnumeratedValueClass*. For instance a *FileType* can only be one element of a predefined list (e.g. CSV, TXT, TSV).

The RDFS standard is used to define the *subClassOf* relationships between the

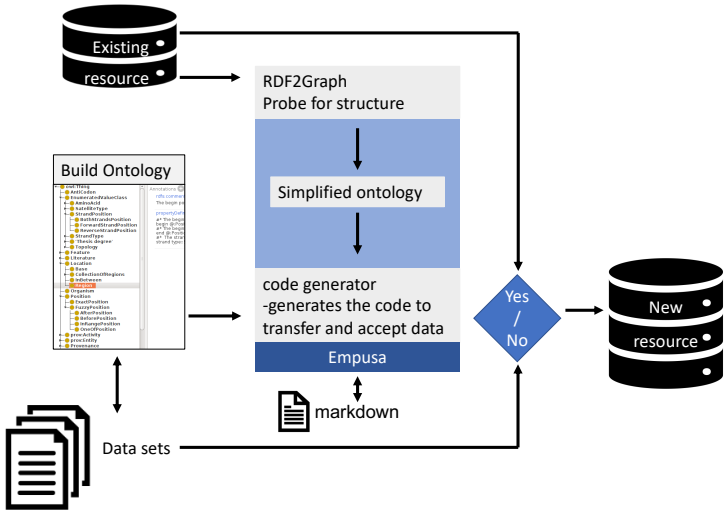

Figure 2.1: ***Simplified overview of the workflow to manage consistent integration of new diverse data with existing resources.*** *Empusa enables error control as it compares the intended content, described in the ontology, with the actual content of the resource. For this, Empusa checks whether or not Subjects and Objects have the properties that the ontology demands. Empusa builds upon RDF2Graph(J. C. v. Dam et al., 2015), a tool to automatically recover the structure of an RDF resource, to generate a visualisation, ShEx file, and/or an OWL ontology thereof.*

classes, whereas the ShEx standard is used to define the properties of each class. Properties of the class are defined through the annotation property *propertyDefinitions* as shown in figure 2.2. For each property the multiplicity and the expected type of the target value can be defined. The multiplicity can either be: *0..1* indicating that the property is optional and at most one reference is allowed; *1..1* indicating that one and only reference is allowed; *0..N* for optional properties with multiple allowed references; and *1..N* for properties that must have at least one reference. The '=' and '∼' sign can be used to define the references to be stored as an ordered or numbered list to ensure that the elements are numbered. Target value types can also be defined. The type of the target value can be either: A simple value (String, Integer or Double, among others); Another class (for example a Protein); Or an IRI, referencing an external resource or ontology or to a sub-ontology (value set). Within the ontology, sub- ontologies (value sets) can be defined under the *EnumeratedValue* class. Every sub-class of *EnumeratedValue* class represents one sub ontology. All

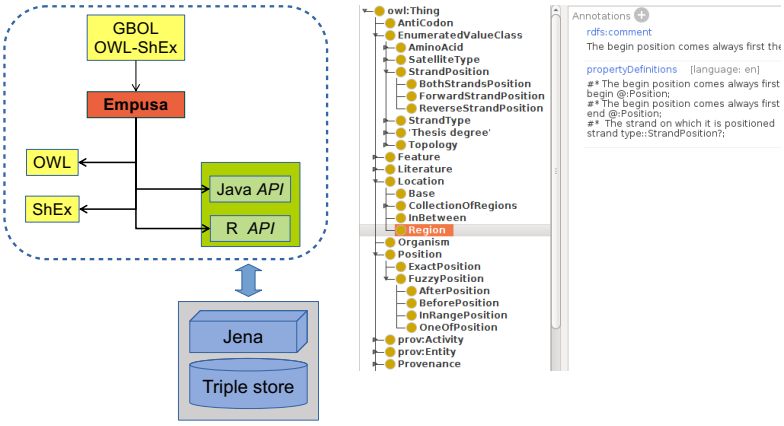

Figure 2.2: ***Empusa file definition****. Left: The input definition file (combining OWL and SHeX) is used to provide an ontology (here the GBOL(J. C. J. v. Dam et al., 2017) ontology is used as example). Empusa generates as output: an OWL file definition, a ShEx file that can be used for instance validation, the corresponding documentation in Markdown format, and R and Java APIs. Right Example input file. Properties within a class can be defined with the propertyDefinitons annotation property. As an example, the Region class has been highlighted. Value sets (sub-ontologies) can be defined under the EnumeratedValueClass class, for example the StrandPosition value set.*

subsequent sub-classes are elements of the sub-ontology of which it is sub-classed from. A class/sub-class structure can be defined for these elements within the sub-ontology.

The Empusa code generator uses this definition to generate: (*i*) An OWL file definition. It should be noted that the OWL file definition is generated as it remains general consensus within the field of semantics that these files are created for each ontology. (*ii*) A full ShEx file that can be used to validate a data set containing information that is encoded with the ontology. (*iii*) An R and Java API, which one can use to generate the data with the encoding of the defined ontology. This API ensures that the multiplicities and referenced types are correct and prevents many errors in the data export. (*iv*) A full documentation of the ontology based on *mkdocs*. The rdfs:label and skos:description properties can be used within the ontology to add a description about the classes and a comment line above each property definition in the simplified ShEx definition and can be used to add a description to each property.

Code availability

Empusa is written in Java with Gradle as build system. Empusa codebase is available at `http://www.gitlab.com/Empusa` under the MIT license. Documentation and tutorials can be found at associated website `http://empusa.org`.

Discussion

Empusa was developed primarily to help develop ontologies focusing on their function as a database schema for RDF resources. The design principles "modularity", "human readability", and "annotation" are followed to ensure that the so generated ontology can be easily extended (Bizer, Heath, and Berners-Lee, 2009). Empusa can automatically and consistently generate an OWL and a ShEx definition, ontology documentation in Markdown, an API, a JSON-LD framing file and a visualisation. Empusa uses parts of the RDF2Graph tool (J. C. v. Dam et al., 2015) to generate a representation that can be subsequently used to generate a visualisation within Cytoscape (Shannon et al., 2003). This allows users to browse the complete ontology intuitively.

Development of Empusa was closely related to the development of the GBOL stack (J. C. J. v. Dam et al., 2017) and the associated tool SAPP (Koehorst et al., 2018). GBOL enables interoperable genome annotation, as it deploys and extends existing ontologies to represent genomic entities, their properties and associated provenance. The GBOL stack contains over 80.000 lines of R and Java code, OWL and ShEx definition files, and documentation files (mkdocs format). Generating such a large amount of code would entail 1 year of manual work (considering an efficiency of 50 lines per hour) (Nawrocki and Wojciechowski, 2001).

Moreover, during the development of the GBOL ontology countless updates were made to correctly encapsulate all the data and associated provenance. Most of these updates were based on insights gained through the data encoding process. Manually updating the code, without using the supporting Empusa tool, would have entailed so much work that it would still be an on-going process. Thus, the Empusa code generator can serve to reduce the time (and costs) associated to development of ontologies and tools.

In conclusion, the Empusa code generator can be used to develop new ontologies combined with automatic generation of API and documentation. This reduces the complexity and time to extend and develop ontologies and tools able to exploit the full potential of Semantic Web technologies for heterogeneous data integration. Moreover, Empusa enables the validation of the generated resources and the verification of the consistency of the exported data thereby bridging the gap between the intended and the actual content of RDF resources.

Acknowledgements

This work has received funding from the Research Council of Norway, No. 248792 (DigiSal) and from the European Union FP7 and H2020 under grant agreements No. 305340 (INFECT), No. 635536 (EmPowerPutida), No. 634940 (MycoSynVac), No. 730976 (IBISBA 1.0), and the Netherlands Organisation for Scientific Research funded UNLOCK project (NRGWI.obrug.2018.005).

Author contributions

JvD was the primary developer of Empusa, explored the use cases and applications and drafted the manuscript. JK participated in code development and testing, explored the use cases and applications and revised the manuscript. PS explored the use cases and applications and revised the manuscript. MS-D explored the use cases and applications and revised the manuscript. All authors critically read, revised and approved the manuscript. All authors had full access to the underlying code and data.

Competing financial interests

The author(s) declare no competing financial interests.

Bibliography

Aranda, Carlos Buil et al. (2013). *SPARQL 1.1 Overview*. URL: https://www.w3.org/TR/sparql11-overview/ (visited on 10/06/2018).

Bao, Jie et al. (2012). *OWL 2 Web Ontology Language Document Overview (Second Edition)*. URL: https://www.w3.org/TR/owl2-overview/ (visited on 10/06/2018).

Bateman, Alex et al. (2017). "UniProt: The universal protein knowledgebase". In: *Nucleic Acids Research* 45.D1, pp. D158–D169. ISSN: 13624962. DOI: 10.1093/nar/gkw1099. arXiv: 1611.06654.

Belleau, François, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette (2008). "Bio2RDF: towards a mashup to build bioinformatics knowledge systems". In: *Journal of Biomedical Informatics* 41.5, pp. 706–716. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2008.03.004.

Berners-Lee, Tim, James Hendler, and Ora Lassila (2001). "The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities". In: *Scientific American* 284.5, pp. 34–43. ISSN: 0036-8733. DOI: 10.1038/scientificamerican0501-34. arXiv: 1204.6441.

Bizer, Christian, Tom Heath, and Tim Berners-Lee (2009). "Linked data-the story so far". In: *Semantic services, interoperability and web applications: emerging concepts*, pp. 205–227.

Boneva, Iovka et al. (2014). "Validating RDF with Shape Expressions". In: *arXiv:1404.1270 [cs]*. URL: http://arxiv.org/abs/1404.1270.

Brickley, Dan and R V Guha (2004). *RDF Vocabulary Description Language 1.0: RDF Schema*. URL: http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

Brickley, Dan, R V Guha, and Brian McBride (2014). *RDF Schema 1.1*. URL: http://www.w3.org/TR/rdf-schema/.

Croft, David et al. (2014). "The Reactome pathway knowledgebase". In: *Nucleic Acids Research* 42.

Dam, Jesse C. J. van, Jasper J. Koehorst, Jon Olav Vik, Peter J. Schaap, and Maria Suarez-Diez (2017). "Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining". In: *bioRxiv*.

Dam, Jesse CJ van, Jasper J Koehorst, Peter J Schaap, Vitor Ap Martins Dos Santos, and Maria Suarez-Diez (2015). "RDF2Graph a tool to recover, understand and validate the ontology of an RDF resource". In: *Journal of biomedical semantics* 6.

Gruber, John (2004). "Daring Fireball: Markdown". In: *Récupéré le* 3.04, p. 2011.

Hastings, Janna et al. (2013). "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013". In: *Nucleic Acids Research* 41.D1, pp. D456–D463. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gks1146.

Jupp, Simon et al. (2014). "The EBI RDF platform: linked open data for the life sciences". In: *Bioinformatics* 30.9, pp. 1338–1339. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt765.

Koehorst, Jasper J et al. (2018). "SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles". In: *Bioinformatics* 34.8. Ed. by John Hancock, pp. 1401–1403. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx767.

Musen, Mark A (2015). "The protégé project: a look back and a look forward". In: *AI matters* 1.

Nawrocki, Jerzy and Adam Wojciechowski (2001). "Experimental evaluation of pair programming". In: *European Software Control and Metrics (Escom)*, pp. 99–101. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.19.1689%5C&rep=rep1%5C&type=pdf.

organisation, W3C (2014). *RDF 1.1 Concepts and Abstract Syntax*. URL: http://www.w3.org/TR/rdf11-concepts/.

Prud'hommeaux, Eric, Jose Emilio Labra Gayo, and Harold Solbrig (2014). "Shape expressions: an RDF validation and transformation language". In: *Proceedings of the 10th International Conference on Semantic Systems*. ACM, pp. 32–40.

Prud'hommeaux, Eric and Andy Seaborne (2008). *SPARQL Query Language for RDF*. URL: http://www.w3.org/TR/rdf-sparql-query/.

Shannon, Paul et al. (2003). "Cytoscape: A software Environment for integrated models of biomolecular interaction networks". In: *Genome Research* 13.11, pp. 2498–2504. ISSN: 10889051. DOI: 10.1101/gr.1239303.

Solbrig, Harold and Eric Prud'hommeaux (2014). *Shape Expressions 1.0 Definition*. URL: http://www.w3.org/Submission/2014/SUBM-shex-defn-20140602/.

3

Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining

Jesse C.J. van Dam*, Jasper J. Koehorst*, Jon Olav Vik,
Peter J. Schaap, Maria Suarez-Diez
* Equal contribution

bioRxiv 184747; doi: https://doi.org/10.1101/184747

Abstract

A standard structured format is used by the public sequence databases to present genome annotations. A prerequisite for a direct functional comparison is consistent annotation of the genetic elements with evidence statements. However, the current format provides limited support for data mining, hampering comparative analyses at large scale.

The provenance of a genome annotation describes the contextual details and derivation history of the process that resulted in the annotation. To enable interoperability of genome annotations, we have developed the Genome Biology Ontology Language (GBOL) and associated infrastructure (GBOL stack). GBOL is provenance aware and thus provides a consistent representation of functional genome annotations linked to the provenance. GBOL is modular in design, extendible and linked to existing ontologies. The GBOL stack of supporting tools enforces consistency within and between the GBOL definitions in the ontology (OWL) and the Shape Expressions (ShEx) language describing the graph structure. Modules have been developed to serialise the linked data (RDF) and to generate a plain text format files.

The main rationale for applying formalised information models is to improve the exchange of information. GBOL uses and extends current ontologies to provide a formal representation of genomic entities, along with their properties and relations. The deliberate integration of data provenance in the ontology enables review of automatically obtained genome annotations at a large scale. The GBOL stack facilitates consistent usage of the ontology.

Introduction

Advances in sequencing technologies have turned genomics into a data-rich scientific discipline in which the total assembled and subsequently annotated sequence data doubles every 30 months (ENA, 2017). To support the growth in data throughput, automated annotation algorithms have become an indispensable supplement to manual annotation (NCBI, 2015; Seemann, 2014) and currently, automatic annotations in the UniProt database outnumber manual annotations 100-fold (Consortium et al., 2014).

Functional genome comparison has been used to identify diagnostic markers, to develop effective treatments, and to understand genotype -phenotype associations (Dutilh et al., 2013; Cooper et al., 2013; Alföldi and Lindblad-Toh, 2013). The volume and heterogeneity of genome annotation data has created a unique type of big data challenge, namely how to transform computational predicted annotations into actionable knowledge. Tapping into these available resources is only efficiently done by computational means and requires a consistent interlinking of data so that data becomes Findable, Accessible, Interoperable and Reusable (FAIR) (Wilkinson et al., 2016).

The format for sharing of public genome sequence annotation data has been developed and is maintained by the International Nucleotide Sequence Database Collaboration (INSDC) a long-standing foundational initiative that operates between the DDBJ, EMBL-EBI and NCBI public repositories. However, tradeoffs between simplicity, human readability and representational power left little support for interoperability, i.e. the ability of computer systems to directly make use of information. The */inference* qualifier provides a structured description of evidence that supports feature identification or assignment. Thus, within the standard formats, data provenance of computational annotations could be stored under this optional inference tag but this tag is not designed to be used for contextual, element-wise provenance.

Currently, most annotations rely on computational predictions of structure or function, and the choice of thresholds for confidence scores becomes a key consideration. Tracking the provenance of genome annotations becomes essential for scien-

tific reproducibility and to enable critical reexamination of analyses. However, such meta-analysis is currently very time-consuming. Efficient meta-analysis would require a framework able to accommodate the various types of annotations (e.g. gene prediction, homology, protein domains) directly linked to the supporting statistical evidence. Presently, no machine-readable infrastructure exists to directly query genome annotations linked to the historical and contextual provenance. The World Wide Web consortium provides the Semantic Web and the Resource Description Framework (RDF) data model, supporting these requirements. For RDF, ontologies are essential as they provide consistency in the meaning of data elements and in the relationship between them (Hoehndorf, Schofield, and Gkoutos, 2015).

In this respect, ontologies already exist for various aspects of biology (Bard and Rhee, 2004). The Sequence Ontology (SO) (Galdzicki et al., 2014) was presented over 12 years ago and was designed as a complete terminology of unambiguous terms related to genetics. However, it was never intended to function as a file format or database schema, and provides no support for linked sets of data attributes. Furthermore, it has limited support for storing based-on provenance except for some experimental codes. FALDO's (Bolleman et al., 2014) only purpose is to unambiguously store genetic locations on a sequence. The Synthetic Biology Open Language (SBOL) (Galdzicki et al., 2014) was successfully designed to describe complete synthetic constructs and the interactions between each of the elements. None of these standards were designed to consistently store feature predictions with evidence provenance and therefore none of these tools provides a complete representation of the genomic information linked to the provenance it is based on.

To meet the requirements and to ensure interoperability of computational predictions, we developed an extendable provenance-centered infrastructure for interoperable genome annotations. The here presented infrastructure consists of two main elements; Firstly, the Genome Biology Ontology Language (GBOL), which directly integrates evidence provenance for the whole dataset and for each included element (dataset- and element- wise provenance). Secondly, the "GBOL stack" of enforcing tools facilitates the consistent usage of ontologies. GBOL is modular in design, extendible and linked to existing ontologies. Empusa has been developed as part of the GBOL stack to ensure consistency within and between ontology (OWL),

the API and the Shape Expressions (ShEx) describing the graph structure. This enables the use of SPARQL queries to include contextual details in large scale functional analyses. Modules have been developed to serialise the linked data (RDF) and to generate a plain text format files.

Results

Ontology structure

GBOL is a genome annotation ontology developed for the application of semantic web technologies in genome annotation and mining. As such GBOL provides the means to consistently describe computationally inferred genome annotations of biological objects typically found in a genome sequence annotation data file in the public repositories. Additionally, it can describe the linked data provenance of the extraction process of genetic information from genome sequences.

An overview of the structure of GBOL is shown in Figure 3.1. The ontology contains 251 classes that can be categorised into 6 broad domains (Table 3.1). In GBOL, sequences have features, which in turn have genomic locations on the sequence. The authority of this relationship is derived from the data provenance that captures both the statistical basis of each individual annotation (element-wise provenance) as well as the programs and parameters used for the complete set of sequences under study (dataset-wise provenance). All annotations for a given sequence can be packed into a single entity called a document.

Table 3.1: Overview of domains, classes and properties described by the the GBOL ontology. Note that some properties might be in multiple sub domains.

Sub domain	Classes	Properties	Value sets
Genomic locations	16	17	1
Genes transcripts and features	114	133	17
Document structure	27	107	7
Dataset-wise provenance	22	54	0
Element-wise provenance	5	9	0
BIBO	59	90	2

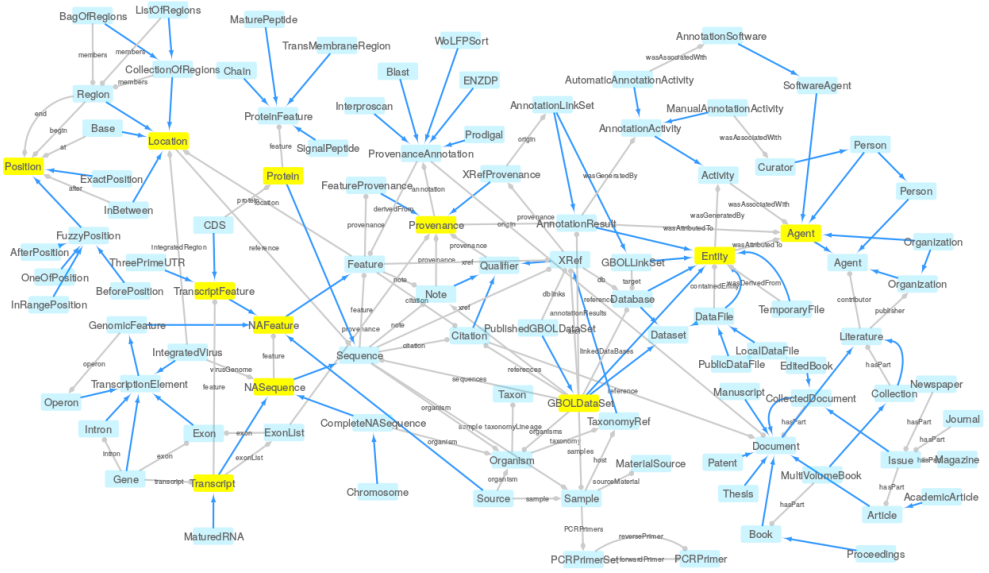

Figure 3.1: ***The GBOL ontology structure:*** *Network based view generated using RDF2Graph (Dam et al., 2015) the GBOL core ontology. Nodes represent types. Blue edges represent subClassOf relationships whereas grey edges represent unique type links. A unique type link is defined as a unique tuple: type of subject, predicate, (data)type of object. Arrow heads indicate the forward multiplicity of the unique type links: 0..1 and 1..1 multiplicities are indicated by diamonds; 0..N and 1..N multiplicities are indicated by circles. Neighbourhood of nodes marked in yellow is further expanded in Figures 4-8*

Design principles

GBOL was developed focussing on its function as as file format and as database schema and has the following design principles: modularity, human readability, and annotation. These principles ensure that the ontology can be easily extended (Bizer, Heath, and Berners-Lee, 2009).

Modularity: The number of classes in the main class tree is kept as small as possible and elements within the data are described with attributes when possible. Furthermore, classes are included in the main class tree only when there are unique properties in a class or in one of the sibling classes. This approach ensures that sub-ontologies can be managed as separate entities within the main ontology and that we can use existing ontologies. As an example the class *RegulationSite* has an attribute *regulatoryClass*, which denotes the type of regulation with a separate set of classes of which all are instances of the regulatory class.

To further simplify the ontology, every attribute is defined as a direct property within the class that links to either a string, an integer, another object or a class in an enumeration set. For each class in which the attribute is used, an 'all values from' axiom is used, with an optional minimal and/or maximal cardinality constraint. The 'all values from' axiom enforces all referenced objects to be of the expected type, which is not the case with the 'some values of' axiom and therefore we excluded the use of the 'some values of' axiom. This approach is fundamentally different from the principle used in the SO, in which attributes are defined using the 'has quality' property in combination with the 'some values of' axiom that references to a class.

Human Readability: All names within the ontology adhere to a set of basic principles to increase (human) readability of the ontology. All class names represent the underlying biological concept as closely as possible avoiding the use of unreadable numbers. All classes start with uppercase whereas properties start with lowercase. All words are spelled out, and white spaces are left out of the names, instead the next word starts with uppercase. In this way, the class 'exact position' becomes 'ExactPosition' and the property 'regulatory class' becomes 'regulatoryClass'. Furthermore, where possible, the names are shortened with abbreviations, as long as they remain understandable for a human reader (e.g. XRef instead of CrossReference).

Annotation: All classes and terms within the ontology are annotated with a short definition; an optional comment with additional usage information; an optional editorial comment relating to the development of the ontology itself; an optional *ddbj* label indicating the presence in the GenBank standard; and an optional SKOS (Miles et al., 2005) exact match to relate classes to terms in existing ontologies.

The GBOL infrastructure

An infrastructure enabling interoperable genome annotations integrated with provenance requires the following characteristics: *i*) An OWL (Antoniou and Van Harmelen, 2004) encoded definition of an ontology. *ii*) An infrastructure to enhance and simplify its usage, consisting of an interface (API) that allows to use Java and R. *iii*) A file format that can be obtained from serialising the linked data (RDF) using a lightweight Linked Data format (JSON-LD) (Sporny, Kellogg, and Lanthaler, 2013) which is subsequently serialised as YAML (Ben-Kiki, Evans, and Net, 2009).

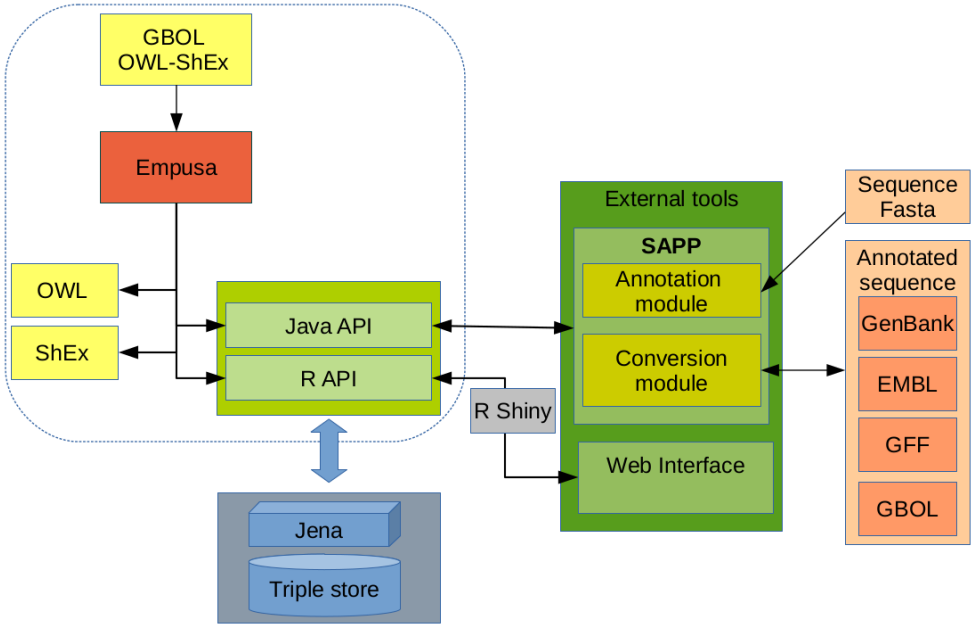

Figure 3.2: ***Schematic of an interoperable provenance centered genome annotation pipeline.*** *The GBOL stack (dashed box) provides the Genome Biology Ontology Language (GBOL) (Yellow) and associated infrastructure to keep it consistent and extendable (Empusa). The SAPP module functions as an interface for (standardly used) genome annotation tools. Using the JAVA API, SAPP retrieves raw genome data from the triple store, runs genome annotation tools in batch and uses the GBOL ontology to automatically store their predictions and associated data provenance directly as RDF triples in the triple store database (Blue). Stored predicted functional annotations, data provenance and linked meta-data can be queried within JAVA and R with SPARQL and by using a web interface (Green). Parsers have been developed for conversion of annotation files in standardly used formats (Orange).*

This format mimics the layout of the current format for sharing of public genome sequence annotation data, but has integrated support to add additional information. *iv*) A ShEx definition for data conformance validation to enhance data consistency (Prud'hommeaux, Labra Gayo, and Solbrig, 2014). And *v*) a tool to convert existing GenBank and EMBL format files into the GBOL format.

GBOL data can be stored in any of the linked data formats (RDF), such as Turtle. The generated API can be used to access the genomic information encoded within the GBOL format, which includes a data consistency validation. The API directly reads from and directly modifies the RDF data structure upon usage of any of the data model functions. This enables the usage of SPARQL within the client code,

which can run a SPARQL query and directly use the resulting objects nodes in the API. Moreover, the RDF data can be structured into a tree with the JSON-LD framing API into JSON-LD, which, in turn, can be further serialized as YAML resulting in a human readable format for sharing of public genome sequence annotation data. By addition of standard annotation tools, the GBOL stack can be at the core of a provenance-centered genome annotation framework (Figure 3.2).

Embedding with other ontologies

GBOL is embedded in the corpus of currently developed web technologies and when possible we have integrated existing ontologies such as: FALDO (Bolleman et al., 2014), PROV-O (Lebo et al., 2013), SO (Eilbeck et al., 2005), SBOL (Galdzicki et al., 2014), BIBO (Giasson et al., 2008), WikiData (Mitraka et al., 2015), FOAF (Brickley and Miller, 2007), Gene ontology (GO) (Ashburner et al., 2000) and the Evidence ontology (Chibucos et al., 2014) as depicted in Figure 3.3. Annotation of genomic location is inspired by FALDO ontology, although several elements had to be modified. The PROV-O ontology was used and extended to store data provenance. Whenever applicable, we added a cross-link to exact matching terms within the FALDO, SO and SBOL ontologies. Identification of persons and institutions is done through the FOAF ontology and BIBO is used to identify publications.

GBOL does not represent a vocabulary to describe genetic, molecular or cellular functions. Instead, terms can be cross-referenced to the many vocabularies that provide functional descriptions to the (products of) genetic elements, such as Gene Ontology, Enzyme commission (EC) numbers, and the CHEBI and RHEA databases (Degtyarenko et al., 2008; Alcántara et al., 2012), among others.

Key GBOL classes

Common elements in genome annotations include different classes of DNA molecules such as chromosomes, plasmids and contigs, genes, transcripts, exons, introns, proteins, protein domains and functional annotations. The following sections summarize the key classes of the ontology. An extensive description for each element can be found in the documentation available at `http://gbol.life/0.1/`.

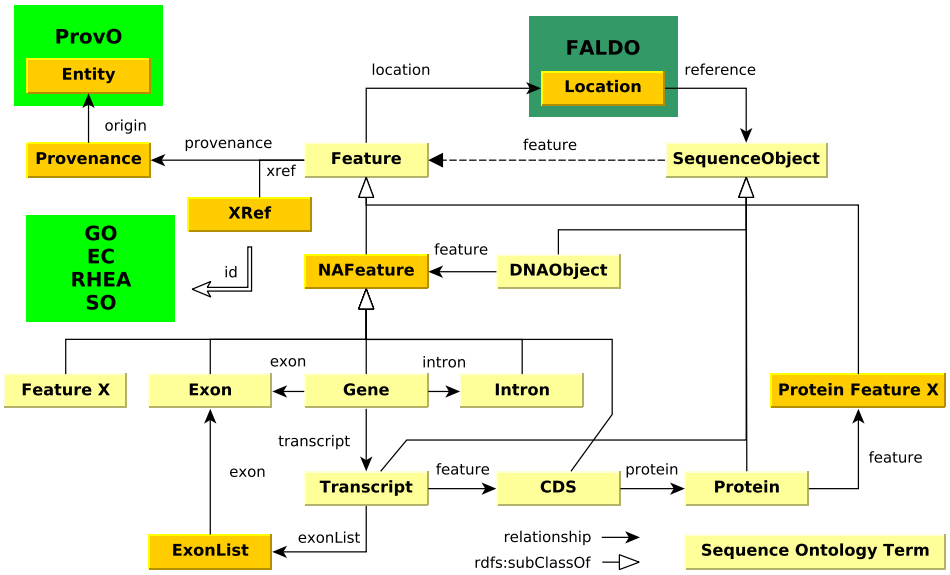

Figure 3.3: Embedding of the GBOL ontology with already existing ontologies. FALDO, ProvO, GO, EC, RHEA and SO are existing ontologies. Classes are in yellow and an explanation is provided in the main text.

Genomic locations: Genomic locations of all features in GBOL is captured with the *Location*, *Position* and *StrandPosition* classes, which are inspired by the FALDO ontology and represented in Figure 3.4. The *Location* and its subclasses together with the *StrandPosition* define an interval on the Sequence, whereas *Position* defines a single position in a sequence. A location can be either: *i*) A region which has begin and end positions; *ii*) A collection of regions (ordered or unordered); *iii*) A single base at a given position; or *iv*) an *InBetween* location denoting a location between two bases after the base of which the position is given. Each region, base and in-between location can be defined to be located on the forward, reverse or both strands, although no strand should be specified if the sequence is a single stranded DNA sequence or a protein sequence. It should be noted that elements of a collection of regions can be located on different sequences. This can be used to encode cases in which an otherwise indistinguishable genetic element is located on multiple chromosomes.

Exactly known positions can be indicated using the *ExactPosition* class containing the *position* property. Otherwise a not exactly known position, also called fuzzy

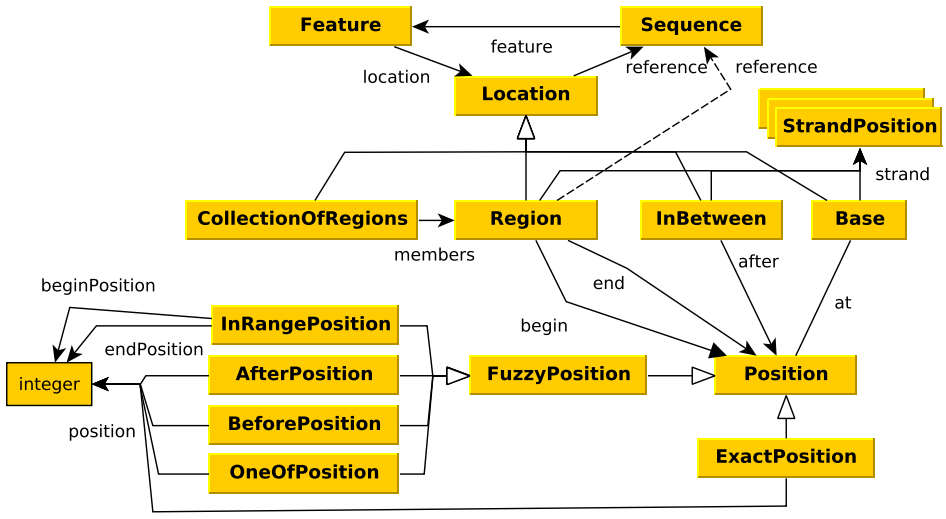

Figure 3.4: Graphical view of the GBOL ontology for genomic locations. An explanation of the classes is provided in the main text.

position, can be indicated using either the *BeforePosition* class containing the *position* property, the *AfterPosition* class containing the *position* property, the *InRangePosition* class containing the *beginPosition* and *endPosition* properties or the *OneOfPosition* class containing multiple *position* properties.

Genes, transcripts and other commonly encountered genomic features: GBOL has a consistent model for storing genes, exons, (alternatively spliced) transcripts, coding sequences and proteins. Central to this model is the *Sequence* class that can have multiple annotations represented in the *Feature* class. An overview is provided in Figure 3.5.

In GBOL a sequence can be specified as a nucleic acid (NA) or a protein sequence. The sequence is attached to the *Sequence* class via the *sequence* property, provided in the DNA, RNA or protein encoding standard. NA-sequences can represent transcripts or other elements such as chromosomes, plasmids, scaffolds, contigs or reads. No distinction is made between DNA and RNA and the *strandType* denotes that it is either a double or single stranded DNA or RNA. As indicated in Figure 3.5 the type of sequence determines the features it might be associated to (*ProteinFeature*, *NAFeature* or *TranscriptFeature*),

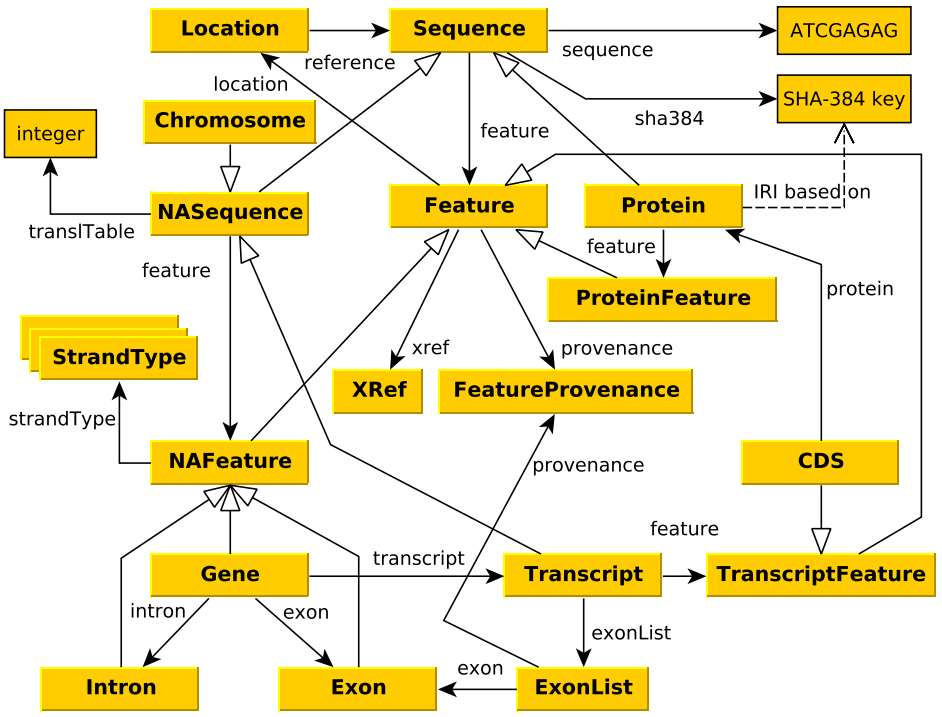

Figure 3.5: Graphical view of the GBOL ontology for genes, transcripts and other commonly encountered genomic features. An explanation of the classes is provided in the main text

Typically, each *GBOLDocument* contains one or more *NASequences* (e.g. *Chromosome*, *Contig*, *mRNA*), which can have multiple features including all gene, exon, intron, sequence variations, and structural, regulatory and repeat annotations. Each gene is linked to its associated exons, introns and transcripts. Due to alternative splicing a gene can have multiple transcripts. Each transcript has its own unique list of exons, which is linked through the *exonList* and associated *exonList* class to all associated exons. A transcript can be either a mRNA, ncRNA, rRNA, tmRNA, tRNA, precursor RNA or a miscellaneous RNA. The type of transcript determines the associated features: mRNA transcripts can have features linked to coding sequence (CDS), 5'-UTR, 3'- UTR and poly A tail.

The mRNA translation table is defined with the *translTable* property from the parent sequence. The association between CDS and the encoded protein is preserved and information about the translation is stored if it is different from the default

translation (for example, use of alternative stop codons).

Each protein has a unique IRI (http://gbol.life/0.1/protein/<SHA-384>) based on the SHA-384 hash of its sequence. This makes it possible to combine protein information from heterogeneous sources, as a protein can be associated to several CDS features. All information related to the protein which is unique to the genome (such as location) should be stored in the CDS feature. Protein annotation features may include, among other, conserved regions, protein domains, binding sites, 3D structure, signal peptides, transmembrane regions, and immunoglobulin regions. Operons can be defined with the *Operon* feature, to which other genomic features, such as genes, can be associated. Additionally, viral genome integration can be denoted using the *IntegratedVirus* feature.

Provenance related classes

Three types of provenance can be distinguished. Metadata refers to the owners of the samples, the biological origin, culture conditions etc. Dataset- and element-wise provenance pertain to the annotation process. All data within a single data collection stored in GBOL is based on the *GBOLDataSet*, which holds among other, references to all included samples, sequences, organisms, annotation results and linked databases. An overview of the document structure is given in Figure 3.6.

A sequence originates from a sample and samples are related to one or multiple organisms. The *sample* property which links to the *Sample* class describes where, when, how, by whom and from what the sample was collected. The fields follow the GenBank format. The *organism* property describes the taxonomic reference, its scientific name and its taxonomic lineage.

All annotations made within the *GBOLDataSet* have associated provenance and should originate from one of the listed annotation results, so that correspondence with originating databases is preserved. The *Database* and the *GBOLDataSet* classes are both sub classed from the void ontology, *Dataset* class contains a general description, including among other title, description, comment, license, version, data download address, SPARQL endpoint URI, and URL encoding.

Dataset-wise provenance: Storage of the dataset-wise provenance is based on the PROV-O ontology in which the *Entity*, *Agent* and *Activity* classes are central. An

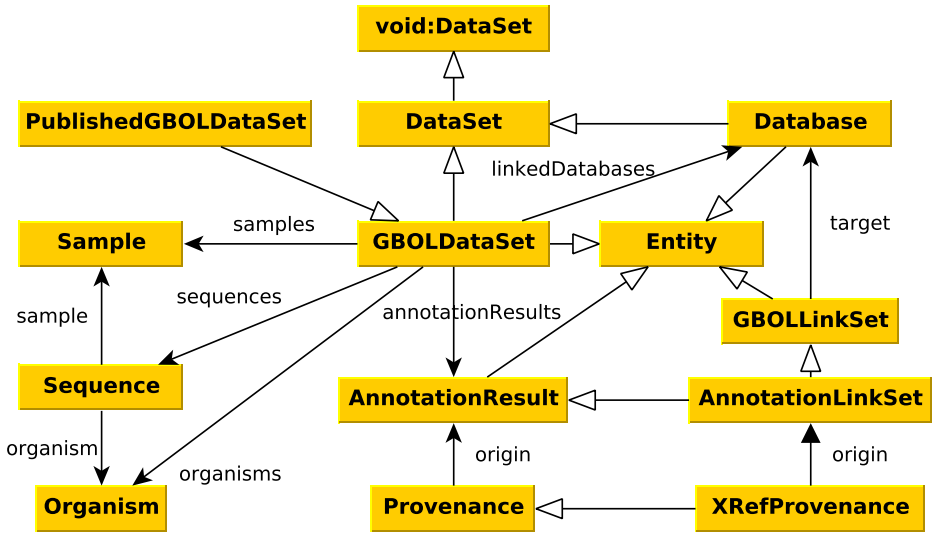

Figure 3.6: Graphical view of the GBOL Document structure. An explanation of the classes is provided in the main text

activity can use and generate entities, which are executed (*wasAssociatedWith*) by an agent. As a result, an entity can be attributed to an agent. The GBOLDataset, AnnotationResult, *GBOLLinkSet* and *Database* classes (indicated in Figure 3.6 and 3.7) are subclasses from the PROV-O ontology *Entity* class, so that for each of these objects provenance on how, when and by whom they were created can be associated.

In GBOL an *Entity* is either a file or an annotation result. The annotation result is a set of triples contained within a GBOL document, whereas a file represents a physical file either on a computer or network. An *agent* can either be a curator, person, organisation or annotation software. For the annotation software a version and code repository with associated commit identifier is included to enable univocal identification. For a curator, an ORCID (Haak et al., 2012) must be specified so that each curator can be uniquely identified together with his/her organisation. Both *Person* and *Organization* are sub-classed from the FOAF ontology to include additional information such as name and email address.

Within GBOL, each activity is an annotation activity, which can be either an automatic process or a manual curation activity, with a start and end time. An automatic annotation must be associated with a software agent and the set of parameters used

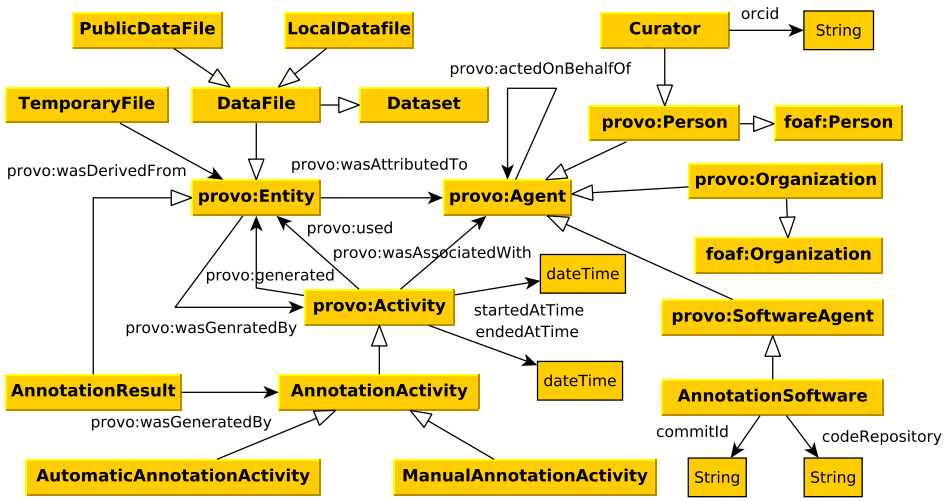

Figure 3.7: Graphical view of the GBOL Dataset-wise provenance. An explanation of the classes is provided in the main text

must be specified including the corresponding input and/or output files. Finally, manual curation must be associated with a curator.

Element-wise provenance and qualifiers: In addition to the dataset-wise provenance, GBOL is able to capture an additional layer of element-wise provenance, as the provenance of all the annotation in GBOL is captured per property per feature with the *FeatureProvenance*, as shown in Figure 3.8. For properties that could have items from multiple sources, we have defined the *Qualifiers*, each with its own associated provenance. A qualifier can either be a *citation*, *note* or cross reference (indicated by *xref*). A citation can hold a reference to literature encoded with the BIBO ontology.

Annotations are linked to the provenance object either through the *provenance* property of the qualifiers or the *onProperty* property of the *Provenance* feature. The provenance object links to both the dataset-wise provenance and the element-wise provenance. The *origin* links the provenance with the dataset-wise provenance (*AnnotationResult*), which includes among other the creation time, identity of the creating agent and the used parameters, as previously mentioned. The *annotation* links to the element-wise provenance (*ProvenanceAnnotation*), which includes: A free text

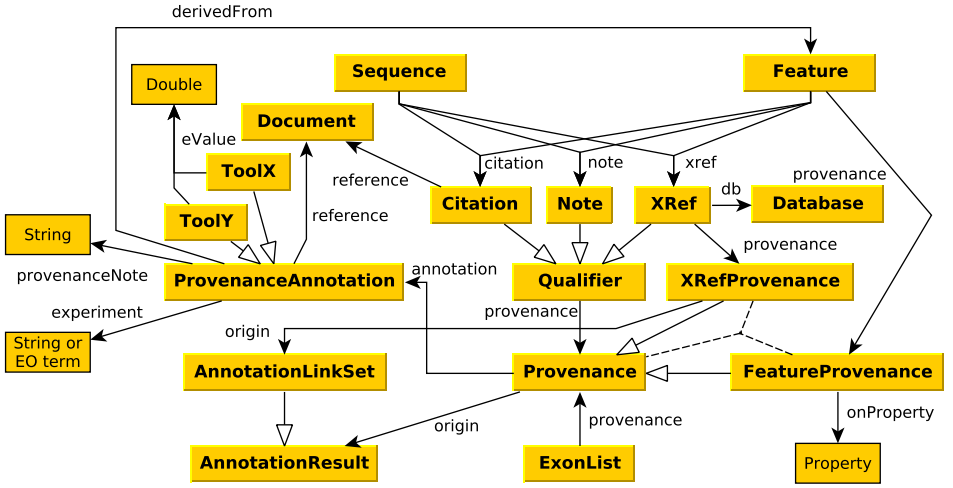

Figure 3.8: Graphical view of the GBOL element-wise provenance. An explanation of the classes is provided in the main text

note to describe the annotation; A list of references supporting the note; An experimental code, preferably from the Evidence Ontology to qualify the evidence supporting the conclusion; An optional *derivedFrom* that links to other features on which it is based.

Finally, each annotation tool generates its own evidence statements, often embedded in a statistical framework, characteristic of the algorithmic approach taken, such as p-values, bit scores, matching regions or any other scoring system. To store tool specific confidence scores, a subclass of the *ProvenanceAnnotation* class can be created. Some example classes include *Blast*, *HMM* and *SignalP* associated with the output of corresponding tools (Camacho et al., 2009; Rabiner and Juang, 1986; Petersen et al., 2011) However, these classes are not part of the GBOL ontology itself.

Empusa

During the development of the standard, difficulties were encountered in managing the large set of properties and structures in the OWL and ShEx definitions and the API needed to encode the annotation information in conjunction with the associated provenance. Moreover, Analyses of various public repositories have shown

that inconsistent, non-enforced usage of ontologies leads to mismatches between the descriptive OWL file and the actual content (Dam et al., 2015). In order to shorten the development cycle and to maintain consistency within and between the OWL and ShEx definitions and the API, a standalone tool was developed named Empusa. The input definition of Empusa is a combination between OWL and a simplified version of ShEx, which can be edited within Protégé (Musen, 2015). The classes are defined in OWL, whereas the properties are defined in each class under the annotation property 'propertyDefinitions' encoded within a simplified format of the ShEx standard. Additionally predefined value sets (for example all regulatory types) can be defined by adding a subclass to the EnumeratedValueClass. Each subclass of the value set is represented as one element within the value set. As standalone tool, Empusa can automatically and consistently generate an OWL and a ShEx definition, ontology documentation in markdown, an API, a JSON-LD framing file and a visualisation. Empusa uses parts of the RDF2Graph tool (Dam et al., 2015) to generate a representation that can be subsequently used to generate a visualisation within Cytoscape (Shannon et al., 2003). This allows users to browse the complete ontology intuitively.

Discussion

Comparative genome analysis is essential to understand the mechanisms underlying evolution and adaptation. Ideally, comparative genomics should be performed at the functional level, as this is highly scalable and more resistant to phylogenetic distances (Koehorst, Saccenti, et al., 2016). However, as functional annotation is performed in a non consistent manner the current practical level of interoperability is at the sequence level. Many tools exists to obtain orthologous clusters which are shaped by a generalised acceptance threshold for similarity and alignment length which is a trade-off between sensitivity and false discovery (Fouts et al., 2012; Li, Stoeckert, and Roos, 2003). At large scales these analysis are hampered by the high computational cost for finding bi-directional best matches. We have shown (Koehorst, Saccenti, et al., 2016) that functional comparison, based on consistently annotated protein domains, provides a fast, efficient and scalable alternative.

The prerequisite of a direct comparative functional analysis is consistent annotation of the genetic elements with evidence statements. Recording the provenance allows class-specific cut-offs for each individual annotation. Element-wise provenance enhances the re-usability of the annotations, and allows the development of methods to combine evidence statements, often derived from complex statistical frameworks, into confidence statements. Element-wise provenance also enables a quick re-evaluation of evidence, for instance by using a tunable cut-off score.

GBOL has been developed to explore available genome sequences using the mining possibilities of linked data. As a result, GBOL has evolved to consistently capture annotation data generated by the Semantic Annotation Platform with Provenance (SAPP), available at `http://semantics.systemsbiology.nl`. Previous versions of the GBOL ontology have been used to compare 432 *Pseudomonas* strains through integration of genomic, functional, metabolic and expression data (Koehorst, Van Dam, et al., 2016). Here GBOL was essential to capture, store and interlink the genomic and functional annotation data. Strikingly, over 432 *Pseudomonas* strains, consistent *de-novo* annotation yielded 838 additional GO-terms and 146 additional protein domains which would not have been identified using the original gene predictions. In addition to determining the functional pan- and core genome of a species, comparative genomics also enables the investigation of genotype-phenotype associations. In (Kamminga et al., 2017) we consistently functionally annotated and compared 80 publicly available mycoplasma genomes. The resulting semantic framework allowed us to efficiently query for functional differentiation of various mycoplasma species in relation to host specificity and phylogenetic distance.

Consistent functional annotation within a semantic framework requires a standardised ontology for the annotated elements and the associated based-on provenance. Linked data ensures that queries can be performed, mining multiple sequences at once, thereby providing a scalable alternative for large scale genome comparisons. The GBOL stack provides the ontology and corresponding API that enables the incorporation of functional annotation and provenance reducing complexity and is the outcome of efforts in a number of studies related to functional comparative genomics. Currently the GBOL stack is being used in various col-

laborative projects to handle genomic data of organisms across all domains of life (DIGISAL, 2017; INFECT, 2017; MycoSynVac, 2017; EmPowerPutida, 2017).

GBOL has been primarily designed to handle genomic annotation. However, it has been designed in a modular and extensible manner so that in the future it can be extended to host other omics data types as proteomics and transcriptomics. The modular design of GBOL ensures that other ontologies can be incorporated and managed as separate entities. For instance, the majority of the feature and sequence classes within GBOL can be connected with those from the Sequence Ontology and are therefore linked with the *skos:exactMatch* predicate. The major difference between GBOL and SO is that SO has been defined as vocabulary of terms related to genetic elements, whereas the GBOL classes have been designed to describe genetic annotation and elements located on a sequence and is inspired on the principles of the GenBank format. However, still a number of features in the SO are not currently available in GBOL and future work should focus on including them. Another possible extension would be to link to other Minimum Information Standards like MIGS and extensions thereof (MIMARKS, MIxS) (Field et al., 2008; Yilmaz et al., 2011) and cross domain experiment reporting standards like ISA-tab (Rocca-Serra et al., 2011). Other possible extensions relate to the development of the sub-ontologies GBOL links to. For instance, BIBO is used to store information on literature references, however the OWL ontology file of BIBO has to be further improved, as it does not specify to which classes all of the properties should belong. Therefore we have chosen to include a less consistent representation of the properties by adding all properties to the root class *bibo:Document*.

Empusa, a core part of the GBOL stack, ensures the correct usage of the ontology through the provided R and JAVA API. We have ensured that Empusa can be used independently of GBOL (documentation available at `http://gbol.life`) and therefore can be used to develop new ontologies combined with an automatically generated API and documentation. This reduces the complexity and time to extend and develop ontologies with corresponding API's and ensures consistent and correct usage of a defined ontology.

Conclusions

Large scale analysis of heterogeneous biological data is hampered by lack of interoperability. To improve the exchange of information formalised information models are required. GBOL provides a formal representation of genomic entities, their properties and relations. The GBOL Stack provides a framework to enforce consistent and correct usage of GBOL. The semantic basis and the integration of provenance enables FAIR genome annotations, thereby unlocking the potential of functional genome annotation data.

Methods

The GBOL ontology is OWL encoded and a ShEx schema is provided. All supporting software (Java and R API, Empusa) are written in Java with Gradle as build system. We use JENA (Jena, 2013) for handling and loading the RDF data into a triple store. Protégé was used for editing the ontology (Musen, 2015).

Storage of the genomic location is inspired by FALDO, although several elements had to be modified e.g. to account for features that start and end on different sequences. Differences include: *i*) *StrandPosition* is not subclassed from *Position*. Instead, an additional property is added to the region, *base* and *InBetween* location, this is done because these location object types can have both a strand position and an index position on the sequence. *ii*) The *reference* property is not part of a Position, but of a Location, because a location that starts on one sequence and ends on another sequence is an undefined sequence. *iii*) The *BaseLocation* and the *InBetweenLocation* classes have been added to the ontology. *iv*) The *BaseLocation*, *InBetweenLocation*, *CollectionOfRegions* and *Region* are children of the *Location* class, such that the rest of the ontology can incorporate these classes. *v*) The *before* and *after* positions have been explicitly defined to include their semantics. *vi*) The classes sub-classed from *FuzzyPosition* have an integer to denote the position and do not point to another position object, which could allow for arbitrary complex location denotations. *vii*) The N- and C-terminal positions have been removed and all indexes are counted from the N-terminal side. Counting from the C-terminal side can be calculated based on the sequence length. *viii*) The reflective properties *beginOf* and *endOf* have been

removed, because a position can also be referenced by the added base location. For consistency we have redefined all FALDO elements within our own namespace.

Cross-links to exact matching terms from other ontologies (such as FALDO, SO and SBOL) where added using skos:exactMatch. Additionally, several properties within the ontology point to existing ontologies, for instance: *i*) The *signalTarget* property of SignalPeptide, the *modificationFunction* of *ModifiedResidue* and the *organelle* of *Sample* are interlinked with GO terms. *ii*) The *experiment* property of ProvenanceAnnotation, which denotes upon which evidence the annotation is based on, should point, where possible, to a term within the Evidence Ontology. *iii*) The *residue* property of *ModifiedResidue* must point to a term within the Protein Modification Ontology (Montecchi-Palazzi et al., 2008). *iv*) GBOL includes the GO terms for *tissueType* of the Sample class and points, when possible, to a term within the BRENDA Tissue and Enzyme Source Ontology (Schomburg et al., 2004).

The source file of the ontology encoded in the Empusa and associated generated OWL definition, ShEx schema and visualization for Cytoscape available at `http://www.gitlab.com/GBOL` under the MIT license. The generated Java and R API are available at `https://gitlab.com/gbol/GBOLapi` and `https://gitlab.com/gbol/RGBOLApi` under the MIT license. The conversion module, which is part of SAPP, is available at `http://www.gitlab.com/SAPP/conversion` under the MIT license. The supporting Empusa code generator is available at `http://www.gitlab.com/Empusa` under the MIT license. All projects are coded in Java and are based on the Gradle build system. All terms are resolvable and can be browsed for at the associated website `http://gbol.life/0.1/`.

Acknowledgements

We thank Benoit Carreres for helpful design discussions. This work has received funding from the Research Council of Norway, No. 248792 (DigiSal) and from the European Union FP7 and H2020 under grant agreements No. 305340 (INFECT), No. 635536 (EmPowerPutida) and No. 634940 (MycoSynVac).

Bibliography

Alcántara, Rafael et al. (2012). "Rhea - A manually curated resource of biochemical reactions". In: *Nucleic Acids Research* 40. DOI: 10.1093/nar/gkr1126.

Alföldi, Jessica and Kerstin Lindblad-Toh (2013). "Comparative genomics as a tool to understand evolution and disease." In: *Genome research* 23. DOI: 10.1101/gr.157503.113.

Antoniou, Grigoris and Frank Van Harmelen (2004). "Web ontology language: Owl". In: *Handbook on ontologies*.

Ashburner, M et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." In: *Nature genetics* 25. DOI: 10.1038/75556.

Bard, Jonathan B L and Seung Y Rhee (2004). "Ontologies in biology: design, applications and future challenges". In: *Nature Reviews Genetics* 5. DOI: 10.1038/nrg1295.

Ben-Kiki, Oren, Clark Evans, and Ingy dot Net (2009). "YAML Ain't Markup Language (YAML) Version 1.2". In: *yaml. org*, *Tech. Rep*.

Bizer, Christian, Tom Heath, and Tim Berners-Lee (2009). "Linked data-the story so far". In: *Semantic services, interoperability and web applications: emerging concepts*.

Bolleman, J. et al. (2014). "FALDO: A semantic standard for describing the location of nucleotide and protein feature annotation." In: *Journal of Biomedical Semantics*. DOI: DOI10.1186/s13326-016-0067-z.

Brickley, Dan and Libby Miller (2007). *FOAF vocabulary specification 0.91*.

Camacho, Christiam et al. (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10. DOI: 10.1186/1471-2105-10-421.

Chibucos, Marcus C et al. (2014). "Standardized description of scientific evidence using the Evidence Ontology (ECO)". In: *Database* 2014.

Consortium, UniProt et al. (2014). "UniProt: a hub for protein information". In: *Nucleic acids research*.

Cooper, David N, Michael Krawczak, Constantin Polychronakos, Chris Tyler-Smith, and Hildegard Kehrer-Sawatzki (2013). "Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced pen-

etrance in human inherited disease". In: *Human Genetics* 132. DOI: 10.1007/s00439-013-1331-2.

Dam, Jesse CJ van, Jasper J Koehorst, Peter J Schaap, Vitor Ap Martins Dos Santos, and Maria Suarez-Diez (2015). "RDF2Graph a tool to recover, understand and validate the ontology of an RDF resource". In: *Journal of biomedical semantics* 6.

Degtyarenko, Kirill et al. (2008). "ChEBI: A database and ontology for chemical entities of biological interest". In: *Nucleic Acids Research* 36. DOI: 10.1093/nar/gkm791.

DIGISAL (2017). "DigiSal, Towards the Digital Salmon: From a reactive to a preemptive research strategy in aquaculture". In:

Dutilh, Bas E et al. (2013). "Explaining microbial phenotypes on a genomic scale: GWAS for microbes". In: *Briefings in Functional Genomics* 12. DOI: 10.1093/bfgp/elt008.

Eilbeck, Karen et al. (2005). "The Sequence Ontology: a tool for the unification of genome annotations." In: *Genome biology* 6. DOI: 10.1186/gb-2005-6-5-r44.

EmPowerPutida (2017). "EmPowerPutida, Exploiting native endowments by refactoring, re-programming and implementing novel control loops in Pseudomonas putida for bespoke biocatalysis". In:

ENA (2017). "ENA European Nucleotide Archive Statistics". In:

Field, Dawn et al. (2008). "The minimum information about a genome sequences (MIGS) specification". In: *Nat Biotechnol.* 26. DOI: 10.1038/1360.

Fouts, Derrick E., Lauren Brinkac, Erin Beck, Jason Inman, and Granger Sutton (2012). "PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species." In: *Nucleic Acids Res* 40. DOI: 10.1093/nar/gks757.

Galdzicki, Michal et al. (2014). "The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology". In: *Nature Biotechnology* 32. DOI: 10.1038/nbt.2891.

Giasson, Frederick, Bruce D'Arcus, Bruce D'Arcus, Bruce D'Arcus, and Bruce D'Arcus (2008). *Bibliographic ontology*. Tech. rep. Technical report.

Haak, Laurel L, Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner (2012). "ORCID: a system to uniquely identify researchers". In: *Learned Publishing* 25.

Hoehndorf, Robert, Paul N Schofield, and Georgios V Gkoutos (2015). "The role of ontologies in biological and biomedical research: a functional perspective". In: *Briefings in Bioinformatics* 16. DOI: 10.1093/bib/bbv011.

INFECT (2017). "INFECT, Systems medicine to understand severe soft tissue infections". In:

Jena, Apache (2013). *Apache jena*. URL: http://jena.apache.org/.

Kamminga, Tjerko et al. (2017). "Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life". In: *Frontiers in cellular and infection microbiology* 7.

Koehorst, Jasper J, Edoardo Saccenti, Peter J Schaap, Vitor AP Martins dos Santos, and Maria Suarez-Diez (2016). "Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics". In: *F1000Research* 5.

Koehorst, Jasper J, Jesse CJ Van Dam, et al. (2016). "Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data". In: *Scientific reports* 6.

Lebo, Timothy et al. (2013). "Prov-o: The prov ontology". In: *W3C recommendation* 30. URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/.

Li, Li, Christian J Stoeckert, and David S Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." In: *Genome research* 13. DOI: 10.1101/gr.1224503.

Miles, Alistair, Brian Matthews, Michael Wilson, and Dan Brickley (2005). "SKOS Core: Simple knowledge organisation for the Web". In: *International Conference on Dublin Core and Metadata Applications* 0.

Mitraka, Elvira et al. (2015). "Wikidata: A platform for data integration and dissemination for the life sciences and beyond". In: *bioRxiv*. DOI: 10.1101/031971.

Montecchi-Palazzi, Luisa et al. (2008). "The PSI-MOD community standard for representation of protein modification data". In: *Nature biotechnology* 26.

Musen, Mark A (2015). "The protégé project: a look back and a look forward". In: *AI matters* 1.

MycoSynVac (2017). "MycoSynVac, Engineering Mycoplasma pneumoniae as a broad-spectrum animal vaccine". In:

NCBI (2015). *The Bacterial Genome Submission Guide*. URL: http://www.ncbi.nlm.nih.gov/genbank/genomesubmit.html (visited on 05/25/2015).

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen (2011). "SignalP 4.0: discriminating signal peptides from transmembrane regions." In: *Nature methods* 8. DOI: 10.1038/nmeth.1701.

Prud'hommeaux, Eric, Jose Emilio Labra Gayo, and Harold Solbrig (2014). "Shape expressions: an RDF validation and transformation language". In: *Proceedings of the 10th International Conference on Semantic Systems*. ACM.

Rabiner, Lawrence and B Juang (1986). "An introduction to hidden Markov models". In: *ieee assp magazine* 3.

Rocca-Serra, Philippe et al. (2011). "ISA software suite: Supporting standards-compliant experimental annotation and enabling curation at the community level". In: *Bioinformatics*. Vol. 27. DOI: 10.1093/bioinformatics/btq415.

Schomburg, Ida et al. (2004). "BRENDA, the enzyme database: updates and major new developments". In: *Nucleic acids research* 32.

Seemann, Torsten (2014). "Prokka: Rapid prokaryotic genome annotation". In: *Bioinformatics* 30. DOI: 10.1093/bioinformatics/btu153.

Shannon, Paul et al. (2003). "Cytoscape: A software Environment for integrated models of biomolecular interaction networks". In: *Genome Research* 13. DOI: 10.1101/gr.1239303.

Sporny, Manu, Gregg Kellogg, and Markus Lanthaler (2013). *JSON-LD 1.0 -A JSON-based Serialization for Linked Data*. URL: http://www.w3.org/TR/2013/CR-json-ld-20130910/.

Wilkinson, Mark D et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.

Yilmaz, P. et al. (2011). "Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications". In: *Nature Biotechnology* 29. DOI: 10.1038/nbt.1823.

4 SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles

Jasper J. Koehorst, Jesse C. J. van Dam, Edoardo Saccenti,
Vitor A. P. Martins dos Santos, Maria Suarez-Diez and Peter J. Schaap

Bioinformatics; doi: https://doi.org/10.1093/bioinformatics/btx767

Abstract

To unlock the full potential of genome data and to enhance data interoperability and reusability of genome annotations we have developed SAPP, a Semantic Annotation Platform with Provenance. SAPP is designed as an infrastructure supporting FAIR *de novo* computational genomics but can also be used to process and analyse existing genome annotations. SAPP automatically predicts, tracks and stores structural and functional annotations and associated dataset- and element-wise provenance in a Linked Data format, thereby enabling information mining and retrieval with Semantic Web technologies. This greatly reduces the administrative burden of handling multiple analysis tools and versions thereof and facilitates multi-level large scale comparative analysis.

Availability

SAPP is written in JAVA and freely available at
`https://gitlab.com/sapp` and runs on Unix-like operating systems. The documentation, examples and a tutorial are available at
`https://sapp.gitlab.io`.

Introduction

Managing the genomic data deluge puts specific emphasis on the ability of machines to automatically find and use the data. To meet this demand and to extract maximum benefit from research investments, digital objects should be Findable, Accessible, Interoperable and Reusable (i.e. FAIR) (Wilkinson et al., 2016).

Genome annotation data is usually findable and accessible through public repositories in which the data is linked to metadata providing detailed descriptions of the data acquisition and generation process. Interoperability reflects the potential for seamless integration of data from independent sources. Currently, genome comparisons usually involve a laborious process of data retrieval, modification and standardisation (canonicalization). Reusability requires rich metadata with provenance for each annotation. Current standard formats (GenBank, EMBL or GFF3) retain the output of the prediction tools (for example for gene identification) but only when they score better than a predefined, often pragmatic, prediction threshold. Detailed information of the actual prediction scores is lost. This hampers critical re-examination of the results.

Because existing genome annotation data is hard to be made FAIR and managing of FAIR genome annotation data requires a considerable administrative load, we developed SAPP, a semantic framework for large scale comparative functional genomics studies. SAPP can automatically annotate genome sequences using standard tools. The unique characteristic of SAPP is that the annotation results and their provenance are stored in a Linked Data format, thus enabling the deployment of mining capabilities of the Semantic Web. As the automatic annotations are incorporated into a dynamic framework, SAPP supports periodic querying, comparison and linking of diverse annotation sources, resulting in up-to-date genome annotations. By interrogating metadata as part of a digital annotation object, annotation data becomes interoperable as the extraction procedure requires no additional standardisation process.

Implementation

SAPP accepts annotated and non-annotated sequence files which are converted into an RDF data structure using the GBOL ontology (Dam et al., 2017). Within SAPP, structural and functional annotation is performed using add-on modules incorporating existing standard annotation tools such as Prodigal and Augustus (Hyatt et al., 2010; Stanke and Morgenstern, 2005). Modules for tRNA, tmRNA, rRNAs, protein domain and CRISPR repeats annotation are also available. New modules can be added. Annotation data and metadata are stored in a compressed graph database (Fernández et al., 2013), as shown in Figure 4.1A.

Genome annotations can be exported to standard formats. All data can be directly queried and compared using the SPARQL endpoint or via the GBOL API (Java/R). Complex queries can be performed on multiple genomes while simultaneously taking meta-data into account. A SPARQL query example is provided in Figure 4.1B. Examples to query SAPP from R, Java or Python, a tutorial and a list of publications in which SAPP was used can be found at `http://sapp.gitlab.io`.

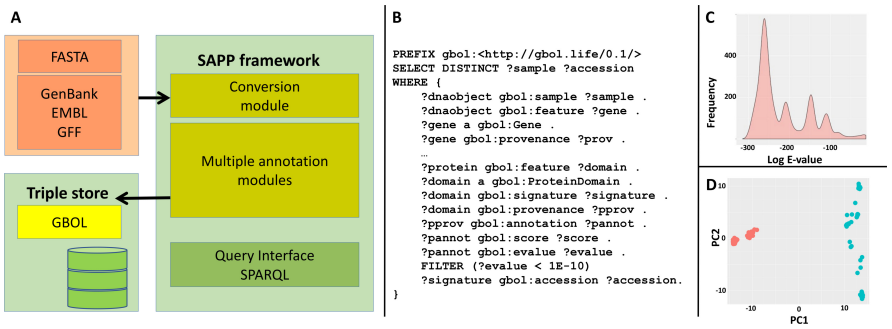

Figure 4.1: ***A:*** *The conversion module imports genome sequences in common formats. Annotation modules perform common tasks such as gene, tRNA, protein and protein domain annotation. Results are stored as Linked Data and consistency is ensured by the GBOL stack.* ***B:*** *SPARQL query to retrieve the E-value score of the instances of the protein domain PF00465 across multiple bacterial genomes.* ***C:*** *Distribution of E-values for protein domain PF00465 across multiple bacterial genomes: note the multimodality of the distribution.* ***D:*** *Principal component analysis of functional similarities of 100 bacterial genomes from the Streptococcus (blue) and the Staphylococcus (orange) genera. PC1 and PC2 account for 51.4% and 10.1% of the variance in the dataset respectively.*

Results and Discussion

Reproducible computational research requires a management system that links data with data provenance. Interoperability requires a strictly defined ontology. Using and sharing Linked Data based on controlled vocabularies and ontologies ensures the interoperability and reusability of the data. SAPP functionalities are unique since none of the existing *de novo* annotation pipelines implement Semantic Web technologies. SAPP generated data fulfil the applicable requirements for data FAIRness proposed by (Wilkinson et al., 2016).

For input and output, these tools interact directly with the database thereby forcing automatic linkage of data and provenance. In this way there is no need to work with predefined thresholds on the parameters controlling the annotation output. SAPP uses a controlled vocabulary to describe genome annotations. Consistency is ensured through the GBOL Stack (Dam et al., 2017).

The GBOL ontology enables consistent genome annotation while integrating dataset-wise and element-wise provenance. The element-wise provenance is the statistical basis or score of each individual annotation, whereas the dataset-wise provenance refers to the programs, versions thereof and parameters used for the complete annotation of the (set of) sequences under study.

GBOL makes use of existing ontologies: PROV-O for activity capturing (Lebo et al., 2013); FOAF for agent information (Brickley and Miller, 2007); BIBO for article information stored within the annotation files (Giasson et al., 2008); SO for sequence information (Eilbeck et al., 2005); FALDO for genomic location (Bolleman et al., 2014), among many others. We refer the reader to (Dam et al., 2017) for detailed information on the integrated ontologies and the data model.

Annotations can be evaluated through critical examination of the provenance. The use of SPARQL allows complex queries across data annotated with SAPP and in direct comparison of these annotations with external resources, such as UniProt. Additionally for specific questions, likelihood values can be integrated, normalized or corrected for multiple testing. For instance, study of E-value distribution on instances of a protein domain across multiple genomes can inform optimal threshold selection, as shown in Figure 4.1C. SAPP implements existing tools: consistency of

SAPP annotation and a comparison with deposited annotations is shown and discussed in (Koehorst, Van Dam, et al., 2016).

By querying multiple consistently annotated genomes simultaneously, large scale functional comparisons can be performed without additional conversion steps (Figure 4.1D and (Koehorst, Saccenti, et al., 2016)).

These examples demonstrate that by adopting FAIR principles to genome annotation, knowledge discovery is facilitated.

Acknowledgements

This work has received funding from the Research Council of Norway, No. 248792 (DigiSal) and from the European Union FP7 and H2020 under grant agreements No. 305340 (INFECT), No. 635536 (EmPowerPutida), Synthetic Biology Investment Theme (KB-32) from Wageningen University & Research, and No. 634940 (MycoSynVac).

Bibliography

Bolleman, J. et al. (2014). "FALDO: A semantic standard for describing the location of nucleotide and protein feature annotation." In: *Journal of Biomedical Semantics*. DOI: DOI10.1186/s13326-016-0067-z.

Brickley, Dan and Libby Miller (2007). *FOAF vocabulary specification 0.91*.

Dam, Jesse C. J. van, Jasper J. Koehorst, Jon Olav Vik, Peter J. Schaap, and Maria Suarez-Diez (2017). "Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining". In: *bioRxiv*.

Eilbeck, Karen et al. (2005). "The Sequence Ontology: a tool for the unification of genome annotations." In: *Genome biology* 6. DOI: 10.1186/gb-2005-6-5-r44.

Fernández, Javier D., Miguel A. Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias (2013). "Binary RDF representation for publication and exchange (HDT)". In: *Journal of Web Semantics* 19. DOI: 10.1016/j.websem.2013.01.002.

Giasson, Frederick, Bruce D'Arcus, Bruce D'Arcus, Bruce D'Arcus, and Bruce D'Arcus (2008). *Bibliographic ontology*. Tech. rep. Technical report.

Hyatt, Doug et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11. DOI: 10.1186/1471-2105-11-119.

Koehorst, Jasper J, Edoardo Saccenti, Peter J Schaap, Vitor AP Martins dos Santos, and Maria Suarez-Diez (2016). "Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics". In: *F1000Research* 5.

Koehorst, Jasper J, Jesse CJ Van Dam, et al. (2016). "Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data". In: *Scientific reports* 6.

Lebo, Timothy et al. (2013). "Prov-o: The prov ontology". In: *W3C recommendation* 30. URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/.

Stanke, Mario and Burkhard Morgenstern (2005). "AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints". In: *Nucleic acids research* 33.

Wilkinson, Mark D et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.

5

Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics

Jasper J. Koehorst, Edoardo Saccenti, Peter J. Schaap, Vitor A. P. Martins dos Santos, Maria Suarez-Diez

F1000 Research; doi: http://dx.doi.org/10.12688/f1000research.9416.3

Abstract

A functional comparative genome analysis is essential to understand the mechanisms underlying bacterial evolution and adaptation. Detection of functional orthologs using standard global sequence similarity methods faces several problems; the need for defining arbitrary acceptance thresholds for similarity and alignment length, lateral gene acquisition and the high computational cost for finding bidirectional best matches at a large scale. We investigated the use of protein domain architectures for large scale functional comparative analysis as an alternative method. The performance of both approaches was assessed through functional comparison of 446 bacterial genomes sampled at different taxonomic levels. We show that protein domain architectures provide a fast and efficient alternative to methods based on sequence similarity to identify groups of functionally equivalent proteins within and across taxonomic boundaries, and it is suitable for large scale comparative analysis. Running both methods in parallel pinpoints potential functional adaptations that may add to bacterial fitness.

Introduction

Comparative analysis of genome sequences has been pivotal to unravel mechanisms shaping bacterial evolution like gene duplication, loss and acquisition (Puigbò et al., 2014; J Peter Gogarten, W. F. Doolittle, and Lawrence, 2002), , and helped in shedding light on pathogenesis and genotype-phenotype associations (Dutilh et al., 2013; Pallen and Wren, 2007).

Comparative analysis relies on the identification of sets of orthologous and paralogous genes and subsequent transfer of function to the encoding proteins. Technically, orthologs are defined as best bi-directional hits (BBH) obtained via pairwise sequence comparison among multiple species and thus exploits sequence similarity for functional grouping. Sequence similarity-based (SB) methods present a number of shortcomings. First, a generalised minimal alignment length and similarity cut-off need to be arbitrarily selected for all, which may hamper proper functional grouping. Second, sequence and function might differ across evolutionary scales. Protein sequences change faster than protein structures and proteins with same function but with low sequence similarity have been identified (Joshi and Xu, 2007; Kuipers et al., 2009). SB methods may fail to group them hampering a functional comparison. This limitation becomes even more critical when comparing either phylogenetically distant genomes or gene sequences that were acquired with horizontal gene transfer events. Recent technological advancements are resulting in thousands of organisms and billions of proteins being sequenced (Goodwin, McPherson, and McCombie, 2016), which increases the need of methods able to perform comparisons at the larger scales.

To overcome these bottlenecks, protein domains have been suggested as an alternative for defining groups of functionally equivalent proteins (Yang, R. F. Doolittle, and Bourne, 2005; L.-G. Snipen and Ussery, 2013; Koehorst et al., 2017) and have been used to perform comparative analyses of *Escherichia coli* (L.-G. Snipen and Ussery, 2013), *Pseudomonas* (Koehorst et al., 2017), *Streptococcus* (Saccenti et al., 2015) and for protein functional annotation (Addou et al., 2009; Thakur and Guttman, 2016). A protein domain architecture describes the arrangement of domains contained in a protein and is exemplified in Figure 5.1. As protein domains

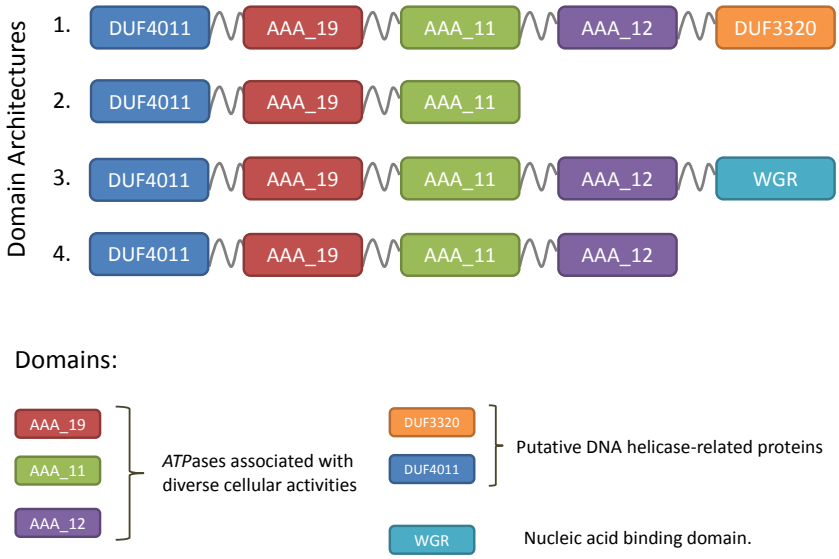

Figure 5.1: ***Domain architecture as a formal description of functional equivalence****. Although the proteins obviously share a common core, four distinct domain architectures involving six protein domains were observed in (1) Enterobacteriacee, (2) Helicobacter pylori, (3) Pseudomonas and (4) Cyanobacteria.*

capture key structural and functional features, protein domain architectures may be considered to be better proxies to describe functional equivalence than a global sequence similarity (Ponting and Russell, 2002). The concept of using the domain architecture to precisely describe the extent of functional equivalence is exemplified in Figure 5.2. Moreover, once the probabilistic domain models have been defined, mining large sets of individual genome sequences for their occurrences is a considerably less demanding computational task than an exploration of all possible bi-directional hits between them (Eddy, 1998; Van Domselaar et al., 2005).

Domain architectures have been shown to be preserved at large phylogenetic distances both in prokaryotes and eukaryotes (Koonin, Wolf, and Karev, 2002; Kummerfeld and Teichmann, 2009). This lead to the use of protein domain architectures to classify and identify evolutionarily related proteins and to detect homologs even across evolutionarily distant species (Björklund et al., 2005; Fong et al., 2007; Song,

Sedgewick, and Durand, 2007; B. Lee and D. Lee, 2009). Structural information encoded in domain architectures has also been deployed to accelerate sequence search methods and to provide better homology detection. Examples are CDART (Geer et al., 2002) which finds homologous proteins across significant evolutionary distances using domain profiles rather than direct sequence similarity, or DeltaBlast (Boratyn et al., 2012) where a database of pre-constructed position-specific score matrix is queried before searching a protein-sequence database. Considering protein domain content, order, recurrence and position has been shown to increase the accuracy of protein function prediction (Messih et al., 2012) and has led to the development of tools for protein functional annotation, such as UniProt-DAAC (Doğan et al., 2016) which uses domain architecture comparison and classification for the automatic functional annotation of large protein sets. The systematic assessment and use of domain architectures is enabled by databases containing protein domain information such as UniProt (The UniProt Consortium, 2015), Pfam (Robert D. Finn et al., 2016), TIGRFAMs (Haft, Selengut, and White, 2003), InterPro (Mitchell et al., 2015), SMART (Letunic, Doerks, and Bork, 2015) and PROSITE (Sigrist et al., 2012), that also provide graphical view of domain architectures.

Building on these observations we aim at exploring the potential of domain architecture-based (DAB) methods for large scale functional comparative analysis by comparing functionally equivalent sets of proteins, defined using domain architectures, with standard clusters of orthogonal proteins obtained with SB methods. We compared the SB and DAB approach by analysing *i*) the retrieved number of singletons (*i.e.* clusters containing only one protein) and *ii*) the characteristics of the inferred pan- and core-genome size considering a selection of bacterial genomes (both gram positive and negative) sampled at different taxonomic levels (species, genus, family, order and phylum). We show that the DAB approach provides a fast and efficient alternative to SB methods to identify groups of functionally equivalent/related proteins for comparative genome analysis and that the functional pan-genome is more closed in comparison to the sequence based pan-genome. DAB approaches can complement standardly applied sequence similarity methods and can pinpoint potential functional adaptations.

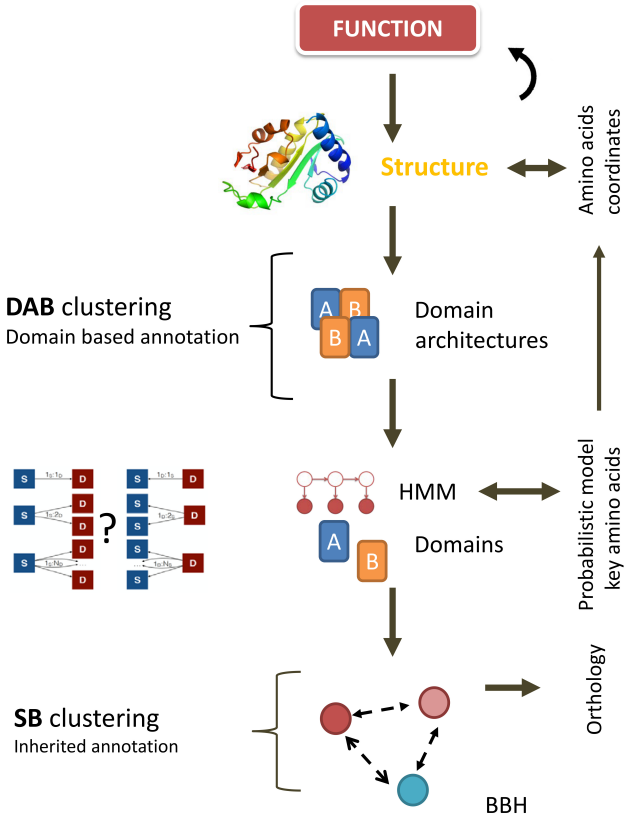

Figure 5.2: ***Relationship between Domain Architecture Based (DAB) and Sequence Similarity based (SB) clustering with respect to functional annotation.*** *Domains are probabilistic models of amino acids coordinates obtained by hidden Markov modelling (HMM) built from (structure based) multiple sequence alignments. Domain architectures are linear combinations of these domains representing the functional potential of a given protein sequence and constitute the input for DAB clustering. SB-orthology clusters inherit functional annotations via best bi-directional hits above a predefined sequence similarity cut-off score.*

Table 5.1: ***Comparison between DAB and SB clustering****. DAB has been performed using HMM from Pfam (29.0) and InterPro (interproscan-5.17-56.0). Fraction refers to the fraction of proteins with at least one (InterPro or PFAM) protein domain. Core- and pan- indicate the sizes of the core- and pan- genomes (based on the sample) and singletons refer to the number of clusters with only one protein.*

		Fraction		DAB	Pfam		DAB	InterPro		SB		
Taxon	Name	InterPro	PFAM	Core-	Pan-	Singletons	Core-	Pan-	Singletons	Core-	Pan-	Singletons
Species	*H. pylori*	0,82 ± 0,01	0,81 ± 0,01	724	1334	142	534	2888	853	1036	1503	295
Species	*L. monocytogenes*	0,89 ± 0,01	0,88 ± 0,02	1333	2142	309	1414	3415	847	2294	2937	746
Genus	*Bacillus*	0,87 ± 0,03	0,85 ± 0,03	792	5984	1474	342	16349	6745	885	9903	5505
Genus	*Pseudomonas*	0,88 ± 0,02	0,87 ± 0,02	1113	6572	1554	646	19387	7444	1453	12204	4838
Genus	*Streptococcus*	0,87 ± 0,02	0,85 ± 0,02	535	3435	845	244	8265	3276	716	4468	2116
Family	Enterobacteriaceae	0,91 ± 0,04	0,90 ± 0,05	146	6690	1664	20	19590	8173	197	10899	6715
Order	Corynebacteriales	0,83 ± 0,05	0,80 ± 0,06	475	6022	1719	130	22558	10554	605	12632	9087
Phylum	Cyanobacteria	0,77 ± 0,04	0,74 ± 0,05	400	9752	4428	120	27421	16140	511	10575	11154

Methods

Bacterial species were chosen on the basis of the availability of fully sequenced genomes in the public domain: two species (*Listeria monocytogenes* and *Helicobacter pylori*), three genera (*Streptococcus*, *Pseudomonas*, *Bacillus*), one family (Enterobacteriaceae), one order (Corynebacteriales), and one phylum (Cyanobacteria) were selected. For each, 60 genome sequences were considered, except for *L. monocytogenes* for which only 26 complete genome sequences were available. Maximal diversity among genome sequences was ensured by sampling divergent species (when possible) at each taxonomic level. Genome sequences were retrieved from the European Nucleotide Archive database (`www.ebi.ac.uk/ena`). A full list of genomes analysed is available in the Data availability section.

De novo genome annotation

To avoid bias due to different algorithms used for the annotation of the original deposited genome sequences, all genomes were *de novo* re-annotated using the SAPP framework (1.0.0) (Koehorst et al., 2017). In particular, the FASTA2RDF, GeneCaller (implementing Prodigal 2.6.2) (Hyatt et al., 2010) and InterPro (implementing interproscan-5.17-56.0) (Jones et al., 2014) modules were used to handle and re-annotate the genome sequences, and to store the results in the RDF data model. This resulted in 446 annotated genomes (7×60 genomes and 1×26 genomes) with provenance. For each annotation step, the provenance information (E-value cut off, score, originating tool or database) was stored together with annotation infor-

mation in a graph database (RDF-model) and can be reproduced through the SAPP framework (`http://semantics.systemsbiology.nl`).

Retrieval of domain architecture

The positions (start and end on the protein sequence) of domains having Pfam (Robert D. Finn et al., 2016), TIGRFAMs (Haft, Selengut, and White, 2003) and InterPro (Mitchell et al., 2015) identifiers were extracted through SPARQL querying of the graph database and domain architectures were retrieved for each protein individually. InterPro aggregates protein domain signatures from different databases. Here no pruning for redundancies has been done. Identification of domains was done using the intrinsic InterPro cut-off that represents in each case the e-values and the scoring systems of the member databases (Mitchell et al., 2015). The domain starting position was used to assess relative position in the case of overlapping domains; alphabetic ordering was used to order domains with the same starting position or when the distance between the starting position of overlapping domains was < 3 amino acids.

Labels indicating N-C terminal order of identified domains were assigned to each protein using the starting position of the domains: the same labels were assigned to proteins sharing the same domain architecture.

Sequence similarity based clustering

To make a direct comparison possible, only protein sequences containing at least one protein domain signature were considered for analysis. BBH were obtained using Blastp (2.2.28+) with an E-value cutoff of 10^{-5} and -max_target_seqs of 10^5. OrthaGogue (1.0.3) (Ekseth, Kuiper, and Mironov, 2013) combined with MCL (14-137) (Dongen, 2000) was used to identify protein clusters on the base of sequence similarity.

Domain architecture based clustering

Domain architecture based clusters were built by clustering proteins with the same labels using bash terminal commands (sort, awk). The number of proteins sharing a

given domain architecture in each genome was stored in a 446×21054 (genomes × domain architectures) matrix. From this matrix a binarised presence-absence matrix was obtained and used solely for principal component analysis.

Heaps' law fitting and pan-genome openness assessment

A Heaps' law model was fit to the abundance matrices using 5×10^3 random genome ordering permutations and the micropan R package (L. Snipen and Liland, 2015).

Software

SAPP, a Semantic Annotation Pipeline with Provenance, stores results in a graph database (Koehorst et al., 2017) used for genome handling and annotation and is available at `http://semantics.systemsbiology.nl`. Matrix manipulations and multivariate analysis were performed using the R software (3.2.2).

Results

SB and DAB approaches were compared by considering eight sets of genome sequences sampled at different taxonomic levels, from species to order, preserving phylogenetic diversity (see Table 5.1). Each set contained 60 genome sequences, except for *Listeria monocytogenes* for which only 26 complete genomes were publicly available. To facilitate the comparison between DAB and SB clusters only protein sequences that contained at least one domain were considered. On average, 85% of the protein sequences contain at least one domain from the InterPro database (see Table 5.1). Values range from 77±4% for Cyanobacteria to 91 ± 4% for Enterobacteriaceae (which include *E. coli*). Since the overall results were the same for gram negative and gram positive bacteria, we will show and comment only on results for the latter. Results obtained for gram negative bacteria are shown in the *Data availability section*.

Cluster formation based on sequence similarity

A standard BBH workflow was used to obtain SB protein clusters for the eight sets. We started by calculating the total number of clusters, corresponding to the pan-

genome size, as shown in Table 5.1. Then we considered protein cluster persistence, that is the number of genomes where at least one member of the cluster is present, divided by the total number of genomes considered. Results are shown in Figure 5.3.

The ratio between the size of the core-genome (clusters with persistence of 1, *i.e.* present in all genomes) and the number of singletons decreased with evolutionary distance (see Table 5.1). It ranged from 3.51 and 3.07 at species level (*H. pylori* and *L. monocytogenes* respectively) to 0.05 and 0.06 when considering members of the same order (Corynebacteriales) and phylum (Cyanobacteria) respectively. A similar pattern is observed when directly comparing the sizes of the pan- and core- genomes of the sampled genomes. Within the gram negative bacteria this ratio ranges from 0.69 for members of the same species (*H. pylori*) to 0.05 for members of the same phylum (Cyanobacteria) with intermediate values (0.12) for sequences from the same genus (*Pseudomonas*).

Cluster formation based on domain architectures

Domain architectures directly rely on the definition of protein domain models: those were retrieved from Pfam, InterPro and TIGRFAMs databases. However, TIGRFAMs results were not further considered due to a lower coverage (Table 5.1). As expected partly overlapping results were obtained when different domain databases were used. The number of singletons was larger when using InterPro rather than Pfam and for the latter we also observed larger core-genome size. These discrepancies can be due to the fact that InterPro aggregates different resources (including Pfam and TIGRFAMs) and domain signatures arising from different databases are integrated with different identifiers in InterPro. In light of this we focused on results obtained by using Pfam whose current release (30.0) contains hidden Markov models for over 16300 domain families. Size and persistence of groups of functionally equivalent proteins obtained using Pfam domains are presented in Figure 5.4.

Similar to what has been observed in the SB case we observed a decrease of the ratio between the size of the core genome and the number of singletons when higher taxonomic levels are considered. For organisms of the same species (*H. pylori* and *L. monocytogenes*) the ratio was 5.09 and 4.30, respectively, while for member of the

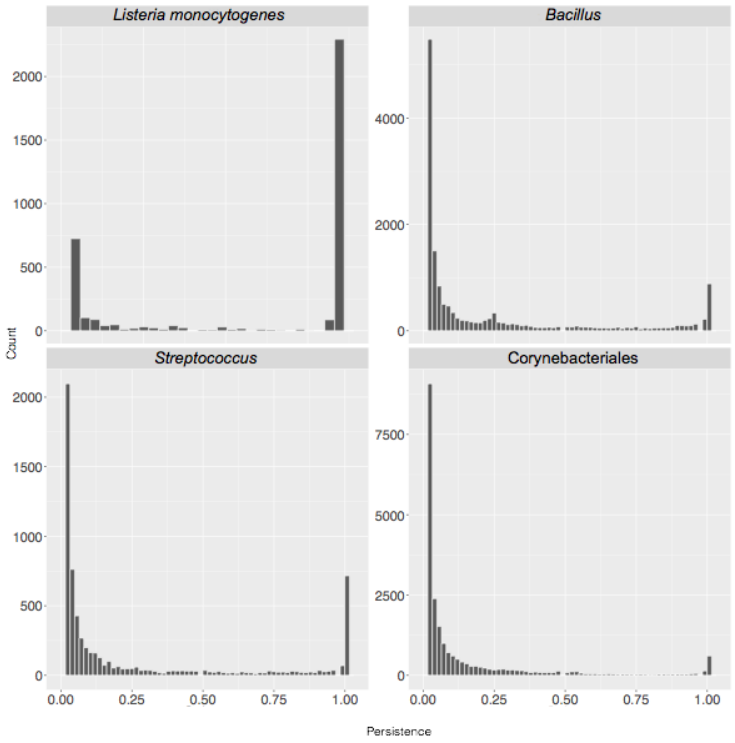

Figure 5.3: ***Persistence of sequence similarity based (SB) clusters.*** *Cluster persistence is defined as the relative number of genomes with at least one protein assigned to the cluster. The frequency of SB clusters according to their persistence is shown.*

same order (Corynebacteriales) and phylum (Cyanobacteria) it was 0.55 and 0.009 respectively. Similarly, also the ratio between the size of the core- and pan-genome decreases as higher taxonomic levels are considered, ranging from 0.54 for *H. pylori* to 0.04 for Cyanobacteria.

Comparison of DAB and SB clusters

We compared the clusters obtained using both approaches and the proteins assigned to them. The number of one-to-one relationships (indicating a complete agreement) between SB and DAB clusters is indicated in Table 5.2 and ranges from 648 (for *H. pylori*) to 1680 (in *Pseudomonas*) corresponding to 50% and 25% of the pan-genome. This indicates that results of SB and DAB clustering tend to be more similar when working at closer phylogenetic distances. However, more complicated cases occur

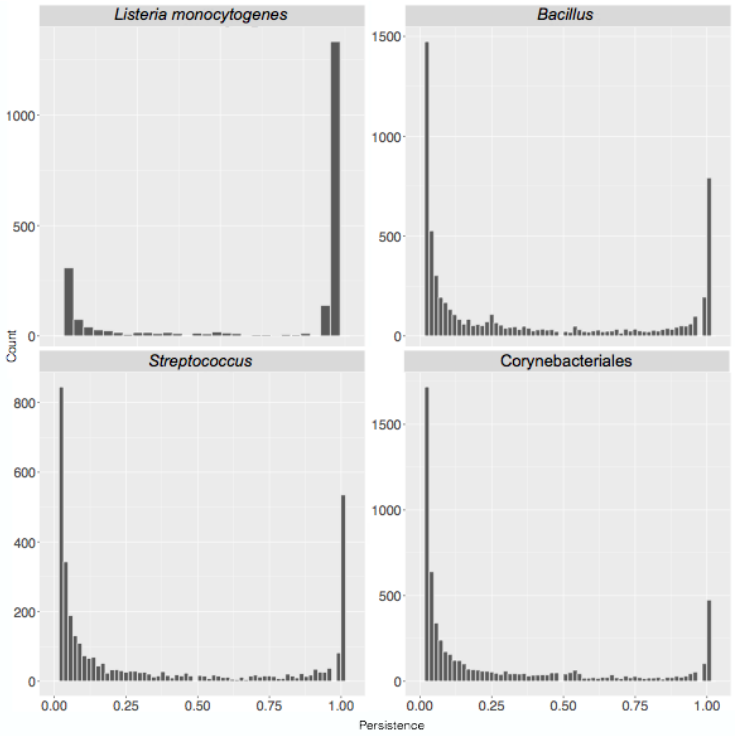

Figure 5.4: ***Persistence of domain architecture based (DAB) clusters.*** *The frequency of DAB clusters according to their persistence is shown*

Table 5.2: Number of identical clusters found with SB and DAB.

Group	Clusters
H. pylori	648
L. monocytogenes	1085
Bacillus	1439
Pseudomonas	1680
Streptococcus	961
Enterobacteriaceae	1649
Corynebacteriales	1034
Cyanobacteria	1127

when proteins in a single SB cluster are assigned to various DAB clusters including singletons and vice versa. An overview of the possible mismatches between SB and DAB clusters is in Figure 5.5. The observed frequency of the different types of cluster mismatches are given in Figure 5.6. We observed that single domain architectures predominated the one-to-one clusters as is shown in Table 5.3

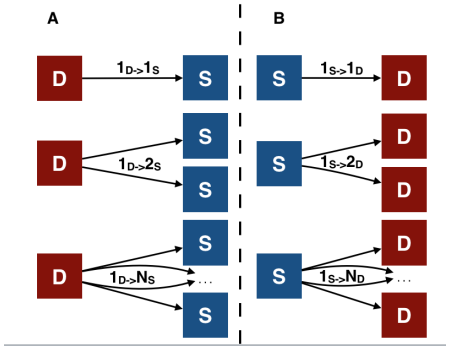

Figure 5.5: ***Summary of possible mismatches between DAB and SB clusters.*** *Mismatches of SB and DAB derived clusters (marked by S and D respectively) can occur in two directions. Panel A: possible cases of mismatch when counting the number of SB clusters the sequences in a DAB cluster are assigned to. $1d \rightarrow 1s$ denotes that all sequences from the D cluster are assigned to the same S cluster. $1d \rightarrow Ns$ denotes that sequences in a single D cluster are assigned to N distinct S clusters with $N \geq 1$. Similarly, (panel B) $1s \rightarrow Nd$ denotes that sequences in a single S cluster are assigned to N distinct D clusters with $N \geq 1$*

Table 5.3: Composition in terms of domains (#domains) of domain architecture found within identical (one-to-one) SB and DAB clusters

#Domains	*H. pylori*	*L. monocytogenes*	*Bacillus*	*Pseudomonas*	*Streptococcus*	Enterobacteriaceae	Cyanobacteria	Corynebacteriales
1	463	768	1119	1185	734	1312	867	772
2	133	207	229	333	164	246	182	192
3	40	76	65	107	43	64	57	45
4	8	23	18	37	13	15	14	16
5	3	9	3	10	5	6	4	5
6	0	2	2	5	1	3	3	4
7	1	0	1	3	1	3	0	0
8	0	0	1	0	0	0	0	0
9	0	0	1	0	0	0	0	0

For *L. monocytogenes* we found 378 $1d \rightarrow 1s$ DAB cluster mismatches, (Figure 5.5, panel A, top case) meaning that in those cases sequences in a DAB cluster are a subset of the sequences in the corresponding SB cluster. This lower number of sequences in the DAB cluster could be due to, for instance an insertion or expansion of a domain, leading to SB clustered sequences with partly overlapping but distinct domain architectures as is depicted in Figure 5.1. Similarly, there are 399 $1s \rightarrow 1d$ clusters. Each of these cases represent a sequence cluster where all the sequences share the same domain architecture, though other sequences exist with the same architecture that have not been included in the cluster due to a too low similarity score. The low similarity between sequences with the same domain architecture could be due to a horizontal acquisition of the gene or to a fast protein evolution at the sequence level. Genes acquired from high phylogenetic distances can greatly

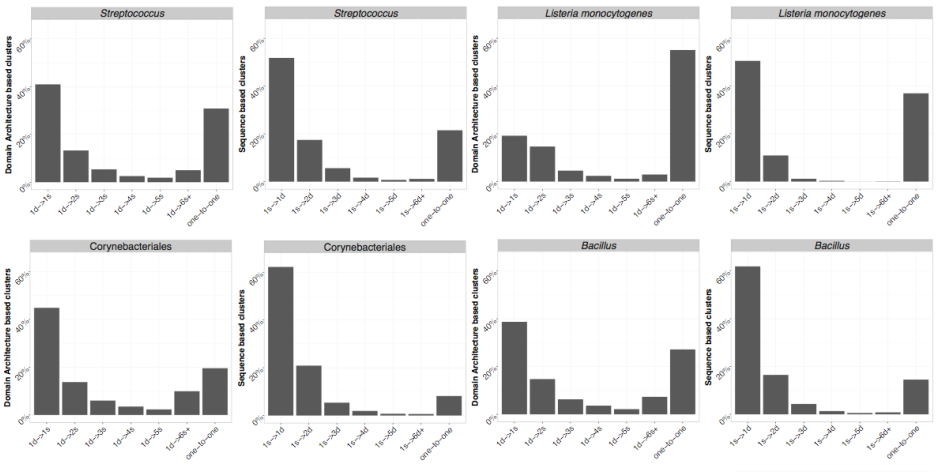

Figure 5.6: ***Comparison between DAB and SB clusters****. On the left DAB is used as a reference and each bar represents the relative frequency of one DAB cluster containing sequences assigned to* $\{1, 2, \ldots, 5\}$ *and 6 or more SB clusters and one-to-one represents the relative frequency of identical cluster. Similarly, on the right SB is used as a reference. Axis labels follow notation in Figure 5.5*

vary in sequence while presenting the same domain architecture.

Proteins contained in a single DAB cluster but assigned to multiple SB clusters contain mostly ABC transporters-like (PF00005) or Major Facilitator Superfamily (MFS, PF07690) domains. This is not surprising considering that such generic functions are usually associated with a high sequence diversity. Conversely, ABC transporters are found in multiple DAB clusters. However, many of them are grouped into a single SB cluster with ATPase domain containing proteins ($1s \rightarrow Nd$ case).

We observed distinct architectures with one of two very similar domains, the GDSL-like Lipase/Acylhydrolase and the GDSL-like Lipase/Acylhydrolase *family* domain (PF00657 and PF13472 respectively) and those architectures were often seen clustered using a SB approach. However, architectures containing both domains were also identified, pointing to a degree of functional difference as a result of convergent or divergent evolution. Still, the corresponding sequences remain similar enough as to be indistinguishable when a SB approach is used.

For SB clustering we also observed the case of identical protein sequences not clustered together, probably because of the tie breaking implementation when BBH are scored.

In all cases we found the size of both the pan- and the core-genome to be larger when a SB approach is used to identify gene clusters and SB approaches lead to a larger number of singletons than DAB ones. This indicates that DAB clusters are assigned to several SB clusters, many of them consisting of just one protein.

When going from species to phylum level, the ratio between the number of DAB and SB singletons changes from 0.48 and 0.41 (for *H. pylori* and *L. monocytogenes* respectively) to 0.19 and 0.40 when considering organisms of a higher taxonomic level (Corynebacteriales and Cyanobacteria respectively).

We investigated the predicted size of the pan-genome upon addition of new sequences. Heaps' law regression can be used to estimate whether the pan-genome is open or closed (Tettelin et al., 2005) through the fitting of the decay parameter α; $\alpha < 1$ indicates openness of the pan-genome (indicating that possibly many clusters remain to be identified within the considered set of sequences), while $\alpha > 1$ indicates a closed one; the α values are given in Table 5.4. In all cases the pan-genome is predicted to be open; however, α values obtained using DAB clusters (α_{DAB}) are systematically closer to one than the α_{SB} obtained with the standard sequence similarity approach.

The α_{DAB} value retrieved for *L. monocytogenes* is strikingly low. Heaps law regression relies on the selected genomes providing a uniform sampling of selected taxon, here species. Analysis of the domain content of the selected genomes shows a divergent behaviour of strain LA111 (genome id GCA_000382925-1). This behaviour is clear in Figure 5.7, where GCA_000382925-1 appears as an outlier of the *L. monocytogenes* group. Removal of this outlier leads to $\alpha_{DAB} = 1.04$ and $\alpha_{SB} = 0.64$, which emphasizes the need for uniform sampling prior to Heaps regression analysis.

DAB comparison across multiple taxa

DAB clusters can be labelled by their domain architecture and since this is a formal description of functional equivalence, results of independently obtained analyses can be combined. Figure 5.7 shows the results of a principal component analysis of the combined DAB clusters for selected genomes from eight taxa. The first

Table 5.4: Decay parameter α of the Heaps regression model using DAB and SB clustering

	α_{DAB}	α_{SB}
H. pylori	0.95	0.42
L. monocytogenes	0.77 (1.04*)	0.50 (0.64*)
Bacillus	0.93	0.59
Pseudomonas	0.94	0.61
Streptococcus	0.87	0.72
Enterobacteriaceae	0.99	0.74
Cyanobacteria	0.64	0.58
Corynebacteriales	0.88	0.52

$\alpha < 1$ indicates an open pan-genome.
*Values obtained upon removal of sequence GCA_000382925.1

two components account for a relatively low explained variance (29%) still grouping of genomes from the same taxa is apparent. High functional similarity among genomes of the same species (*H. pylori* and *L. monocytogenes*) is reflected by the compact clustering, while phylogenetically more distant genomes appear scattered in the functional space defined by the principal components.

Discussion

We have shown that domain architecture-based methods can be used as an effective approach to identify clusters of functionally equivalent proteins, leading to results similar to those obtained by classical methods based on sequence similarity.

To assess whether DAB results were consistent with those of SB methods, we have chosen OrthaGogue as a representative of the latter class. Several tools such as COGNITOR (David M. Kristensen et al., 2010) and MultiPARANOID (Alexeyenko et al., 2006) are available that implement different algorithm solutions to identify homologous sequences. However, despite different implementations, they all rely on sequence similarity as a proxy for functional equivalence. Here we considered SB methods as a golden standard for functional comparative genomics, especially when organisms within close evolutionary proximity are considered. Our aim was to investigate whether using HMMs instead of sequence similarity would yield similar results, thereby justifying their use for large scale functional genome comparisons. Regarding domain architectures, we have explored different alternatives, as we have

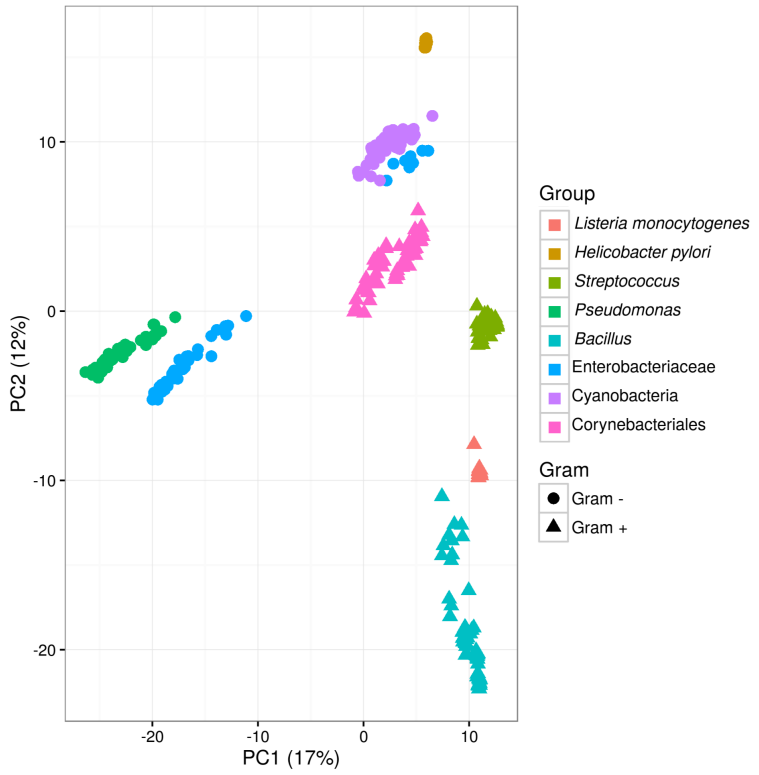

Figure 5.7: ***Large scale functional comparison of species.*** *Principal component analysis of functional similarities of 446 genomes based on the presence/absence of domain architectures on the corresponding genomes. The variance explained by the first two components is indicated on axes labels.*

seen that the chosen database or set of reference domains plays a critical role; for example, the low coverage of TIGRFAM prevents the obtaining of reasonable clusters.

The DAB approach takes advantage of the large computational effort that has already been devoted to the identification and definition of protein domains in dedicated databases such as Pfam. Protein domain models are build using large scale sequence comparisons which is an extremely computationally intensive task. However, once the domain models are defined, mining a sequence for domain occurrences is a much less demanding task. The task with the higher computational load (the definition of the domains) is performed only once and results can be stored and re-used for further analysis. This provides an effective scalable approach for large

scale functional comparisons which by and large is independent of phylogenetic distances between species.

The chosen set of domain models and the database used as a reference greatly impact the results. InterPro aggregates protein domain signatures from different databases, which leads to redundancy of the domain models. This redundancy causes overlaps between the entries and an increase of the granularity of the clusters retrieved. This can bias downwards the size of the pan-genome and upwards the size of the core- genome, as shown in Table 1. In InterPro this redundancy is taken into account by implementing a hierarchy of protein families and domains. The entries at the top of these hierarchies correspond to broad families or domains that share higher level structure and/or function. The entries at the bottom correspond to specific functional subfamilies or structural/functional sub-classes of domains (Mitchell et al., 2015). Using InterPro for DAB clustering would require taking into account the hierarchy of protein families and domains. However, this would pose challenges of its own and would require discrimination of the functional equivalence of different signatures within the same hierarchy.

Another source of redundancy are functionally equivalent domains from distantly related sequences. Pfam represents this through related families, termed clans, where relationships may be defined by similarity of sequence, structure or profile-HMM. Clans might contain functionally equivalent domains; however, it is not clear whether this is always the case as the criteria for clan definition include functional similarity but not functional equivalence (Robert D Finn et al., 2006).

Members of a clan have diverging sequences and very often SB approaches would recognise the evolutionary distance between the sequences and group them in different clusters. If we were to assume that members of a clan are functionally equivalent and collect them in the same DA cluster, we will have a higher number of cases where a single DA cluster is split in multiple sequence clusters 1d→Ns. In addition, there would be a higher number of cases of sequence clusters with the same DA though none exactly matching the DA clusters (1s→1d cases).

In many cases a one-to-one correspondence could be established between DAB and SB clusters indicating that often the sequence can be used as a proxy for function. At first this may seem a trivial result, however it has a profound implication.

Domain model databases (in this case Pfam) contain enough information, encoded by known domain models, to represent the quasi totality of biological function encoded in the bacterial sequences analysed in this study. However, it is important to stress that the comparisons have been performed considering sequences with known domains, representing currently around 85% of the genome coding content, a number that will only increase in the future.

A significant advantage of the DAB method over the SB method is that the domain architecture captured within a cluster can be used as a formal description of the function. Currently, more than 20% of all separable domains in the Pfam database, are so-called domains of unknown function (DUF). Although in bacterial species they are often essential (Goodacre, Gerloff, and Uetz, 2013). With the DAB method they are formally included and often semantically linked to one or more domains of known function.

The starting position of the domains was used to generate labels indicating N-C terminal order of identified domains. The labels were used only for clustering as proteins sharing the same labels were assigned to the same clusters. When using the mid-point or the C-terminal position, labelling could be affected however, it does not affect the obtained clusters.

A content-wise formal labelling of DAB clusters makes a seamless integration of multiple independently performed DAB analysis possible. This allows for a comparison of potential functionomes across taxonomic boundaries, as presented in Figure 5.7, while new genomes can be added at a computational cost $O(n)$, with n the number of genomes to be analysed. On the other hand, addition of a new genome using an SB approach requires a new set of all-against-all sequence comparisons which comes at a $O(n^2)$ computational cost. However, approaches have been proposed to overcome these shortcomings of SB methods, such as COGNITOR which reduces the computational to $O(n)$ by using pre-computed databases. In this respect, the DAB approach is similar to the approach implemented in COGNITOR, by searching against existing databases of domain architectures.

The bimodal shape of the distributions presented in Figures 5.3 and 5.4 indicates the relative role of horizontal gene transfer and vertical descent when shaping bacterial genomes. The first peak accounts for sequences (or functions) only present

in a small number of genome sequences which have been likely acquired by horizontal gene transfer. The second peak accounts for high persistence genetic regions representing genes (or functions) belonging to the taxon core which has been likely acquired by vertical descent.

A measure of the impact of vertical descent and horizontal gene transfer is provided by the ratio between the core- and pan- genome sizes. The number of singletons provides a measure of the number of genes horizontally acquired from species outside the considered group.

Two of the most prominent differences between the two approaches are the number of retrieved singletons and the core- to pan- genome size ratio. Multiple members of the same taxon might acquire the same function through horizontal gene transfer (Soucy, Huang, and Johann Peter Gogarten, 2015). This is likely to occur given that they would have similar physiological characteristics, hence they would tend to occupy a similar niche or, at least, more similar than when comparing species from different taxa. As the origin of the horizontally acquired genes may vary for each organism, an SB approach will correctly recognise the heterologous origin of the corresponding sequences and those will be assigned to singletons. However, the probabilistic hidden Markov models used for domain recognition are better at recognising the functional similarity of the considered sequences and clusters them together.

Another indication of the relative impact of horizontal and vertical gene acquisition events is provided by the openness or closedness of the genome. Values for the decay parameter α in Table 5.3 indicate a relatively large impact by horizontal gene transfer. Within the considered taxa we observed $\alpha_{DAB} > \alpha_{SB}$, meaning that the sequence diversity is larger than the functional diversity. Upon addition of new genomes to the sample the rate of addition of new sequence clusters appears higher than the rate of the addition of new functions.

Limitations of DAB approaches

We have shown that domain architecture-based methods can be used as an effective approach to identify clusters of functionally equivalent proteins, leading to results similar to those obtained by classical methods based on sequence similarity. How-

ever, whether DAB methods are more accurate than SB methods to assess functional equivalence will require further analysis. In this light, results of functional conservation for both approaches could be compared in terms of GO similarity and/or EC number (Altenhoff and Dessimoz, 2009; David M Kristensen, 2016). Partial domain hits might arise as a result of alignment, annotation and sequence assembly artefacts. To reduce the number of partial domain hits additional pruning could be implemented to distinguish these cases. However, this is an open problem that requires caution as it could influence the functional capacity of an organism and clustering approaches using DA.

The performance of DAB methods may be sub-optimal when dealing with newly sequenced genomes that are not yet well-characterised enough to have all of their domains present in domain databases, since DAB methods will be unable to handle unknown architectural types. Around 15% of the genome coding content corresponds to sequences with no identified protein domains. DAB approaches can be complemented with SB methods to consider these sequences or even protein sequences with low domain coverage, possible indicating the location of protein domains yet to be identified. Since DAB methods rely on the constant upgrading of public resources like UniProt and Pfam databases, an initial assessment of domain coverage appears as a sine qua non condition for application of these methods. DAB approaches could be used to assess the consistency of existing orthologous groups in terms of their domain architectures, at least when domain architectures are expected to be completely known in advance (for instance in the case of micro-evolutionary variations within a species where mutational events may disrupt a protein's function). For other purposes, such as the discovery of a new phyla of cellular life that contains radically different domain architectures, global similarity methods may be preferred (David M Kristensen, 2016).

Conclusions

As protein domain databases have evolved to the point where DAB and SB approaches produce similar results in closely related organisms, the DAB approach provides a fast and efficient alternative to SB methods to identify groups of func-

tionally equivalent/related proteins for comparative genome analysis. The lower computational cost of DAB approaches makes them the better choice for large scale comparisons involving hundreds of genomes.

Highly redundant databases, such as InterPro, are best suited for domain based protein annotation, though are as not effective for DAB clustering if the goal is to identify clusters of functionally equivalent proteins. To enable DAB approaches for highly structured databases, such as InterPro, the hierarchy of protein families and domains within has to be explicitly considered. Currently Pfam is a better alternative for this task.

Differences between DAB and SB approaches increase when the goal is to study bacterial groups spanning wider evolutionary distances. The functional pan-genome is more closed in comparison to the sequence based pan-genome. Both methods have a distinct approach towards horizontally transferred genes, and the DAB approach has the potential to detect functional equivalence even when sequence similarities are low.

Complementing the standardly applied sequence similarity methods with a DAB approach pinpoints potential functional protein adaptations that may add to the overall fitness.

Author contributions

JJK, MSD, ES, PJS participated in the set-up of the research. JJK and MSD were responsible for the analysis. JJK, ES, PJS, MSD and VdS wrote the manuscript. All authors critically revised the manuscript.

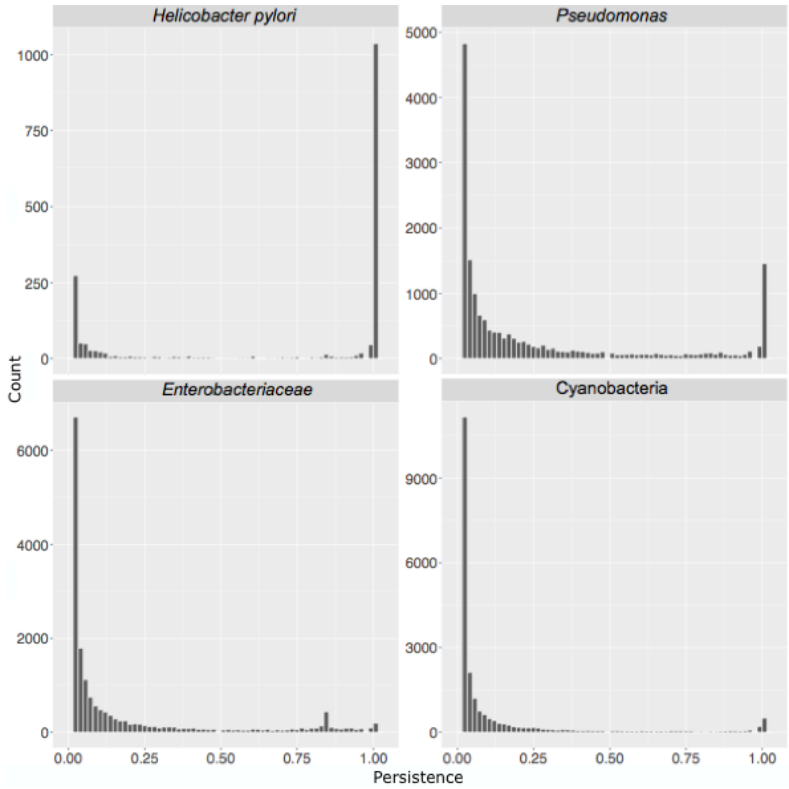

Figure 5.8: ***Persistence of Sequence Based (SB) clusters****. Cluster persistence is defined as the relative number of genomes with at least one protein assigned to the cluster. The plots show frequency of SB clusters according to their persistence. Publicly available and complete genome sequences assigned to each taxon were selected so that phylogenetic diversity within the taxon was preserved, as described in materials and methods. 60 distinct genome sequences were considered for each of the depicted taxa.*

Acknowledgements

This work was partly supported by the European Union's Horizon 2020 research and innovation programme (EmPowerPutida, Contract No. 635536, granted to Vitor A P Martins dos Santos).

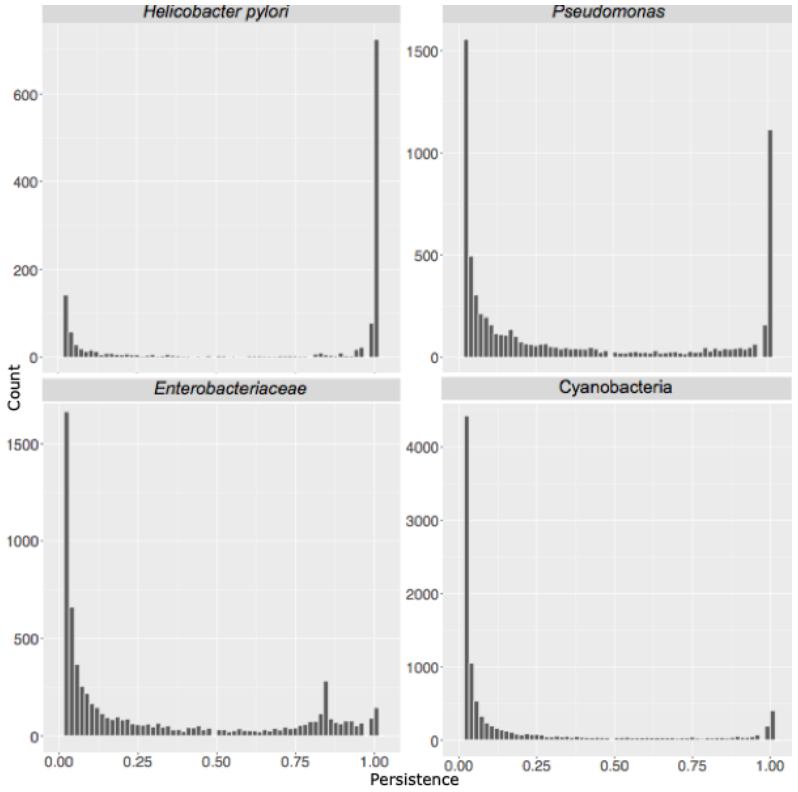

Figure 5.9: ***Persistence of Domain Architecture Based (DAB) clusters****. The plots show the frequency of DAB clusters according to their persistence.*

Bibliography

Addou, Sarah, Robert Rentzsch, David Lee, and Christine A. Orengo (2009). "Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer". In: *Journal of Molecular Biology* 387. DOI: 10.1016/j.jmb.2008.12.045.

Alexeyenko, Andrey, Ivica Tamas, Gang Liu, and Erik L L Sonnhammer (2006). "Automatic clustering of orthologs and inparalogs shared by multiple proteomes". In: *Bioinformatics*. Vol. 22. DOI: 10.1093/bioinformatics/btl213.

Altenhoff, Adrian M and Christophe Dessimoz (2009). "Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods". In: *PLoS Comput Biol* 5. DOI: 10.1371/journal.pcbi.1000262.

Björklund, Åsa K, Diana Ekman, Sara Light, Johannes Frey-Skött, and Arne Elofsson (2005). "Domain rearrangements in protein evolution". In: *Journal of molecular biology* 353.

Boratyn, Grzegorz M et al. (2012). "Domain enhanced lookup time accelerated BLAST". In: *Biology Direct* 7. DOI: 10.1186/1745-6150-7-12.

Doğan, Tunca et al. (2016). "UniProt-DAAC: Domain Architecture Alignment and Classification, a New Method for Automatic Functional Annotation in UniProtKB". In: *Bioinformatics*.

Dongen, Stijn van (2000). "Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht". In:

Dutilh, Bas E et al. (2013). "Explaining microbial phenotypes on a genomic scale: GWAS for microbes". In: *Briefings in Functional Genomics* 12. DOI: 10.1093/bfgp/elt008.

Eddy, Sean R (1998). "Profile hidden Markov models." In: *Bioinformatics* 14.

Ekseth, Ole Kristian, Martin Kuiper, and Vladimir Mironov (2013). "OrthAgogue: an agile tool for the rapid prediction of orthology relations". In: *Bioinformatics* 30. DOI: 10.1093/bioinformatics/btt582.

Finn, Robert D et al. (2006). "Pfam: clans, web tools and services". In: *Nucleic Acids Research* 34. DOI: 10.1093/nar/gkj149.

Finn, Robert D. et al. (2016). “The Pfam protein families database: Towards a more sustainable future”. In: *Nucleic Acids Research* 44. DOI: 10.1093/nar/gkv1344.

Fong, Jessica H, Lewis Y Geer, Anna R Panchenko, and Stephen H Bryant (2007). “Modeling the Evolution of Protein Domain Architectures Using Maximum Parsimony”. In: *Journal of molecular biology* 366. DOI: 10.1016/j.jmb.2006.11.017.

Geer, Lewis Y, Michael Domrachev, David J Lipman, and Stephen H Bryant (2002). “CDART: protein homology by domain architecture”. In: *Genome Research* 12. DOI: 10.1101/gr.278202.

Gogarten, J Peter, W Ford Doolittle, and Jeffrey G Lawrence (2002). “Prokaryotic evolution in light of gene transfer”. In: *Molecular Biology and Evolution* 19.

Goodacre, Norman F, Dietlind L Gerloff, and Peter Uetz (2013). “Protein domains of unknown function are essential in bacteria.” In: *MBio* 5. DOI: 10.1128/mBio.00744-13.

Goodwin, Sara, John D. McPherson, and W. Richard McCombie (2016). “Coming of age: ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* 17. DOI: 10.1038/nrg.2016.49.

Haft, Daniel H., Jeremy D. Selengut, and Owen White (2003). *The TIGRFAMs database of protein families*. DOI: 10.1093/nar/gkg128.

Hyatt, Doug et al. (2010). “Prodigal: prokaryotic gene recognition and translation initiation site identification”. In: *BMC Bioinformatics* 11. DOI: 10.1186/1471-2105-11-119.

Jones, Philip et al. (2014). “InterProScan 5: Genome-scale protein function classification”. In: *Bioinformatics* 30. DOI: 10.1093/bioinformatics/btu031.

Joshi, Trupti and Dong Xu (2007). “Quantitative assessment of relationship between sequence similarity and function similarity”. In: *BMC genomics* 8.

Koehorst, Jasper J. et al. (2017). “SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles”. In: *Bioinformatics* 1.

Koonin, Eugene V, Yuri I Wolf, and Georgy P Karev (2002). “The structure of the protein universe and genome evolution”. In: *Nature* 420. DOI: 10.1038/nature01256.

Kristensen, David M (2016). “Referee Report For: Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics [version 1; referees: 1 approved, 2 approved with

reservations]". In: *F1000Research* 5:1987. DOI: 10.5256/f1000research.10140.r15678.

Kristensen, David M. et al. (2010). "A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches". In: *Bioinformatics* 26. DOI: 10.1093/bioinformatics/btq229.

Kuipers, Remko K P et al. (2009). "Correlated mutation analyses on super-family alignments reveal functionally important residues". In: *Proteins: Structure, Function, and Bioinformatics* 76.

Kummerfeld, Sarah K and Sarah A Teichmann (2009). "Protein domain organisation: adding order". In: *BMC Bioinformatics* 10. DOI: 10.1186/1471-2105-10-39.

Lee, Byungwook and Doheon Lee (2009). "Protein comparison at the domain architecture level". In: *BMC Bioinformatics* 10. DOI: 10.1186/1471-2105-10-S15-S5.

Letunic, Ivica, Tobias Doerks, and Peer Bork (2015). "SMART: recent updates, new developments and status in 2015". In: *Nucleic acids research* 43.

Messih, Mario Abdel, Meghana Chitale, Vladimir B Bajic, Daisuke Kihara, and Xin Gao (2012). "Protein domain recurrence and order can enhance prediction of protein functions". In: *Bioinformatics* 28. DOI: 10.1093/bioinformatics/bts398.

Mitchell, Alex et al. (2015). "The InterPro protein families database: The classification resource after 15 years". In: *Nucleic Acids Research* 43. DOI: 10.1093/nar/gku1243.

Pallen, Mark J and Brendan W Wren (2007). "Bacterial pathogenomics". In: *Nature* 449. DOI: 10.1038/nature06248.

Ponting, Chris P and Robert R Russell (2002). "The Natural History of Protein Domains". In: *Annual Review of Biophysics and Biomolecular Structure* 31. DOI: 10.1146/annurev.biophys.31.082901.134314.

Puigbò, Pere, Alexander E Lobkovsky, David M Kristensen, Yuri I Wolf, and Eugene V Koonin (2014). "Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes". In: *BMC biology* 12.

Saccenti, Edoardo, David Nieuwenhuijse, Jasper J Koehorst, Vitor AP Martins dos Santos, and Peter J Schaap (2015). "Assessing the metabolic diversity of streptococcus from a protein domain point of view". In: *PloS one* 10.

Sigrist, Christian J A et al. (2012). “New and continuing developments at PROSITE”. In: *Nucleic acids research*.

Snipen, Lars and Kristian Hovde Liland (2015). “micropan: an R-package for microbial pan-genomics”. In: *BMC Bioinformatics* 16. DOI: 10.1186/s12859-015-0517-0.

Snipen, Lars-Gustav and David W Ussery (2013). “A domain sequence approach to pangenomics: applications to Escherichia coli”. In: *F1000Research* 1. DOI: 10.12688/f1000research.1-19.v2.

Song, N, R D Sedgewick, and D Durand (2007). “Domain architecture comparison for multidomain homology identification”. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 14. DOI: 10.1089/cmb.2007.A009.

Soucy, Shannon M, Jinling Huang, and Johann Peter Gogarten (2015). “Horizontal gene transfer: building the web of life”. In: *Nature Reviews Genetics* 16.

Tettelin, HervÃ© et al. (2005). “Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102. DOI: 10.1073/pnas.0506758102.

Thakur, Shalabh and David S Guttman (2016). “A De-Novo Genome Analysis Pipeline (DeNoGAP) for large-scale comparative prokaryotic genomics studies”. In: *BMC bioinformatics* 17.

The UniProt Consortium (2015). “UniProt: a hub for protein information.” In: *Nucleic acids research* 43. DOI: 10.1093/nar/gku989.

Van Domselaar, Gary H et al. (2005). “BASys: a web server for automated bacterial genome annotation.” In: *Nucleic acids research* 33. DOI: 10.1093/nar/gki593.

Yang, Song, Russell F Doolittle, and Philip E Bourne (2005). “Phylogeny determined by protein domain content”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102. DOI: 10.1073/pnas.0408810102.

6

Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data

Jasper J. Koehorst, Jesse C. J. van Dam, Ruben G. A. van Heck, Edoardo Saccenti, Vitor A. P. Martins Dos Santos, Maria Suarez-Diez, Peter J. Schaap

Bioinformatics; doi: https://doi.org/10.1093/bioinformatics/btx767

Abstract

Pseudomonas is a highly versatile genus containing species that can be harmful to humans and plants while others are widely used for bioengineering and bioremediation.

We analysed 432 sequenced *Pseudomonas* strains by integrating results from a large scale functional comparison using protein domains with data from six metabolic models, nearly a thousand transcriptome measurements and four large scale transposon mutagenesis experiments.

Through heterogeneous data integration we linked gene essentiality, persistence and expression variability. The pan-genome of *Pseudomonas* is closed indicating a limited role of horizontal gene transfer in the evolutionary history of this genus. A large fraction of essential genes are highly persistent, still non essential genes represent a considerable fraction of the core-genome.

Our results emphasize the power of integrating large scale comparative functional genomics with heterogeneous data for exploring bacterial diversity and versatility.

Introduction

The *Pseudomonas* genus exhibits a broad spectrum of traits and *Pseudomonas* species show a remarkable adaptability to the biochemical nature of the large variety of environments, often extreme, they thrive in (Wu et al., 2011; Timmis, 2002). The genus currently includes almost 200 recognised species, which have been clustered into seven groups and into lineages on the basis of a limited set of loci (Loper et al., 2012). Some species are well-studied because they are human or plant pathogens, like *P. aeruginosa* or *P. syringae*, or because they are considered harmless and possess interesting biodegradation properties while others can produce a variety of extraordinary secondary metabolites with anti-microbial properties (Gross and Loper, 2009). *P. putida* KT2440 is even Generally Recognized as Safe (GRAS-certified) for expression of heterologous genes and has been transformed into a genetically accessible laboratory and industrial workhorse (Nelson et al., 2002).

A number of comparative genomics studies have been performed in the past (Wu et al., 2011; Loper et al., 2012; Baltrus et al., 2011) but the number of available *Pseudomonas* genomes quadrupled in the last five years due to the widespread use and the advancement of high-throughput sequencing technologies. As of December 2015, the complete and draft genomes of 432 strains distributed over 33 species are publicly available (see Supplementary Figure S1). This plethora of data entitles an in-depth comparative re-analysis of *Pseudomonas* genomes to explore their metabolic and ecological diversity.

Large scale functional comparison based on sequence similarity is challenged by methodological problems, such as the need of of defining arbitrarily generalized minimal alignment length and similarity cut-off for all sequence to be analyzed, and it is hampered by the high computational cost, since time and memory requirements scale quadratically with the number of genome sequences to be compared (Koehorst et al., 2016). Many bacterial proteins consist of two or more domains and fusion/-fission events are the major drivers of modular evolution of multi-domain bacterial proteins (Pasek, Risler, and Brezellec, 2006). Interspecies domain variation can thus give rise to an annotation transfer problem: sequence based functional annotation methods use a consecutive alignment to identify common ancestry and therefore

may miss domain insertion/deletion, exchange or repetition events, which may lead to functional shifts and promiscuity. Comparisons at protein sequence level should therefore be complemented with comparisons at the protein domain level (Koehorst et al., 2016). In addition, in order to avoid technical biasses a biologically meaningful functional comparison requires consistent and up-to-date annotations. Instead, the biological information available in public databases varies in quality due to the use of different databases and annotation pipelines that include different methods and may assign different names, acronyms and aliases to the same protein. Re-interpretation of these predictions in most cases requires reverse engineering as data provenance is usually not available.

In this paper 432 Pseudomonas genome sequences were *de novo* re-annotated and the generated annotation information was integrated through a semantic platform with data from six metabolic models, nearly a thousand transcriptome measurements and four large scale transposon mutagenesis experiments. We identified phylogenetic relationships among different species using protein domains and performed extensive analysis of the core- and pan-genomes of the *Pseudomonas* genus and considered the habitat factor while analyzing the pan/core-genome. Finally, we linked domain content and domain variability of persistent and essential genes and their transcriptional regulation.

Results

De novo annotation of *P. putida* KT2440 as a minimal working example

P. putida KT2440 (Nelson et al., 2002) is one of the best-characterized *Pseudomonas* strains. A *de novo* annotation obtained using an in-house annotation pipeline, the annotation deposited in GenBank (NC_002947) and an alternative annotation obtained using RAST (Aziz et al., 2008) were compared, see Table 6.1.

The total number of genes identified using three gene calling methods, Prodigal 2.6 (in our pipeline), Glimmer3 (RAST), and Glimmer (GenBank) are very similar, differing less than 4%. However, as each of these algorithms have an intrinsic false discovery rate in start-site prediction, significant differences in the start position of the identified genes were found. The number of exact matches in gene start-sites is

Table 6.1: Annotation results for P. putida KT2440. GenBank refers to the original deposited annotation (available at NCBI), whereas RAST and SAPP refer respectively to their annotation.

	#Genes	#Unique start/end positions	#Unique GO	Unique domains	Unique EC
GenBank	5350	170	0	3574	443
RAST	5531	62	726	3631	447
SAPP	5555	252	1403	3636	447

only 73% (4073 genes) confirming previous observations (Tripp et al., 2015). These 5' variations in gene identification can result in a putative gain or loss of biological functions; however, since different naming conventions are used in the different annotation protocols applied, a direct functional comparison to spot possible differences is not possible (Figure 6.1).

The use of controlled vocabularies overcomes this issue, so that functional comparison can be performed using gene ontology (GO) terms, Enzyme Commission (EC) numbers and InterPro identifiers. For the GenBank deposited annotation no GO information was available but the difference observed between the RAST and the *de novo* annotation is striking. This minimal working example shows that even for a single genome a comparative analysis of functional annotations derived from three work-flows is almost impossible by computational means due to lack of standardization and data provenance. This example further emphasizes that comparative genomic analysis requires homogeneous annotation.

Comparison of the genomic potential of *Pseudomonas* species

Since for a comparative genomics study a consistent and standardized genome annotation is a prerequisite, we evaluated the impact by comparing the functional annotations of 432 *Pseudomonas* genomes with a *de novo* annotation. We used both complete and draft genomes. According to the quality metric defined by Cook and Ussery, almost 30% of the available draft genomes were of low quality (Cook and Ussery, 2013). This was mostly due to a high number of contigs and not to the quality of the assemblies in itself, so they were included in the analysis.

GenBank files were converted into RDF, extracting genome sequences and gene-calls. Genomes were structurally and functionally re-annotated. The originally de-

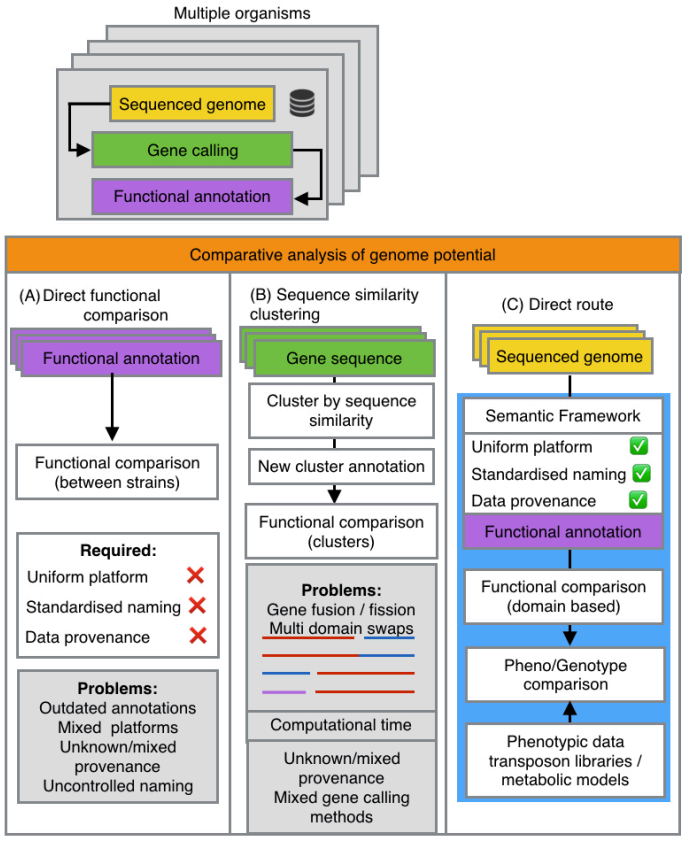

Figure 6.1: ***Alternatives for functional genome comparison****: **A) Direct comparison of genome potential using existing annotation** is often hampered by lack of standardization of gene calling and annotation tools, mixed and unknown data provenance and inconsistent naming of function. **B) Sequence similarity clustering** bypasses inconsistent functional annotations. Computational time scales quadratically with the number of genome sequences and gene fusion/fission events might be overlooked. **C) Usage of standardised annotation tools** ensures uniform genome annotation prior to comparison; annotation provenance is stored for all steps.*

posited gene-calls were functionally re-annotated as well and a pairwise comparison of GO terms, and EC identifiers assigned to the originally deposited and the *de novo* gene-calls was performed at gene and protein domain level. Figure 6.2 summarizes the results for the available 58 complete genomes. Differences in annotations were observed at all functional levels. Per genome on average 38 new genes were predicted while a functional re-annotation of the set of complete genomes yielded 838 additional GO-terms and 146 additional domains (For a more detailed overview see Supplementary Data S2). Considering the full set of 432 genomes, on average a

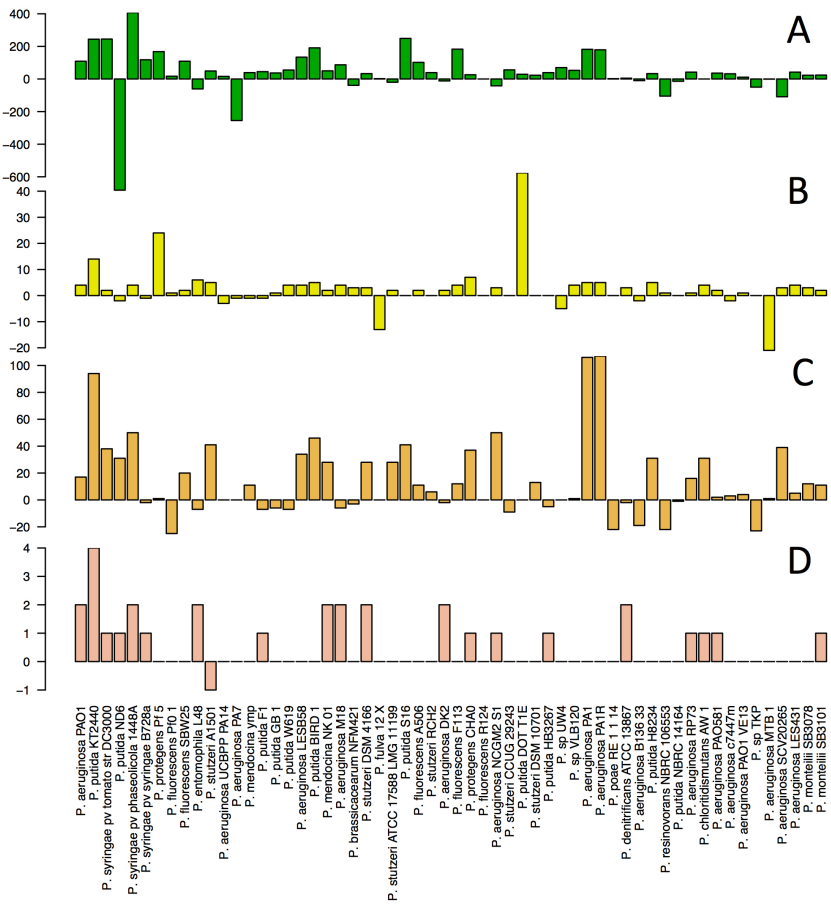

Figure 6.2: ***De novo annotation of Pseudomonas genomes.*** *Comparison between the original and de novo annotations of 58 completely sequenced Pseudomonas genome sequences. Barplots indicate differences in the number of retrieved genome features terms between the de novo annotations and the original deposited annotations. A) gene abundance; B) protein domains; C: GO terms, and D: EC identifiers. The genomes are ordered from left to right by deposition date in the NCBI database (from oldest to newest).*

difference of 153 genes per genome was detected. The results advocate for routine implementation of consistent gene-calling methods combined with an up-to-date functional annotation before performing comparative genomic analyses, as many of these differences will results in gain or loss of biological functions.

Sequence and function based comparative genomics of *Pseudomonas*

Genome-wide comparative analysis usually relies on sequence similarity clustering based on a blast-based all-against-all bidirectional best hit (BBH) heuristic approach. There are several limitations to this approach. Firstly, the runtime increases quadratically with the number and complexity of the species involved. Secondly, clustering is strongly context-dependent as it dramatically depends on chosen cut-off values to define statistical significance of sequence similarity. Problems may arise with in-paralogous sequences that evolve at very similar rates resulting from recent duplication events (Notebaart et al., 2005). Thirdly, protein fusion and fission events are difficult to detect using alignments and thus critical information might be lost.

An alternative approach, already employed in a comparative genomics study of *Escherichia coli* (L.-G. Snipen and Ussery, 2012), consists of grouping of proteins on the base of domain architectures with a fixed N-C terminal order (B. Lee and D. Lee, 2009). Clustering based on domain order is highly scalable and moreover, most protein domains represent structural folds that can be directly linked to function. Here, both approaches were compared. Protein sequence similarity clusters were identified in a BBH approach using orthAgogue (Ekseth, Kuiper, and Mironov, 2013). Due to runtime constraints, protein clustering was limited to the analysis of the 58 complete genomes leading to the identification of 14757 protein clusters. For each protein found within a cluster the domain content and N-C terminal domain order ranked by the position of the first detected amino acid of the domain (domain start) in the protein sequence (domain architectures) was analysed and is summarized in Figure 6.3A. 5515 sequence based protein clusters (37%) present a one-to-one correspondence to domain architectures, whereas 3134 (21%) can be associated to two distinct domain architectures. Overall, 93% of the identified clusters can be associated to 4 or less distinct domain architectures. Figure 6.3A also shows the number of proteins in each orthologous cluster. 3162 clusters (21%) contain proteins lacking established domains and almost 75% of them contain less than 10 sequences. These clusters correspond, in their vast majority, to hypothetical proteins. Regarding the core genome, 1618 clusters (11%) were found to be present in all 58 genomes. From these 1618 protein clusters, 242 contained duplication events

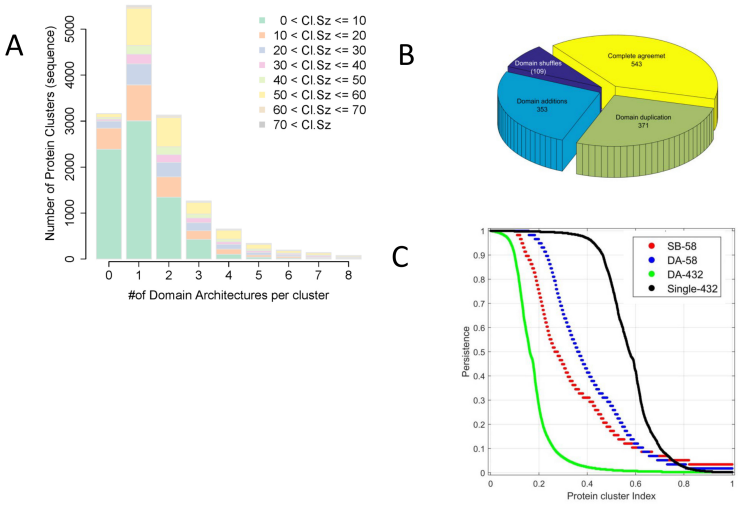

Figure 6.3: ***Domain architectures in sequence based clusters of orthologous proteins*** *A) Number of distinct domain architectures per cluster B) Variability in domain architectures per gene cluster in core-genome. Complete agreement indicates a unique domain architecture shared by all members of the cluster; For the cases where multiple domain architectures were found in a sequence cluster, the number of cases corresponding to domain duplications, additions and shuffles are indicated. (For A and B only 58 complete genome sequences considered). C) Persistence analysis within the Pseudomonas genus. The curves indicate the persistence of each of the cluster. Clusters have been arranged by decreasing persistence values and the x-axis has been scaled to 0-1 range, in this way the cluster with the highest persistence have an x value of 0 and the cluster with the lowest persistence has an x value of 1. The y-axis indicates the persistence of a given cluster (see Equation 1): for instance a persistence of 0.8 indicates that 80% of the analyzed genomes contain sequences in that given cluster. SB-58 refers to the use of sequence based cluster considering the 58 complete genomes; DA-58 and DA-432 refers to the use of protein domains, for 58 and 432 genomes respectively; Single-432 reproduces the analysis for single domain proteins found in the full set 432 genome sequences.*

leaving 1376 distinct single copy gene protein clusters common to all 58 genomes. 543 of those clusters showed a single domain architecture whereas the rest contained domain architecture variations as summarized in Figure 6.3B. We noted that such variability was mainly due to swapping or inversion in domains order. In a sequence based approach domain order variation can potentially lead to false negatives, broken clusters and even reduction of the core genome when more genomes are added to the analysis.

The analysis of 58 complete genome sequences showed that domain architectures retain enough information for functional characterization and that they can be used

as a fingerprint for a functional cluster. Since the computational cost for obtaining protein domain identification scales linearly with number of genomes and can be easily distributed over multiple machines, we used these functional fingerprints to extend the analysis to all 432 *Pseudomonas* genomes. Over two million (2,704,339) genes were identified coding for over one million (1,196,884) unique protein sequences of which 85.6% (1,024,877) contain known protein domains. Figure 6.3C shows the results of persistence analysis, reporting the fraction of the total number of analysed genomes in which the corresponding cluster/protein domain/domain architecture was found; 40% the protein domains are persistent in the genus, showing that the functional information at domain level is preserved.

Classification of *Pseudomonas* strains based on genome potential

Patterns of protein domain presence/absence can provide an alternative and complementary way for assessing strain diversity (S. Yang, Doolittle, and Bourne, 2005; Alako et al., 2006). There are still many unclassified *Pseudomonas* strains and there is a continuous development on assessing the phylogeny using various approaches (Bertels et al., 2014). Figure 6.4 shows a distance tree of genome potential based on presence/absence of protein domains for the 58 complete *Pseudomonas* genomes. We found excellent agreement between this distance tree and the taxonomic classification based on 16S sequences indicating that binary patterns of protein domains retain enough information to reconstruct evolutionary history. The positioning of *Pseudomonas* sp. UW4 within the clade of *P. fluorescence*, confirms a previous observation based on 16S and three housekeeping genes (*gyrB*, *rpoB* and *rpoD*) (Duan et al., 2013). *P. aeruginosa* and *P. stutzeri* clades are conserved while *P. putida* and *P. fluorescence* clades shows the addition of different species.

We further extended the domain based distance analysis to include all 432 *Pseudomonas* strains (see Supplementary Figure S3). The majority of the strains cluster in accord with their taxonomic classification. Many of the unclassified strains could be classified either in *P.aeruginosa* (4) or *P. putida* (13).

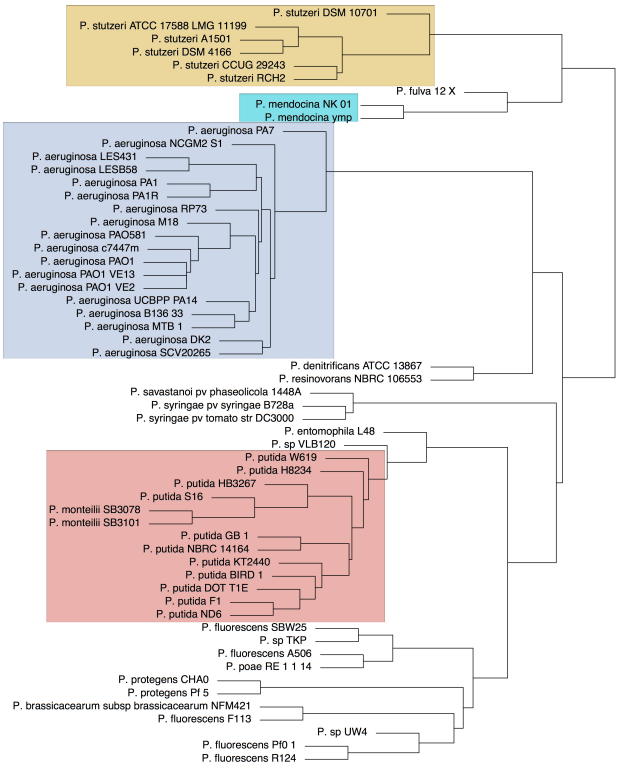

Figure 6.4: ***Domain based distance tree of 58 Pseudomonas strains*** *The tree was build considering the pattern of presence/absence of protein domains using an average clustering approach. Only completely sequenced genomes are considered. The phylogenetic clusters corresponding to the most abundant species (P. stutzeri, P. mendocina, P. aeruginosa and P.putida) are colour-shadowed.*

Exploring the pan- and core-genome of *Pseudomonas* at protein domain level

The core-genome of a taxon level is defined as the genes persistently present in the population, while the pan-genome is essentially the amount of different genes found within a population at the specified taxonomic level (L. Snipen, Almøy, and Ussery, 2009). The currently available genomes allow to measure the pan- and core-genome sizes, however these sizes change upon the addition of new sequences. The core-genome is usually reduced and the pan-genome increases mostly due to the discovery of novel accessory genes that accumulate by lateral transfer, forming new trait combinations until saturation has been reached. Saturated pan-genomes with a stable core-genome are called closed. From the currently available genomes an esti-

mation can be made, using mathematical modelling (L. Snipen, Almøy, and Ussery, 2009), of the size of the pan- and core- genomes that are expected if the sequences of every existing strain were to be included in the analysis. We refer to these estimations as estimated pan- and core- genome sizes.

Genome potential of the genus *Pseudomonas* is reflected in its metabolic diversity which allows individual species to inhabit a wide variety of environments. With the current set of 432 (draft) genomes we studied whether the observed diversity in genome potential reflects a closed pan-genome. We initially considered the 58 complete genomes. Observed core-genome of 2687 protein domains was to be confronted with an estimated size of 2681. For the pan-genome we found 6472 protein domains (observed) versus 6541 (estimated). Since these measures depend on the number of genomes considered, we explored how these measures vary by using a different number of genomes (from 5 to 58). This was achieved by applying a 10-fold random re-sampling from the 58 genomes to obtain an indication of the possible variability (Figure 6.5). As expected the size of the core-genome of the genus decreases with the number of genomes considered while that of the pan-genome increases. The observed and estimated sizes of both the pan- and core-genome are rather stable with respect to the number of genomes used in the calculation, except for small sample size (< 15).

Including draft genomes in the calculations resulted in a dramatic reduction, up to the 73%, of the size of the core-genome both observed and estimated, which dropped to 726 and 720 protein domains architectures, respectively. Interestingly, this reduction does not lead to a loss of functional information since single domains are highly persistent as previously stated (40%).

We observed a large variability for both measures. The reduction of the core size and its variability can be partly explained due to the inclusion of draft genomes with a high number of gaps containing non-sequenced genes. The difference between observed and estimated sizes reduced to only one protein domain for both the pan- and core-genome, indicating saturation. Addition of new genome sequences to the analysis will most likely not lead to the identification of a significant set of new domains within this genus. This saturation effect does not depend on the particular estimation model used. Saturation of the pan-genome was also seen through a heap

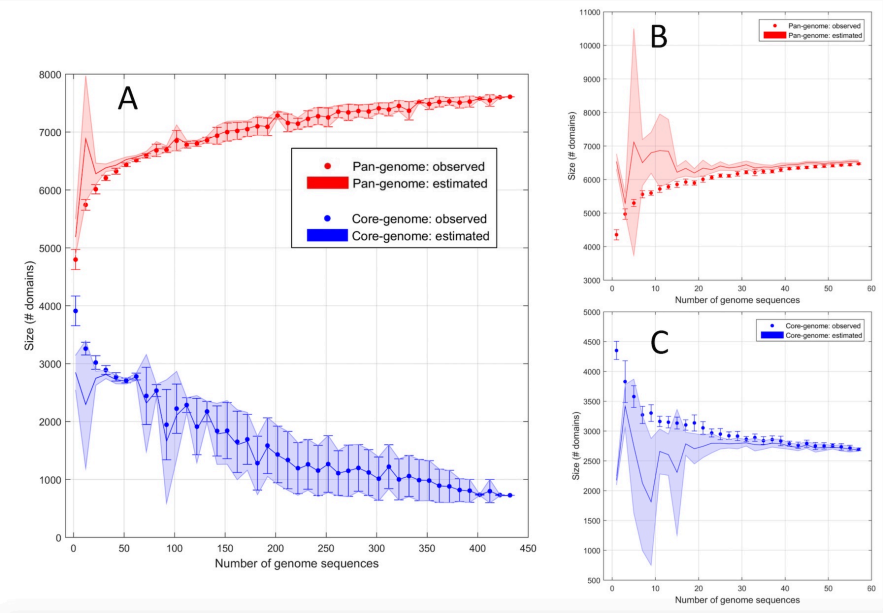

Figure 6.5: ***Pseudomonas pan- and core-genome defined on the base of protein domains*** *A) Complete overview of the distribution of the size of the pan- and core- distribution of protein domains. Error bars correspond to standard deviations based on 10 measured random realizations of the indicated number of genomes whereas the shadowed area is the estimated standard deviation using the same approach. B) Pan-genome of the 58 fully circular genomes. C) Core-genome of the 58 fully circular genomes.*

model ($\alpha = 1.30 \pm 0.05$). In this analysis values > 1 indicate a closed pan-genome (Tettelin et al., 2008).

Essentiality analysis of domains in the core-genome

From a functional point of view, the core-genome of a genus is most likely enriched in essential genes necessary for (long term) viability and adaptation to ever changing environmental conditions. Since persistence can be used to identify genes required for survival (Medini et al., 2005; Acevedo-Rocha et al., 2013), a positive correlation between persistence (the number of genomes sharing a given gene) and essentiality can be hypothesized. To verify this hypothesis we combined gene essentiality measures with gene persistence in the genus. Gene essentiality was defined from experimental results available for two *P. aeruginosa* strains (PAO1 and PA14) (S. A. Lee et al., 2015; Liberati et al., 2006) and from *in silico* predictions. For the latter, we

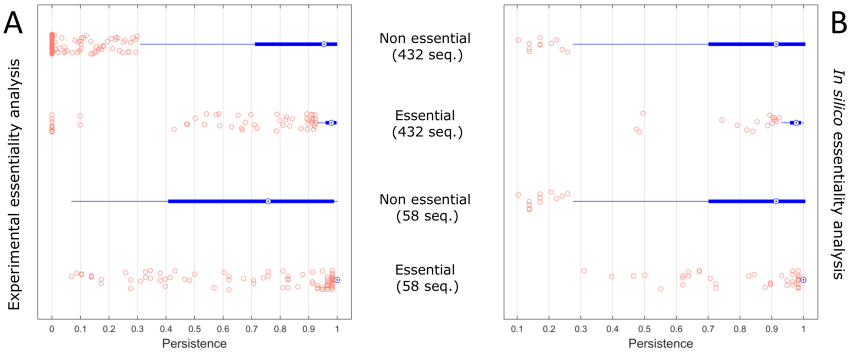

Figure 6.6: ***Persistence of (non) essential genes****. A) Persistence of essential and non-essential genes as derived by experimental investigations. B) Persistence of essential and non-essential genes as derived by in silico modelling using genome based constrained metabolic modelling. Results shown pertain the use of the iMO1086 model for P. aeruginosa PAO1. In both cases persistence is calculated using the 58 completely sequenced Pseudomonas genomes and the complete set of 432 genomes sequences. Magenta (circle) dots indicate outliers.*

considered 6 genome-scale constraint-based metabolic models which rely on functional annotation to uncover the metabolic potential of biological systems and are able to accurately predict gene essentiality in a large variety of growth conditions (Orth, Thiele, and Palsson, 2010).

We observed that essential genes show higher persistence values than non essential ones: this relationship is conserved when persistence is computed either using a sequence similarity based approach on 58 completely sequenced genomes or for 432 genomes by using a domain architecture approach as shown in Figure 6.6A.

A comparison of gene persistence and essentiality for the two strains showed that 65% of genes found to be essential for PA14 growth on LB are also essential for growth of PAO1 on either LB, minimal with pyruvate or sputum agar, but only 39% of genes reported to be essential for PAO1 growth were found to be essential for PA14 (See Supplementary Figure S4). This difference could be due to the smaller set of tested conditions. We used a less stringent cut-off for persistence: 0.95 instead of 1 to allow for non-sequenced genes due to incomplete draft genomes. Therefore, we observed that a small fraction of persistent genes is present in only one of the two strains (0.016% and 0.025% for PA14 and PAO1, corresponding to 75 and 47 genes respectively) which are likely to have been lost through evolution.

Table 6.2: Conditional gene essentiality predictions using six metabolic models from three Pseudomonas species.

Organism	P. aeruginosa		P. putida			P. fluorescens
Model	iMO1056	iMO1086	iJN746	iJP815	iJP962	iSB1139
Medium sources						
#Carbon	49	51	60	40	43	44
#Nitrogen	32	33	22	25	27	19
#Sulfur	4	1	10	1	1	6
#Phosphor	2	2	1	1	1	2
Genes						
#Essential/persistent*	115/106	149/132	118/104	112/100	162/148	117/95
#Conditional/persistent*	591/278	601/278	389/170	113/64	495/252	615/290
#Non-essential	348	336	253	593	305	407
#Overlapping genes	95			68		

*Persistence was computed for each essential and conditional essential genes over the 58 *Pseudomonas* genomes

Analysis of the complete pan-genome revealed that 1252 single copy genes are persistent. Of these, almost one third (404) were found to be essential *in vivo* under three growth conditions (LB, minimal-pyruvate or sputum agar) for *P. aeruginosa* PAO1 strain (S. A. Lee et al., 2015). Similar ratios were observed for strain PA14.

1112 unique domains were identified in the 404 essential persistent genes and 1340 unique domains in the non-essential but persistent genes. 203 domains were shared between essential and non-essential persistent genes. Essential genes contain a larger repertoire of unique, single copy domains: 404 essential persistent genes contained, on average, 1.53 single copy domains whereas for non essential persistent genes, the average was 0.82.

In vivo essentiallity analysis were limited to four conditions. Using metabolic models a wider range of conditions can be explored albeit the analysis is restricted to metabolic genes. We considered six genome scale constraint based metabolic models describing the metabolism of *P. aeruginosa* PAO1 (models iMO1056 (M. a. Oberhardt et al., 2008) and iMO1086 (M. A. Oberhardt et al., 2011)), *P. fluorescens* SBW25 (iSB1139 (Borgos et al., 2013)) and *P. putida* KT2440 (iJN746 (Nogales et al., 2008), iJP815 (Puchalka et al., 2008), and iJP962 (M. A. Oberhardt et al., 2011)).

We explored a wide range of growth conditions with varying carbon, nitrogen, phosphorus and sulphur sources and for each medium composition, gene essentiality predictions were performed using Flux Balance Analysis and are summarized in Table 6.2.

Figure 6.6B shows results for *P. aeruginosa* model iMO1086, confirming what

was observed for experimental data. Of the 750 essential metabolic genes that were identified under 3366 media compositions for iMO1086, 169 genes were identified to be essential under experimental conditions whereas 42 genes were essential but not *in silico* (25%). Average persistence over the 58 complete genomes was 0.96±0.14 for predicted essential genes and 0.85±0.24 for non-essential, which we found to be significant (p-value < 0.01 for a Wilcoxon test). When considering the 432 genomes, we still observed difference in the persistence of predicted essential and non essential genes 0.95±0.12 versus 0.89±0.21, p-value < 0.01). Similar results were also obtained when using essentiality predictions for the other metabolic models.

Using metabolic models to simulate media compositions we identified additional genes that were essential in a number of conditions, retrieving on average 1.47 single copy domains per gene, consistently with what observed for essentiality experiments. We further combined the models' predictions and we inspected genes predicted to be essential in all the tested conditions. For *P. putida*, the three models showed an overlap of 68 essential genes. Interestingly, these genes contained 2.53 single copy domains on average, underpinning previous results. Non-essential genes contain domains that are shared with other genes. This can result in the presence of isozymes or of potentially moonlighting enzymes which can step in for essential functions in the case of deletions or mutations.

Variability of gene expression and its association to persistence and essentiality in *Pseudomonas*

Associations between gene essentiality and low variation in protein abundance have been observed in *E. coli* (Taniguchi et al., 2010). We hypothesized the existence of an association between gene persistence and expression level variation. We analysed gene expression variability in *P.aeruginosa* using a gene expression compendium containing over 900 samples and 100 datasets regarding *P.aeruginosa* PAO1 genes (Tan et al., 2016). Each gene was assigned a score, Variability, for transcriptional variation. Persistent genes tend to show significantly lower degree of variation in expression level than non persistent ones (p-value < 0.01); this holds true also for essential genes (Figure 6.7). Similar results are obtained when analysing a more lim-

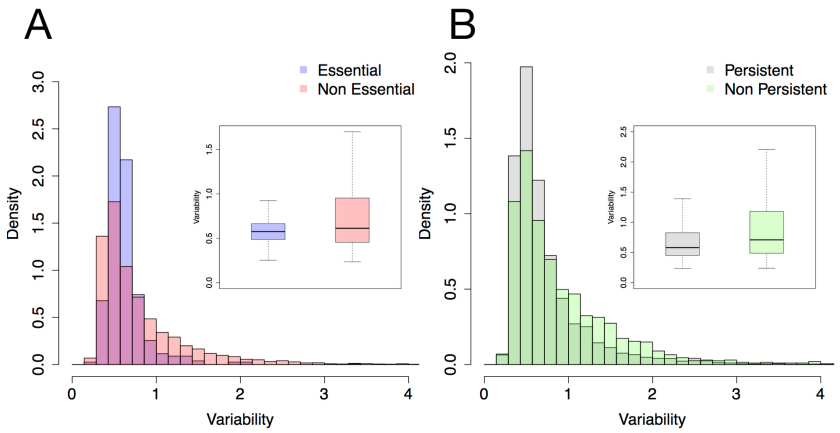

Figure 6.7: ***Variability of gene expression levels and its association with persistence and essentiality*** *A) Distribution of Variability score for (non) persistent genes (genes with persistence lower or higher than 0.95, respectively). Box plots show Variability values for both groups. Difference between mean values is significant (p-val < 0.01). B) Distribution of Variability score for essential and non-essential genes with gene essentiality derived experimentally (S. A. Lee et al., 2015). Box plots show Variability values for both groups. Difference between mean values is significant (p-val < 0.01).*

ited dataset containing RNAseq measurements of *P.aeruginosa* PA14 in 14 growth conditions (Dötsch et al., 2015) (see Supplementary Methods S5) This association between low expression variability and persistence/essentiality could indicate that expression of genes in the core-genome is likely to be buffered and independent from environmental growth conditions. To the best of our knowledge such associations have never been established on such large scale due to the limitations associated to comparing hundreds of genome sequences.

Discussion

For our analysis we did not rely on previously existing annotations, but we performed a consistent re-annotation of all the sequences using a standardised approach that ensured coherence and uniformity. A sequenced based approach was used for a prior comparative analysis to define clusters of orthologous proteins in the smaller dataset of 58 complete genomes. Due to polynomial growth of computational time, this approach is not feasible for large data sets. Mining a gene se-

quence for domain occurrences is less computationally demanding, which provides an effective scalable approach.

Sequence based approaches are used to identify clusters of orthologous proteins, however the analysis of domain architectures is targeted towards the identification of groups of functionally equivalent proteins. Protein domains provide a standardised way to assess sequence variation and its impact in function, since every amino acid has a characteristic weight in the domain model. Protein domains are more strongly associated to protein structure than protein sequences, thereby providing a closer link to function that can bridge over larger evolutionary distances, which is essential to comparative functional analysis. Still there is a need for improving how protein domain are defined to accommodate similar models arising from, possibly different, databases and to take into account positional variations that might lead to spurious domain inversions.

When applied to the inferred proteomes of the 58 complete genomes, both clustering methods yield similar results. The same clusters were obtained in 40% of the cases meaning that each of these clusters contained an equal number of proteins, captured the same strains and shared the same domain architectures. In 20% of the cases, very similar but numerically distinct clusters were obtained, as a given sequence similarity cluster had captured two distinct domain architectures. In most of these cases variability in domain architecture were caused by changes in domain order due to small variations in the start position of overlapping domains. Approximately, 20% of identified proteins have no recognizable functional domains. As most of these proteins are hypothetical they were not considered for functional analysis. When only proteins containing domains are considered, over 90% of the clusters identified using sequence comparisons contain 4 or less distinct architectures.

The differences in the persistence curves shown in Figure 6.3C show that the way the clusters are defined, either using sequence similarity or protein domains, impacts the calculation of gene persistence: this has repercussions on the definition of the core genome and its size. We found these differences to be larger when more genomes are considered. This is more likely linked to the broader range of phylogenetic distances among considered genomes: this is explored in more detail in

Koehorst *et al.*(Koehorst et al., 2016)

Our analysis resulted in the identification of the pan- and core-domainome of 432 *Pseudomonas* which is closed according to the heap model as also recently noted for the *P. aeruginosa* species (Mosquera-Rendón et al., 2016). This suggests that sequencing additional strains will fail to add new genes to the pan-genome: however, this is likely an oversimplification. Here, we understand closeness of the pan-genome as measure of the genus ability to acquire exogenous genes and as a proxy for the ratio between vertical and horizontal gene transfer indicating that horizontal gene transfer has not played a major role in shaping the genome content of the genus.

Key characteristics of *Pseudomonas* must be located in the genus core-genome, however comparison with metabolic models shows that identified core is not autonomously functional. Not all the genes in the core-genome seem to be essential (under given tested conditions), however essential genes represent $\approx$ 40% of the core-genome, in agreement with previously reported ratios for other species/genus (X. Yang et al., 2016). The remaining 60% contain unique features defining the genus.

We found a strong association between gene essentiality and protein domain properties. We observe an inverse correlation between the number of proteins in the genome containing the considered domain and essentiality, with average number of domains uniquely present in the considered protein going from 1.5 to 0.8 when non essential/essential genes in the core-genome are considered. The average number of single copy domains per gene further increases when stricter criteria for gene essentiality are applied, namely that genes should be essential in all the simulated media.

Accurate algorithms to predict gene essentiality from genomic features have been also developed and domain enrichment score has been shown to have a high predictive power (Deng, 2015) which is computed based on the ratio of occurring frequencies of a particular domain between essential genes and the total genes in the whole genome of already characterized species. Here we have established a link between the number of copies of a domain in a genome and gene essentiality that can be used to complement essentiality predictions.

The extensive use of metabolic reconstructions allowed us to identify conditionally essential genes, and a large number of single copy domains is also observed in these genes. This supports the idea that protein domains are the driving force behind gene essentiality which is preserved through protein domains rather than through the conservation of entire genes (Deng et al., 2011).

We have shown that lower fluctuations in gene expression are associated to essential and/or persistent genes. Further work is required to clarify the overlap and intertwining between both gene categories (essential/persistent) and to clarify the (possibly different) regulatory mechanisms stabilizing their expression levels.

Methods

Genome retrieval

Genbank files containing genome sequences and existing annotations for 58 circular genomes and 374 draft genomes of the *Pseudomonas* genus were downloaded from the GenBank database in June 2015. Annotation of *Pseudomonas* KT2440 was also downloaded from RAST (Aziz et al., 2008). A detailed list of the included strains is available (see Supplementary Figure S1 and Supplementary Data S2).

Genome de novo annotation

To perform the re-analysis of the 432 gemomes sequences we used a in-house pipeline for annotation and data storage (Koehorst et al., 2016). Likewise existing annotation pipelines such Prokka (Seemann, 2014), it relies on external feature prediction tools to identify the coordinates of genomic features within genomics sequences. The pipeline consists of a number of python modules that execute annotation applications and convert results and provenance directly into the RDF data model with a self defined ontology (the complete description of the implemented ontology can be obtained using RDF2Graph (Dam et al., 2015)) using the RDFLib library. For genetic elements determination a variety of tools is implemented such as Prodigal (Hyatt et al., 2010) for gene prediction. The main difference is that results are stored as Turtle files (Beckett and Berners-Lee, 2008) containing an RDF model which allows simultaneous exploration of annotation data of multiple genome se-

quences, greatly facilitation multiple comparison and the integration of heterogeneous source of information. Since it deploys semantic features allowing the storage of data provenance, we refer to it as SAPP (semantic annotation pipeline with provenance). Annotation can be exported to other formats for downstream processing with other tools such as Roary (Page et al., 2015)

Each genome sequence was converted to the RDF data model using the EMBL/GBK to RDF module. The FASTA2RDF, GeneCaller (a semantic wrapper for Prodigal 2.6 (Hyatt et al., 2010)) and InterPro (a wrapper for InterProScan (Jones et al., 2014)) modules were used to handle and annotate the genome sequences. Results were retrieved with SPARQL queries.

Protein domain presence and phylogenetic analysis

A SPARQL query was used to extract the presence of protein domains for all 432 genomes. Data were stored in a 432 (genomes) by 7608 (protein domains) binary matrix (0/1 for absence/presence). Protein domains were identified by their INTERPRO identifiers. Phylogenetic trees based on protein domains were created taking as input the domain presence/absence matrix. The R package `pvclust` was implemented in R (version 3.3.1) (Team, 2013) with a binary distance and average clustering approach with a bootstrap value of 10 (Suzuki and Shimodaira, 2006).

Protein domain architecture based clustering

The positions (start and end on the protein sequence) of domains having InterPro (Jones et al., 2014) identifiers were used to extract domain architectures (*i.e.* combinations of protein domains). Protein domains were retrieved for each protein individually. The domain starting positions were used to assess relative position in the case of overlapping domains; alphabetic ordering was used in the case of domains with the same starting position. Labels indicating N-C terminal order of identified domains were assigned to each protein so that the same labels were assigned to proteins sharing the same domain architecture. Here we have followed a strict approach and two domain architectures were considered different whenever they had different domains or they appeared in different order. For more details see Koehorst et al.,

2016.

Estimation of pan- and core-genome size

The estimated number of domains in the pan- and core-genomes expected if the sequences of every existing strain were to be included in the analysis were computed using binomial mixture models as implemented in the `micropan` R package (L. Snipen and Liland, 2015) using the domain presence/absence matrix previously defined and default values for the parameters. *Pan-* and *core-* analysis was initially performed on the 87 genomes with a maximum of 3 contigs to avoid bias due to incomplete genome sequences. Analysis was extended to the remaining 374 draft genome sequences available. To obtain an indication of the variability of these measures as function of the number of sequences used, these were calculated by a 10 fold random sampling from the full set. Heap analysis as implemented in the `micropan` R package was used to estimate openness or closeness of the pan-genome (Tettelin et al., 2008) using 500 genome permutations and repeating the calculation 10 times. Final measure is given as the mean ± standard error.

Orthologous gene detection

Orthologous genes were calculated initially for the set of 58 completely sequenced genomes. Protein sequences predicted using Prodigal 2.6 were extracted using a SPARQL query and used in a Best Bidirectional Hit approach (Tatusov, Koonin, and Lipman, 1997): using an all-versus-all BLASTP comparison and an E-value threshold of 10^{-5} and a maximum target sequence of 10^5. OrthAgogue (Ekseth, Kuiper, and Mironov, 2013) was used to convert BLAST results into a weighted graph. The MCL (Dongen, 2000) clustering algorithm was applied, using an inflation value of 1.5, on the graph to define protein clusters. The results were then extrapolated to the full set of 432 genomes using cluster specific domain fingerprints. Specifically, the sequence clusters obtained through MCL clustering on the 58 complete genomes were used to define sets of protein domains (each sequence cluster was mapped to a set of domains). The remaining genomes were then looked for any given domain set defined on the 58 genomes to define their presence/absence in the draft genomes.

Persistence and essentiality analysis

The persistence of a gene can be defined as

$$Persistence = \frac{N(\text{orth})}{N}$$

where $N(orth)$ is the number of genomes carrying a given orthologue and N is the number of genomes searched (Fang, Rocha, and Danchin, 2005). For the 58 completely sequenced genomes, orthologous genes were inferred using a BBH approach. For the full set of 432 sequenced genomes orthologous genes were inferred by making use of protein domain arrangements.

Locus tags for predicted proteins were inferred from the original annotation through SPARQL. Locus tags were linked to gene essentiality as defined in experimental studies available for *P. aeruginosa* PAO1 (S. A. Lee et al., 2015) and PA14 (Liberati et al., 2006). For each of the predicted proteins with inferred locus tag the corresponding protein cluster was initially calculated for the 58 genomes. The domain architecture corresponding to each cluster was extracted and subsequently scanned against all 432 available sequences. We used the MCL clusters as a reference set for the identification of domain architecture variations which were then extrapolated over the 432 genomes. The persistence for each locus tag was calculated and compared against the essentiality score obtained from two experimental studies.

Metabolic model essentiality analysis

We considered six genome scale constraint based metabolic models describing the metabolism of *P. putida* KT2440 (models iJN746 (Nogales et al., 2008), iJP815 (Puchalka et al., 2008), and iJP962 (M. A. Oberhardt et al., 2011)), *P. aeruginosa* PAO1 (models iMO1056 (M. a. Oberhardt et al., 2008) and iMO1086 (M. A. Oberhardt et al., 2011)) and *P. fluorescens* SBW25 (model iSB1139 (Borgos et al., 2013)). For each genome-scale metabolic model we performed a single gene essentiality analysis in a large number of growth media varying in carbon (C), nitrogen (N), phosphorus (P) and sulphur (S) source. To define the growth media we first identified candidate C, N, P, and S sources in each model independently. Because chemical sum formulas were not always available, we considered each compound for which an exchange reaction was present as a candidate C, N, P and S sources. We changed

the *in silico* medium composition to a minimal salts medium containing glucose as C source, ammonia as N source, phosphate as P source, sulphate as S source, in addition to oxygen, water, H^+, and a variety of salts depending on the particular model considered. The potential of each candidate C, N, P, and S source was then evaluated by adding it to the *in silico* medium while omitting the default C, N, P, or S sources. Growth predictions were performed using Flux Balance Analysis (Orth, Thiele, and Palsson, 2010) as implemented in the Matlab COBRA Toolbox (Schellenberger et al., 2011). This provided 4 lists of compounds that were suitable as C, N, P or S sources which were then combined into a single list of growth media by taking all combinations of compounds from the 4 lists. For each medium, we then used the *singleGeneDeletion* function from the COBRA toolbox to determine the growth rate of the mutant strains. If a gene knock-out reduced the *in silico* growth rate below 10^{-6} we considered the gene as essential. Models and Matlab scripts used in this analysis are available in Supplementary Data S6.

Comparison of gene expression profiles

A publicly available gene expression compendium for *P. aeruginosa* was retrieved (Tan et al., 2016). Briefly, this dataset contains a collection of gene expression datasets (950 individual samples pertaining 109 distinct datasets) measured using Affymetrix platform GPL84 and processed using a common normalization and background correction protocol. The final dataset contains expression measurements (in a log_2 scale) for 5549 genes from *P. aeruginosa* PAO1. For every gene we considered its expression profile in this compendium and a Variability value was calculated as the ratio between the standard deviation and the mean.

Availability of Data and Materials

The annotation pipeline framework is distributed under the MIT license. The pipeline all genomic data, data provenance and computational results associated with this study are freely available at `http://semantics.systemsbiology.nl`. Additionally, the data associated to this study are provided in turtle format as an RDF serialized dump. This dataset is made available under the Open Database License:

Acknowledgements

This work was supported by the European Commission-funded FP7 project INFECT (contract number: 305340). This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation.

Author contributions

JJK, JvD, VMdS and PJS participated in the conception and design of the study. JJK and JvD were responsible for the code and design of the semantic framework. RvH performed model-based essentiality analysis. MSD performed the integration of expression data. JJK, ES, VMdS, MSD, and PJS wrote the manuscript. All authors critically revised the manuscript.

Bibliography

Acevedo-Rocha, Carlos G, Gang Fang, Markus Schmidt, David W Ussery, and Antoine Danchin (2013). "From essential to persistent genes: a functional approach to constructing synthetic life". In: *Trends in Genetics* 29.

Alako, Blaise T F, Daphne Rainey, Harm Nijveen, and Jack A M Leunissen (2006). "TreeDomViewer: a tool for the visualization of phylogeny and protein domain structure". In: *Nucleic acids research* 34.

Aziz, Ramy K et al. (2008). "The RAST Server: rapid annotations using subsystems technology." In: *BMC genomics* 9. DOI: 10.1186/1471-2164-9-75.

Baltrus, David a et al. (2011). "Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 \textitPseudomonas syringae isolates." In: *PLoS pathogens* 7. DOI: 10.1371/journal.ppat.1002132.

Beckett, David and Tim Berners-Lee (2008). *Turtle - Terse RDF Triple Language*. URL: http://www.w3.org/TeamSubmission/turtle/.

Bertels, Frederic, Olin K Silander, Mikhail Pachkov, Paul B Rainey, and Erik van Nimwegen (2014). "Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads". In: *Molecular biology and evolution* 31.

Borgos, Sven E F et al. (2013). "Mapping global effects of the anti-sigma factor MucA in Pseudomonas fluorescens SBW25 through genome-scale metabolic modeling." In: *BMC systems biology* 7. DOI: 10.1186/1752-0509-7-19.

Cook, Helen and David W Ussery (2013). "Sigma factors in a thousand E. coli genomes". In: *Environmental Microbiology* 15.

Dam, Jesse CJ van, Jasper J Koehorst, Peter J Schaap, Vitor Ap Martins Dos Santos, and Maria Suarez-Diez (2015). "RDF2Graph a tool to recover, understand and validate the ontology of an RDF resource". In: *Journal of biomedical semantics* 6.

Deng, Jingyuan (2015). "Gene Essentiality: Methods and Protocols". In: ed. by Jason Long Lu. New York, NY. Chap. An Integra.

Deng, Jingyuan et al. (2011). "Investigating the predictability of essential genes across distantly related organisms using an integrative approach". In: *Nucleic Acids Research* 39.

Dongen, Stijn van (2000). "Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht". In:

Dötsch, Andreas et al. (2015). "The Pseudomonas aeruginosa Transcriptional Landscape Is Shaped by Environmental Heterogeneity and Genetic Variation". In: *mBio* 6.

Duan, Jin, Wei Jiang, Zhenyu Cheng, John J. Heikkila, and Bernard R. Glick (2013). "The Complete Genome Sequence of the Plant Growth-Promoting Bacterium Pseudomonas sp. UW4". In: *PLoS ONE* 8. Ed. by Mark Willem John van Passel. DOI: 10.1371/journal.pone.0058640.

Ekseth, Ole Kristian, Martin Kuiper, and Vladimir Mironov (2013). "OrthAgogue: an agile tool for the rapid prediction of orthology relations". In: *Bioinformatics* 30. DOI: 10.1093/bioinformatics/btt582.

Fang, Gang, Eduardo Rocha, and Antoine Danchin (2005). "How essential are nonessential genes?" In: *Molecular biology and evolution* 22.

Gross, Harald and Joyce E Loper (2009). "Genomics of secondary metabolite production by Pseudomonas spp." In: *Natural product reports* 26. DOI: 10.1039/b817075b.

Hyatt, Doug et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11. DOI: 10.1186/1471-2105-11-119.

Jones, Philip et al. (2014). "InterProScan 5: Genome-scale protein function classification". In: *Bioinformatics* 30. DOI: 10.1093/bioinformatics/btu031.

Koehorst, Jasper J, Edoardo Saccenti, Peter J Schaap, Vitor AP Martins dos Santos, and Maria Suarez-Diez (2016). "Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics". In: *F1000Research* 5.

Lee, Byungwook and Doheon Lee (2009). "Protein comparison at the domain architecture level". In: *BMC Bioinformatics* 10. DOI: 10.1186/1471-2105-10-S15-S5.

Lee, Samuel A et al. (2015). "General and condition-specific essential functions of Pseudomonas aeruginosa." In: *Proceedings of the National Academy of Sciences of the United States of America* 112. DOI: 10.1073/pnas.1422186112.

Liberati, Nicole T et al. (2006). "An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.

Loper, Joyce E et al. (2012). "Comparative genomics of plant-associated Pseudomonas spp.: insights into diversity and inheritance of traits involved in multitrophic interactions." In: *PLoS genetics* 8. DOI: 10.1371/journal.pgen.1002784.

Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli (2005). "The microbial pan-genome." In: *Current opinion in genetics & development* 15. DOI: 10.1016/j.gde.2005.09.006.

Mosquera-Rendón, Jeanneth et al. (2016). "Pangenome-wide and molecular evolution analyses of the Pseudomonas aeruginosa species". In: *BMC genomics* 17.

Nelson, K. E. et al. (2002). "Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440". In: *Environmental Microbiology* 4. DOI: 10.1046/j.1462-2920.2002.00366.x.

Nogales, Juan, Ines Palsson Bernhard Øand Thiele, Bernhard Ø Palsson, and Ines Thiele (2008). "A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory." In: *BMC systems biology* 2. DOI: 10.1186/1752-0509-2-79.

Notebaart, Richard A., Martijn A. Huynen, Bas Teusink, Roland J. Siezen, and Berend Snel (2005). "Correlation between sequence conservation and the genomic context after gene duplication". In: *Nucleic Acids Research* 33. DOI: 10.1093/nar/gki913.

Oberhardt, Matthew a, Jacek Puchałka, Kimberly E Fryer, Vítor a P Martins dos Santos, and Jason a Papin (2008). "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1." In: *Journal of bacteriology* 190. DOI: 10.1128/JB.01583-07.

Oberhardt, Matthew A et al. (2011). "Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis". In: *PLoS computational biology* 7. DOI: 10.1371/journal.pcbi.1001116.

Orth, Jeffrey D, Ines Thiele, and Bernhard Ø O Palsson (2010). "What is flux balance analysis?" In: *Nat Biotech* 28.

Page, Andrew J. et al. (2015). "Roary: Rapid large-scale prokaryote pan genome analysis". In: *Bioinformatics* 31. DOI: 10.1093/bioinformatics/btv421.

Pasek, Sophie, J.-L. Risler, and P. Brezellec (2006). "Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins". In: *Bioinformatics* 22. DOI: 10.1093/bioinformatics/btl135.

Puchalka, J et al. (2008). "Genome-Scale Reconstruction and Analysis of the Pseudomonas putida KT2440 Metabolic Network Facilitates Applications in Biotechnology". In: *Plos Computational Biology* 4. DOI: 10.1371/journal.pcbi.1000210.

Schellenberger, Jan et al. (2011). "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0". In: *Nature protocols* 6.

Seemann, Torsten (2014). "Prokka: Rapid prokaryotic genome annotation". In: *Bioinformatics* 30. DOI: 10.1093/bioinformatics/btu153.

Snipen, Lars, Trygve Almøy, and David W Ussery (2009). "Microbial comparative pan-genomics using binomial mixture models". In: *BMC Genomics* 10. DOI: 10.1186/1471-2164-10-385.

Snipen, Lars and Kristian Hovde Liland (2015). "Micropan: An R-package for microbial pan-genomics". In: *BMC bioinformatics* 16. DOI: 10.1186/s12859-015-0517-0.

Snipen, Lars-Gustav and David W Ussery (2012). "A domain sequence approach to pangenomics: applications to Escherichia coli". In: *F1000Research*. DOI: 10.3410/f1000research.1-19.v1. URL: http://f1000research.com/articles/1-19/v1.

Suzuki, Ryota and Hidetoshi Shimodaira (2006). "Pvclust: an R package for assessing the uncertainty in hierarchical clustering". In: *Bioinformatics* 22.

Tan, Jie, John H Hammond, Deborah A Hogan, and Casey S Greene (2016). "ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions". In: *mSystems* 1.

Taniguchi, Yuichi et al. (2010). "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells". In: *Science* 329.

Tatusov, Roman L, Eugene V Koonin, and David J Lipman (1997). "A genomic perspective on protein families". In: *Science* 278.

Team, R Core (2013). "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria." In: URL: http://www.r-project.org/.

Tettelin, Hervé, David Riley, Ciro Cattuto, and Duccio Medini (2008). "Comparative genomics: the bacterial pan-genome". In: *Current Opinion in Microbiology* 11. DOI: 10.1016/j.mib.2008.09.006.

Timmis, Kenneth N (2002). "Pseudomonas putida: a cosmopolitan opportunist par excellence". In: *Environmental Microbiology* 4. DOI: 10.1046/j.1462-2920.2002.00365.x.

Tripp, H James et al. (2015). "Toward a standard in structural genome annotation for prokaryotes". In: *Standards in genomic sciences* 10.

Wu, Xiao et al. (2011). "Comparative genomics and functional analysis of niche-specific adaptation in Pseudomonas putida". In: *FEMS Microbiology Reviews* 35. DOI: 10.1111/j.1574-6976.2010.00249.x.

Yang, Song, Russell F Doolittle, and Philip E Bourne (2005). "Phylogeny determined by protein domain content". In: *Proceedings of the National Academy of Sciences of the United States of America* 102. DOI: 10.1073/pnas.0408810102.

Yang, Xiaowen et al. (2016). "Analysis of pan-genome to identify the core genes and essential genes of Brucella spp." In: *Molecular Genetics and Genomics*.

7

Expected and observed genotype complexity in prokaryotes: correlation between 16S-rRNA phylogeny and protein domain content

Jasper J. Koehorst, Edoardo Saccenti, Vitor Martins dos Santos, Maria Suarez-Diez*, Peter J. Schaap*

* jointly supervised

bioRxiv; doi: https://doi.org/10.1101/494625

Abstract

The omnipresent 16S ribosomal RNA gene (16S-rRNA) is used to identify and classify bacteria though it does not take into account the distinctive functional characteristics of taxa. We explored functional domain landscapes of over 5700 complete bacterial genomes, representing a wide coverage of the bacterial tree of life, and investigated to what extent the observed protein domain diversity correlates with the expected evolutionary diversity, using 16S-rRNA as metric for evolutionary distance.

Domain analysis showed that 83% of the bacterial genes code for at least one of the 9722 domain classes identified. By comparing clade specific and global persistence scores, candidate horizontal gene transfer and signifying domains could be identified. 16S-rRNA and functional domain content distances were used to evaluate and compare species divergence and overall a sigmoid curve is observed. Already at close 16S-rRNA evolutionary distances, high levels of functional diversity can be observed. At a larger 16S-rRNA distance, functional differences accumulate at a relatively lower pace.

Analysis of 16S-rRNA sequences in the same taxa suggests that, in many cases, additional means of classification are required to obtain reliable phylogenetic relationships. Whole genome protein domain class phylogenies correlates with, and complements 16S-rRNA sequence-based phylogenies. Moreover, domain-based phylogenies can be constructed over large evolutionary distances and provide an in-depth insight of the functional diversity within and among species and enables large scale functional comparisons. The increased granularity obtained, pave way for new applications to better predict the relationships between genotype, physiology and ecology.

Introduction

The most commonly used method to classify bacteria and to identify new isolates is the analysis of the omnipresent 16S ribosomal RNA (16S-rRNA) gene (Weisburg et al., 1991; Yarza et al., 2014) by a direct comparison of the gene sequence with highly curated 16S-rRNA gene sequence databases (Yilmaz et al., 2014; Quast et al., 2013; Yoon et al., 2017; McDonald et al., 2012; Q. Wang et al., 2007; Hinchliff et al., 2015).

Using only the 16S-RNA gene for taxonomic characterisations presents limitations and disadvantages. First, arbitrary minimal sequence similarity thresholds are used as working boundaries for differentiating between taxonomic ranks. Although these thresholds prove to be very useful for classification purposes, they are subject to progressive insights and are limited as there is no biological meaning attached to it (Gupta, 2016). For instance, the minimal sequence similarity threshold for species delineation, proposed for the 16S-rRNA gene, has changed over time from 97% to 98.7% (Stackebrandt and Goebel, 1994; Kim et al., 2014) and even at this updated stringency level, for some phylogenetic groups, the resolution is too limited for a definite species classification (Janda and Abbott, 2007). Second, a restriction to the analysis of sequence variations in a single gene does not take into account the distinctive functional characteristics of the different prokaryotic taxa nor can it explain the genotypic, and the consequently phenotypic, differentiation observed between closely related species due to events such as gene loss or acquisition.

Alternative, inter-genomic BlastN-based sequence similarity methods exist that take into account full genome sequences. Examples are Average Nucleotide Identity (ANI) (Konstantinidis and Tiedje, 2005), Genome Blast Distance Phylogeny (GBDP) (Meier-Kolthoff et al., 2013) or a combination of 16S-rRNA sequence similarity and ANI values (Chun et al., 2018). These methods help to increase taxonomic coherence at the smaller evolutionary distances, but are less suitable to monitor the impact of mutation, gene loss and horizontal gene transfer (HGT).

To better understand the impact of gene loss and HGT and to improve the characterisation of functional diversity, the analysis needs to be performed beyond genome sequence similarity comparison by considering protein function. Protein encoding

genes reveal a modular design, with domains forming distinct globular structural and functional units. Bacterial innovation is in part driven by gain, loss, duplication and rearrangement of these functional units, resulting in the emergence of proteins with new domain combinations (Basu, Poliakov, and Rogozin, 2009; Zmasek and Godzik, 2012). Thus, a direct comparison of protein domain content should be able to reconstruct bacterial phylogeny independent of gene sequence similarity (Yang, Doolittle, and Bourne, 2005) and as such may serve as a better indicator of shared physiology and ecology (Jasper J Koehorst, Saccenti, et al., 2016; LG Snipen and D. Ussery, 2013).

In this study we present an exhaustive exploration of the functional landscape of over 5700 complete bacterial genomes representing a wide coverage of the bacterial tree of life and investigated to what extent protein domain diversity correlates with taxonomic diversity using the 16S-rRNA gene sequence as metrics for evolutionary distances.

Results

We analysed 5713 fully sequenced publicly available bacterial genomes corresponding to a wide range of different bacterial lineages (57 classes, 243 families, 818 genera and multiple strains of 1330 species), providing a good representation of the bacterial diversity observed in nature (See supplementary file S1 for more information). Genome sizes varied from 0.1 Mbp up to 13 Mbp. To avoid technical bias due to the use of different annotation strategies, all genomes were *de-novo* re-annotated with SAPP (Jasper J. Koehorst et al., 2017) (see Methods section for details). The total number of genes varied from 167 (*Candidatus Tremblaya princeps*) to 9968 (*Streptomyces bingchenggensis* BCW-1). We analysed 5713 fully sequenced publicly available bacterial genomes corresponding to a wide range of different bacterial lineages (57 classes, 243 families, 818 genera and multiple strains of 1330 species), providing a good representation of the bacterial diversity observed in nature (See supplementary file S1 for more information). Genome sizes varied from 0.1 Mbp up to 13 Mbp. To avoid technical bias due to the use of different annotation strategies, all genomes were *de-novo* re-annotated with SAPP (Jasper J. Koehorst et al., 2017) (see Methods

section for details). The total number of genes varied from 167 (*Candidatus Tremblaya princeps*) to 9968 (*Streptomyces bingchenggensis* BCW-1).

16S-rRNA variability within and between species

From the 5713 completely sequenced genomes, 25098 complete 16S-rRNA genes could be retrieved. On average the predicted length of the 16S-rRNA gene was 1531 ± 94 nt (See supplementary Figure Supplementary file S1) and 84% of the completed genomes (4772) contained between two and fifteen copies of the 16S-rRNA gene (Figure 7.1). The 16S-rRNA genes from phylogenetic groups of at least 50 strains were further analysed at family level. As can be seen in Figure 7.1B, among different families there is a diverse variation in copy number. While in some families the 16S-rRNA copy number is largely restricted to a single copy gene, in *Bacillaceae* the copy number ranged from 1 to 15 copies. Furthermore, 52% of the analysed genomes contained two or more non-identical copies of the 16S-rRNA gene. Intragenomic sequence variation reflected an overall sequence identity of 99.6 (+0.4 / -2)%, which is higher than the currently accepted 98.7% threshold for species delineation.

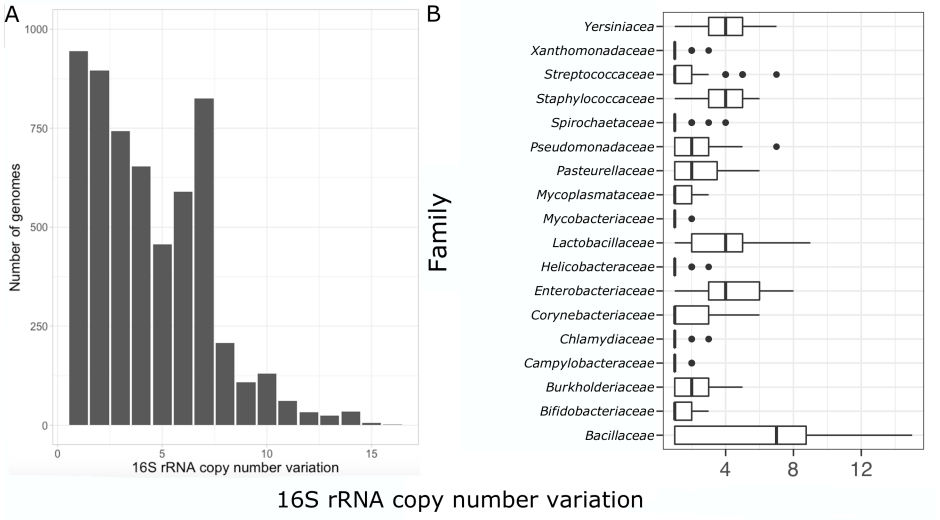

Figure 7.1: ***16S-rRNA copy number variation****. A) 16S-rRNA gene copy number variation in the complete set B) Copy number variation at family level; families represented by more than 50 strains were analysed.*

For the complete set of genomes, a species network based on pair-wise 16S-rRNA sequence similarity scores was built. In this network, nodes represent genomes and edges were drawn between nodes when the 16S-rRNA similarity was at least 98.7%. Network connectivity analysis identified 2025 connected components (subnetworks). For further study, 294 subnetworks were selected linking ten or more nodes. In thirty-two of these subnetworks, taxonomic inconsistencies were observed as they linked genomes of two or more species. The majority (30) of these inconsistent subnetworks linked species belonging to the same genus. However, two subnetworks were identified that linked species from different genera. The first subnetwork contained species of the *Escherichia* and *Shigella* genera. The second subnetwork was even more diverse and contained members of the *Citrobacter*, *Enterobacter*, *Klebsiella*, *Kosakonia*, *Raoultella* and *Salmonella* genera (Figure 7.2), both subnetworks eventually belong to the *Enterbacteriaceae* family. Overall, network analysis suggested that in many cases additional means of classification are required to obtain reliable phylogenetic relationships.

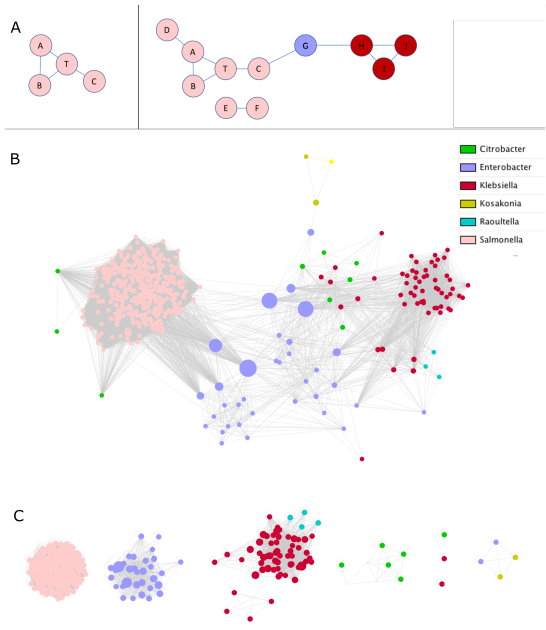

Figure 7.2: ***Topology of the similarity subnetwork of Enterbacteriaceae.*** *Nodes represent genomes and edges are drawn if the 16S-rRNA similarity >98.7%. A) Network topologies with colours indicating different species groups. Left panel, unambiguous species assignment; strains A,B and C are directly connected to type strain T. Right panel: Observed topology. Leave node strain D is in the cluster but has no direct link with type strain T. 16S-rRNA sequences of strain E and strain F are below the set similarity threshold and form an unlinked subnetwork. Strain G of the blue species functions as an articulation point linking the pink and red species subnetworks. B) Subnetwork linking six different genera based on the 16S-rRNA gene sequences using a sequence similarity threshold >98.7%. Size of each node is dependent on the betweenness centrality. Enterobacter is the main component that connects the different genera as no direct linkage between Salmonella and Klebsiella is observed. Three strains of Citrobacter have a direct connection to Salmonella and are disconnected from other Citrobacter strains. One Enterobacter (Enterobacter sp. R4-368) is isolated from the rest and is only connected to Kosakonia. The Raoultella genera have a close similarity to some of the Klebsiella strains. C) Topology of domain-class content subnetworks of the same strains using as threshold a binary distance ≤0.1. Distinct subnetworks are observed. Salmonella is now completely separated from the other genera; Enterobacter, Klebsiella and Citrobacter also form distinct clusters with a few members forming separate subnetworks.*

Protein domain architectures

By breaking proteins into domains and using precomputed profile hidden Markov models (pHMM) to classify these domains, a semantically consistent classification of encoded protein functions can be obtained (Jasper J Koehorst, Saccenti, et al., 2016). As a pHMM gives greater weight to matches at conserved sites they are also better

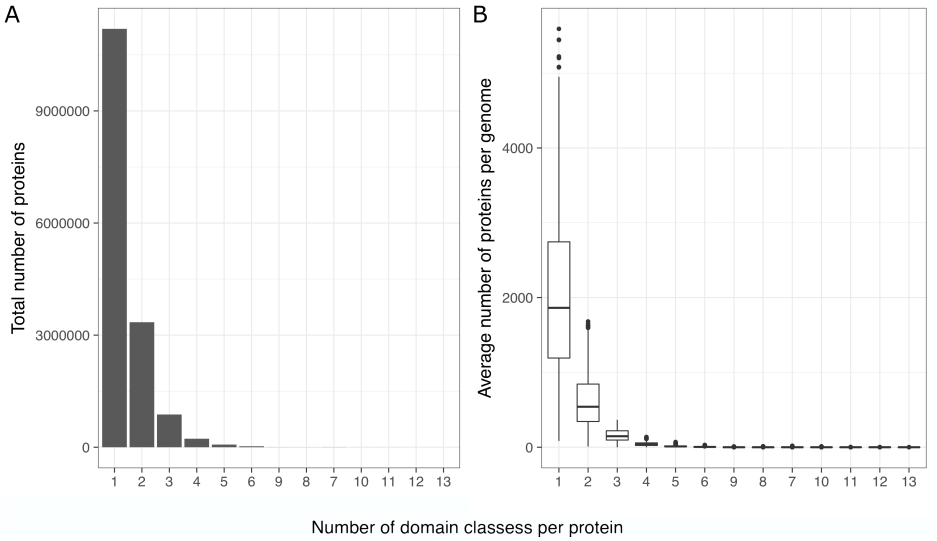

Figure 7.3: ***Frequency distribution of*** $1, 2, ..., 13$ ***domain classes*** *A) In the full dataset. B) In each genome*

for remote homology detection than standard sequence similarity-based methods (Sonnhammer et al., 1998). To obtain such protein classification the 18949996 inferred protein sequences were scanned for the presence of Pfam domains (Robert D. Finn, Coggill, et al., 2016). A total of 15747648 protein sequences were found to contain at least one domain instance (83.1%) and in total 9722 distinct protein domain classes were detected (See supplementary file S1 for more details). Two Pfam domains were discovered in 17.7% (3345544) and three or more domains in 6.4% (1205997) of these proteins (Table 7.1). Thus, the majority of the bacterial proteins appear to be single domain proteins (Figure 7.3A). Moreover, we observed that most multiple domain proteins appear to contain domain repetitions. Similar domain distributions were obtained when individual genomes were analysed, indicating that this is a general property of the architecture of bacterial genomes (Figure 7.3B).

Table 7.1: ***Overview of the number of proteins and corresponding protein domain content.*** *The majority of the proteins (83.1%) contained at least a single domain and only a few (1.74%) contained more than 3 domains.*

	Number of proteins	Fraction
All proteins	18949996	100%
>0 domains	15747648	83.1%
1 domain	11196108	59.1%
2 domains	3345544	17.7%
3 domains	875863	4.6%
>3 domains	330133	1.74%
>10 domains	15457	0.08%
>50 domains	208	0.0011%

Genome distribution of protein domains

The distribution of the domain classes across the studied genomes is shown in Figure 7.4. Panel A shows that there is a direct correlation between the genome size and the total number of domains detected. A sublinear relationship is observed between the total number of protein domains and the total number of protein domain classes indicating that domain copy numbers, but not so much the number of domain classes, increase in the larger genomes (Figure 7.4 panel B). On average, we counted 2.02 domain copies per genome. This copy number, however, showed a large variability, ranging from 1.07 copies for *Carsonella ruddii* (strain PV) (Nakabachi et al., 2006) to 4.58 copies for *Streptomyces bingchenggensis* (strain BCW-1) (X. J. Wang et al., 2010).

Domain persistence and analysis of the pan- and core-domainomes

In total, 9722 domain classes were detected. The overall persistence (the fraction of the genomes sharing a given domain class) is shown in Figure 7.5. Only 324 domain classes were ubiquitous in over 95% of the analysed genomes. Three domains, PF00009, (GTP-binding elongation factor family), PF01479, (S4 domain) and PF03144 (Elongation factor Tu domain 2) were shown to persist in all genomes. Additionally, a small number of domains were found to be present in over 99.9% of the studied genomes, PF00012 (Hsp70 protein), PF00318 (Ribosomal protein S2), PF00380 (Ribosomal protein S9/S16), PF00679 (Elongation factor G C-terminus), PF01926 (50S ribosome-binding GTPase), PF02811 (PHP domain), PF07733 (Bacte-

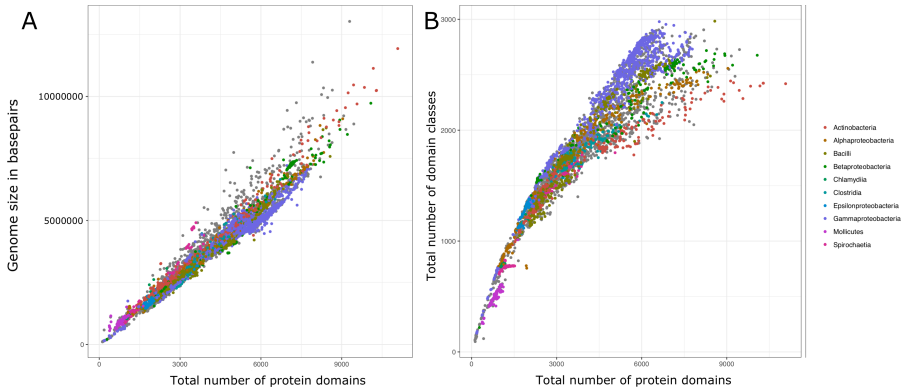

Figure 7.4: ***Distribution of the domain classes across the studied genomes*** *A) Correlation between genome size and number of protein domains. B) Correlation between the total number of domains and total number of domain classes. A sublinear relation is observed, suggesting that in the larger genomes an increase in domain copy number is favoured over an increase in domain classes.*

rial DNA polymerase III alpha subunit) and PF14492 (Elongation Factor G, domain II). Among the studied genomes there are domain classes with a high copy number. The domain with the highest copy number is PF00005, representing the ATP-binding domain of ABC transporters, with on average 62.9 copies per genome, yet the domain is absent in twelve small-sized genomes.

Accurate measurements of the pan- and core- domainome sizes would entail knowledge of the functional content of every single organism in the corresponding group. We have estimated their respective sizes for the 18 families that contained more than 50 members each (Figure 7.6A). The largest observed pan-domainome was of *Bacillaceae* with 4783 protein domain classes. The largest core was observed for *Yersiniaceae* (1844 domain classes) (Figure 7.6B).

When analysing the genomes of the *Chlamydiaceae* family, 78% of the protein domain classes are conserved. In contrast, the core of *Enterobacteriaceae* only covers 7% of the in total 4444 domains (Figure 7.6C). This is mostly due to the size of the genomes from the *Moranella*, *Riesia*, *Blochmannia* and *Ishikawaella* (Nikoh et al., 2011), genera as they are smaller than 1 Mbp, encoding as low as 444 genes, whereas the average genome size of *Enterobacteriaceae* is 4.8 Mbp, encoding on average 4510 genes. When excluding the small sized genomes, the core increases to 938 protein domains with a slightly smaller pan-domainome of 4441 yielding a 21% ra-

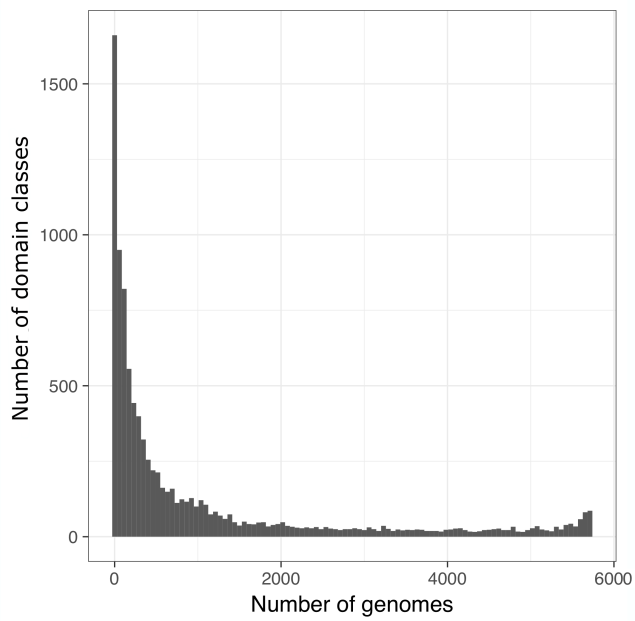

Figure 7.5: ***Distribution of domain classes over 5713 genomes***

.

tio between the core and pan-domainome. This shows the impact of including or excluding specific genomes in the analysis, as a single or few genomes can reduce the core significantly, thereby possibly eluting important information.

Openness of the pan-domainome provides another indication of the relative impact of horizontal acquisition and vertical transmission in shaping the domainome. Fitting a Heap's law, we estimated whether the pan-domainome for each of the largest families was either open or closed by fitting the decay parameter of a Heap's law function, α. The pan-domainome is closed when α >1.0 and open when α <1.0. The majority of the bacterial families here considered showed a closed pan-domainome (Figure 7.6D). For *Enterobacteriaceae* the Heap's parameter dropped from α=1.21 to α=1.17 upon removal of the previously indicated smaller genomes.

Signifying domains and horizontal domain transfer

Log persistence scores (log-P) were calculated for each of the domain classes present in the pan-domainomes from the five most abundant monophyletic species groups,

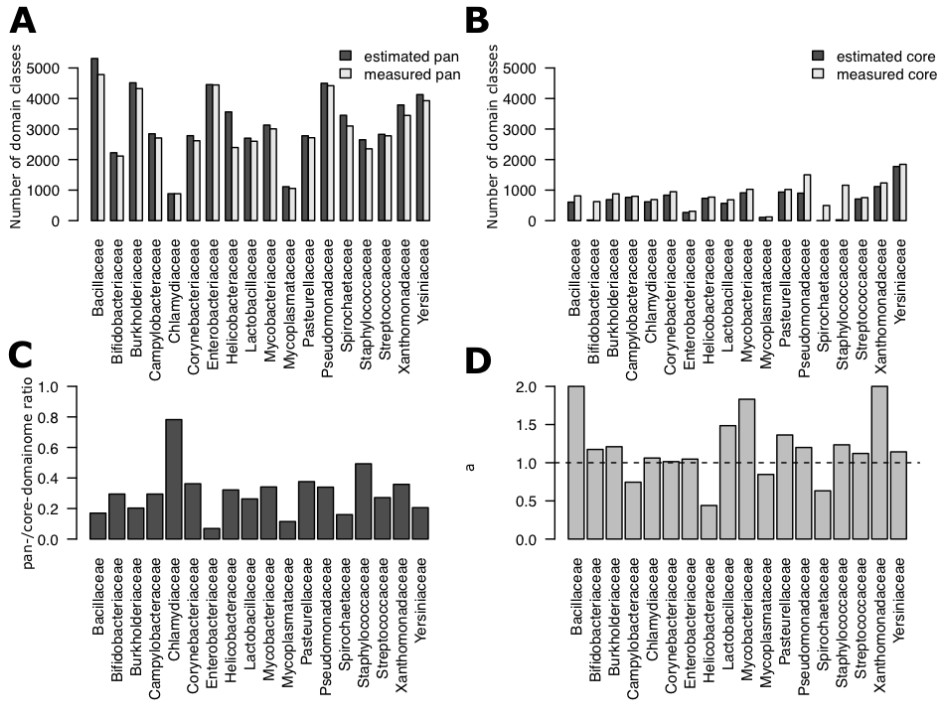

Figure 7.6: ***Persistence analysis of families with more than 50 members****. The estimated pan-domainome (Panel A) and estimated core (Panel B) shows a large degree of variability ranging from 78% for Chlamydiaceae and 7% for Enterobacteriaceae. The conservation ratio of the pan/-core (Panel C) shows that in only Chlamydiaceae more than half of the protein domain content is conserved. The family pan-genome is closed (Panel D) when $\alpha>1$.*

namely *Chlamydia trachomatis* (74), *Escherichia coli* (105), *Helicobacter pylori* (65), *Salmonella choleraesuis* (350) and *Staphylococcus aureus* (74). As null-model we consider the persistence in the full set of 5713 genome sequences.

For a small set of domain classes high (log-P) scores were obtained and are likely signifying domain classes (Figure 7.7 and Supplementary Table S3 logP). On the other end of the scale we find a large amount of domain classes with negative log-P scores. These incidental domains have a low to very low intra-species persistence which suggests that they may have been acquired by horizontal gene transfer. Unlike the high scoring domains most of them have been assigned a molecular, often metabolic, function.

Table 7.2: Salmonella choleraesuis top 25 signifying and incidental domains

.

PFAM	log-P	Global persistence	Description
PF09460	4.025	0.056	Saf-pilin pilus formation protein
PF06767	4.017	0.062	Sif protein
PF05364	4.013	0.062	Salmonella type III secretion SopE effector N-terminus
PF07108	4.013	0.062	PipA protein
PF16583	4.012	0.061	Zinc-regulated secreted antivirulence protein C-terminal domain
PF16728	3.991	0.061	Domain of unknown function (DUF5066)
PF15942	3.976	0.064	Domain of unknown function (DUF4751)
PF07824	3.969	0.064	Type III secretion chaperone domain
PF09052	3.965	0.064	Salmonella invasion protein A
PF05775	3.950	0.059	Enterobacteria AfaD invasin protein
PF11047	3.941	0.065	Salmonella outer protein D
PF05925	3.914	0.066	Enterobacterial virulence protein IpgD
PF08052	3.906	0.066	PyrBI operon leader peptide
PF13998	3.873	0.068	MgrB protein
PF02510	3.858	0.069	Surface presentation of antigens protein
PF13979	3.852	0.068	SopA-like catalytic domain
PF02090	3.840	0.070	Salmonella surface presentation of antigen gene type M protein
PF04741	3.815	0.071	InvH outer membrane lipoprotein
PF07487	3.811	0.071	SopE GEF domain
PF09119	3.801	0.072	SicP binding
PF05688	3.794	0.071	Salmonella repeat of unknown function (DUF824)
PF03433	3.759	0.074	EspA-like secreted protein
PF09599	3.759	0.074	Salmonella-Shigella invasin protein C (IpaC_SipC)
PF10940	3.737	0.074	Protein of unknown function (DUF2618)
PF05689	3.727	0.074	Salmonella repeat of unknown function (DUF823)
...			
PF13442	-7.502	0.518	Cytochrome C oxidase, cbb3-type, subunit III
PF09424	-7.522	0.525	Yqey-like protein
PF01769	-7.533	0.529	Divalent cation transporter
PF06750	-7.546	0.534	Bacterial Peptidase A24 N-terminal domain
PF09084	-7.555	0.537	NMT1/THI5 like
PF03309	-7.557	0.538	Type III pantothenate kinase
PF06271	-7.573	0.544	RDD family
PF12802	-7.610	0.558	MarR family
PF01628	-7.642	0.571	HrcA protein C terminal domain
PF01593	-7.680	0.586	Flavin containing amine oxidoreductase
PF10397	-7.689	0.589	Adenylosuccinate lyase C-terminus
PF00355	-7.732	0.607	[2Fe-2S] domain
PF01220	-7.743	0.612	Dehydroquinase class II
PF14693	-7.769	0.623	Ribosomal protein TL5, C-terminal domain
PF01809	-7.769	0.623	Haemolytic domain
PF02616	-7.771	0.624	Segregation and condensation protein ScpA
PF04079	-7.781	0.629	Segregation and condensation complex subunit ScpB
PF03448	-7.815	0.644	MgtE intracellular N domain
PF07521	-7.823	0.647	Zn-dependent metallo-hydrolase RNA specificity domain
PF01883	-7.868	0.668	Iron-sulfur cluster assembly protein
PF02686	-7.969	0.716	Glu-tRNAGln amidotransferase C subunit
PF02637	-8.019	0.741	GatB domain
PF02934	-8.026	0.744	GatB/GatE catalytic domain
PF01425	-8.173	0.824	Amidase
PF00825	-8.196	0.838	Ribonuclease P
			Number of strains analysed 350; α=0.89

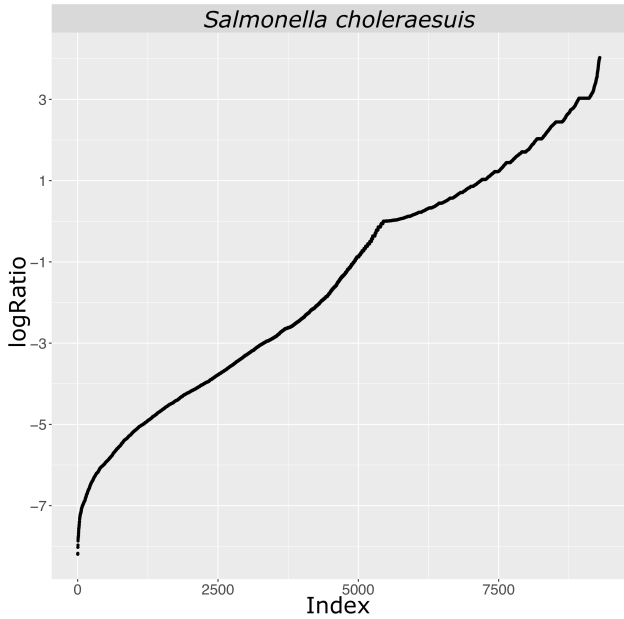

Figure 7.7: ***Persistence scores of Salmonella choleraesuis protein domain classes****. For each domain class present in the S. choleraesuis pan-domainome, persistence scores are compared with the pan-domainome persistence scores obtained from the complete set of 5713 genomes.*

Co-evolution of bacterial 16S-rRNA and whole genome domain content

Protein domains provide a formal description of genome encoded functionalities each contributing to bacterial genotypic complexity. The functional relatedness of an arbitrary pair of genomes can thus be determined by finding the fraction of encoding domain classes in common relative to the the number of domain classes present in each of these genomes. Through inclusion of the 16S-rRNA data the co-evolution of bacterial 16S-rRNA gene sequences with genotypic complexity can be studied (Figure 7.8). In panel A the distribution of domain based distances is plotted using a binary dissimilarity score. Likewise in panel D the distribution of 16S-rRNA sequence distances is plotted. Panel C shows a pairwise comparison between 16S-rRNA distances and functional distances for the analysed genomes. Finally, panel B, presents a schematic representation of the relationship between the two methods.

Overall, a good agreement is found between both approaches to evaluate species

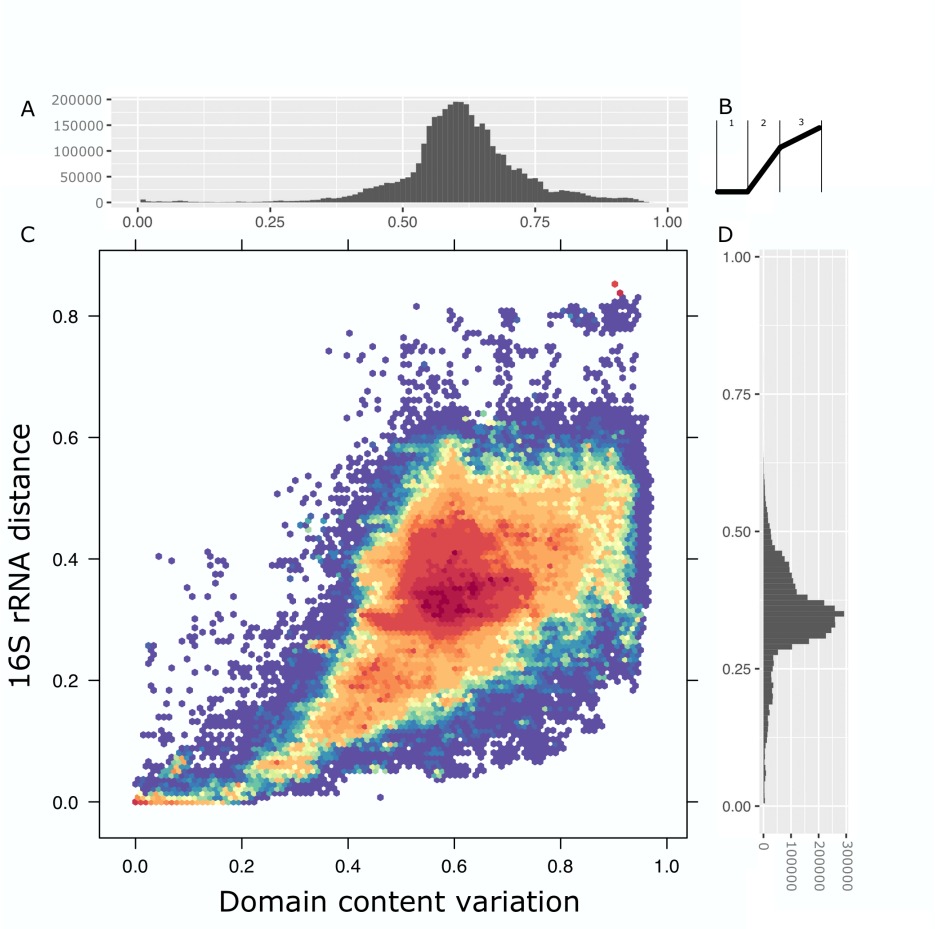

Figure 7.8: ***Distance comparison of the 16S-rRNA gene with the functional diversity****. A) Distribution of domain based distances. B) Schematic representation of the three stages of diversification. 1) a fast-short-term evolution, as evolutionary distances measured by 16S-rRNA remain small, while functional diversification has already taken place. 2) long-term evolution, in which functional diversification occurs at a scale compatible with diversification by 16S-rRNA sequence evolution. 3) The distance of the 16S-rRNA remains behind the functional diversity as the 16S-rRNA distance can only diverse so far without loss of function. C) Comparison between pairwise 16S-rRNA distances and pairwise functional distances. D) Distribution of 16S-rRNA based distances.*

divergence. Analysis of the 16S-rRNA distances shows a marked differentiation in the [0.3, 0.35] interval, which appears as a steep increase in the abundance of instances of these distance values (Figure 7.8D). These differentiations correspond to lineage boundaries (specifically class and phylum differences). This increased density corresponds to the higher density in the center of the plot (Figure 7.8C), that reflects that most of the performed comparisons involve members distantly related in the evolutionary scale. This is also apparent on the higher number of instances of functional differences in the [0.6, 0.7] interval (Figure 7.8A), however functional differences accumulate more gradually, and no steep increase is observed.

The relationship between the two methods to evaluate species differences can be approximated through a sigmoid curve and three regimes can be distinguished (Figure 7.8B). Species at close evolutionary distances show a broad range of functional similarity (Figure 7.8B region 1). A high diversity is observed, so that genomes with high similarity regarding their 16S-rRNA can show high functional diversity. The second region shown in Figure 7.8B, region 2, corresponds to regions of relatively large genetic differentiation (class differences) that accumulate functional differences at a relatively lower pace. Finally, the third region (region 3) corresponds to very distant species that as expected, have a large degree of functional differentiation.

In addition to functional similarities between evolutionary close strains, Figure 7.8C also indicates the presence of functionally very similar but evolutionary distant genomes. These are to be found in the region with low domain content variation (<0.05) and a large 16S-rRNA distance (>0.4). *Gluconacetobacter diazotrophicus* PAl 5, *Moraxella catarrhalis* BBH18 and *Pseudomonas aeruginosa* 39016 are some examples. Similar results are obtained when the analysis is repeated considering all available genomes. The presence of more than one copy of the 16S-rRNA gene may introduce a larger variability, however the overall agreement of 16S-rRNA classification remains the same.

Discussion

For several decades 16S-rRNA sequence similarity scores provided a good working metric for prokaryotic taxonomic classifications, but the ever-expanding sequence databases and taxonomic complexity now pinpoint at the limitations. Here, we have used a set of 5713 complete genomes to evaluate the predictive power of pair-wise 16S-rRNA sequence similarity scores on the diversity and taxonomic classification of these genomes.

We observed intragenomic variation of 16S-rRNA gene sequences, but further analysis showed that within the selection, this variation is limited and well above the currently advised species threshold of 98.7%, meaning that regardless of the selected copy, the same taxonomic classification should be obtained.

A network approach was subsequently used to study pair-wise 16S-rRNA sequence similarities between the 5713 sequenced strains (Figure 7.2). By using the currently accepted 98.7% minimal sequence similarity threshold, optimally this approach should lead to 1330 separate species networks, each containing all sequenced strains of a defined single species and each individual node within such subnetwork should at least have a direct link to the node that represents the reference or type strain (Figure 7.2 panel A). However, many more subnetworks were obtained and what was observed is that strains of the same species are in separate subnetworks. Additionally, strains with intermediate 16S-rRNA sequences were present functioning as articulation points merging what should have been independent species subnetworks. (Figure 7.2 panel B and C). With the continuous addition of new 16S-rRNA sequences it is likely that species amalgamation will become more frequent. In the light of this, a more appropriate approach would be to consider the similarity threshold as a confidence level. In this way, there is a high probability that two sequences with a 16S-rRNA sequence identity below the selected threshold belong to different species. This provides a probabilistic interpretation to the threshold.

We used Pfam protein domain-class content to study strain diversity. Protein domains are considered to be distinct functional units and as such responsible for a particular function or interaction. The Pfam 30 protein family database consists of 16306 domain families or classes (Robert D Finn et al., 2006) of which 9721 were

present in the studied dataset. Furthermore, we found that approximately 83% of the protein-encoding genes harbour at least one Pfam domain suggesting that the encoded domain-class content may provide a good metric to study strain diversity.

The core-genome of a taxonomic group contains genes that are present in all members of that group whereas the pan-genome contains all the different genes that can be found in any member of the population (Lars Snipen, Almøy, and David W Ussery, 2009). Here we extended the idea to protein domain classes, as has been previously reported (LG Snipen and D. Ussery, 2013; Jasper J Koehorst, Van Dam, et al., 2016). We observed that most domain classes have a low persistence overall (Figure 7.5), but as shown in Figure 7.6, by adding taxonomic information, distinct sets of domain classes accumulate in the core domainomes of the various clades suggesting that these core sets are somehow contributing to the physiology and ecology of these clades.

At family level, the pan/core domainome ratio is observed to be on average below 0.4 (Figure 7.6), but at lower taxonomic ranks this ratio increases. For *C. trachomatis* this ratio was determined to be 0.96, for *Escherichia coli* 0.58, for *Helicobacter pylori 0.83* and for *Staphylococcus aureus* 0.76. We assumed that species core domainomes would consist of signifying or even species-specific domain classes and domain-classes representing essential metabolic functions. We expected that signifying domain-classes are only highly persistent within a clade but that domain-classes representing metabolic functions would be widely spread. For each domain class present in the pan-domainome of five selected species we calculated the ratio between clade specific persistence and global persistence (log-P scores) using a null-model that assumes that domain-classes are evenly distributed over the strains. The analysed species contributed to 6.2% or less of the total number of strains.

Top log-P scoring domains mostly corresponded to domains of unknown function (DUF) or domains involved in signal transduction whereas, being omnipresent, metabolic functions were underrepresented. Of the 25 top scoring domains, 6 in *Salmonella choleraesuis*, 15 in *Chlamydia trachomatis*, 8 in *Escherichia coli*, 3 in *Helicobacter pylori* and 11 in *Staphylococcus aureus* corresponded to a DUF class. For the *Mycoplasma* species it has been established that many DUFs are essential for growth (Kamminga et al., 2017; Hutchison et al., 2016) and at least four of the DUFs in the

present study, two specific for *Escherichia coli* (PF07041 and PF10897) and two for *Helicobacter pylori* (PF12033 and PF10398) indeed have been characterised as being essential (Goodacre, Gerloff, and Uetz, 2014). Between these five species, top scoring domains also show no significant overlap suggesting that they are evolutionary conserved and may have a prominent role in shaping the species. Protein domain classes with the lowest persistence ratio's are likely HGT candidates. Functionally, most of them represent a metabolic function suggesting as has been reported (Ochman, Lawrence, and Groisman, 2000; Dutta and Pan, 2002) that horizontal gene transfer is an important source of metabolic diversity.

The impact of the presence of signifying domains in the core domainome is demonstrated in Figure 7.2C. Nodes from the Enterobacteriaceae subnetwork (Panel B) were re-analysed using pair-wise domain-class content distance analysis. A similarity threshold of 90% resulted in clade specifc domain-class subnetworks for *Salmonella*, *Enterobacter* and to a lesser extent for *Klebsiella*. Note that by adopting a whole-genome domainome approach, the history of every domain-class present in the pan-domainome, is taken into account. However, signifying domain classes are the main contributors and similar to what has been observed in Ochman et al. (Ochman, Lerat, and Daubin, 2005), we observed that the many incidental HGT candidate domain classes appear to have little impact on whole-genome domainome based phylogenetic reconstructions.

The ratio between the core- and pan-domainome size of groups of organisms at different phylogenetic levels provided a good estimate for beta-diversity. A relatively low ratio between the core and pan-domainome reduces the functional assignments that can be inferred from the 16S-rRNA classification. Conversely, a high ratio gives more certainty that functionalities are present. Overall the majority of the analysed families showed a low ratio indicating that only a reduced functional landscape can be extrapolated using 16S-rRNA analysis and the ratio can differ significantly among families. For example, *Chlamydiaceae* shows a large ratio whereas *Enterobacteriaceae* has the lowest observed ratio, indicating that the *Chlamydia* genus which consists mostly of pathogenic bacteria that are obligate intracellular parasites have evolved through simplification instead of complexification and are therefore less diverse (Wolf and Koonin, 2013). Whereas *Enterobacteriaceae* is a diverse family

consisting of members that are part of the gut flora and also contains a wide range of pathogenic species, showing a more diverse functional landscape.

Combining the information from the functional landscape with 16S-rRNA sequences, allowed us to relate the functional diversity with evolutionary distances (Figure 7.8). This analysis revealed that three stages of diversification can be defined (Plata, Henry, and Vitkup, 2015). The first stage represents a fast-short-term evolution, as 16S-rRNA evolutionary distances remain small, though functional diversification has already taken place. This happens in closely, near identical, related strains where gene acquisition could play a significant role in functional diversity. The second stage represents a long-term evolution, in which functional diversification occurs at a scale compatible with evolutionary time, as reflected by 16S-rRNA evolution. In the third stage diversification of the functional landscape continues but, due to 16S-rRNA genetic constraints, does not align well with 16S-rRNA sequence distances.

Conclusions

16S-rRNA similarity scores can still be used as a metric for taxonomic classification but we propose a more probalistic interpretation as its performances will be better at higher taxonomic levels.

Whole genome protein domain phylogenies correlate with, and complement 16S-rRNA sequence-based phylogenies. Moreover, domain-based phylogenies reveal rapid functional diversification, allowing for large scale functional comparisons between clades and can be constructed over large evolutionary distances.

Protein domain persistence ratio's highlight both signifying domain classes and HGT candidates. The increased granularity obtained will pave the way for new applications to better predict the relationships between genotype, physiology and ecology.

Methods

Genome annotation

A total of 5713 publicly available complete bacterial genomes were downloaded from the NCBI repository (November 2016) (Agarwala et al., 2016). To prevent technical bias due to the use of different annotation tools and pipelines and different thresholds for assessing the significance of the inferred genetic elements, genomes were consistently structurally and functionally *de-novo* annotated using SAPP (Jasper J. Koehorst et al., 2017), an annotation platform implementing a strictly defined ontology (Dam et al., 2017).

16S-rRNA prediction was performed using RNAmmer 1.2 (Lagesen et al., 2007). Genes were predicted using Prodigal (2.6.3) (Hyatt et al., 2010) and the identified proteins were functionally annotated using the Pfam library (version 30.0) within InterProScan (version 5.21-60.0) (Robert D. Finn, Attwood, et al., 2017; Robert D. Finn, Coggill, et al., 2016). Annotations were automatically converted into RDF according to the GBOL ontology (Dam et al., 2017) and loaded into a semantic database for high-throughput annotation and analysis. For the retrieval of information, SPARQL was used (See supplementary file S5 for all queries used).

Quality analysis

Scaling laws have been identified in the genomic distribution of protein domains (De Lazzari et al., 2017). These laws result in linear relationships in the number of domain classes with n copies and the total number of domain classes in a genome (See supplementary Figure S5). We have verified the linear relationships in the analysed genomes. These indicators have been used here to further verify the integrity of the assembled genomes (Cosentino Lagomarsino et al., 2009). Overall, the previously reported scaling laws also hold true when a higher number of genomes is studied.

Estimation of pan- and core-domainome size

The estimated number of domain classes in the pan- and core-genomes expected, if the sequences of every existing strain were to be included in the analysis, were computed using binomial mixture models as implemented in the micropan R package (Lars Snipen and Liland, 2018) using default values for the parameters. Heap's analysis as implemented in the micropan R package was used to estimate openness or closeness of the pan-genome using 500 genome permutations and repeating the calculation 10 times.

Domain persistence

The following formulas were used to calculate persistence ratios

$$Persistence = \frac{number\ of\ genomes\ encoding\ the\ domain}{total\ number\ of\ considered\ genomes}$$

$$log\text{-}P = log_2 \frac{clade\ specific\ persistence}{overall\ persistence}$$

16S-rRNA distance calculations

From the *de-novo* annotation, 16S-rRNA sequences were obtained from the semantic database through a SPARQL query (See supplementary file S6 for all queries used). In total 25098 16S-rRNAs were retrieved. rRNA's that were of low quality (containing N's) or differed in size greater than the standard deviation were removed from the analysis. Duplicated 16S-rRNAs were merged into a single copy for the multiple alignment. For each 16S-rRNA the orientation was validated using OrientationChecker (Ashelford et al., 2006). The complete gene was used for calculation of pairwise alignment distances using the clustal omega suite for all possible 16S-rRNA pairs (Dataset 1 aligned). The resulting matrix was binarized using 98.7% sequence similarity as a cutoff. The binary matrix was then represented as networks using igraph (Csardi and Nepusz, 2006) in R (Team, 2013).

Domain based distance calculations

Genome distances based on protein domain class content were computed using the asymmetric binary method in which vectors are regarded as binary bits. Non-zero

elements are on and zero elements are off. The distance is the proportion of bits in which only one is on amongst those in which at least one is on (dist function in R). A similarity cutoff of ≤ 0.1 was used.

Statistical software

Statistical analysis and visualisations were performed using R and the following packages, data.table (Dowle and Srinivasan, 2018), reshape2 (Wickham, 2007), plotly (Sievert, 2018), Biostrings (Pagès et al., 2017), devtools (Wickham, Hester, and Chang, 2018), micropan (Lars Snipen and Liland, 2018), gridExtra (Auguie, 2017), hexbin (Carr et al., 2018) and RColorBrewer (Neuwirth, 2014).

Supplementary files

All supplementary files can be found at `https://doi.org/10.4121/uuid:37346f1e-f8cf-4112-ba4b-1223a3e4edda`. This resource contains the entire RDF resource that was used in this study.

Author contributions

JJK, PJS and MSD participated in the conception and design of the study. JJK was responsible for the analysis. JJK, ES, PJS and MSD wrote the manuscript. All authors critically revised the manuscript.

Acknowledgements

This work was carried out on the Dutch national e-infrastructure with the support of the SURF foundation. This work was partly supported by the European Union's Horizon 2020 research and innovation programme (EmPowerPutida, Contract No. 635536, granted to Vitor A P Martins dos Santos) and the Netherlands Organisation for Scientific Research funded UNLOCK project (NRGWI.obrug.2018.005) and has received funding form the European Union's Horizon 2020 research and innovation programme under grant agreement 730976 (IBISBA 1.0).

Bibliography

Agarwala, Richa et al. (2016). "Database resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 44. DOI: 10.1093/nar/gkv1290.

Ashelford, Kevin E., Nadia A. Chuzhanova, John C. Fry, Antonia J. Jones, and Andrew J. Weightman (2006). "New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras". In: *Applied and Environmental Microbiology* 72. DOI: 10.1128/AEM.00556-06.

Auguie, Baptiste (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3. URL: https://CRAN.R-project.org/package=gridExtra.

Basu, Malay Kumar, Eugenia Poliakov, and Igor B. Rogozin (2009). "Domain mobility in proteins: Functional and evolutionary implications". In: *Briefings in Bioinformatics*. DOI: 10.1093/bib/bbn057.

Carr, Dan, ported by Nicholas Lewin-Koh, Martin Maechler, and contains copies of lattice functions written by Deepayan Sarkar (2018). *hexbin: Hexagonal Binning Routines*. R package version 1.27.2. URL: https://CRAN.R-project.org/package=hexbin.

Chun, Jongsik et al. (2018). "Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes". In: *International Journal of Systematic and Evolutionary Microbiology*. DOI: 10.1099/ijsem.0.002516.

Cosentino Lagomarsino, Marco, Alessandro L Sellerio, Philip D Heijning, and Bruno Bassetti (2009). "Universal features in the genome-level evolution of protein domains." In: *Genome biology* 10. DOI: 10.1186/gb-2009-10-1-r12.

Csardi, Gabor and Tamas Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal* Complex Systems. URL: http://igraph.org.

Dam, Jesse C. J. van, Jasper J. Koehorst, Jon Olav Vik, Peter J. Schaap, and Maria Suarez-Diez (2017). "Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining". In: *bioRxiv*.

De Lazzari, Eleonora, Jacopo Grilli, Sergei Maslov, and Marco Cosentino Lagomarsino (2017). "Family-specific scaling laws in bacterial genomes". In: *Nucleic Acids Research* 45. DOI: 10.1093/nar/gkx510.

Dowle, Matt and Arun Srinivasan (2018). *data.table: Extension of 'data.frame'*. R package version 1.11.4. URL: https://CRAN.R-project.org/package=data.table.

Dutta, Chitra and Archana Pan (2002). "Horizontal gene transfer and bacterial diversity". In: *Journal of biosciences* 27.

Finn, Robert D., Teresa K. Attwood, et al. (2017). "InterPro in 2017-beyond protein family and domain annotations". In: *Nucleic Acids Research* 45. DOI: 10.1093/nar/gkw1107.

Finn, Robert D., Penelope Coggill, et al. (2016). "The Pfam protein families database: Towards a more sustainable future". In: *Nucleic Acids Research* 44. DOI: 10.1093/nar/gkv1344.

Finn, Robert D et al. (2006). "Pfam: clans, web tools and services". In: *Nucleic Acids Research* 34. DOI: 10.1093/nar/gkj149.

Goodacre, Norman F., Dietlind L. Gerloff, and Peter Uetz (2014). "Protein Domains of Unknown Function Are Essential in Bacteria". In: *mBio* 5. Ed. by Claire M. Fraser. DOI: 10.1128/mBio.00744-13. URL: https://mbio.asm.org/content/5/1/e00744-13.

Gupta, Radhey S. (2016). "Impact of genomics on the understanding of microbial evolution and classification: The importance of Darwin's views on classification". In: *FEMS Microbiology Reviews*. DOI: 10.1093/femsre/fuw011.

Hinchliff, Cody E. et al. (2015). "Synthesis of phylogeny and taxonomy into a comprehensive tree of life". In: *Proceedings of the National Academy of Sciences* 112. DOI: 10.1073/pnas.1423041112.

Hutchison, Clyde A. et al. (2016). "Design and synthesis of a minimal bacterial genome". In: *Science*. DOI: 10.1126/science.aad6253.

Hyatt, Doug et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11. DOI: 10.1186/1471-2105-11-119.

Janda, J. Michael and Sharon L. Abbott (2007). *16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls*. DOI: 10.1128/JCM.01228-07.

Kamminga, Tjerko et al. (2017). "Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life". In: *Frontiers in cellular and infection microbiology* 7.

Kim, Mincheol, Hyun Seok Oh, Sang Cheol Park, and Jongsik Chun (2014). "Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes". In: *International Journal of Systematic and Evolutionary Microbiology*. DOI: 10.1099/ijs.0.059774-0.

Koehorst, Jasper J, Edoardo Saccenti, Peter J Schaap, Vitor AP Martins dos Santos, and Maria Suarez-Diez (2016). "Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics". In: *F1000Research* 5.

Koehorst, Jasper J, Jesse CJ Van Dam, et al. (2016). "Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data". In: *Scientific reports* 6.

Koehorst, Jasper J. et al. (2017). "SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles". In: *Bioinformatics* 1.

Konstantinidis, Konstantinos T. and James M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes". In: *Proceedings of the National Academy of Sciences* 102. DOI: 10.1073/pnas.0409727102.

Lagesen, Karin et al. (2007). "RNAmmer: consistent and rapid annotation of ribosomal RNA genes." In: *Nucleic Acids Research* 35. DOI: 10.1093/nar/gkm160.

McDonald, Daniel et al. (2012). "An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea". In: *The ISME Journal* 6. DOI: 10.1038/ismej.2011.139.

Meier-Kolthoff, Jan P, Alexander F Auch, Hans-Peter Klenk, and Markus Göker (2013). "Genome sequence-based species delimitation with confidence intervals and improved distance functions". In: *BMC bioinformatics* 14.

Nakabachi, A. et al. (2006). "The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella". In: *Science* 314. DOI: 10.1126/science.1134196.

Neuwirth, Erich (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. URL: https://CRAN.R-project.org/package=RColorBrewer.

Nikoh, Naruo, Takahiro Hosokawa, Kenshiro Oshima, Masahira Hattori, and Takema Fukatsu (2011). "Reductive evolution of bacterial genome in insect gut environment". In: *Genome Biology and Evolution* 3. DOI: 10.1093/gbe/evr064.

Ochman, Howard, Jeffrey G Lawrence, and Eduardo A Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation". In: *nature* 405.

Ochman, Howard, Emmanuelle Lerat, and Vincent Daubin (2005). "Examining bacterial species under the specter of gene transfer and exchange". In: *Proceedings of the National Academy of Sciences* 102.

Pagès, H, P Aboyoun, R Gentleman, and S DebRoy (2017). "Biostrings: Efficient manipulation of biological strings". In: *R Package Version* 2.

Plata, Germán, Christopher S Henry, and Dennis Vitkup (2015). "Long-term phenotypic evolution of bacteria". In: *Nature* 517.

Quast, Christian et al. (2013). "The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools". In: *Nucleic Acids Research* 41. DOI: 10.1093/nar/gks1219.

Sievert, Carson (2018). *plotly for R*. URL: https://plotly-book.cpsievert.me.

Snipen, Lars, Trygve Almøy, and David W Ussery (2009). "Microbial comparative pan-genomics using binomial mixture models". In: *BMC Genomics* 10. DOI: 10.1186/1471-2164-10-385.

Snipen, Lars and Kristian Hovde Liland (2018). *micropan: Microbial Pan-Genome Analysis*. R package version 1.2. URL: https://CRAN.R-project.org/package=micropan.

Snipen, LG and DW Ussery (2013). "A domain sequence approach to pangenomics: applications to Escherichia coli [version 2; referees: 2 approved]". In: *F1000Research* 1.19. DOI: 10.12688/f1000research.1-19.v2.

Sonnhammer, Erik LL, Sean R Eddy, Ewan Birney, Alex Bateman, and Richard Durbin (1998). "Pfam: multiple sequence alignments and HMM-profiles of protein domains". In: *Nucleic acids research* 26.

Stackebrandt, E. and B. M. Goebel (1994). "Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology". In: *International Journal of Systematic and Evolutionary Microbiology* 44. DOI: 10.1099/00207713-44-4-846.

Team, R Core (2013). “R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.” In: URL: http://www.r-project.org/.

Wang, Qiong, George M. Garrity, James M. Tiedje, and James R. Cole (2007). “Na??ve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy”. In: *Applied and Environmental Microbiology* 73. DOI: 10.1128/AEM.00062-07.

Wang, Xiang Jing et al. (2010). *Genome sequence of the milbemycin-producing bacterium Streptomyces bingchenggensis*. DOI: 10.1128/JB.00596-10.

Weisburg, W. G., S. M. Barns, D. A. Pelletier, and D. J. Lane (1991). “16S ribosomal DNA amplification for phylogenetic study”. In: *Journal of Bacteriology* 173. DOI: 10.1128/JB.173.2.697-703.1991.

Wickham, Hadley (2007). “Reshaping Data with the reshape Package”. In: *Journal of Statistical Software* 21. URL: http://www.jstatsoft.org/v21/i12/.

Wickham, Hadley, Jim Hester, and Winston Chang (2018). *devtools: Tools to Make Developing R Packages Easier*. R package version 1.13.6. URL: https://CRAN.R-project.org/package=devtools.

Wolf, Yuri I. and Eugene V. Koonin (2013). “Genome reduction as the dominant mode of evolution”. In: *BioEssays* 35. DOI: 10.1002/bies.201300037.

Yang, S., R. F. Doolittle, and P. E. Bourne (2005). “Phylogeny determined by protein domain content”. In: *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.0408810102.

Yarza, Pablo et al. (2014). “Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences”. In: *Nature Reviews Microbiology* 12. DOI: 10.1038/nrmicro3330.

Yilmaz, Pelin et al. (2014). “The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks”. In: *Nucleic Acids Research* 42. DOI: 10.1093/nar/gkt1209.

Yoon, Seok Hwan et al. (2017). “Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies”. In: *International Journal of Systematic and Evolutionary Microbiology* 67. DOI: 10.1099/ijsem.0.001755.

Zmasek, Christian M. and Adam Godzik (2012). "This Déjà Vu Feeling-Analysis of Multidomain Protein Evolution in Eukaryotic Genomes". In: *PLoS Computational Biology*. DOI: 10.1371/journal.pcbi.1002701.

8 Systematic function-based genome prospecting for industrial traits applied to 1,3-propanediol production

Jasper J. Koehorst*, Nikolaos Strepis*, Alfons J. M. Stams,

Diana Z. Sousa, Peter J. Schaap

* Equal contribution

Abstract

Due to the success of next-generation sequencing, there has been a vast build-up of sequenced microbial genomes in the public repositories. For bioprospecting of this huge genomic potential for biotechnological benefiting, new efficient and flexible methods need to be developed. In this study Semantic Web techniques are applied to develop a function-based genome mining approach following a knowledge and discovery in database protocol. Focusing on the industrial important trait of 1,3-propanediol production 178 new candidate species were identified. Furthermore, the genetic architecture of the trait was resolved, and essential domains identified. Three newly identified non-pathogenic strains were successfully tested for 1,3-propanediol production.

Introduction

Next generation sequencing technologies (NGS) have turned the publicly available genome repositories into data-rich scientific resources that currently contain structural and functional genome information of thousands of bacterial genomes (Silvester et al., 2018). For biotechnological benefits, these repositories are excellent resources to mine for new and alternative cell factories, industriophilic traits and enzymes.

While the biotechnology field has embraced Omics technologies impacting biotech innovation in multiple ways (Gates, 2000), setting up large scale functional screenings in these genomics resources for industriophilic traits is still challenging. NGS data generation has caused biocuration to be outpaced rapidly and currently more than 99% of the functional predictions in UniProt are based on automated computational predictions (UniProt Consortium, 2018). The quality of computationally inferred functional genome annotations varies due to lack of data and element-wise provenance, the use of different annotation pipelines, the continuous updates of the reference databases used, generic, non-standardized, annotation acceptance thresholds and an inconsistent naming of protein functions, all adding to a lower degree of interoperability (Jasper J Koehorst, Van Dam, et al., 2016). Since structural annotations, predicted gene and protein sequences are presented in a highly standardized format resulting in a much higher degree of interoperability. The 'bottom-up' approaches starting with a gene or protein sequence in an effort to find a corresponding function in genome(s) of interest are normally used. However, over larger phylogenetic distances, gene and protein sequence-based clustering algorithms are hampered by lateral gene transfer, gene fusion/fission events and domain expansions. Furthermore, at larger scales, they suffer from high computational cost as time and memory requirements scale quadratically with the number of genome sequences to be compared (Wall et al., 2010). A systematic robust function-based genome screening procedure requires a high degree of semantic Interoperability. This means that functional information can be directly compared on the basis of a pre-established syntactic interoperable genome annotation and computational predictions are linked to their provenance. To accomplish this, we recently have devel-

oped SAPP, a semantic annotation infrastructure supporting FAIR computational genomics (Jasper J. Koehorst et al., 2017). SAPP uses the GBOL ontology as syntax (Dam et al., 2017) and automatically predicts, tracks and stores structural and functional genome predictions and associated dataset- and element-wise provenance in a Linked Data format. For a systematic presentation of protein functions, protein domain architectures are used as proxy (Jasper J Koehorst, Saccenti, et al., 2016). Demonstrating a high level of scalability, this set up has been successfully used in an integrated analysis of the functional landscape of 432 Pseudomonas strains (Jasper J Koehorst, Van Dam, et al., 2016).

In this study Interoperable genome annotations are used in a systematic function-based in silico screening for bacterial species that can convert the bio-refinery by-product glycerol into the industrial high-value monomer 1,3-propanediol (1,3-PD) (Saxena et al., 2009; Jiang et al., 2016) 1,3-PD is an important precursor of biomaterials. It is currently used as a monomer for novel polyester and biodegradable plastics, such as polytrimethylene terephthalate (Saxena et al., 2009). 1,3-PD is a typical product of glycerol fermentation and currently very few species, mostly enterobacteria, are known to form it (Barbirato et al., 1996). The underlying metabolic trait, a two-step reductive conversion of glycerol to 1,3-PD regenerates NAD+, required for the oxidative conversion of glycerol to dihydroxyacetone (Figure 8.1).

The first step, dehydration of glycerol to 3-hydroxypropionaldehyde (3-HPA) is mediated by a vitamin B12-dependent glycerol dehydratase although an oxygen sensitive B12-independent alternative enzyme has been reported (Raynaud et al., 2003). The second step reduces 3-HPA to 1,3-PD using the (NADH)+H+ dependent 1,3-propanediol-oxydoreductase (PDOR) regenerating NAD+ (Jiang et al., 2016).

For genome prospecting a collection of 84,300 publicly available bacterial genomes were loaded in the SAPP semantic framework, structurally and functionally annotated, and mined for candidates to produce 1,3-PD using a knowledge discovery in databases approach (KDD) (Ristoski and Paulheim, 2016). Overall the systematic analysis increased our knowledge on the genetic architecture of this metabolic trait, in terms of the overall domain composition, distribution and essentiality. The approach suggested that, compared to some 30 producers in litera-

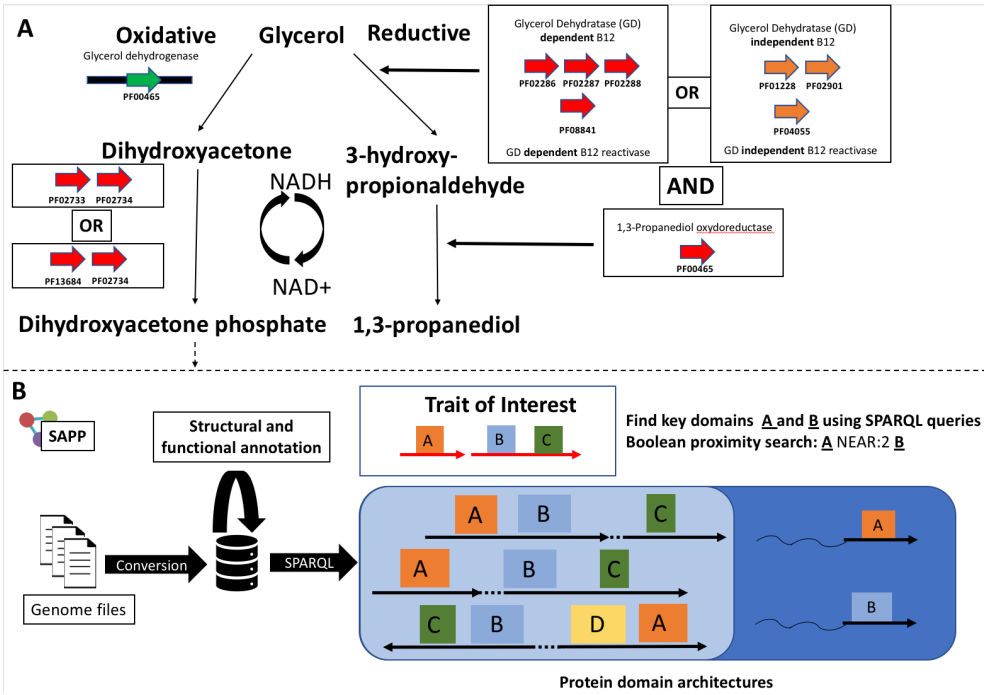

Figure 8.1: ***Search strategy for 1, 3-propanediol candidate species*** *(A) Key domains in the main pathway for B12 dependent bioconversion of glycerol to 1,3-propanediol are indicated in red. Key domains for the alternative B12-independent reductive pathway are indicated in orange. Note the generic iron alcohol domain is used in both the oxidative and reductive branch but not included in the search strategy for the oxidative branch (indicated in green). (B) Generalized functional based search strategy for traits using SAPP. Genome sequence in standard format are converted to an RDF database and complemented with structural and functional annotation. SPARQL, search strategies are deployed to identify domains of interest (dark blue) and complemented with proximity searches (light blue) to find key domain enriched regions.*

ture (Table 8.1), at least 187 genome sequenced species potentially have the trait for 1,3-PD production. Three newly identified non-pathogenic species, *Acetobacterium wieringae*, *Clostridium magnum* and *Carnobacterium Funditum* were experimentally validated for 1,3-PD production. When grown in glycerol, *Clostridium magnum* and *Carnobacterium Funditum* produced 1,3-PD as the main product.

Table 8.1: Strains capable of producing 1,3-Propanediol

Organism	Genome assembly ID	in silico prediction (B12)*	1,3-PD (mol/mol)	Acetate (mol/mol)	Reference
Caloramator viterbensis DSM 13723T			0.69	0.21	(Seyfried et al., 2002)
Citrobacter freundii ATCC 8090**	GCA_000312465.1	Dependent	0.65	0.22	(Fabien Barbirato et al., 1998)
	GCA_000734905.1	Dependent			
Clostridium beijerinckii DSM 791			0.55		(Pradima, Kulkarni, et al., 2017)
Clostridium beijerinckii NRRL B-593	GCA_002006325.1	—	0.59	0	(Gungormusler, Gonen, and Azbar, 2011)
Clostridium butyricum AKR102a			0.52	N.D.	(Wilkens et al., 2012)
Clostridium butyricum CNCM 1211			0.64	0.11	(Fabien Barbirato et al., 1998)
Clostridium butyricum DSM 5431			0.51	N.A.	(Matsumura, Nomura, Sato, et al., 2008)
Clostridium butyricum E4	GCA_000182605.1	—	0.58	0.03	(Petitdemange et al., 1995)
Clostridium butyricum E5			0.54	0.08	(Petitdemange et al., 1995)
Clostridium butyricum F2b			0.63	0.13	(Papanikolaou, Fick, and Aggelis, 2004)
Clostridium butyricum VPI 1718			0.68	0.04	(Chatzifragkou et al., 2011)
Clostridium butyricum VPI3266	GCA_000409755.1	—	0.6	0.01	(González-Pajuelo, Andrade, and Vasconcelos, 2004)
Clostridium diolis DSM 15410	GCA_000409695.1	—	0.58	0.12	(Otte et al., 2009)
Clostridium multifermentans DSM 5431			0.4	0.01	(Biebl et al., 1999)
Clostridium pasteurianum DSM 525	GCA_000330945.1	Dependent	0.5	0.12	(Taconi, Venkataramanan, and Johnson, 2009)
Clostridium perfringens str. 13	GCA_000009685.1	Dependent	N.A	N.A.	(Sun et al., 2003)
Halanaerobium saccharolyticum DSM 6643	GCA_000350165.1	Dependent	0.41	0.26	(Kivistö, Santala, and Karp, 2012a)
Klebsiella michiganensis strain M5al	GCA_000308735.2	Dependent	0.41	0.03	(Yang, Tian, and J. Li, 2007)
Klebsiella pneumoniae ATCC 25955	GCA_000409715.1	Dependent	0.65	0.22	(Fabien Barbirato et al., 1998)
Klebsiella pneumoniae DSM 2026	GCA_000949515.1	Dependent	0.62	0.09	(Mu et al., 2006)
Klebsiella pneumoniae DSM 4799			0.45	N.D.	(Jun et al., 2010)
Klebsiella pneumoniae HR526			0.4	N.D.	(Xu et al., 2009)
Klebsiella pneumoniae XJ-Li			0.61	0.05	(Zhang et al., 2007)
Lactobacillus brevis N1E9.3.3			0.46	0.07	(Vivek, Pandey, and Binod, 2018)
Lactobacillus buchneri B 190			0.27	0.18	(Schütz and Radler, 1984)
Lactobacillus panis PM1					(Kang, Korber, and Tanaka, 2014)
Pantoea agglomerans CNCM-1210			0.61	0.16	(Fabien Barbirato et al., 1998)
Raoultella planticola DR3	GCA_000735435.1	—	N.D.	0.15	(Jarvis, Moore, and Thiele, 1997)
Shimwellia blattae ATCC 33430			0.45		(Rodriguez et al., 2016)
Trichococcus pasteurii	GCA_900112935.1	Dependent	0.??		(Strepis et al., 2016)
Trichococcus sp. ES5	GCA_900067165.1	Dependent	0.??		(Antonie H van Gelder et al., 2012)

* Dependent: Presence of protein domains required for B12-dependent 1,3-PD production; —: Absence for these domains suggesting presence of a B12-independent pathway

** Two independent genome assemblies available for this strain

Results

Development of the genome mining workflow

To cope with the huge amount of biological input data, we followed a Knowledge Discovery in Databases (KDD) process. In this systematic multistep process, the actual 'pattern searching' step is preceded by the equally important steps of 'data preparation' and 'incorporation of prior knowledge' (Dawyndt et al., 2006). For metabolic trait discovery, the KDD process was adjusted to the specific needs in bio-prospecting. Data silo and transformation enabling a high level of interoperability, pattern searching, pattern validation and modification using a training data set and data-mining.

Incorporation of prior knowledge:

In 1,3 PD production two parallel pathways are used for dissimilation of glycerol. In the oxidative pathway glycerol is dehydrogenated to dihydroxyacetone by a NAD+-linked glycerol dehydrogenase, then to dihydroxyacetone phosphate by an ATP-dependent dihydroxyacetone kinase (Forage and Lin, 1982). Additionally, two alternative reductive pathways, either dependent on vitamin B12 or not exist for regenerating NAD+ for 1.3-PD production. For validation of the search strategy and to gain further genetic insights the domains involved in 1,3-PD formation a list of known 1,3 PD producing strains was obtained. From the literature some 30 strains have been reported to produce 1,3-PD (Table 8.1). Genome sequences from 14 different strains were obtained from the EBI-ENA data warehouse (Silvester et al., 2018) as not for all of them genome sequences are available. For Citrobacter freundii ATCC8090 two draft genome sequences of comparable quality could be obtained and initially both were kept increasing the number of genomes to 15. For four strains in this set the 1,3 PD operon has previously been molecularly characterised (Figure 8.2A). In total, 12 strains were selected for the training dataset and *Trichococcus pasteurii* was used for cross-validation (Table 8.1).

Data transformation:

To obtain a high degree of interoperability and for inclusion of data provenance the 13 genome sequences were de novo structurally and functionally annotated through the SAPP framework using the modules, Prodigal (Hyatt et al., 2010) for gene prediction and InterProScan (Finn, Attwood, et al., 2017) for protein annotation. Genome annotations were exported as Linked Data in graph database using the Resource Description Framework (RDF) as a data-metadata model (Klyne and Carroll, 2004; Jasper J. Koehorst et al., 2017). Protein domain architectures were used as proxy for protein function (Jasper J Koehorst, Saccenti, et al., 2016). As no differences was observed between the functional annotations of two *Citrobacter freundii* ATCC8090 draft genome sequences they were treated as one.

Development of a function-based search strategy for the oxidative branch:

The first two enzymes essential for the oxidative pathway involved 1,3 PD production are glycerol dehydrogenase which is dependent on NADH and produces dihydroxyacetone (DHA), and DHA kinase which phosphorylates DHA. The four Pfam domains describing these key reactions are Fe-ADH (PF00465) for glycerol dehydrogenase and DAK1, (PF02733) or DAK1_2, PF13684) both capturing the kinase domain of the dihydroxyacetone kinase family in combination with DAK2, (PF02734), capturing phosphatase domain of the dihydroxyacetone kinase family (Figure 8.1). A SPARQL query (see methods section for details) for DAK2 domain neighbors in the training set database followed by a Boolean DAK1/DAK2 or DAK1_2/DAK2 proximity search in the search results, showed that the genomes of all 1,3 PD producing species in the training data set encoded at least one dihydroxyacetone kinase family protein with a DAK1/DAK2 or DAK1_2/DAK2 protein architecture; five strains coded for DAK1/DAK2 and four strains DAK1_2/DAK2. Three Clostridia strains, C. perfringens str. 13, C. pasteurianum DSM 525 and C. diolis DSM 15410 encode both configurations. Another observation is the high persistency and copy number of the Pfam domain family of Iron-containing alcohol dehydrogenase/aldehyde reductases (PF00465) with on average 16.6 copies per genome. Due to the inherent low discriminative power of this domain, it was decided to only use a

DAK1/DAK2 OR DAK1_2/DAK2 proximity and not to include PF00465 in the large scale functional screening as preselection for the oxidative branch.

Development of a function-based search strategy for the reductive branch:

For 1,3-PD production two alternative pathway are known, a well-studied B12-dependent pathway and a lesser studied oxygen sensitive B12-independent pathway (Figure 8.1). A series of SPARQL queries (see methods section for details) were used to obtain the persistency and copy number of key domains in the two alternative pathways among the 12 genomes of the training set. The iron containing aldehyde reductase/alcohol dehydrogenase domain (PF00465), is present in both the oxidative and reductive branch and functions in the reductive branch as aldehyde reductase. The three dehydratase Pfam domains signifying the B12-dependent pathway, PF02287 (small subunit), PF02288 (medium subunit), and PF02286 (large subunit) have a persistency of 0,6 with on average per genome 1,7 (large and small subunit) or 3,7 copies (medium subunit). The isofunctional dioldehydratase reactivase ATPase-like domain PF08841 has a persistency of 0,6 and a copy number of one. Note that only four strains in the training data set are molecularly characterized. The low level of persistency of key domains of the B12-dependent pathway in the training set therefore immediately suggests that multiple strains actually may not use the B12-dependent pathway. The B12-independent glycerol dehydratase is reported to be involved in 1,3-PD production and consists of a Glycine radical (PF01228) and a pyruvate formate lyase-like domain (PF02901) (Raynaud et al., 2003). Both domains are present in all twelve genomes of the training data set with a relative high copy number of six. The persistency of the Radical SAM superfamily domain, a key part of the B12-independent glycerol dehydratase reactivase enzyme (Demick and Lanzilotta, 2011), was also one with an average copy number of 32 indicating that these three domains are promiscuous and normally also used to support other functionalities (Table 8.2).

As key domains of both the B12-dependent and independent pathway were either too generic indicated by a high copy number or were associated with a low level

Table 8.2: Properties of key domains involved in glycerol dissimilation in 1,3 PD producers

Domain	Mean Copy Number*	Proximity Search query (compounds and distance)
	Oxidative pathway	
PF02733 (DAK1) PF13684 (DAK1_2) PF02734 (DAK2)	2.2 1.6 2.7	DAK1 AND DAK2 OR DAK1_2 AND DAK2 (immediately adjacent)
	B12-dependent reductive Pathway	
PF00465 PF02286 PF02287 PF02288 PF08841	16.6 1.7 1.7 3.7 1.7	All, within 20.000 up or downstream of the B12 dependent dehydratase domains
	B12-independent reductive Pathway	
PF01228 PF02901 PF04055	13.8 14.4 32.4	All, within 20.000 up or downstream of the B12 independent dehydratase domains

of persistency, a Boolean multi-compound proximity search was developed for the reductive branch. In this approach the physical co-localization of key domains in the respective genomes is included in the search. From the list of natural producers with a molecularly characterized 1,3 PD operon, the size of 1,3-PD operon was estimated to encompass approximately 18,000 nucleotides. Furthermore, signifying domains can be found on both strands due to the use of internal promoters (Figure 8.2A). Including this, the search criteria were set such that the region of interest for B12-dependent pathway should contain at least five signature domains (PF00465, PF02286, PF02287, PF02288, PF08841) in a window of 40,000 base pairs extending 20.000 up and downstream of the dehydratase domains. Based on the single characterized B12-independent strain four domains were specified (PF00465, PF01228, PF02901, PF04055) extending 20.000 up and downstream of the dehydratase domains. Furthermore, taking internal promoters into account, domains were allowed to be present on both strands.

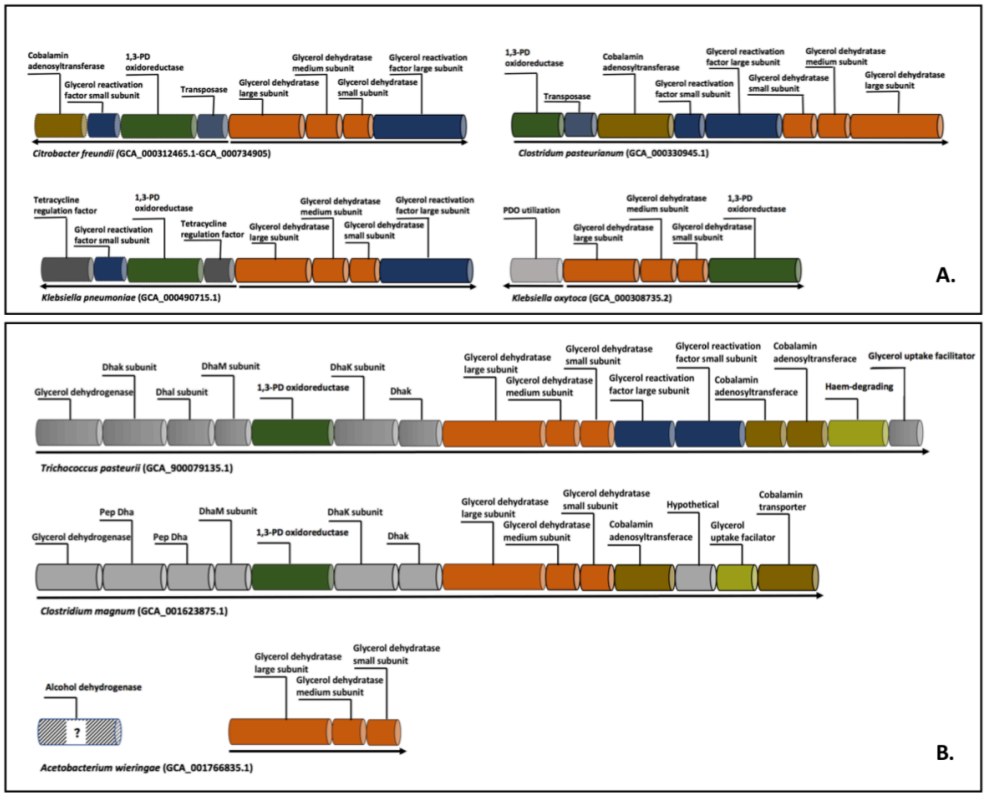

Figure 8.2: ***Representation of operonic structures for the metabolic trait of 1,3-propanediol production.*** *(A) Genetic architecture of the 1,3 PD operon of previously molecularly characterized species. (B) Genetic architecture of the 1,3 PD operon of selected species. The colored blocks represent genes Arrows indicate direction of transcription. Species names are indicted with genome sequences in brackets. Note that A. wieringae operonic structure does not include a 1,3-PD dehydrogenase*

Applying these criteria to the training data set resulted in the identification of three strains containing B12-dependent operon signatures; *Citrobacter freundii* ATCC8090, *Klebsiella pneumoniae* DSM2026, and *Clostridum perfringens*; four strains containing the B12-dependent as well as B12-independent operon signatures; *Clostridum pasteurianum* DSM525, *Halanaerobium saccharolyticum* DSM6643, *Klebsiella michiganensis*, and *Klebsiella pneumoniae* ATCC25955) (Figure 8.3); and five strains solely containing the independent operon signatures; *Clostridium beijerinckii* NRRL B-593, the canonical B12 independent strain *Clostridum butyricum* DSM10702, *Clostridum butyricum* E4, *Clostridum diolis* DSM15410 and *Raoultella*

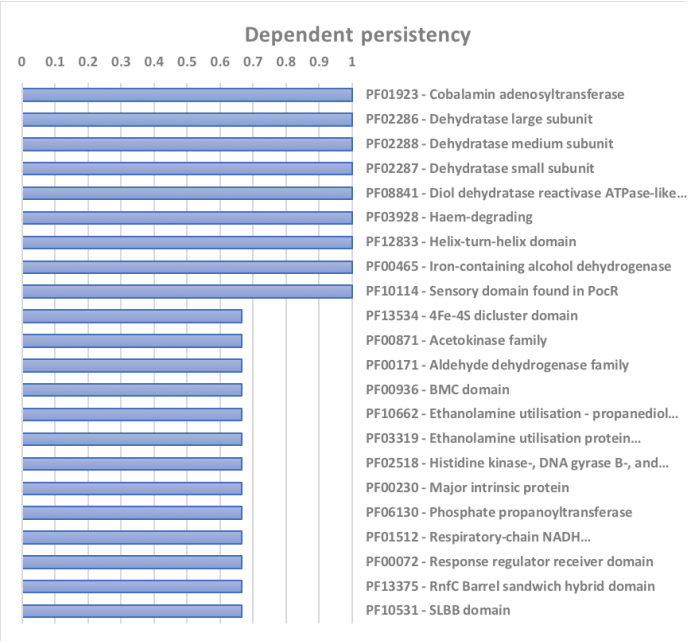

Figure 8.3: ***Overall domain composition of the B12-dependent 1,3 PD operon derived from the training data set.*** *Input strains are in Table 8.1*

planticola DR3 (Figure 8.4). Figure 8.3 displays the overall domain composition of the B12-dependent 1,3 PD operon in the training set. Including the five signifying domains used in the query, the core of the B12-dependent 1,3-PD operon consists of nine domains Other highly persistent domains enriched within the syntenic region are 'Cobalamin adenosyltransferase', (PF01923), 'Haem-degrading' (PF03928),'Helix-turn-helix domain' (PF12833) and 'sensory domain found in PocR' (PF10114).

Figure 8.4 displays the overall domain composition of the B12-independent 1,3 PD operon derived from the training set. Beside the four specified domains the core was expanded by the '4Fe-4S single cluster domain' (PF13353). Other domains highly persistent in B12-dependent operon such as the 'Sensory domain found in PocR' and the 'Helix-turn-helix domain' are also enriched in some of the putative B12-independent operonic structures except in *Clostridium butyricum* DSM 10702 and *Raoultella planticola* DR3.

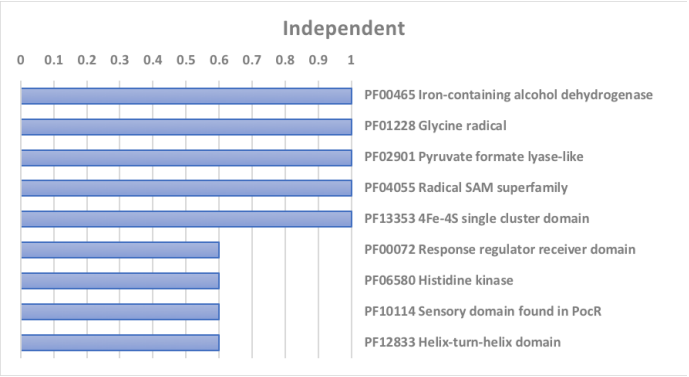

Figure 8.4: ***Overall domain composition of the B12-independent 1,3 PD operon derived from the training data set.*** *Input strains are in Table 8.1*

Data silo

To mine publicly available genomes for the presence of the vitamin B12 1,3-PD operon 84,329 bacterial genomes containing, 51 phyla, 64 classes, 145 orders, 335 families, 1,126 genera and 2,661 species were obtained from the EBI-ENA data warehouse. The SAPP semantic annotation framework was subsequently used for a de novo structural and functional annotation of these genomes resulting in a semantic database of 365,920,933 predicted protein encoding genes linked to the corresponding protein sequences, predicted protein domain architectures and structural and functional prediction provenance. Further analysis was performed on predicted protein encoding genes with a Prodigal confidence score of at least 95%. At this threshold 95.1% of the proteins remain pertaining 98.3% of the assigned protein domains.

Data Mining for candidate 1,3 PD producers using the B12-dependent pathway

Oxidative branch: Following the function-based search strategy outlined above and summarised in Figure 8.2, a proximity search with the two DAK configurations in parallel resulted in a 55% reduction of the search space. The reduced search space consisted of 37.791 genomes with in total 834 different species from 158 genera. The most abundantly present species were Streptococcus pneumoniae, Listeria monocy-

togenes, Klebsiella pneumoniae, each with more than a thousand strains. The reduced search space was used as input for a proximity search of the reductive branch.

Reductive branch: Identification of B12-dependent strains according to the criteria used for known strain validation, resulted in 187 species (4,142 strains) including the two Trychococcus species not included in the training set. Strains of *Listeria monocytogenes* (1,777) and *Klebsiella pneumoniae* (1,605), both pathogenic species were overrepresented. An overview of identified strains and species is Table 8.4.

At species level, domain persistency of the B12-dependent 1,3-PD trait showed a high degree of similarity with the training set. The most frequently observed additional domains showing a local persistency of 0,80 and above are PF03928 (0,97), PF01923 (0,90), PF00936 (0,87), PF00171 (0,84), PF03319 (0,82) (and PF06130 (0,80) (Figure 8.5). PF03928, a haem degrading domain often flanked by PF03319 and PF06130 (in the order of PF03319-PF03928-PF06130), can be of importance for aldehyde reductase which can have different cofactor specificities including haem (Machielsen et al., 2006). PF01923 represents an enzyme catalyzing the conversion of cobalamin (vitamin B12) into one of its coenzyme forms, adenosylcobalamin (coenzyme B12, AdoCbl), which is vital for the functioning of B12-dependent glycerol dehydratase (Knietsch et al., 2003). The compartmentalization domain, PF00936, could be used to encapsulate enzymatic steps important for 1,3-PD production (Zarzycki, Erbilgin, and Kerfeld, 2015). PF00171, has been shown to enhance 1,3-PD production (Lee et al., 2014). The Ethanolamine utilisation protein EutN/carboxysome (PF03319) is involved in the cobalamin-dependent degradation of ethanolamine. PF06130 a signature for phosphate propanoyltransferase is involved in phosphorylation of 3-hydroxypropionyl-CoA to 3-hydroxypropionyl phosphate (Matsakas et al., 2018).

Confirming the *in silico* predictions a number of strains present in Table 8.4 were already reported to be 1,3-PD producers. Using key domains from the B12-dependent reductive pathway ten of these strains were assigned to use the B12-dependent route. Interestingly, for five other strains the *in silico* results suggests that these strains may use the B12-independent route (Table 8.1).

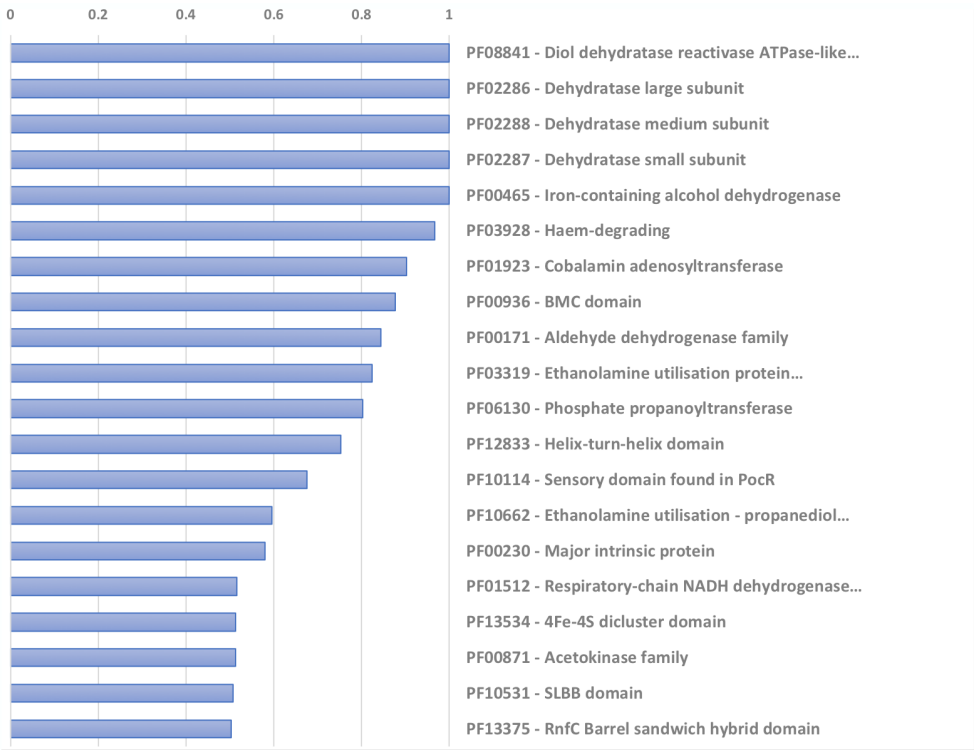

Figure 8.5: DependentPersistencyAll

Experimental validation

Focussing on the B12-dependent pathway, *Acetobacterium wieringae*, *Carnobacterium funditum* DSM 5970 and *Clostridium magnum* DSM 2767 were selected for experimental validation based on their biosafety level. Ca. funditum and Cl. magnum were chosen because they have the complete reductive branch for B12-dependent 1,3-PD production. A. wieringae lacks the key domain PF04055 and therefore the operon presents an incomplete reductive branch as it lacks the aldehyde reductase function. *Trichococcus pasteurii* was selected as a positive control. All strains were grown on glycerol. *Ca. funditum*, *Cl. magnum* and the *T. pasteurii* control strain produced 1,3-PD as the main product with acetate as a byproduct. *A. wieringae* showed a significant reduced growth and produced acetate as a main product (See Table 8.3).

Table 8.3: 1,3-propanediol and acetate yields from glycerol fermentations of selected strains.

Organism	Genome assembly ID	1,3-PD (mol/mol)	Acetate (mol/mol)	OD
Acetobacterium wieringae DSM 1911	GCA_001766835.1	0.18	0.94	0.179
Carnobacterium funditum DSM 5970	GCA_000744185.1	0.33	0.07	0.266
Clostridium magnum DSM 2767	GCA_001623875	0.56	0.03	0.180
Trichococcus pasteurii DSM 2381	GCA_900079135.1	0.66	0.12	0.325

Discussion

Bioprospecting entails the systematic search for economically valuable genetic and biochemical resources from nature (Artuso, 2002). Genome prospecting, the in-silico mining of sequenced genomes and metagenomes for new biotechnologically relevant proteins and enzymes is a relatively new field. In genome prospecting two approaches can be applied, a "top-down" approach that begin by searching for a function and is followed by identification of the corresponding gene(s) and a "bottom-up" approach that start with a gene of interest in an effort to find a corresponding function in genome(s) of interest. For bottom-up approaches many sequence similarity tools such as Blast (Altschul et al., 1997) exist. Bottom-up studies usually start with a selection of (meta)-genome sequences followed by sequence similarity-based clustering and selection of candidate sequences and re-annotation of interesting candidates thereby avoiding ambiguity related problems in current functional annotations. For bacterial species however, a priori, gene fusion-fission events can be expected (Pasek, Risler, and Brezellec, 2006) hampering sequence similarity-based detection and clustering of multi-domain proteins encoding genes, while the same domains may also be present in multiple proteins meaning that first-matches genomic sequences in a sequence similarity search may not encode the wanted function. When searching for polygenic traits these problems are aggravated and a function-based approach searching for key functions may be more effective especially when there is insufficient understanding of the genetic architecture of trait in terms of the minimal number of genes required for the trait and function of the domains encoded by the different genes. Protein domains have been shown to provide an accurate representation of the functional capabilities of a protein (Jasper J Koehorst, Saccenti, et al., 2016; Haft et al., 2017). To overcome ambiguity related problems in functional annotations SAPP identifies and annotates protein function

based on domain architecture. As profile hidden Markov Models (HMM), which favour in their scoring functionally important sites, are used for the identification of protein domains, statistically robust annotation profiles can be obtained over large phylogenetic distances. KDD is a multi-step process involving data preparation and transformation, pattern searching, evaluation and iteration after modification. By using protein domain architectures as proxy for protein functions a high level of standardization is obtained. As protein domain architectures can be directly transformed in highly interoperable strings of Pfam domain identifiers "top-down" functional screenings can be done efficiently. Applying the KDD approach on Linked Data allows for validation of initial results in multiple ways and to iterate after modification. By using molecular knowledge obtained from molecular characterizations of the 1,3 PD operon of four species and validating and iterating the search pattern on a training data set of genome sequences of twelve known 1,3-PD producers, a large collection of 84,300 publicly available bacterial could be efficiently mined in a top-down approach yielding 178 new candidate species and successful experimental verification of three of these species. The additional finding of *Clostridium pasteurianum*, *Halanaerobium saccharolyticum* and *Lactobacillus diolivorans* / *panis* / *reuteri* was reconfirmed through literature (Luers et al., 1997; Kivistö, Santala, and Karp, 2012b; Pflügl et al., 2012; Khan et al., 2013; Amin et al., 2013). Of known producers, such as *Clostridium butyricum*, the 5 strains that were analyzed all were identified as B12-independent. Driven by a strong need to be able to integrate and analyze biodata across databases, there has been a considerable increase in the adoption of Semantic Web technologies in the life-sciences (Cheung et al., 2009). However, a SPARQL endpoint for phenotypic data is only available via WikiData (Mitraka et al., 2015). Other resources for phenotypic data such as BacDive, here used to obtain biosafety levels of candidate species, do provide API's to mine their data but currently cannot be directly queried using SPARQL. With the growing importance of Semantic Web technologies for the life sciences interoperability levels on all aspect will increase, enhancing mining possibilities, aiding the discovery of new traits in an unprecedented pace.

Conclusions

Through transformation of genomic data into a FAIR linked-data format, iterative function-based approaches can be developed to mine the large genome repositories. By presenting functional annotation as unambiguous protein domain architectures a high level of interoperability is obtained allowing for the development of efficient function-based top-down searches not limited to supervised trait identification.

Materials and methods

Data annotation and mining

Bacterial genomes were downloaded from the ENA database using the enaBrowserTools (Silvester et al., 2018). All downloaded microorganisms were converted into a semantic repository using the annotation platform based on functional analysis (SAPP) (Jasper J. Koehorst et al., 2017). All genomes were de-novo structurally re-annotated using Prodigal (Hyatt et al., 2010) and functionally annotated using Pfam from InterProScan (Finn, Bateman, et al., 2014; Finn, Attwood, et al., 2017). Each genome was stored as an individual database and were queried using the SAPP HDTQuery module which is a combination of Apache SPARK, HDT and SPARQL (Fernández et al., 2013; Prud'hommeaux and Seaborne, 2008).

Oxidative branch identification

To identify strains containing the oxidative branch as shown in Figure 8.1, DAK filtering was applied. Genes containing DAK1 (PF02733) or DAK1_2 (PF13684) in combination with DAK2, (PF02734), 95% gene confidence according to prodigal and neighbor linking was applied to identify strains capable for growth on glycerol (Figure 8.1 with 95% cutoff).

Reductive branch identification

Strains that contained the oxidative branch were further investigated for the reductive branch using domains according to the B12 dependent or independent pathway (Figure 8.1 and 8.1). Region identification was performed using PF02287 and 20.000

bp up and downstream. Regions containing all domains of interest were further investigated for domain occurrence and syntheny.

B12 synthesis identification

Phylogenetic information was complemented with phenotypic information through BacDive (Söhngen et al., 2016). The BacDive resource was parsed and each entry record was transformed into RDF allowing integration of genotypic and phenotypic information. SPARQL queries were used to retrieve information with regards to pathogenicity and temperature (Figure 8.2 and 8.3).

Physiological analysis

Clostridium magnum DSMZ 2767 (Uhlig et al., 2016) and *Acetobacterium wieringae* DSMZ 1911 (Poehlein et al., 2016) were obtained from the German Collection for Microorganisms and Cell Cultures (DSMZ, Braunschweig, Germany). *Trichococcus* strain ES5, was previously obtained from a methanogenic bioreactor sludge by (Antonie H. van Gelder et al., n.d.). The anaerobic growth media for C. magnum and A. wierignae was prepared as previously described (A. J.M. Stams et al., 1993) and complemented with 1g/L yeast extract. Additionally, C. magnum required 0,5g/L of cysteine as a reducing agent. The inoculation of all strains was 5% of total serum bottle. *C. magnum*, *A. wieringae* and *Trichococcus* strain ES5 were grown on 20mM glycerol. Yield was measured as 1,3-PD concentration divided by utilized glycerol. Growth measurements were based on optical density (OD) using a spectrometer (Hitachi U-1500, Labstuff, The Netherlands). All soluble substrates and intermediates were measured with an Agilent HPLC system equipped with Agilent Metacarb 67H column (3006.5 mm) (Thermo Fisher Scientific, MA) and a refractive index detector. The OD and HPLC measurements were conducted in time points of 0, 24 and 48 hours.

Table 8.4: Species identified having a B12 dependent reductive branch

Eubacterium hallii
Anaeromusa acidaminophila
Anaerosalibacter massiliensis
Bacillaceae bacterium MTCC 10057
Bacillus azotoformans
Bacillus massiliosenegalensis
Bacillus sp. 7504-2
Bacteroides fragilis
Blautia obeum
Blautia schinkii
Blautia sp. UBA2945
Carnobacteriaceae bacterium UBA7837
Carnobacterium alterfunditum
Carnobacterium funditum
Citrobacter amalonaticus
Citrobacter braakii
Citrobacter freundii
Citrobacter freundii
Citrobacter koseri
Citrobacter portucalensis
Citrobacter rodentium
Citrobacter sp. A1
Citrobacter sp. CFSAN044567
Citrobacter sp. KTE151
Citrobacter sp. KTE30
Citrobacter sp. KTE32
Citrobacter sp. L17
Citrobacter werkmanii
Clostridia bacterium UC5.1-2H11
Clostridiaceae bacterium BRH_c20a
Clostridiales bacterium DRI-13
Clostridiales bacterium mt11
Clostridiales bacterium PH28_bin88
Clostridiales bacterium VE202-03
Clostridium baratii
Clostridium botulinum
Clostridium drakei
Clostridium estertheticum
Clostridium lundense
Clostridium magnum
Clostridium pasteurianum
Clostridium perfringens
Enterobacter asburiae
Enterobacter cloacae
Enterobacter cloacae complex 'Hoffmann cluster IV'
Enterobacter cloacae complex sp. ECNIH6
Enterobacter cloacae complex sp. ECNIH7
Enterobacter cloacae complex sp. GN02468
Enterobacter hormaechei
Enterobacter kobei
Enterobacter sp. 10-1
Enterobacter sp. Bisph2
Enterobacter sp. GN02600
Enterobacter sp. MGH 33
Enterobacter sp. WCHECl-C4
Enterobacteriaceae bacterium UBA2606
Enterobacteriaceae bacterium UBA3163
Enterobacteriaceae bacterium UBA3516
Enterobacteriaceae bacterium UBA4747
Enterobacteriaceae bacterium UBA5935
Enterobacteriaceae bacterium UBA869
Enterobacteriaceae bacterium UBA898
Enterococcus avium
Enterococcus malodoratus
Enterococcus mundtii
Enterococcus pseudoavium
Enterococcus raffinosus
Enterococcus sp. 3H8_DIV0648
Enterococcus sp. HMSC066C04
Enterococcus sp. HMSC072H05
Enterococcus sp. HMSC29A04
Escherichia coli
Escherichia fergusonii
Eubacteriaceae bacterium CHKCI004
Eubacterium sp. 14-2
Eubacterium sp. An11
Eubacterium sp. An3
Eubacterium sp. UBA3279
Eubacterium sp. UBA3326
Eubacterium sp. UBA7134
Firmicutes bacterium UBA3567
Firmicutes bacterium UBA3570
Firmicutes bacterium UBA3573
Flavonifractor plautii
Flavonifractor sp. An306
Flavonifractor sp. An82
Fusobacteriaceae bacterium UBA2433
Fusobacterium hwasookii
Fusobacterium nucleatum
Fusobacterium sp. CM1
Fusobacterium sp. CM22
Fusobacterium sp. HMSC064B11
Fusobacterium sp. HMSC073F01
Fusobacterium sp. OBRC1
Fusobacterium ulcerans
Fusobacterium varium
Geobacillus sp. (strain Y4.1MC1)
Geobacillus thermoglucosidasius
Geosporobacter ferrireducens
Halanaerobium hydrogeniformans
Halanaerobium saccharolyticum
Intestinimonas butyriciproducens
Klebsiella aerogenes
Klebsiella michiganensis
Klebsiella oxytoca
Klebsiella pneumoniae
Klebsiella pneumoniae
Klebsiella quasipneumoniae
Klebsiella sp. 10982
Klebsiella sp. AA405
Klebsiella sp. AS10
Klebsiella sp. HMSC09D12
Klebsiella sp. HMSC16A12
Klebsiella sp. HMSC22F09
Klebsiella sp. KGM-IMP216
Klebsiella sp. KTE92
Klebsiella sp. LTGPAF-6F
Klebsiella sp. M5al
Klebsiella sp. OBRC7
Klebsiella sp. PO552
Klebsiella sp. S1
Klebsiella variicola
Klebsiella variicola CAG:634
Kluyvera intermedia
Lachnospiraceae bacterium 3-1
Lachnospiraceae bacterium 7_1_58FAA
Lachnospiraceae bacterium A2
Lactobacillus brevis
Lactobacillus collinoides
Lactobacillus coryniformis
Lactobacillus curvatus
Lactobacillus diolivorans
Lactobacillus fuchuensis
Lactobacillus ginsenosidimutans
Lactobacillus graminis
Lactobacillus kimchicus
Lactobacillus kisonensis
Lactobacillus mellifer
Lactobacillus namurensis
Lactobacillus panis
Lactobacillus paracollinoides
Lactobacillus pobuzihii
Lactobacillus rapi
Lactobacillus rennini
Lactobacillus reuteri
Lactobacillus rossiae
Lactobacillus silagei
Lactobacillus siliginis
Lactobacillus similis
Lactobacillus spicheri
Lactobacillus versmoldensis
Listeria innocua
Listeria ivanovii
Listeria monocytogenes
Listeria seeligeri
Listeria welshimeri
Metakosakonia massiliensis
Mycobacterium vaccae
Paraclostridium benzoelyticum
Paraclostridium bifermentans
Pediococcus acidilactici
Pediococcus claussenii
Pediococcus pentosaceus
Phytobacter ursingii
Pluralibacter gergoviae
Propionibacterium freudenreichii
Quasibacillus thermotolerans
Salmonella choleraesuis
Sebaldella termitidis
Serratia marcescens
Shigella boydii
Shigella flexneri
Shigella sonnei
Streptococcus australis
Streptococcus sanguinis
Streptococcus sp. AS14
Streptococcus sp. F0442
Streptococcus suis
Thermincola ferriacetica
Thermincola potens
Thermoanaerobacter sp. (strain X513)
Thermoanaerobacter sp. (strain X514)
Thermoanaerobacter sp. A7A
Thermoanaerobacterium thermosaccharolyticum
Tissierellia bacterium S5-A11
Trichococcus pasteurii
Trichococcus sp. ES5

```
PREFIX gbol:<http://gbol.life/0.1/>
SELECT DISTINCT ?accession ?accession2 ?gene ?gene2 ?estrand ?endpos ?
    end2pos ?score ?evalue ?score2 ?evalue2
WHERE {
  VALUES ?accession {'PF02287' 'PF01228'}
  ?dnaobject gbol:sample ?sample .
  ?dnaobject gbol:feature ?gene .
  ?gene a gbol:Gene .
  ?gene gbol:location ?geneloc .
  ?geneloc gbol:end ?end .
  ?end gbol:position ?endpos .
  ?gene gbol:transcript ?transcript .
  ?gene gbol:provenance ?geneprov .
  ?gene gbol:exon ?exon .
  ?exon gbol:location ?elocation .
  ?elocation gbol:strand ?estrand .
  ?geneprov gbol:annotation ?geneannot .
  ?geneannot a <http://semantics.systemsbiology.nl/sapp/0.1/Prodigal> .
  ?geneannot gbol:conf ?geneconf .
  ?transcript gbol:feature ?cds .
  ?cds gbol:protein ?protein .
  ?protein gbol:feature ?domain .
  ?domain a gbol:ProteinDomain .
  ?domain gbol:signature ?signature .
  ?domain gbol:provenance ?pprov .
  ?pprov gbol:annotation ?pannot .
  ?pannot gbol:score ?score .
  ?pannot gbol:evalue ?evalue .
  ?signature gbol:accession ?accession .
```

```
  ?dnaobject gbol:feature ?gene2 .
  ?gene2 a gbol:Gene .
  ?gene2 gbol:provenance ?gene2prov .
  ?gene2prov gbol:annotation ?gene2annot .
  ?gene2annot a <http://semantics.systemsbiology.nl/sapp/0.1/Prodigal> .
  ?gene2annot gbol:conf ?gene2conf .
  FILTER(?gene2conf >= 95)
  ?gene2 gbol:location ?gene2loc .
  ?gene2loc gbol:end ?end2 .
  ?end2 gbol:position ?end2pos .
  FILTER(?end2pos > ?endpos - 20000)
  FILTER(?end2pos < ?endpos + 20000)
  ?gene2 gbol:exon ?exon2 .
  ?exon2 gbol:location ?elocation2 .
  ?elocation2 gbol:strand ?estrand .
  ?gene2 gbol:transcript ?transcript2 .
  ?transcript2 gbol:feature ?cds2 .
  ?cds2 gbol:protein ?protein2 .
  ?protein2 gbol:feature ?domain2 .
  ?domain2 a gbol:ProteinDomain .
  ?domain2 gbol:signature ?signature2 .
  ?domain2 gbol:provenance ?pprov2 .
  ?pprov2 gbol:annotation ?pannot2 .
  ?pannot2 gbol:score ?score2 .
  ?pannot2 gbol:evalue ?evalue2 .
  ?signature2 gbol:accession ?accession2 .
}
```

Figure 8.1: Retrieval of regions (40.000 bp) around PF02287 or PF01228

```
PREFIX csb:<http://csb.wur.nl/genome/>
SELECT DISTINCT ?species ?level (COUNT(?level) AS ?count)
WHERE {
    ?entry csb:taxonomy_name ?taxon .
    ?taxon csb:strains ?strains .
    ?strains csb:species ?species .
    ?entry csb:application_interaction ?interaction .
    ?interaction csb:risk_assessment ?risk .
    ?risk csb:biosafetylevel ?level .
} GROUP BY ?species ?level
ORDER BY DESC(?count)
```

Figure 8.2: Biosafety level retrieval of DSMZ to RDF conversion

```
PREFIX csb:<http://csb.wur.nl/genome/>
SELECT DISTINCT ?species ?temp ?tempType (COUNT(?temp) AS ?count)
WHERE {
    ?entry csb:taxonomy_name ?taxon .
    ?taxon csb:strains ?strains .
    ?strains csb:species ?species .
    ?entry csb:culture_growth_condition ?condition .
    ?condition csb:culture_temp ?cultureTemp .
    ?cultureTemp csb:temperaturerange ?temp .
    ?cultureTemp csb:test_type ?tempType .
} GROUP BY ?species ?temp ?tempType
ORDER BY DESC(?count)
```

Figure 8.3: Temperature group retrieval of DSMZ to RDF conversion

Author contributions

JJK, NS and PJS participated in the conception and design of the study. JJK was responsible for the code of the semantic implementation and phenotypic integration. JJK and NS were responsible for the KDD approach and phenotypic analysis. NS performed the experimental validation. JJK, NS and PJS wrote the manuscript. All authors critically revised the manuscript.

ACKNOWLEDGEMENTS

This research was supported by the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement (323009) and by the Gravitation grant (024.002.002) of the Netherlands Ministry of Education, Culture and Science. The work conducted by the U.S. Department of Energy Joint Genome Institute (DOE-JGI), a DOE Office of Science User Facility, was supported by the Office of Science of the DOE under Contract No. DE-AC02-05CH11231. IBISBA 1.0 - Project ID: 730976 - Funded under: H2020-EU.1.4.1.2. - Integrating and opening existing national and regional research infrastructures of European interest This work was carried out on the Dutch national e-infrastructure with the support of the SURF foundation.

Bibliography

Altschul, Stephen F et al. (1997). *Gapped BLAST and PSI-BLAST: A new generation of protein database search programs*. DOI: 10.1093/nar/25.17.3389.

Amin, Heba M., Abdelgawad M. Hashem, Mohamed S. Ashour, and Rajini Hatti-Kaul (2013). "1,2 Propanediol utilization by Lactobacillus reuteri DSM 20016, role in bioconversion of glycerol to 1,3 propanediol, 3-hydroxypropionaldehyde and 3-hydroxypropionic acid". In: *Journal of Genetic Engineering and Biotechnology* 11. DOI: 10.1016/j.jgeb.2012.12.002.

Artuso, Anthony (2002). "Bioprospecting, benefit sharing, and biotechnological capacity building". In: *World Development* 30.

Barbirato, F, J P Grivet, P Soucaille, and A Bories (1996). "3-Hydroxypropionaldehyde, an inhibitory metabolite of glycerol fermentation to 1,3-propanediol by enterobacterial species." In: *Applied and environmental microbiology* 62.

Barbirato, Fabien, El Hassan Himmi, Thierry Conte, and André Bories (1998). "1, 3-Propanediol production by fermentation: an interesting way to valorize glycerin from the ester and ethanol industries". In: *Industrial Crops and Products* 7.

Biebl, H, K Menzel, A-P Zeng, and W-D Deckwer (1999). "Microbial production of 1, 3-propanediol". In: *Applied microbiology and biotechnology* 52.

Chatzifragkou, Afroditi et al. (2011). "Production of 1, 3-propanediol by Clostridium butyricum growing on biodiesel-derived crude glycerol through a non-sterilized fermentation process". In: *Applied microbiology and biotechnology* 91.

Cheung, K.-H., Eric Prud'hommeaux, Yimin Wang, and Susie Stephens (2009). "Semantic Web for Health Care and Life Sciences: a review of the state of the art". In: *Briefings in Bioinformatics* 10. DOI: 10.1093/bib/bbp015.

Dam, Jesse C. J. van, Jasper J. Koehorst, Jon Olav Vik, Peter J. Schaap, and Maria Suarez-Diez (2017). "Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining". In: *bioRxiv*.

Dawyndt, Peter et al. (2006). "Mining fatty acid databases for detection of novel compounds in aerobic bacteria". In: *Journal of microbiological methods* 66.

Demick, Jonathan M. and William N. Lanzilotta (2011). "Radical SAM activation of the B12-independent glycerol dehydratase results in formation of 5'-deoxy-5'-(methylthio)adenosine and not 5'-deoxyadenosine." In: *Biochemistry* 50. DOI: 10.1021/bi101255e.

Fernández, Javier D., Miguel A. Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias (2013). "Binary RDF representation for publication and exchange (HDT)". In: *Journal of Web Semantics* 19. DOI: 10.1016/j.websem.2013.01.002.

Finn, Robert D., Teresa K. Attwood, et al. (2017). "InterPro in 2017-beyond protein family and domain annotations". In: *Nucleic Acids Research* 45. DOI: 10.1093/nar/gkw1107.

Finn, Robert D., Alex Bateman, et al. (2014). "Pfam: the protein families database". In: *Nucleic Acids Research* 42. DOI: 10.1093/nar/gkt1223.

Forage, R. G. and E C Lin (1982). "DHA system mediating aerobic and anaerobic dissimilation of glycerol in Klebsiella pneumoniae NCIB 418." In: *Journal of bacteriology* 151.

Gates, Phil (2000). "The Biotech Century". In: *Futures* 32. DOI: 10.1016/S0016-3287(00)00041-0.

Gelder, Antonie H van, Rozelin Aydin, M Madalena Alves, and Alfons JM Stams (2012). "1, 3-Propanediol production from glycerol by a newly isolated Trichococcus strain". In: *Microbial biotechnology* 5.

Gelder, Antonie H. van, Rozelin Aydin, M. Madalena Alves, and Alfons J.M. Stams (n.d.). "1,3-Propanediol production from glycerol by a newly isolated Trichococcus strain". In: *Microbial Biotechnology*. DOI: 10.1111/j.1751-7915.2011.00318.x.

González-Pajuelo, M, JC Andrade, and I Vasconcelos (2004). "Production of 1, 3-propanediol by Clostridium butyricum VPI 3266 using a synthetic medium and raw glycerol". In: *Journal of Industrial Microbiology and Biotechnology* 31.

Gungormusler, Mine, Cagdas Gonen, and Nuri Azbar (2011). "Continuous production of 1, 3-propanediol using raw glycerol with immobilized Clostridium beijerinckii NRRL B-593 in comparison to suspended culture". In: *Bioprocess and biosystems engineering* 34.

Haft, Daniel H et al. (2017). "RefSeq: an update on prokaryotic genome annotation and curation". In: *Nucleic acids research* 46.

Hyatt, Doug et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11. DOI: 10.1186/1471-2105-11-119.

Jarvis, GN, ERB Moore, and JH Thiele (1997). "Formate and ethanol are the major products of glycerol fermentation produced by a Klebsiella planticola strain isolated from red deer". In: *Journal of applied microbiology* 83.

Jiang, Wei, Shizhen Wang, Yuanpeng Wang, and Baishan Fang (2016). "Key enzymes catalyzing glycerol to 1,3-propanediol". In: *Biotechnology for Biofuels* 9. DOI: 10.1186/s13068-016-0473-6.

Jun, Sun-Ae et al. (2010). "Microbial fed-batch production of 1, 3-propanediol using raw glycerol with suspended and immobilized Klebsiella pneumoniae". In: *Applied biochemistry and biotechnology* 161.

Kang, Tae Sun, Darren R Korber, and Takuji Tanaka (2014). "Metabolic engineering of a glycerol oxidative pathway in Lactobacillus panis PM1 to utilize bioethanol thin stillage: Potential to produce platform chemicals from glycerol". In: *Applied and environmental microbiology*.

Khan, Nurul H. et al. (2013). "Isolation and characterization of novel 1,3-propanediol-producing Lactobacillus panis PM1 from bioethanol thin stillage". In: *Applied Microbiology and Biotechnology* 97. DOI: 10.1007/s00253-012-4386-4.

Kivistö, Anniina, Ville Santala, and Matti Karp (2012a). "1, 3-Propanediol production and tolerance of a halophilic fermentative bacterium, Halanaerobium saccharolyticum subsp. saccharolyticum". In: *Journal of biotechnology* 158.

— (2012b). "1,3-Propanediol production and tolerance of a halophilic fermentative bacterium, Halanaerobium saccharolyticum subsp. saccharolyticum". In: *Journal of Biotechnology* 158. DOI: 10.1016/j.jbiotec.2011.10.013.

Klyne, Graham and Jeremy J Carroll (2004). "Resource Description Framework (RDF): Concepts and Abstract Syntax". In: *W3C Recommendation* 10.

Knietsch, Anja, Susanne Bowien, Gregg Whited, Gerhard Gottschalk, and Rolf Daniell (2003). "Identification and characterization of coenzyme B12-dependent

glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures". In: *Applied and Environmental Microbiology*. DOI: 10.1128/AEM.69.6.3048-3060.2003.

Koehorst, Jasper J, Edoardo Saccenti, Peter J Schaap, Vitor AP Martins dos Santos, and Maria Suarez-Diez (2016). "Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics". In: *F1000Research* 5.

Koehorst, Jasper J, Jesse CJ Van Dam, et al. (2016). "Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data". In: *Scientific reports* 6.

Koehorst, Jasper J. et al. (2017). "SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles". In: *Bioinformatics* 1.

Lee, Sung Mok et al. (2014). "Enhancement of 1,3-propanediol production by expression of pyruvate decarboxylase and aldehyde dehydrogenase from Zymomonas mobilis in the acetolactate-synthase-deficient mutant of Klebsiella pneumoniae". In: *Journal of Industrial Microbiology and Biotechnology* 41. DOI: 10.1007/s10295-014-1456-x.

Luers, Frauke, Markus Seyfried, Rolf Daniel, and Gerhard Gottschalk (1997). "Glycerol conversion to 1, 3-propanediol by Clostridium pasteurianum: cloning and expression of the gene encoding 1, 3-propanediol dehydrogenase". In: *FEMS Microbiology Letters* 154.

Machielsen, Ronnie, Agustinus R. Uria, S. W M Kengen, and J. van der Oost (2006). "Production and Characterization of a Thermostable Alcohol Dehydrogenase That Belongs to the Aldo-Keto Reductase Superfamily". In: *Applied and Environmental Microbiology* 72. DOI: 10.1128/AEM.72.1.233-238.2006.

Matsakas, Leonidas, Kateřina Hruzova, Ulrika Rova, and Paul Christakopoulos (2018). "Biological Production of 3-Hydroxypropionic Acid: An Update on the Current Status". In: *Fermentation*.

Matsumura, M, Nakao Nomura, Seigo Sato, et al. (2008). "Growth and 1, 3-propanediol production on pre-treated sunflower oil bio-diesel raw glycerol using a strict anaerobe-Clostridium butyricum." In: *Current Research in Bacteriology* 1.

Mitraka, Elvira et al. (2015). "Wikidata: A platform for data integration and dissemination for the life sciences and beyond". In: *bioRxiv*. DOI: 10.1101/031971.

Mu, Ying, Hu Teng, Dai-Jia Zhang, Wei Wang, and Zhi-Long Xiu (2006). "Microbial production of 1, 3-propanediol by Klebsiella pneumoniae using crude glycerol from biodiesel preparations". In: *Biotechnology letters* 28.

Otte, Burkhard, Eike Grunwaldt, Osama Mahmoud, and Stefan Jennewein (2009). "Genome shuffling in Clostridium diolis DSM 15410 for improved 1, 3-propanediol production". In: *Applied and environmental microbiology* 75.

Papanikolaou, Seraphim, Michel Fick, and George Aggelis (2004). "The effect of raw glycerol concentration on the production of 1, 3-propanediol by Clostridium butyricum". In: *Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology* 79.

Pasek, Sophie, J.-L. Risler, and P. Brezellec (2006). "Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins". In: *Bioinformatics* 22. DOI: 10.1093/bioinformatics/btl135.

Petitdemange, E, C Dürr, S Abbad Andaloussi, and G Raval (1995). "Fermentation of raw glycerol to 1, 3-propanediol by new strains ofClostridium butyricum". In: *Journal of industrial microbiology* 15.

Pflügl, Stefan, Hans Marx, Diethard Mattanovich, and Michael Sauer (2012). "1,3-Propanediol production from glycerol with Lactobacillus diolivorans". In: *Bioresource Technology*. DOI: 10.1016/j.biortech.2012.05.121.

Poehlein, Anja, Frank Robert Bengelsdorf, Bettina Schiel-Bengelsdorf, Rolf Daniel, and Peter Dürre (2016). "Genome sequence of the acetogenic bacterium Acetobacterium wieringae DSM 1911T". In: *Genome announcements* 4.6, e01430–16.

Pradima, J, M Rajeswari Kulkarni, et al. (2017). "Review on enzymatic synthesis of value added products of glycerol, a by-product derived from biodiesel production". In: *Resource-Efficient Technologies*.

Prud'hommeaux, Eric and Andy Seaborne (2008). "SPARQL Query Language for RDF". In: *W3C Recommendation* 2009. DOI: citeulike-article-id:2620569. URL: http://www.w3.org/TR/rdf-sparql-query/.

Raynaud, C., P. Sarcabal, I. Meynial-Salles, C. Croux, and P. Soucaille (2003). "Molecular characterization of the 1,3-propanediol (1,3-PD) operon of Clostridium bu-

tyricum". In: *Proceedings of the National Academy of Sciences* 100. DOI: 10.1073/pnas.0734105100.

Ristoski, Petar and Heiko Paulheim (2016). "Semantic Web in data mining and knowledge discovery: A comprehensive survey". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 36. DOI: 10.1016/j.websem.2016.01.001.

Rodriguez, Alberto, Mateusz Wojtusik, Vanessa Ripoll, Victoria E Santos, and F Garcia-Ochoa (2016). "1, 3-Propanediol production from glycerol with a novel biocatalyst Shimwellia blattae ATCC 33430: operational conditions and kinetics in batch cultivations". In: *Bioresource technology* 200.

Saxena, R. K., Pinki Anand, Saurabh Saran, and Jasmine Isar (2009). "Microbial production of 1,3-propanediol: Recent developments and emerging opportunities". In: *Biotechnology Advances* 27. DOI: 10.1016/j.biotechadv.2009.07.003.

Schütz, Helmut and Ferdinand Radler (1984). "Anaerobic reduction of glycerol to propanediol-1.3 by Lactobacillus brevis and Lactobacillus buchneri". In: *Systematic and Applied Microbiology* 5.

Seyfried, Markus, Delina Lyon, Fred A Rainey, and Juergen Wiegel (2002). "Caloramator viterbensis sp. nov., a novel thermophilic, glycerol-fermenting bacterium isolated from a hot spring in Italy." In: *International journal of systematic and evolutionary microbiology* 52.

Silvester, Nicole et al. (2018). "The European Nucleotide Archive in 2017". In: *Nucleic Acids Research* 46. DOI: 10.1093/nar/gkx1125.

Söhngen, Carola et al. (2016). "BacDive - The Bacterial Diversity Metadatabase in 2016". In: *Nucleic Acids Research* 44. DOI: 10.1093/nar/gkv983.

Stams, A. J.M., J. B. Van Dijk, C. Dijkema, and C. M. Plugge (1993). "Growth of syntrophic propionate-oxidizing bacteria with fumarate in the absence of methanogenic bacteria". In: *Applied and Environmental Microbiology* 59.

Strepis, Nikolaos et al. (2016). "Description of Trichococcus ilyis sp. nov. by combined physiological and in silico genome hybridization analyses". In: *International journal of systematic and evolutionary microbiology* 66.

Sun, Jibin, Joop van den Heuvel, Philippe Soucaille, Yinbo Qu, and An-Ping Zeng (2003). "Comparative genomic analysis of dha regulon and related genes for anaerobic glycerol metabolism in bacteria". In: *Biotechnology progress* 19.

Taconi, Katherine A, Keerthi P Venkataramanan, and Duane T Johnson (2009). "Growth and solvent production by Clostridium pasteurianum ATCC® 6013™ utilizing biodiesel-derived crude glycerol as the sole carbon source". In: *Environmental Progress & Sustainable Energy: An Official Publication of the American Institute of Chemical Engineers*.

Uhlig, Ronny, Anja Poehlein, Ralf-Jörg Fischer, Rolf Daniel, and Hubert Bahl (2016). "Genome Sequence of the Autotrophic Acetogen Clostridium magnum DSM 2767". In: *Genome Announcements* 4. DOI: 10.1128/genomeA.00464-16.

UniProt Consortium, The (2018). "UniProt: the universal protein knowledgebase". In: *Nucleic Acids Research* 46. DOI: 10.1093/nar/gky092.

Vivek, Narisetty, Ashok Pandey, and Parameswaran Binod (2018). "An efficient aqueous two phase systems using dual inorganic electrolytes to separate 1, 3-propanediol from the fermented broth." In: *Bioresource technology* 254.

Wall, Dennis P. et al. (2010). "Cloud computing for comparative genomics". In: *BMC Bioinformatics* 11. DOI: 10.1186/1471-2105-11-259.

Wilkens, Erik, Anne Katrin Ringel, Diana Hortig, Thomas Willke, and Klaus-Dieter Vorlop (2012). "High-level production of 1, 3-propanediol from crude glycerol by Clostridium butyricum AKR102a". In: *Applied microbiology and biotechnology* 93.

Xu, Yun-Zhen et al. (2009). "Metabolism in 1, 3-propanediol fed-batch fermentation by ad-lactate deficient mutant of Klebsiella pneumoniae". In: *Biotechnology and Bioengineering* 104.

Yang, Guang, Jiesheng Tian, and Jilun Li (2007). "Fermentation of 1, 3-propanediol by a lactate deficient mutant of Klebsiella oxytoca under microaerobic conditions". In: *Applied Microbiology and Biotechnology* 73.

Zarzycki, Jan, Onur Erbilgin, and Cheryl A. Kerfeld (2015). "Bioinformatic characterization of glycyl radical enzyme-associated bacterial microcompartments". In: *Applied and Environmental Microbiology* 81. DOI: 10.1128/AEM.02587-15.

Zhang, GenLin, BinBin Ma, XiaoLin Xu, Chun Li, and Liwei Wang (2007). "Fast conversion of glycerol to 1, 3-propanediol by a new strain of Klebsiella pneumoniae". In: *Biochemical Engineering Journal* 37.

9 iBioSystems: development of an undergraduate course aimed at closing the gap between computational and experimental studies on biological systems

Jasper J. Koehorst, Philippe G.B. Puylaert, Maria Suarez-Diez and Peter J. Schaap

Abstract

The advent of high-throughput techniques and the increased accuracy and reproducibility of measuring techniques is transforming biology into a fully quantitative science. Computational analyses have become essential to interpret experimental outcomes and to identify numerical anomalies. Even when working in a wet-lab bench oriented research setting, computational programs are used at some stage in research. The ability to select and deploy analysis tools and algorithms thus has become an indispensable research skill and challenges us to train our students to become statistically and computationally fluent. The aim of iBioSystems is to demonstrate to undergraduate students how computational and wet-lab bench studies on bio-systems can be integrated to solve complex biological questions.

Introduction

In this era of genomics, computational analysis has become an integral part of normal 'wet-lab' routines. However, we have seen that for practical oriented undergraduate students, computational methods and predictions are often felt complex, abstract and difficult. To bridge this gap, a course was developed introducing the 'Moist-lab', a modern style of research where computational predictions are combined with and foster hands-on wet-lab experiments. In iBioSystems, an integrative wet-dry cycle of experimenting is followed, to study genotype-phenotype relationships in prokaryotes and simple eukaryotes. As such, the Moist-lab focuses on deriving a deeper understanding of biological systems by uncovering biological meaning from genome scale data through integration with outcomes of wet lab experiments. Specifically, the course teaches *i*) how genome information is translated in function, *ii*) how regulation of microbial metabolic processes can take place and *iii*) how genome data can be used to predict responses of microbial organisms and ecosystems to (a)biotic environmental cues.

A second objective of this course is to make students aware of the importance of the 'FAIR Guiding Principles for scientific data management and stewardship' as an essential component of computational but also wet lab experiments (Wilkinson et al., 2016). During this course, a structured digital lab journal will be used to keep track of the entire process. Additional efforts are made by pointing at the FAIRness of an experiment. As such, the course demonstrates to the students how increasing the FAIRness can improve the quality for both experimental as well as computational experiments by using mandatory minimal information models such as MIxS (Yilmaz et al., 2011) and the use of information frameworks to (automatically) collect and communicate metadata (i.e. sample characteristics, technologies used, type of measurements made), employing a combination of workflows and schemas such as ISA-TAB.

iBioSystems

The main rationale behind the set-up of this course was that biological phenomena should be simultaneously taught from both an experimental and computational per-

spective. Since both fields complement each other in different ways this will improve the understanding of the more abstract parts of both fields. In practice, this can be ambitious as it puts high demands on teachers as they must be experienced in experimental and or computational biology or at least are used to apply key elements of the 'Moist lab' approach in their own research.

Course content

We have adopted an integrated approach to learning, starting with bacterial systems to introduce basic experimental and theoretical concepts. Lectures are integrated with practical courses demonstrating the application of a Moist Systems Biology approach. It expands the basic understanding of biological systems in which Brock is the current de-facto standard (Madigan et al., 2017). To demonstrate the Moist-lab approach we introduced 'discovery practicals', i.e., challenging exercises deliberately stipulated in a free format, where students work with recently sampled and sequenced species of "unknown" origin. A series of SOPs are available allowing students to design, perform and interpret wet-lab and computational experiments thereby fostering creativity, communication and collaborative skills. Besides the opportunity to practice the necessary technical skills in both fields, exercises are meant to bridge the gap between computational predictions and experimental observations / results. Exercises are aimed at all cognitive levels with specific attention to 'the ability to apply knowledge and understanding' through synthesis and evaluation of experimental and computational results (Berg, 2005). The assessment of a student's understanding should ideally entail both fields. Assessment is performed through *i*) the documentation in the electronic lab-notebook and *ii*) a scientifically written report which through a peer-reviewed process is cross-evaluated by fellow students enabling students to critically review each other's article and a final exam. In this cross-evaluation process students are trained on how to critically evaluate the quality of a report and use a digital platform to assess and provide and use this feedback to improve their work (Lesterhuis et al., 2017).

Learning outcomes

Upon successful completion of the course it is expected that students are able to

- formulate research problems such that they can be solved by an integrated experimental/computational biology approach.
- explain the iterative cycle: prediction and experimental verification.
- communicate scientific questions across experimental and theoretical disciplines and to collaborate across disciplinary borders.
- select and apply the type of data generation and bioinformatics approaches that are suitable for a given research problem.
- handle data and experimental designs in a FAIR manner.
- critically assess evidence and scientific argumentation in integrative studies of biological systems based on an understanding of both experimental and computational biology methodologies.

The course has been given twice in 2017-2018 at Wageningen University & Research. A pilot was performed with 6 master students also to test and assess the quality of the developed SOPs and to identify additional wet lab and computational experiments. The improved course was taken by 36 undergraduate students in the academic year 2017-2018. All had a basic understanding in biological systems and experimental procedures. The students were evaluated with an emphasis on FAIR management through the use of an electronic lab notebook which is evaluated on reusability of computational as well as wetlab experiments. Datasets obtained and generated, should be well documented according to the *de facto* standard. To assess their cognitive level, a scientific report was written by each student and evaluated based on their experimental findings and the cross-linking with computational predictions. Learning outcomes were assessed through a final exam.

An example of a learning activity: nitrate reduction

In the following, an example of teaching and learning activities in the iBioSystems course are shown. This activity is performed in-silico in week three and tested on the wet-lab in week four (see Figure 9.2, 9.3 and 9.4). The final goal is that the stu-

dents perform genotype -phenotype associations using computational tools, which are then validated through dedicated wet-lab experiments. Initially, the students are given a general lecture on metabolic pathways and their association to metabolic phenotypes such as auxotrophys or nitrate reduction capabilities followed by an introductory practical in which they learn to use tools such as KAAS, KEGG mapper and how to interpret the results (See Figure 9.2). The students are then offered a description of the nitrate reduction test and its possible outcomes (See Figure 9.3 and 9.4). In the following week, the students go to the wet-lab and perform the test on the strain they have been studying. They are then encouraged to discuss on the agreement (or disagreement) between the computational prediction and the experimental verification.

Course materials

Background material related to biological systems and computational methods is obtained from 'Brock Biology of Microorganisms' (Madigan et al., 2017) and 'Practical Bioinformatics' (Agostino, 2012). Students can store all experimental procedures and results into an Electronic Lab Notebook (such as Labfolder or eLABJournal). Most of the computational work is performed through Galaxy, a data analysis platform were students can use common bioinformatic tools such as SPAdes, Prodigal and BLAST (Goecks et al., 2010; Bankevich et al., 2012; Hyatt et al., 2010; Camacho et al., 2009). Other analysis, such as pathway analysis is performed through web tools such as KEGG mapper (Ogata et al., 1999). Experimental and computational protocols are provided through a digital environment which is continuously complemented with new or updated protocols and maintained by the lecturers.

Future perspective

The Systems Biology portfolio contains two types of courses: those devoted to general tools and techniques and those devoted to concepts and tools specific to Systems and/or Synthetic Biology.

Advanced courses are essential to educate new students of systems and synthetic biologists with a portfolio similar to the one in (Cvijovic et al., 2016). Still, possibly

the greatest challenge lies in increasing the mathematical and computational literacy of biologist from any field and this should be addressed through general courses embedded in interdisciplinary approaches preferably starting already at undergraduate level.

The tackling of multifaceted problems and interdisciplinary approaches in education have been proved effective to increase computational literacy (Rubinstein and Chor, 2014), therefore iBioSystems has been developed in collaboration with the Systems and Synthetic Biology (SSB) and the Microbiology Laboratory in Wageningen University. In this way, we ensure good integration with other courses the biotechnology students follow in their BSc.

Currently, iBioSystems is primarily offered to students in the Biotechnology BSc, although students from other disciplines in the life sciences are encouraged to take it as part of the BSc minor "Systems Biology". In the future, we expect to enlarge the suggested modules with new experiments and computational analysis that would be better aligned with the specific needs of students in other disciplines such as Animal Sciences, Biology, Molecular Life Sciences and Plant Sciences.

Nitrate reduction test

Anaerobic respiration involves the reduction of inorganic molecules (other than oxygen), by using them as terminal electron acceptors. Some of the most common molecules are inorganic nitrogen compounds, including ammonia (NH3), nitrogen gas (N2), nitrous oxide (N2O), nitric oxide (NO), nitrite (NO2-), nitrogen dioxide (NO2), and nitrate (NO3-). Some Gram-negative bacteria (most Enterobacteriaceae) possess the enzyme nitrate reductase, a molybdenum- containing membrane-integrated enzyme that catalyzes the one-step reduction of nitrate to nitrite. Other microorganisms have the ability to further reduce nitrite to nitrogenous gasses, such as NO, N2O, and N2. This process is known as denitrification and is widely used in the sewage treatment to stimulate algal growth. Denitrification is also of global significance, as it converts fixed nitrogen (nitrate) to environmentally significant gaseous nitrogen compounds. Alternatively, some bacteria may convert nitrite to ammonia through a dissimilative process.
Next week (Week 4, Monday) you will perform a nitrate reduction test in the lab. This test is used to test whether the bacteria is able to reduce nitrate to nitrite. Here, we will predict the output of the test based on genomic information. The final goal is that you will link the computational prediction with the wet lab experiment.

- Read the introduction to the wet-lab experiment (see Figure 9.3 and 9.4)
- Read the introduction to the computational experiment (see 9.2).
- Go to KEGG pathways, bear in mind that we are interested in nitrogen reduction, so you will have to go to the map associated to Nitrogen metabolism (which is part of the Energy metabolism).
- Question: Is *Escherichia coli* K-12 MG1655 able to reduce nitrate? What would be the output of the nitrate reduction test in *E.coli*?
- Question: What would be the output of the nitrate reduction test in *Pseudomonas aeruginosa* PAO1?
- Predict the output of the nitrate reduction test in your genome. The answer will become available next week (when your run the wet lab experiment)! If your prediction is correct, great! If the prediction is not correct, then something very interesting might be going on!

Figure 9.1: ***Hand out for the students:*** *Description of the application of the moist-lab cycle applied to the analysis of nitrate reduction in bacteria. In this particular example E. coli and P. aeruginosa have been chosen because they give different outputs in the test (colourless/red and colourless/colourless respectively).*

In the "Gene Prediction" exercise we have predicted and generated a file containing protein sequences in FASTA format. Some of these proteins might have enzymatic functions and there are many ways to unravel the possible biological function of a protein. The most commonly used aproach is through BLAST against a well curated database such as SwissProt. Another such database is called KEGG (Kyoto Encyclopedia of Genes and Genomes) KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
To analyse protein sequences more easily they have developed KAAS. KAAS (KEGG Automatic Annotation Server) provides functional annotation of genes by BLAST or GHOST comparisons against the manually curated KEGG GENES database. The result contains KO (KEGG Orthology) assignments and automatically generated KEGG pathways.
In this exercise we have reduced the number of proteins (a few thousand in your genome) to only six. The sequences are shown here in FASTA format. Please copy them and go to KAAS.

- Use the KAAS job request (SBH method) under partial genome. When analysing your own genome you can use the BBH method under Complete or Draft Genome.
- Past the sequences in the Query sequences box.
- For good practices give the dataset a name under Query name
- Fill in your / WUR email address
- Select under GENES data set (for prokaryotes)
- Click compute
- You will receive an email with a link to start the job!
- After the exercise you can analyse your own genome with KAAS

Figure 9.2: Computational protocol for mapping genome content into KEGG maps

Experimental Protocol

Nitrate reduction can be detected by culturing bacteria in a nitrate broth under anaerobic conditions. After overnight incubation, the ability of the microorganisms to reduce the provided nitrate to nitrite is defined by the addition of sulfanilic acid (Reagent A) and alpha-naphtylamine (Reagent B). More specifically, if nitrate reductase activity is present (nitrate -> nitrite), sulfanilic acid (Reagent A) forms a colourless complex (nitrate-sulfanilic acid), which subsequently reacts with α-naphtylamine (Reagent B) giving a cherry-red precipitate (prontosil) (Figure A). If no colour is observed, the organism does not have the ability to reduce nitrate to nitrite (absence of nitrate reductase activity) or nitrate is completely reduced to ammonia or even molecular nitrogen. Hence, to clarify which of the two possibilities holds true, zinc powder is added, to the colorless cultures that already contain the reagents. In this case, red colour indicates inability of the microorganism to reduce nitrate to nitrite, whilst no change of colour depicts reduction of nitrate to ammonia or molecular nitrogen through nitrite formation.

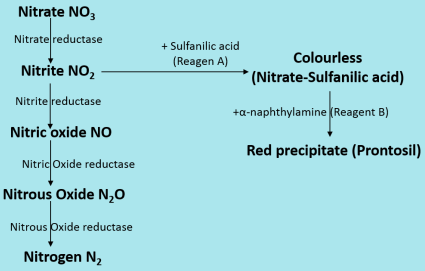

Figure A: Reduction of nitrogenous compounds. The right side of the figure shows the nitrate production test

Figure 9.3: Introduction for the experimental analysis of nitrate reduction

Day 1

- Resuspend a colony in 50 µL of sterile water in a sterile eppendorf tube.
- Inoculate 5 mL Nitrate Broth in a plastic tube with the cell suspension. Do this slowly and carefully to prevent adding oxygen to the medium.
- Add 1 cm of paraffin oil on the surface of the liquid culture (to create anaerobic conditions).
- Incubate at optimal temperature for your organism for 24 to 48 h.

Day 2

Perform a negative control by doing the test also with a tube that only contains Nitrate Broth and no bacteria.

- Add 3 drops of Reagent A and 3 drops of Reagent B. Invert tube several times and observe the colour change after a few minutes. KEEP THUMB ON LID TO PREVENT LEAKAGE
- If the suspension turns pink-red: the reaction is positive (nitrate reduction) and the test is completed. If the suspension remains colourless after the addition of reagents A and B: add a small amount ("sharp knife point") of zinc powder to the medium. Shake the tube vigorously and allow it to stand at room temperature for 10-15 min.

Results

If the medium remains colourless after the addition of Zn powder: the test result is positive.
If the medium turns pink after the addition of Zn powder: the result is negative.
Important notes

- The negative control should also be tested. There should be no pink colour formation after adding reagent A and B and if zinc powder is added the colour should change to pink.
- Addition of too much zinc powder can results in a false-negative reaction.

Figure 9.4: Wetlab protocol for the analysis of nitrate reduction

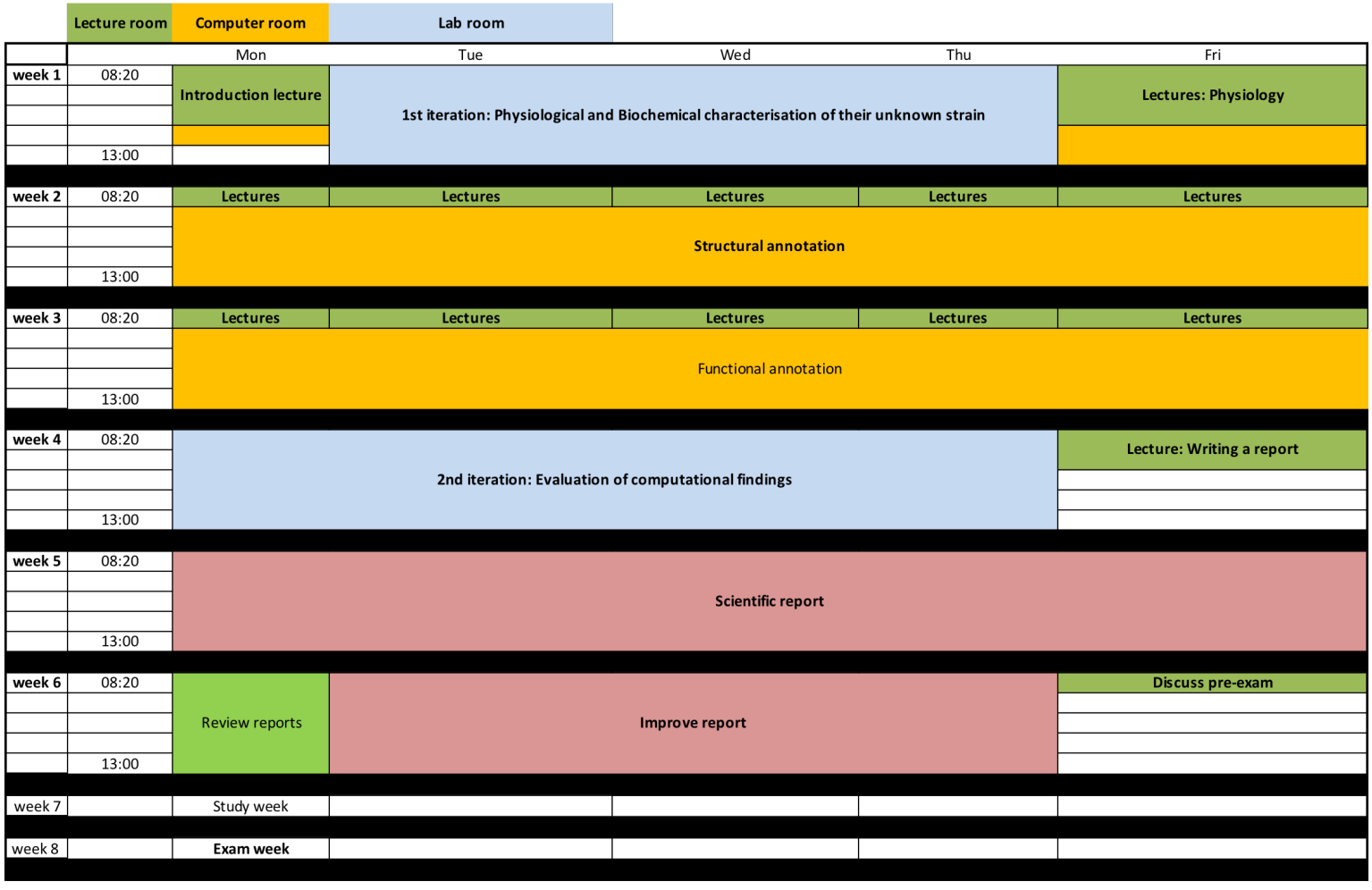

Figure 9.5: Overview of the course schedule as given in 2017-2018

Bibliography

Agostino, Michael (2012). *Practical bioinformatics*.

Bankevich, Anton et al. (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing". In: *Journal of computational biology* 19.

Berg, Mogens (2005). "The Framework for Qualifications of the European Higher Education Area". In: *Chancen und Grenzen eines Qualifikationsrahmens. Eine gemeinsame Veranstaltung der Service-Stelle Bologna der HRK und des Projekts Qualitätssicherung, Berlin, Hochschulrektorenkonferenz. litätssicherung, Berlin, Hochschulrektorenkonferenz*.

Camacho, Christiam et al. (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10. DOI: 10.1186/1471-2105-10-421.

Cvijovic, Marija et al. (2016). "Strategies for structuring interdisciplinary education in Systems Biology: an European perspective". In: *NPJ systems biology and applications* 2.

Goecks, Jeremy, Anton Nekrutenko, James Taylor, and The Galaxy Team (2010). "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences". In: *Genome Biol* 11. DOI: 10.1186/gb-2010-11-8-r86.

Hyatt, Doug et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11. DOI: 10.1186/1471-2105-11-119.

Lesterhuis, Marije, San Verhavert, Liesje Coertjens, Vincent Donche, and Sven De Maeyer (2017). "Comparative judgement as a promising alternative to score competences". In: *Innovative practices for higher education assessment and measurement*.

Madigan, M.T., K.S. Bender, D.H. Buckley, W.M. Sattley, and D.A. Stahl (2017). *Brock Biology of Microorganisms, Global Edition*. ISBN: 9781292235196. URL: https://books.google.nl/books?id=0-1DDwAAQBAJ.

Ogata, Hiroyuki et al. (1999). *KEGG: Kyoto encyclopedia of genes and genomes*. DOI: 10.1093/nar/27.1.29.

Rubinstein, Amir and Benny Chor (2014). "Computational thinking in life science education". In: *PLoS computational biology* 10.

Wilkinson, Mark D et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.

Yilmaz, P. et al. (2011). "Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications". In: *Nature Biotechnology* 29. DOI: 10.1038/nbt.1823.

10 General discussion: Bridging the knowledge gap

ABSTRACT

The aim of this thesis was to increase our understanding on how genome information leads to function and phenotype. One of the biggest hurdles in understanding (microbial) genome information and developing biological systems is to convert the deluge of information available from various heterogeneous data sources into actionable knowledge. Integrated analysis of new and existing biological data, information and knowledge, requires an information retrieval and management system that is not only efficient and extendable but also facilitates reproducibility of data-driven scientific findings. Maintaining a high degree of transparency in scientific reporting is essential and to facilitate knowledge discovery, a set of guiding principles have been defined to make data Findable, Accessible, Interoperable, and Reusable (FAIR). Adopting a Semantic Systems Biology approach for analysis and modelling of microbial ecosystems provides a strong support for FAIR by design experimentation. However, this requires development and implementation of community standards, minimal information models and ontologies for analytical and computational data types, essential for the inference of interactions and emergent properties. In this last chapter I gathered a number of use-cases that provided direct input into the Gap Analysis process that ultimately led to the currently developed platform (**Chapter 4**) which has since been used in a wide range of applications.

GENERAL DISCUSSION

The central dogma of molecular biology concerns "the detailed residue-by-residue transfer of sequential information" and states that: "such information cannot be transferred back from protein to either protein or nucleic acid" (Crick, 1970). Since the introduction of the central dogma, the relationship between DNA, RNA and proteins are apparent.

To study biological systems, we have to take into account three levels of information in which DNA contains the information regarding structural organisation of genetic elements and RNA and proteins provide the dynamic interactions and functional responses to the environment. Furthermore, these three levels are tightly interlinked and regulated. The central dogma also describes a strict semantic relationship between these different "instances"thereby connecting these data types. It is strict and directional in the sense that RNA translates into protein and proteins cannot be transcribed into DNA. This results in a multipartite knowledge graph that is not only descriptive but also predictive in character.

In a top-down approach, having protein at hand, it makes sense to search for the encoding gene. In a bottom-up approach, having found a new protein-encoding gene, the function thereof can be inferred using sequence similarity with genes encoding known functions. These levels of information were managed and manageable in parallel and often unlinked information resources until the development of the high-throughput sequencer, which not only revolutionised the field but also transformed the field in a data-intensive science (Figure 10.1).

To prevent 'drowning in data', various community efforts have been initiated and resulted in the development of platforms and data integration tools such as Seek, RightField, Wikipedia, Wikidata, WikiPathways and BioModels. Seek (Katherine Wolstencroft et al., 2015) enables the exchange of scientific datasets including models, simulations and research outcomes in public/private groups according to the ISA format (Investigation, Study, Association) *(ISA Model & Serialization Specifications 1.0 (October 2016)*. RightField (Katy Wolstencroft et al., 2011), also part of the FAIRDOM consortium (Stanford et al., n.d.), allows to semantically annotate excel sheets using predefined ontologies, making different data sets interoperable and

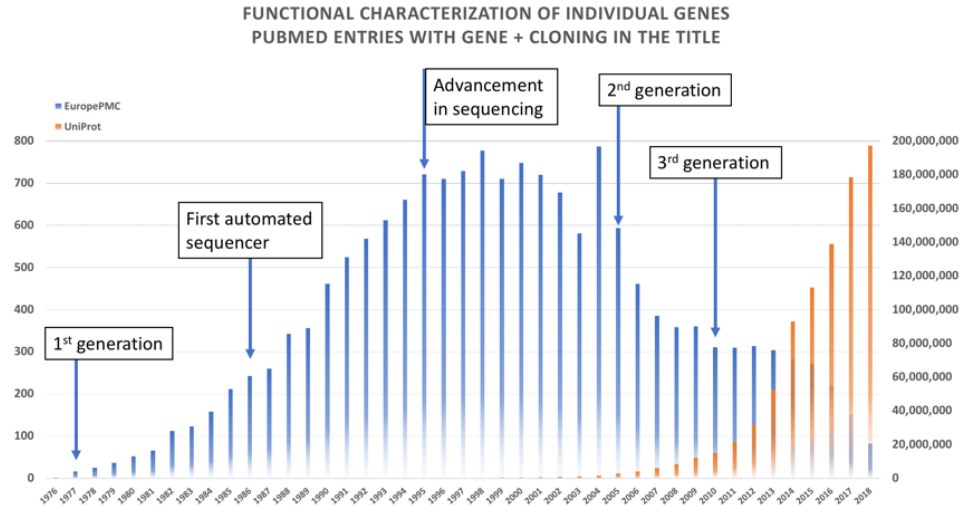

Figure 10.1: ***Number of publications related to the characterisation of individual genes and proteins in contrast to the number of proteins sequenced****. With the advancement of 2nd and 3rd generation sequencers, publications describing the characterisation of individual genes and proteins started to decline since 2000 according to EuropePMC (blue), while the number of electronically inferred gene to protein translations available in UniProt increased significantly and are now primarily characterised through in-silico analyses.*

adhere to a standardised layout (Katy Wolstencroft et al., 2011). Wikipedia, a well-known resource for a collaborative encyclopedia which is directly linked to Wikidata (Mitraka et al., 2015), a central storage for structured information, which can be read by both humans and machines and WikiPathways, a resource for and by scientists, for the development of high quality pathway maps (Slenter et al., 2018). The storage and exchange of finalised models is achieved through BioModels containing a range of models describing processes like signalling, protein-drug interactions, metabolic pathways or epidemic models.

All these initiatives have been developed for the exchange of already fine-grained information. To enable biological data to become comprehensible and manageable, it is required that this information is both human and machine readable. One of the earliest initiatives to standardise biological knowledge for a wide variety of organisms was in the form of a relational database scheme named Chado (Mungall, Emmert, Consortium, et al., 2007). Chado is driven by standardised ontologies and its main drawback is the inflexibility of the technology used with respect to new

heterogenous data-sources.

The Semantic Web was still under development and became an official W3C recommendation on the 15th of January in 2008. With the release of SPARQL 1.1 (2013), introducing support of federated queries allowing users or machines to query multiple resources as if it were a single database, the power of Semantic Web technologies became more apparent. Since Semantic Web became an official standard, UniProt has released its resource in an RDF model (`http://www.uniprot.org/news/2008/06/10/release`) and ontologies specific for the field of life science started to be developed (Bolleman et al., 2014; The Gene Ontology Consortium, 2015; Eilbeck et al., 2005). However, such systematic data integration was often performed by the larger public repositories such as UniProt, Rhea and neXtProt (UniProt Consortium, 2017; Alcántara et al., 2012; Gaudet et al., 2015). Functional applications that could actually help users to directly convert and semantically link their own research data, remained absent.

Chapters 5, 6, 7 and 8 demonstrate the power of semantic system biology approaches. In the following paragraphs, I provide a number of additional examples from my own work further exemplifying some of the major benefits of using consistent annotation and ontologies to link and describe heterogeneous data sources in the RDF data model, and the subsequent analysis of linked data using top-down approaches and SPARQL.

In another study (Worm et al., 2014), the main interest was to understand genomic differences between bacteria with a syntrophic lifestyle and a non-syntrophic lifestyle. The analysis was performed on 19 different strains of which five were syntrophic sort chain fatty acid (SCFA) degraders, two non-synthrophic SCFA degraders and twelve sulfate reducers that were never tested for syntrophic growth.

The understanding of genomic differences among the different strains could have been achieved in a classic bottom-up approach through identification of sequence clusters and afterwards unravel the functional capacity of each cluster of interest. However, this would have required an extensive sequence-based comparison and, in addition, in-depth manual biocuration to characterise the sequence clusters obtained, which also requires expert knowledge. To make the biocuration process FAIR, it would require documentation of the reasoning substantiating the associ-

ations made.

Instead an alternative top-down route was undertaken, using already present functional descriptions, which in principle could automate the classification process. However, protein descriptions are currently not very well standardised and due to this low level of interoperability cannot be used for direct comparisons.

A more defined and standardised functional description of proteins can be achieved through the identification of conserved motifs or domains in proteins. One of the most well-known patterns is the Zinc finger motif (Miller, McLachlan, and Klug, 1985). Since then dedicated databases have been initiated with a main focus on the detection of motifs with specific functional capabilities.

To understand the specific functional properties with regards to syntrophy, InterProScan searches were incorporated in the data framework. InterProScan is a bundled resource of various databases such as PFAM, Superfamily, Gene3D and others. Initiating a high degree of interoperability at functional level this, new top-down, approach immediately revealed specific properties for syntrophs and also detected additional strains that revealed close similarity to syntrophic organisms as can be seen in Figure 10.2 (originated from (Worm et al., 2014).

The efficiency of the top-down approach immediately led to new interest in the analysis of various organisms. In (Visser et al., 2014) a more in-depth analysis of the carboxydotrophic sulfate-reducers was performed using a combination of existing resources complemented with protein domain annotation. This approach was later also applied to Archaea in order to find additional differences between archaeal methanotrophs and related methanogens (Timmers et al., 2017). Due to the integration of functional annotation on already existing genomes as was performed in (Worm et al., 2014; Visser et al., 2014; Timmers et al., 2017) we realised that for a number of problems standardised top-down approaches can be very efficient.

With more genomes being sequenced, the number of genomes used in comparative genomics were expected to grow continuously, thereby increasing scaling problems encountered in sequence based bottom-up approaches. The idea of a platform capable of handling and standardising genome annotation was therefore initiated.

In 2015, a preliminary version of the semantic annotation platform that was later to be called SAPP (Semantic Annotation Platform with Provenance), was developed,

		Syntrophomonas wolfei	*Syntrophus aciditrophicus*	*Syntrophothermus lipocalidus*	*Syntrophobacter fumaroxidans*	*Pelotomaculum thermopropionicum*	*Desulfotomaculum kuznetsovii*	*Desulfobulbus propionicus*	*Desulfobulbus japonicus*	*Desulfatibacillum alkenivorans*	*Desulfatirhabdium butyrativorans*	*Desulfobacterium autotrophicum HRM2*	*Desulfospira joergensenii*	*Desulfatiglans balticum*	*Desulfomonile tiedjei*	*Desulfarculus baarsii*	*Desulfosporosinus meridiei*	*Desulfotalea psychrophila*	*Desulfatibacillum aliphaticivorans*	*Desulfotomaculum gibsoniae*
Growth on butyrate [¥]																				
Growth on propionate [¥]										✓										
Extra-cytoplasmic FDH alpha subunit	IPR006443	1	2	1	3	1	0	0	0	2	1	5	2	0	2	0	1	0	0	1
FdhE-like protein	IPR024064	4	6	2	5	3	0	0	0	0	0	4	4	0	2	1	2	0	0	2
FDH accessory protein	IPR006452	2	3	1	2	1	0	0	0	0	0	2	2	0	1	0	1	0	0	1
Capsule synthesis protein, CapA	IPR019079	2	2	4	2	1	0	0	0	0	0	4	2	0	2	0	4	4	0	2
Cell cycle, FtsW / RodA / SpoVE,	IPR018365	1	2	2	2	1	0	0	0	2	1	3	0	1	1	2	2	0	0	0
Ribonuclease P, conserved site	IPR020539	1	1	1	1	1	0	0	0	2	1	1	0	0	1	1	1	1	0	0

Figure 10.2: ***Domain based genome comparison of syntrophic and non-syntrophic butyrate and/or propionate degraders****. Domains present in genomes of all butyrate and/or propionate-degrading syntrophs and absent in those of non-syntrophs are listed and domain abundance is indicated. Syntrophs are shaded orange, non-syntrophs are shaded blue and sulfate reducers that were never tested for syntrophic growth are shaded green. The pale colour green corresponds to draft genomes and the darker colours (orange, blue, green) correspond to complete genomes.*

tested and in-house used for the analysis of *Bacillus smithii* and immediately yielded new insights in this industrial organism (Bosma et al., 2016). Initially, *B. smithii* was annotated with RAST, which is a widely accepted approach for genome annotation (Aziz et al., 2008). This was then further complemented with an analysis based on protein domains as applied in previous studies. This showed to be a powerful approach for the identification of previously unknown protein functions. In an in-depth comparison of domain-based annotations with a manually curated RAST annotation, 142 genes could be additionally functionally annotated, all of which except 4 were previously marked as hypothetical proteins. Using RAST, the methylglyoxal pathway was identified only towards D-lactate but additional analysis using protein domains revealed the presence of all genes necessary for L-lactate production via methylglyoxal (Figure 10.3).

The first real comparative genomics study using SAPP was in (Saccenti et al.,

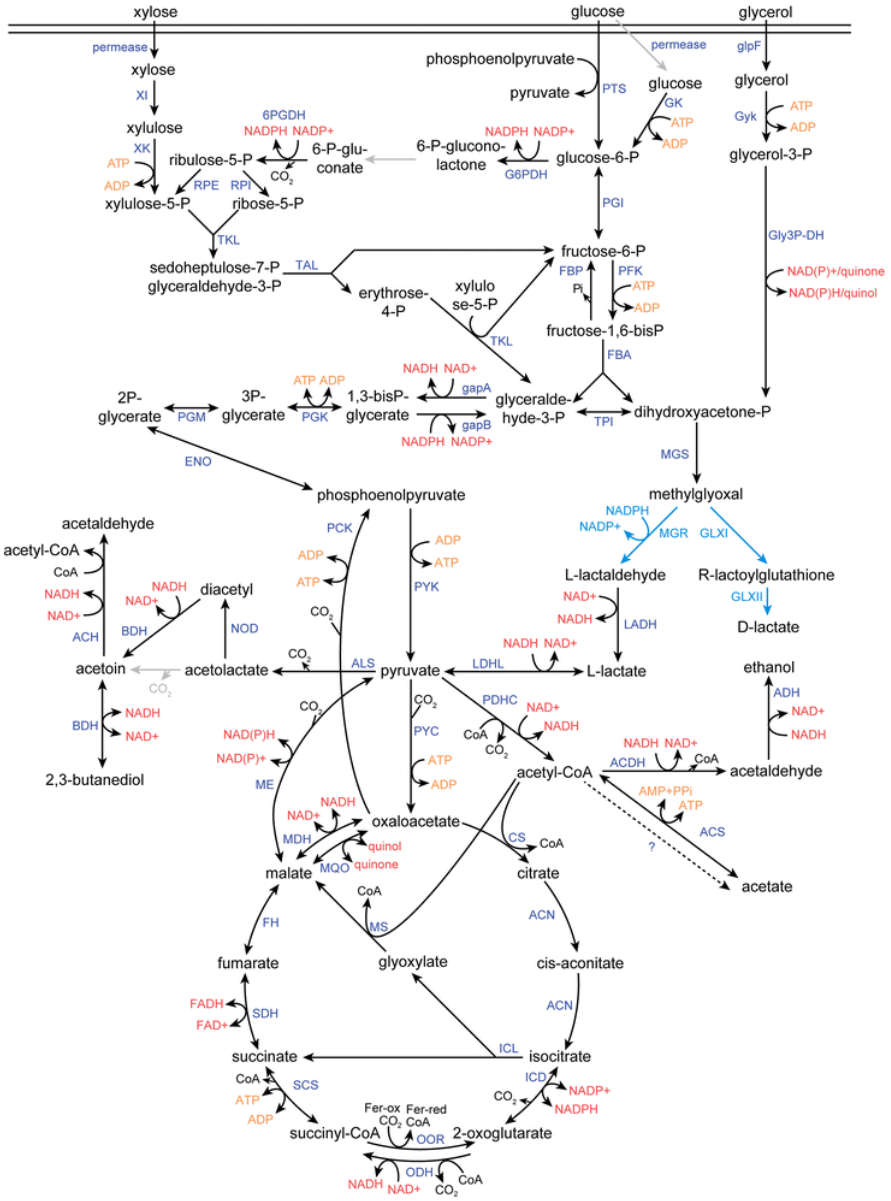

Figure 10.3: ***Reconstruction of central carbon metabolism of Bacillus smithii DSM 4216T.*** *Blue lines indicate pathways based on EC-number identified only via domainome analysis; grey lines indicate pathways found not to be present both by RAST annotation and domainome analysis*

2015). In this paper, an in-depth top-down investigation was performed on the metabolic diversity of *Streptococcus*. Here we also observed that for *Streptococcus* the

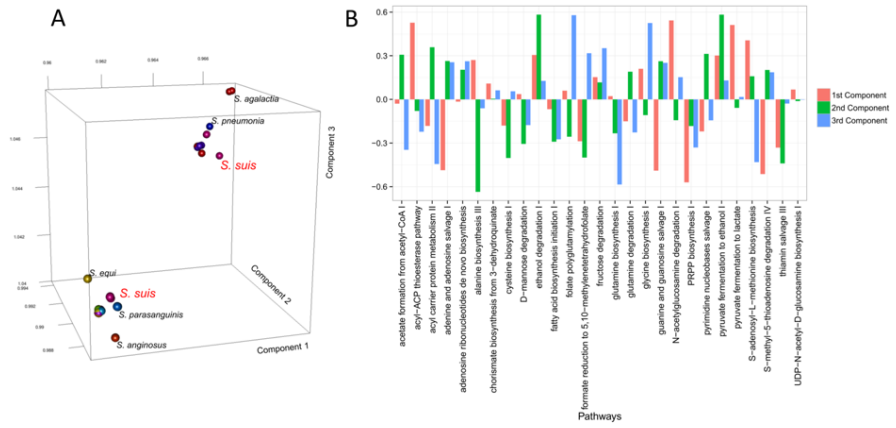

Figure 10.4: A) 3D score plot for the INDSCAL model on the (dis)similarity matrices of the 27 core metabolic pathways of the 121 Streptococcus strains. B) Loadings of the INDSCAL model.

impact of *de novo* structural and functional re-annotation was significant and could yield up to nearly 800 additional domains when current, state-of-the-art, gene prediction methods were applied. Through these domains, enzymatic functions could be inferred and metabolic pathways could be reconstructed resulting in a detailed overview of pathways shared among different numbers of *Streptococcus* strains. For example, generic pathways related to the biosynthesis of ATP, GTP, UTP or amino-acid conversion were found among all analysed strains. More interestingly is the absence of pathways such as the Glutamate degradation I, Glutamate degradation X, Glutamate biosynthesis II in specific species groups, and Glutamate biosynthesis pathway III, which were absent in the clinically important *Streptococcus pyogenes* but were present in all other *Streptococcus* species. Pathway similarity analysis showed that metabolically, *Streptococcus suis* strains, an important pathogen of pigs, can be separated in two main clusters. In depth investigation showed that based on pathway dissimilarity we were able to distinguish serotype 2 from serotype 1/2, 1, 3, 7, 9 and 14 showing a greater distinction than was previously reported (Zhang et al., 2011).

In (Kamminga et al., 2017), we expanded the functional analysis towards *Mycoplasma* species, which are among the smallest parasites known to date. In this study, we investigated the functional diversity among 80 *Mycoplasma* species sequenced strains of this genus. We identified and separated *Mycoplasma* species with

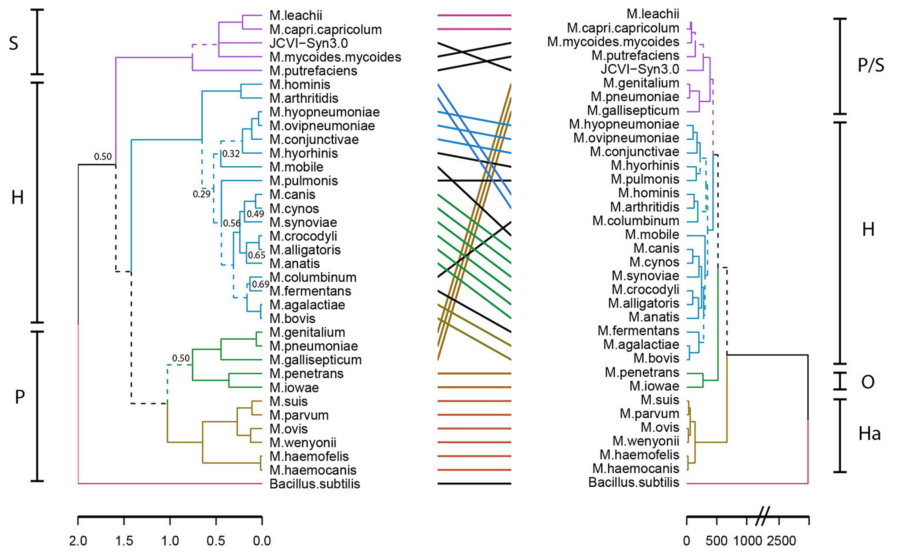

Figure 10.5: ***Niche-driven functional evolution****. Accelerated functional evolution causes separation of haemoplasma species and several other Mycoplasma species when phylogenetic clusters are compared to functional clusters. Dashed lines indicate distinct branches. Left: Standard phylogenetic tree using 16S-rRNA (maximum likelihood, 500x bootstrapped). Right: Functional clustering based on Manhattan distance calculated from the presence/absence matrix of domains. Groups indicated are: S, Spiroplasma; H, Hominis; P, Pneumoniae; Ha, Haemoplasma; and O, Other.*

the capability of infecting the blood or tissue (Figure 10.5). For some groups it was even possible to predict whether the strains were capable of infecting the ruminant, pig or human. The capability of such predictions enables the possibility to zoom in into the functional landscape and identify the genotypic characteristics responsible for these particular phenotypes. Comparative functional genomics was also applied using the "minimal synthetic bacterium JCVI-Syn3.0"as a reference point to identify proteins and corresponding protein domains for minimal life, leading to the interesting observation that not all protein domains essential for JCVI-Syn3.0 "minimal life"were persistent in all of the 80 *Mycoplasmas* species. This suggests that the different *Mycoplasma* species have alternative domain configurations replacing the currently known JCVI-Syn3.0 minimum requirements that solve the issue of life, This suggests that there are many alternative versions of minimal life which are currently not understood.

Having established a protocol for functional comparative genomics, SAPP was

applied to study a set of *Pseudovibrio* species to obtain insights into the symbiotic capacities and metabolic flexibilities of this genus. Through comparative genomics, in which SAPP played a major role in the re-annotation and standardisation of the data, 56 samples (31 *Pseudovibrio* strains and 25 sponge isolates) were functionally analysed (Versluis et al., 2018).

In **chapter 6**, the semantic comparison could be scaled up to hundreds of genomes in which 432 *Pseudomonas* strains were analysed in-depth, which allowed us to identify essential genes and corresponding protein domains based on metabolic models and transposon libraries. These essential domains showed a higher persistence in contrast to non-essential domains. Persistence was further investigated in **chapter 7** where we identified protein domain classes that were specific for a given species. These domains are mostly domains of unknown function, making them interesting targets for further studies to understand their role in a species. We also show that enzymatic functions are most likely to be propagated to different species groups through horizontal gene transfer. This information, which domains are most likely to be found within a given species and only in that species, could be applied for the identification of species in meta-genomic samples. To identify more precisely to which species a novel strain belongs to, the creation of phylogenetic trees based on protein domains allows to be a rapid and scalable approach to determine this. Previous research has shown that this metric corresponds well to the current standard in which 16S-rRNA sequence similarity is used and even with a higher resolution. This could lead the way to more precisely determine species characteristics as well as their position in a phylogenetic tree.

The infrastructure (SAPP) can also be used to scan genomes for proteins or traits of interest. This methodology was applied in (Peng et al., 2017) where in a top-down approach bacterial genomes of high quality (complete chromosomal assembly) from the ENA repository were functionally screened for specific protein domains. These domains, encoding for haloacid dehalogenases, are key enzymes for bacterial strains making them capable of using haloalkanoates as the sole source of carbon and energy. The approach resulted in the identification of *Pseudomonas*, *Xanthobacter* and *Methylobacterium* strains.

As this approach has proven to be successful it was then applied for the identi-

fication of strains capable of producing 1,3-propanediol (**Chapter 8**). Through the use of a repertoire of 80.000 functionally annotated genomes, 187 species were predicted to be capable for the production of this compound of interest. Through the use of federated queries, resources containing information related to pathogenicity could be accessed to identify non-pathogenic species. This resulted in the validation of several species which prove to be capable of producing 1,3-propanediol.

FUTURE PERSPECTIVE

Semantic Systems Biology and model-based Systems Biology are data integration and analysis approaches that strive to achieve complementary goals. Model-based Systems Biology uses mathematical modelling to analyse biological data. Based on prior knowledge, model-based Systems Biology aims to provide either biologically acceptable explanations for this new data, or to develop new hypotheses based on the integration of new and modelled data. Integration and sharing of data, information and knowledge is in the realm of Semantic Systems Biology. The deliberate exploitation of Semantic Web technologies for integration and sharing of heterogeneous bio-data sources with computational predictions and associated meta-data will not only lead to the development of new, testable hypotheses but the ability to directly link data and data provenance, also opened new ways for computational support in quality checking of computationally inferred annotations.

It is essential to accept that data growth in life sciences is not only exponentially but is also becoming more heterogeneous by nature. This trend is reducing the amount of time that can be spent on research and data mining, since most of the efforts now lie in data formatting. Translating it into a format that can communicate with other computational and experimental data is essential as without it, new discoveries will remain to be discovered. In this thesis, we have shown an approach on making genetic data FAIR with the development of SAPP and GBOL, and how it can be used for biotechnological implementations. The GBOL-stack is the fundamental basis on which we can expand and incorporate other experimental resources and ontologies (**Chapter 3**).

The incorporation of biological data into GBOL is continuously explored. For plant studies, the incorporation of QTL and GWAS data from various crops would allow a direct semantic linkage between these data sets with functional and structural annotation. This direct link enables the possibility to study correlations between variants and its impact on the structural and functional landscape.

Similar to QTL and GWAS data, semantic integration of expression data obtained through RNA sequencing techniques (RNAseq) would allow to quickly identify variation in expression between sets of genes, regions within a genome or pathway

fluxes, based on RNA expression levels. RNAseq and other biological data, such as gene knock-out studies for the identification of essentiality or the linkage to phenotypic properties, can be used in both semantic and model-based systems biology data integration and analysis approaches. Tight integration of this data in a semantic framework can be beneficial for the continuous development of biological models, thereby gaining further insights into the inner workings and capabilities of an organism.

The first step towards a higher level of data quality is to make it FAIR. It should be common sense to have your data in a *Findable*, *Accessible*, *Interoperable* manner, which should result in *Reusable* data, even if it is only to be used internally or for a selected group of people.

For each of the biological data-types, consistency checks and evaluations are required to ensure that a data set is updated to the highest possible quality according to the proposed FAIR metrics (Wilkinson et al., 2016). In this thesis, we have shown the pivotal role of Empusa in the development of GBOL and corresponding GBOL-Stack allowing SAPP to incorporate biological data in a consistent and high quality way. When expanding GBOL to incorporate experimental data, expression or GWAS and QTL information these quality checks are essential to ensure that those resources obtain the same level of quality. For other resources, such as genome scale metabolic models, support for quality checks have been recently developed in the form of a quality reporting system, allowing models to be re-evaluated or evaluated during its development process (Lieven et al., 2018).

It is fair to state that Semantic Systems Biology has great potential but is still underexposed and in development. Currently, we have only scratched the surface of what is possible with integration and (re-)usage of biological data. A large proportion of biological resources remain yet to be unlocked and is not used to its full potential.

Bibliography

Alcántara, Rafael et al. (2012). "Rhea - A manually curated resource of biochemical reactions". In: *Nucleic Acids Research* 40. DOI: 10.1093/nar/gkr1126.

Aziz, Ramy K et al. (2008). "The RAST Server: rapid annotations using subsystems technology." In: *BMC genomics* 9. DOI: 10.1186/1471-2164-9-75.

Bolleman, J. et al. (2014). "FALDO: A semantic standard for describing the location of nucleotide and protein feature annotation." In: *Journal of Biomedical Semantics*. DOI: DOI10.1186/s13326-016-0067-z.

Bosma, Elleke F et al. (2016). "Complete genome sequence of thermophilic Bacillus smithii type strain DSM 4216 T". In: *Standards in genomic sciences* 11.

Crick, Francis (1970). "Central dogma of molecular biology". In: *Nature* 227.

Eilbeck, Karen et al. (2005). "The Sequence Ontology: a tool for the unification of genome annotations." In: *Genome biology* 6. DOI: 10.1186/gb-2005-6-5-r44.

Gaudet, Pascale et al. (2015). "The neXtProt knowledgebase on human proteins: current status". In: *Nucleic acids research* 43.D1, pp. D764–D770.

ISA Model & Serialization Specifications 1.0 (October (2016). URL: http://isa-specs.readthedocs.io/en/latest/ (visited on 07/13/2018).

Kamminga, Tjerko et al. (2017). "Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life". In: *Frontiers in cellular and infection microbiology* 7.

Lieven, Christian et al. (2018). "Memote: A community-driven effort towards a standardized genome-scale metabolic model test suite". In: *bioRxiv*.

Miller, J, AD McLachlan, and A Klug (1985). "Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes." In: *The EMBO journal* 4.

Mitraka, Elvira et al. (2015). "Wikidata: A platform for data integration and dissemination for the life sciences and beyond". In: *bioRxiv*. DOI: 10.1101/031971.

Mungall, Christopher J, David B Emmert, FlyBase Consortium, et al. (2007). "A Chado case study: an ontology-based modular schema for representing genome-associated biological information". In: *Bioinformatics* 23. DOI: 10.1093/bioinformatics/btm189.

Peng, Peng et al. (2017). "Concurrent haloalkanoate degradation and chlorate reduction by Pseudomonas chloritidismutans AW-1T". In: *Applied and environmental microbiology* 83.

Saccenti, Edoardo, David Nieuwenhuijse, Jasper J Koehorst, Vitor AP Martins dos Santos, and Peter J Schaap (2015). "Assessing the metabolic diversity of streptococcus from a protein domain point of view". In: *PloS one* 10.

Slenter, Denise N et al. (2018). "WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research". In: *Nucleic acids research* 46.D1. DOI: 10.1093/nar/gkx1064.

Stanford, Natalie J et al. (n.d.). "FAIRDOM: Reproducible Systems Biology through FAIR Asset Management". In:

The Gene Ontology Consortium (2015). "Gene Ontology Consortium: going forward". In: *Nucleic Acids Research* 43. DOI: 10.1093/nar/gku1179.

Timmers, Peer HA et al. (2017). "Reverse methanogenesis and respiration in methanotrophic archaea". In: *Archaea* 2017.

UniProt Consortium, The (2017). "UniProt: the universal protein knowledgebase". In: *Nucleic acids research* 45.D1. DOI: 10.1093/nar/gkw1099.

Versluis, Dennis et al. (2018). "Comparative genomics highlights symbiotic capacities and high metabolic flexibility of the marine genus Pseudovibrio". In: *Genome biology and evolution* 10.

Visser, Michael et al. (2014). "Genome analyses of the carboxydotrophic sulfate-reducers Desulfotomaculum nigrificans and Desulfotomaculum carboxydivorans and reclassification of Desulfotomaculum caboxydivorans as a later synonym of Desulfotomaculum nigrificans". In: *Standards in genomic sciences* 9.

Wilkinson, Mark D et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.

Wolstencroft, Katherine et al. (2015). "SEEK: A systems biology data and model management platform". In: *BMC Systems Biology*. DOI: 10.1186/s12918-015-0174-y.

Wolstencroft, Katy et al. (2011). "RightField: Embedding ontology annotation in spreadsheets". In: *Bioinformatics* 27. DOI: 10.1093/bioinformatics/btr312.

Worm, Petra et al. (2014). "A genomic view on syntrophic versus non-syntrophic lifestyle in anaerobic fatty acid degrading communities". In: *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1837.

Zhang, Anding et al. (2011). "Comparative genomic analysis of Streptococcus suis reveals significant genomic diversity among different serotypes". In: *BMC Genomics* 12. DOI: 10.1186/1471-2164-12-523. URL: https://doi.org/10.1186/1471-2164-12-523.

11 Summary and acknowledgements

Summary

The aim of this thesis was to increase our understanding on how genome information leads to function and phenotype. To address these questions, I developed a semantic systems biology framework capable of extracting knowledge, biological concepts and emergent system properties, from a vast array of publicly available genome information. In **chapter 2**, Empusa is described as an infrastructure that bridges the gap between the intended and actual content of a database. This infrastructure was used in **chapters 3** and **4** to develop the framework. **Chapter 3** describes the development of the Genome Biology Ontology Language and the GBOL stack of supporting tools enforcing consistency within and between the GBOL definitions in the ontology (OWL) and the Shape Expressions (ShEx) language describing the graph structure. A practical implementation of a semantic systems biology framework for FAIR (*de novo*) genome annotation is provided in **chapter 4**. The semantic framework and genome annotation tool described in this chapter has been used throughout this thesis to consistently, structurally and functionally annotate and mine microbial genomes used in **chapter 5-10**. In **chapter 5**, we introduced how the concept of protein domains and corresponding architectures can be used in comparative functional genomics to provide for a fast, efficient and scalable alternative to sequence-based methods. This allowed us to effectively compare and identify functional variations between hundreds to thousands of genomes. In **chapter 6**, we used 432 available complete *Pseudomonas* genomes to study the relationship between domain essentiality and persistence. In this chapter the focus was mainly on domains involved in metabolic functions. The metabolic domain space was explored for domain essentiality and persistence through the integration of heterogeneous data sources including six published metabolic models, a vast gene expression repository and transposon data. In **chapter 7**, the correlation between the expected and observed genotypes was explored using 16S-rRNA phylogeny and protein domain class content as input. In this chapter it was shown that domain class content yields a higher resolution in comparison to 16S-rRNA when analysing evolutionary distances. Using protein domain classes, we also were able to identify signifying domains, which may have important roles in shaping a species.

To demonstrate the use of semantic systems biology workflows in a biotechnological setting we expanded the resource with more than 80.000 bacterial genomes. The genomic information of this resource was mined using a top down approach to identify strains having the trait for 1,3-propanediol production. This resulted in the molecular identification of 49 new species. In addition, we also experimentally verified that 4 species were capable of producing 1,3-propanediol.

As discussed in **chapter 10**, the here developed semantic systems biology workflows were successfully applied in the discovery of key elements in symbiotic relationships, to improve functional genome annotation and in comparative genomics studies. Wet/dry-lab collaboration was often at the basis of the obtained results.

The success of the collaboration between the wet and dry field, prompted me to develop an undergraduate course in which the concept of the "Moist" workflow was introduced (**Chapter 9**)

Dankwoord / Acknowledgements

It has only been so many years since I started my PhD...

Ten eerste wil ik mijn ouders bedanken die mij altijd gesteund hebben en mij alle mogelijkheden hebben geven om niet een, niet twee maar drie studies te kunnen doen om te zijn waar ik nu sta. Het begon misschien allemaal wat hobbelig op de middelbare school maar sinds ik de roeping van Life Sciences gevonden heb ging het als een speer en heb ik altijd met veel plezier gestudeerd en gewerkt.

Lisa, lieve lieve schat, zonder jouw steun en toeverlaat zou het onmogelijk zijn geweest om mijn thesis "op tijd" af te krijgen. Deze periode heeft je gelukkig niet nog gekker gemaakt dan dat je al bent want zo heb ik je het liefst. Van begin af aan heb je me gestimuleerd om ook aan mijn eigen onderzoek te werken en ik moet toegeven dat dit wel eventjes geduurd heeft. Tevens ben je ook de beste grammer checker die ik tot nu toe heb meegemaakt. Hoe vaak je wel niet 's avonds lekker in de tuin zat met een wetenschappelijk artikel in plaats van een mooi boek... Bedankt voor al je steun de afgelopen jaren en we gaan er nu samen van genieten.

Peter, I still remember that I was emailing you from my internship in London to ask if you knew anyone or websites which I could use to search for a job. Instead you offered me this PhD and I have not left since. It has been a wonderful journey with you not only in the department or one of the many cities we visited in project meetings but also our wonderful time in Thailand, who can say that he went on holiday with his boss, and lives to tell the tale! It was always nice to help out during the many classess of Bioinformation Technology and thank you for listening to my complains about missing a course that combines both computational as well as experimental work to which you replied, then why don't you develop one? And so we did...

Maria, I cannot remember the number of times you helped me out with finalizing an article that was on the bench for too long or with R technicalities in which

Stackoverflow could not help me out. It is really a pleasure to work with and discuss all kinds of aspects with you and **Edoardo**, our amazing statistican who happens to live in the wrong century (his words, not mine) but helped me out on many occassions.

Vitor, our traveling salesman, we never know where you are but you bring along so many connections and possibilities that make the group as it is today!

Everybody I have met and worked with during the many years at SSB, it has been fantastic to get to know you all! **Jesse**, without you, the Semantic World would be an empty place. The time we spent together on grinding new ideas, tackling world problems and the development of a new basis on which this department and my thesis has been mostly based upon. Together with **Niels** we have had crazy ideas, many walks and looked into the world of Enterpeneurship. **Niels**, beside being a wonderful colleague and paranymph, it is also nice to have you as my nextdoor neighbour to borrow eggs from and drink a nice bottle of wine. While we were devouring pizzas together with **Ruben** and **Dorett**, **Jesse** and **Michael** were tinkering on our little robots to get it all to work. **Benoit** you were always in for nice discussions mostly based on semantics, and a wonderful addition to the Team and eventually we will finalise TBOL. The largest analysis was of course performed with **Nikolaos**, mining 80.000 genomes just to find some strains that can produce 1,3-PDO. It has been a great puzzle and it was nice working together on this topic. **Wasin**, you arrived when **Jesse** was leaving and you gracefully picked up the latest development of NG-Tax 2.0 and improved it even further. It has been great to work with you on finding out what was inside the code and the fun talks we had in the office while working side by side. **Erika**, you sure know how to organise a nice dinner in the city, the drinks were good and the bill was long and you should organise them more often. **Niru**, a big thanks for helping out with the INFECT project and the nice lunches. It still makes me laugh about how easily you can lose your appetite, I will try to talk about more general topics during the lunch breaks just so you can finish your plate. The Christmas dinner has passed by now but the award for the most happy person at SSB of course goes to **Linde**,

it is great to work with you on *Pseudomonas* and to find out if and how long it can hold its breath. **Bastian**, thank you for all the wonderful times and enthusiasm in our office. It was really great to work with you and discuss about the wonderful world of metagenomics. **Tjerko**, it has been great working with you on one of the smallest organisms, and **Peter** keeps on talking about risk assesments and lean project management since you have passed our department. **Bart**, thank you for being my paranymph and what would SSB be without you. The servers would have burned up and all the data would have been lost. Thank you for keeping everything under control, making it a little less like the Wild West. **Nong**, thank you for the wonderful tour around your country where you have shown us beautiful areas and where we ate at places that I would have never been able to find. **Tom**, **Rita**, thank you for helping out in the iBioSystems course, I sure hope you are having as much fun as I am.

Of course the field of Systems Biology and micro-organisms are intwined on so many levels and therefore I am more than grateful that I have had wonderful oppertunities to collaborate with the bright minds from Microbiology. First of all, a big thanks to **Philippe**, together we designed the fundamentals for iBioSystems and this course has been a great success and we are even expanding outside the WUR! Of course I cannot teach alone and with the wonderful help of **Irene** we are making it all happen. During my Master thesis and just when I was starting my PhD I had the amazing / crazy opportunity to work with **Raymond** on the CRISPR systems together with **John** from which many projects emerged and this collaboration will hopefully remain in the future. Other collaborations with **Janneke**, **Elleke**, **Petra**, **Peer** and **Dennis** amongst others, it was really nice to work with you on your projects.

The students I supervised over the years, **Tristan**, **David**, **Janneke**, **Marco**, **Peter**, **William**, **Guus**, **Wiebe**, **Henk**, **Alex** and **Valerie**. It was great to see how you picked up new techniques, helped me out in the analysis or the development thereof and even published some work.

Last but not least, I would like to thank all the teachers from MLO-OSS, HLO Nijmegen and Wageningen University who provided me with the knowledge and guidance over all those years!

ABOUT THE AUTHOR

Since I was young, I have always been fascinated by the fields of nature and computational systems. It was during an open day at the Sint Jans Lyceum, where I came in contact with a teacher from MLO-OSS who told me all about oil yields in peanuts and all the techniques that came with it. It was fascinating to see how all kinds of techniques were applied and the decision to start my educational career there was made before the day was over. These 4 years were amazing in which I learned all kinds of aspects, from fundamental chemistry to the genetic makeup of organisms but I also realised that my learning experience was far from over. I moved to Nijmegen where I followed a three-year program to obtain my Bachelor in the Life Sciences, where beside studying, I was also attending many of the festivities in and around the city. In the last year, which I spend in Brisbane Australia, I was often drawn towards the computational analysis of the data I generated during the experimental work. This made me realise I wanted to know more about this field, which resulted in following a Master in Bioinformatics at Wageningen University after which I started my scientific career at the Laboratory of Systems and Synthetic Biology. Since then I have worked on numerous projects, joined the iGem team, designed a new course combining computational and experimental work and had the opportunity to teach at a beautiful place in Bangkok. I will continue my scientific career at Wageningen University, and I am looking forward to see what the future will bring.

EDUCATION AND TRAINING

Overview of completed training activities	**Year**
Preparing PhD research proposal	2013
EBI Roadshow (Amsterdam)	2013
Project and time management (Wageningen)	2014
Techniques for Writing and Presenting a Scientific Paper (Wageningen)	2014
Entrepreneurship in and outside science (Wageningen)	2014
Speed reading (Wageningen)	2015
HPC workshop (Amsterdam)	2015

Conferences	**Year**
NBIC (Lunteren)	2013
Symposium Systems Biology for Food, Feed, and Health (Wageningen)	2013
Pattern recognition (Leiden)	2013
Managing and Integrating Information in the Life Sciences (Leiden)	2013
NBIC (Lunteren)	2014
Life science e-infrastructure workshop (Amsterdam)	2014
SBNL Symposium (Maastricht)	2015
Easy-M (Berlin)	2016
Applied Bioinformatics in Life Sciences (Leuven)	2016
Biocuration conference (Geneve)	2016
BioSB conference (Lunteren)	2018
Norway systems biology (Bergen)	2018

Project meetings	**Year**
iGem (Amsterdam)	2012
INFECT (Stockholm)	2013
INFECT (Ericeira)	2013
INFECT (Braunsweich)	2014
INFECT (Braunsweich)	2015
INFECT (Copenhagen)	2015
INFECT (Berlin)	2016
INFECT (Copenhagen)	2017
INFECT Kick-out meeting (Saltsjöbaden)	2017
SAFE-AQUA (Bangkok)	2018
DigiSal (Bergen)	2018
Elixir (Athens)	2018
Elixir (Utrecht)	2018
IBISBA (Barcelona)	2018
NETTAB, Building a FAIR Bioinformatics environment (Genoa)	2018
Human Genetics (Leiden)	2018

Teaching activities	**Year**
Bioinformation Technology (Period 1/5)	2012-2017
Guest lecturer Data Managment (Wageningen)	2014-2015
Toolbox (Period 4)	2016-2017
Big Data Management (VLAG)	2017
iBioSystems (Period 2)	2017-2018
BioTec (Bangkok)	2018

List of publications

Bosma, Elleke F et al. (2016). "Complete genome sequence of thermophilic Bacillus smithii type strain DSM 4216 T". In: *Standards in genomic sciences* 11.

Dam, Jesse C. J. van, Jasper J. Koehorst, Jon Olav Vik, Peter J. Schaap, and Maria Suarez-Diez (2017). "Interoperable genome annotation with GBOL, an extendable infrastructure for functional data mining". In: *bioRxiv*.

Dam, Jesse CJ van, Jasper J Koehorst, Peter J Schaap, Vitor Ap Martins Dos Santos, and Maria Suarez-Diez (2015). "RDF2Graph a tool to recover, understand and validate the ontology of an RDF resource". In: *Journal of biomedical semantics* 6.

Fiebig, Anne et al. (2015). "Comparative genomics of Streptococcus pyogenes M1 isolates differing in virulence and propensity to cause systemic infection in mice". In: *International Journal of Medical Microbiology* 305.

Hesselman, Matthijn C et al. (2012). "The Constructor: a web application optimizing cloning strategies based on modules from the registry of standard biological parts". In: *Journal of biological engineering* 6.

Kamminga, Tjerko et al. (2017). "Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life". In: *Frontiers in cellular and infection microbiology* 7.

Kazakoff, Stephen H et al. (2012). "Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree Pongamia pinnata". In: *PLoS One* 7.

Koehorst, Jasper J, Edoardo Saccenti, Peter J Schaap, Vitor AP Martins dos Santos, and Maria Suarez-Diez (2016). "Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics". In: *F1000Research* 5.

Koehorst, Jasper J, Jesse CJ Van Dam, et al. (2016). "Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data". In: *Scientific reports* 6.

Koehorst, Jasper J. et al. (2017). "SAPP: functional genome annotation and analysis through a semantic framework using FAIR principles". In: *Bioinformatics* 1.

Lieven, Christian et al. (2018). "Memote: A community-driven effort towards a standardized genome-scale metabolic model test suite". In: *bioRxiv*.

Mehboob, Farrakh et al. (2016). "Genome and proteome analysis of Pseudomonas chloritidismutans AW-1T that grows on n-decane with chlorate or oxygen as electron acceptor". In: *Environmental microbiology* 18.

Ouwerkerk, Janneke P et al. (2017). "Complete Genome Sequence of Akkermansia glycaniphila Strain PytT, a Mucin-Degrading Specialist of the Reticulated Python Gut". In: *Genome announcements* 5.

Peng, Peng et al. (2017). "Concurrent haloalkanoate degradation and chlorate reduction by Pseudomonas chloritidismutans AW-1T". In: *Applied and environmental microbiology* 83.

Quax, Tessa EF et al. (2013). "Differential translation tunes uneven production of operon-encoded proteins". In: *Cell reports* 4.

Saccenti, Edoardo, David Nieuwenhuijse, Jasper J Koehorst, Vitor AP Martins dos Santos, and Peter J Schaap (2015). "Assessing the metabolic diversity of streptococcus from a protein domain point of view". In: *PloS one* 10.

Staals, Raymond HJ, Yoshihiro Agari, et al. (2013). "Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus". In: *Molecular cell* 52.

Staals, Raymond HJ, Yifan Zhu, et al. (2014). "RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus". In: *Molecular cell* 56.

Swarts, Daan C, Jasper J Koehorst, Edze R Westra, Peter J Schaap, and John van der Oost (2015). "Effects of Argonaute on gene expression in Thermus thermophilus". In: *PLoS One* 10.

Timmers, Peer HA et al. (2017). "Reverse methanogenesis and respiration in methanotrophic archaea". In: *Archaea* 2017.

Versluis, Dennis et al. (2018). "Comparative genomics highlights symbiotic capacities and high metabolic flexibility of the marine genus Pseudovibrio". In: *Genome biology and evolution* 10.

Visser, Michael et al. (2014). "Genome analyses of the carboxydotrophic sulfate-reducers Desulfotomaculum nigrificans and Desulfotomaculum carboxydivo-

rans and reclassification of Desulfotomaculum caboxydivorans as a later synonym of Desulfotomaculum nigrificans". In: *Standards in genomic sciences* 9.

Worm, Petra et al. (2014). "A genomic view on syntrophic versus non-syntrophic lifestyle in anaerobic fatty acid degrading communities". In: *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1837.

Colophon

The research described in this thesis was financially supported by the European Union Horizon2020 projects, INFECT (Project reference: 321529), EmPowerPutida (Project reference: 635536) and MycoSynVac (Project reference: 634942) and by the NWO project SafeChassis (Project reference: 15814). Printed by: Proefschrift maken