



QfO-7

QFO-7
Novel Challenges in the Quest for
Orthologs
Sitges, Spain
17.-18. September 2022



SENCKENBERG
world of biodiversity

GOETHE 
UNIVERSITÄT
FRANKFURT AM MAIN

Table of Contents

| | |
|----------------------------------|----|
| Program | 3 |
| Abstracts – Talks | 5 |
| Abstracts – Posters | 25 |
| Information for Presenters | 37 |
| List of Participants | 38 |
| Conference Notes | 43 |
| Addresses and Maps..... | 44 |
| Stay Safe..... | 45 |

Program

Day 1 - Saturday, September 17th 2022

08:00 Open Registration

09:00 Welcome Ingo Ebersberger / Christophe Dessimoz

Session 1 Chair: Paul Thomas

09:10 Natasha Glover Bringing orthology to the public in the light of evolution

09:35 Yannis Nevers Multifacet quality assessment of gene repertoire annotation with OMArk

10:00 Diego Fuentes PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies

10:30 Coffee Break

Session 2 - Chair: T.B.A.

11:00 Ivana Pilizota A new approach for efficient storage and retrieval of homology data

11:25 Dannie Durand Modeling the Evolution of Multidomain Architectures

11:50 David Moi Reconstructing protein interactions across time using phylogeny-aware graph neural networks

12:15 Lunch Break

Round Table Discussion 1

13:30

13:50

14:10

14:30 Wrap up - RTD1

14:50 Coffee Break

Session 3 - Chair: Salvatore Cosentino

15:20 Damian Szklarczyk Comprehensive interactome and functional annotation using STRING database.

15:45 Paul Thomas A complete draft human functionome as determined by the Gene Ontology Phylogenetic Annotation Project

16:10 Jaime Huerta Cepas Functional and evolutionary significance of unknown genes from uncultivated taxa

16:35 Ikuo Uchiyama Recent developments in MBGD and its application to genomic functional inference

17:00 Poster session

19:00 Conference Dinner

Day 2 - Sunday, September 18th 2022

08:00 Open Registration

09:00 Welcome and General Announcements

Session 4 - Chair: Michael Hiller

09:10 Salvatore Cosentino SonicParanoid enhanced by machine learning allows fast de novo orthology inference of huge MAG datasets

09:35 Victor Rossier Eliminating the bottleneck of orthology inference with OMAmer unleashes the full potential of comparative genomics

10:00 Sina Majidian BIOQA: toward a representative benchmark dataset of biological questions/answers involving orthology, gene expression, and complementary omics data

10:30 Coffee Break

Round Table Discussion 2

11:00

11:20

11:40

12:00 Wrap up RTD2

12:20 Lunch

Session 5 - ECCB2022 Workshop NTB-11 Chair: Ingo Ebersberger (GU)

14:00 Welcome and opening of the workshop

14:05 Nicola Bordin (UCL) AlphaFold 2, ultra-fast structural comparisons and deep learning to distinguish relatives

14:40 Claire Hu (Harvard) DIOPT: an integrative resource of ortholog/paralog prediction

15:15 Felix Langschieb (U. Frankfurt) NcOrtho: Accurate identification of microRNA orthologs

15:40 Coffee Break

16:10 Luis Pedro Coelho (Fudan University) Big catalogs and small genes. Finding structure in prokaryotic genes using metagenomics

16:45 Christian Zmasek (JCVI) Classifying Viral Proteins into Strict Ortholog Groups Using Domain-architecture Aware Inference of Orthologs

17:20 Thomas Richards (Oxford) The Darwin tree of life project and linking large inventories of genome data to understanding orthologue groups and phenotype evolution

17:55 Wrap up QfO-7 Ingo Ebersberger / Christophe Dessimoz

18:00 End of Conference

Abstracts – Talks

| Talk No. | Name | Institute | Country | Titel |
|----------|---------------------|---|----------------|---|
| 01 | Natasha Glover | University Oxford | Switzerland | Bringing orthology to the public in the light of evolution |
| 02 | Yannis Nevers | University of Lausanne | Switzerland | Multifacet quality assessment of gene repertoire annotation with OMArk |
| 03 | Diego Fuentes | IRB Barcelona - Institute for Research in Biomedicine | Spain | PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies |
| 04 | Ivana Pilizota | EMBL-EBI | United Kingdom | A new approach for efficient storage and retrieval of homology data |
| 05 | Dannie Durand | Carnegie Mellon University | USA | Modeling the Evolution of Multidomain Architectures |
| 06 | David Moi | University of Lausanne DBC | Switzerland | Reconstructing protein interactions across time using phylogeny-aware graph neural networks |
| 07 | Damian Szklarczyk | UZH / SIB Swiss Institute of Bioinformatics | Switzerland | Comprehensive interactome and functional annotation using STRING database. |
| 08 | Paul Thomas | University of Southern California | USA | A complete draft human functionome as determined by the Gene Ontology Phylogenetic Annotation Project |
| 09 | Jaime Huerta Cepas | CBGP (UPM-INIA/CSIC) | Spain | Functional and evolutionary significance of unknown genes from uncultivated taxa |
| 10 | Ikuo Uchiyama | National Institute for Basic Biology | Japan | Recent developments in MGD and its application to genomic functional inference |
| 11 | Salvatore Cosentino | University of Tokyo | Japan | SonicParanoid enhanced by machine learning allows fast de novo orthology inference of huge MAG datasets |
| 12 | Victor Rossier | University of Lausanne | Switzerland | Eliminating the bottleneck of orthology inference with OMAmer unleashes the full potential of comparative genomics |
| 13 | Sina Majidian | University of Lausanne | Switzerland | BIOQA: toward a representative benchmark dataset of biological questions/answers involving orthology, gene expression, and complementary omics data |
| 14 | Nicola Bordin | University College London | United Kingdom | AlphaFold 2, ultra-fast structural comparisons and deep learning to distinguish relatives |
| 15 | Claire Hu | Harvard Medical School | USA | DIOPT: an integrative resource of ortholog/paralog prediction |
| 16 | Felix Langschieid | Goethe University Frankfurt | Germany | NcOrtho: Accurate identification of microRNA orthologs |
| 17 | Luis Pedro Coelho | Fudan University | China | Big catalogs and small genes. Finding structure in prokaryotic genes using metagenomics |
| 18 | Christian Zmasek | J Craig Venter Institute | USA | Classifying Viral Proteins into Strict Ortholog Groups Using Domain-architecture Aware Inference of Orthologs |
| 19 | Thomas Richards | University Oxford | United Kingdom | The Darwin tree of life project and linking large inventories of genome data to understanding orthologue groups and phenotype evolution |

Bringing orthology to the public in the light of evolution

Natasha Glover, Marie-Claude Blatter, Monique Zahn, and Christophe Dessimoz

Swiss Institute of Bioinformatics

There is an increasingly large gap between scientists and the people who's research it's meant to benefit - the public. The decline in **science** literacy, specifically in biology and evolution, has become especially apparent during the past few years.

Here, I present "In the Light of Evolution," a project for science communication centered around various topics in evolution. This 3-year project aims to foster the interest and curiosity of a public of all ages about the evolution of species. The main output is a website, <https://lightofevolution.org/en/>, which hosts a series of stories (i.e. easy-to-understand articles) about evolution, genomics, phylogenetics, as well as related interactive workshops.

Some concrete questions which are currently addressed in the Light of Evolution project are: How many genes do humans and bananas have in common? Why is it important to study the evolution of coronaviruses? What is the link between tyrannosaurus and chicken? And, how can we use ancient DNA to investigate the migration of ancient peoples? These articles are written in a simplified, pedagogic way, and are geared towards engaging and creating a dialog with the public. The hands-on and online activities reinforce the concepts learned.

As orthology is a cornerstone for evolutionary studies, several of the stories and activities in the Light of Evolution project are centered around orthology. I will present examples of these stories and activities with the hope that orthology researchers and users can get ideas for outreach activities. These ideas are useful for educational resources and obtaining publicly-funded grants focused on orthology.

Multifacet quality assessment of gene repertoire annotation with OMArk

Yannis Nevers, Victor Rossier, and Christophe Dessimoz

University of Lausanne

Assessing the quality of protein-coding gene repertoires inferred from genome annotations (i.e. proteome) has become critical in an era of increasingly abundant genome sequences for a widening diversity of species. State-of-the-art quality assessment tools such as BUSCO measure the completeness of a gene repertoire - using a limited set of conserved genes - but are blind to other types of errors in genome annotation, such as over-prediction of protein-coding genes from non-coding or contaminant genomic regions.

To overcome these limitations, we developed OMArk, a software relying on fast, alignment-free sequence comparisons with precomputed ortholog groups across the tree of life to assess not only the completeness, but also the consistency of a given proteome with those of related species. Specifically, we compare the test proteome with an expected model obtained from ancestral proteome reconstruction by OMA on ~2500 reference proteomes across the tree of life. Completeness is measured as the proportion of the conserved gene families that are found in the proteome in one or multiple copies and consistency as the proportion of genes with clear homologs in gene families known to exist in the target lineage. Moreover, by comparing each target sequence to its homologous counterparts, OMArk also reports which genes are likely to be fragmented or to have divergent gene structure. Finally, OMArk evaluates contamination based on the taxonomic distribution of gene families with homologs in the proteome.

We validate OMArk with simulated data, then perform a global analysis of a publicly available dataset of 1805 publicly available eukaryotic proteomes—identifying examples of data quality issues.

OMArk is available on GitHub (<https://github.com/DessimozLab/OMArk>) and will soon be available as an interactive online tool at <https://omark.omabrowser.org>.

PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies

Diego Fuentes and Toni Gabaldón

Barcelona Supercomputing Center (BSC) / Institute for Research in Biomedicine (IRB)

Gene phylogenies represent the evolutionary relationships across genes in different species. These phylogenetic trees are commonly used to aid in the inference of homology relationships (i.e. orthology and paralogy) as well as of evolutionary relevant events such as family expansions, recombination and horizontal gene transfer. The plurality of evolutionary histories of genes encoded by an organism's genome is best represented by a genome-wide collection of phylogenetic trees (i.e.: a phylome).

PhylomeDB is a free, accessible and comprehensive web-server of genome-wide collections of gene phylogenies. It was first described in 2006 and has been regularly updated and expanded over the years. The current version, PhylomeDB v5, hosts over 8 million maximum likelihood trees in 534 public phylomes, being the largest public repository of gene trees. Moreover, it also provides homology relationships for genes in over 6000 species through seamless MetaPhOres integration.

A new approach for efficient storage and retrieval of homology data

Kevin Gao, Ivana Pilizota, Thiago Genez, Cristina Guijarro-Clarke, Botond Sipos, Thomas Walsh, David Thybert, and Fergal Martin

EMBL-EBI

Homology inference lies at the core of comparative genomics and is necessary for downstream analyses, including gene function prediction, studying gene family evolution and species tree reconstruction. With the large-scale sequencing initiatives such as the Darwin Tree of Life (<https://www.darwintreeoflife.org>), storing, managing and accessing homology data has become a pressing issue for comparative genomics data resources and services. In general, the number of putative homologous relationships increases quadratically with the number of species. Hence, new efficient and scalable approaches for storage and retrieval of homology data are highly desired in order to provide the data to the users. For example, in Ensembl (www.ensembl.org) the homology relationships between genes are stored in a database table where one entry represents a putative homologous relationship between two genes. Increasing the number of species from ~200 to ~300, increased the number of homology entries from ~400 million to >1 billion. Thus, storing all pairwise homologous relationships will not be feasible for tens of thousands of genomes.

To solve this problem, we took advantage of the hierarchical structure of a gene tree and avoided storing the details of all putative homologous relationships. This allowed us to reduce the complexity of homology storage from $O(n^2)$ to $O(n \log n)$. Furthermore, we indexed the trees using an interval-based labelling of tree nodes in order to parse the trees and extract homology information in a timely manner. Python implementation of the labelling scheme yielded up to 5x speedup on test datasets queried with a Python API. We are currently working on C++ implementation and storing the gene trees in a binary format to further optimise the storage and data retrieval.

Modeling the Evolution of Multidomain Architectures

Xiaoyue Cui, Maureen Stolzer, and Dannie Durand

Carnegie Mellon University

Accurate ortholog prediction requires a detailed understanding of the evolutionary forces that shape homologous protein families. Multidomain protein families pose special challenges for orthology prediction. These families evolve by domain insertions, duplications, and deletions, resulting in variations in domain architecture (the sequence of domains in N- to C-terminal order) within a single family. Only a tiny fraction of possible domain combinations is observed in nature. While this suggests that domain order and co-occurrence are highly constrained, these constraints are poorly understood.

Here, we present new methods for evolutionary analysis of multidomain families. First, we introduce a stochastic model of domain architecture evolution that reflects the stringent constraints on domain architecture composition. This model is implemented in DomArchov, a simulator of domain architecture evolution that uses data-driven transition probabilities to capture lineage-specific forces acting on domain gain and loss.

Second, we introduce a machine learning framework to assess how well DomArchov recapitulates the properties of genuine domain architectures. Domain architectures are represented as points in a high dimensional space that places architectures with similar domain content and order in close proximity. Our model uses neural networks from natural language processing to learn representations that capture these properties. Sets of domain architectures can then be compared by superimposing the corresponding sets of points. Using this framework, we demonstrate that the agreement between genuine and simulated domain architectures exceeds chance expectation. This framework promises broad applicability beyond simulator performance assessment. We are currently investigating the use of this framework to compare the sets of domain architectures across genomes and across functional classes.

Reconstructing protein interactions across time using phylogeny-aware graph neural networks

David Moi and Christophe Dessimoz

University of Lausanne DBC

Genes which are involved in the same biological processes tend to co-evolve. Thus, metabolic pathways, protein complexes, and other kinds of protein-protein interactions can be inferred by looking for correlated patterns of gene retention and loss across the tree of life—a technique called phylogenetic profiling. Recent methodological developments on phylogenetic profiling have focused on scalability improvements to take advantage of the rapidly accumulating genomic data. However, state-of-the-art methods assume that the correlation resulting from co-evolving proteins is uniform across all species considered. This is reasonable for interactions already present at the root of the species considered, but less so for ones that emerge in more recent lineages. To address this challenge and take advantage of recent developments in deep learning methods, we introduce a phylogenetic profiling method based on orthology data which processes large gene co-phylogenies using neural networks. We show that post-processing conventional phylogenetic profiles using deep neural networks can improve predictions, but requires onerous training on specific phylogenies. Overcoming this limitation by taking the topology of the species tree as an input, Graph Neural Networks are shown to outperform all other methods when interaction detection is not centered on just one species of interest, while also predicting when interactions appeared and in which taxa they are present.

Comprehensive interactome and functional annotation using STRING database.

Damian Szklarczyk, Radja Hachilif, and Christian von Mering

SIB / University of Zurich

In-depth understanding of the protein function has to consider the available functional annotations, such as GO terms or pathway membership and its interaction neighbourhood. STRING is the database of protein-protein interaction and the enrichment analysis tool that aims to fully functionally describe the proteins of more than 14000 fully sequenced proteomes. It integrates known and predicted protein-protein associations from multiple sources and represents them as a single benchmarked score for each interaction link giving a complete overview of the protein's role in the organism. Such networks are, among other ways, explorable through a user friendly web-interface, accessible via REST API and downloadable in the form of the flat-files for further analysis.

Every day new genomes are sequenced and existing genomes are re-sequenced and re-annotated. In the new version of STRING the user can submit any fully sequenced genome for complete network and functional annotation. To do so. will require only a minimal input from the user in a form of proteome in a fasta format or a genbank format (GBFF) and, if known, a taxonomical clade of the given genome. Utilizing hierarchical orthology relationships from the eggNOG database precomputed for more than 1600 taxonomical clades the STRING algorithm chooses the most precise orthologous group for each protein and leverages all the data already included in the database for this group and its parent groups to predict both the functional neighbourhood and the functional annotation of each protein in the user input. The resulting dataset is explorable in the same manner as any other organism present in the STRING database: it's sharable, with fully working user-interface, extensive REST API access, set of enrichment analysis tools and clustering methods, and with simple flat-files available for download.

A complete draft human functionome as determined by the Gene Ontology Phylogenetic Annotation Project

Paul Thomas¹, Marc Feuermann², Huaiyu Mi¹,
Pascale Gaudet², Anushya Muruganujan¹, and Dustin Ebert¹

¹ University of Southern California

² Swiss Institute of Bioinformatics

The roadmap for the GO Phylogenetic Annotation project was laid out in 2011. After more than 10 years of work, we are announcing the completion of the first phase of the project: the annotation of all “annotatable” gene families containing human genes. By combining two methods of knowledge unification— ontological and evolutionary— as well as expert biocuration, we have created a “phylogenetically selected” set of consistent GO annotations for human genes (and related genes in other organisms). The process includes selecting not only among annotations with direct experimental support, but also high confidence homology inferences using explicit evolutionary modeling. We have created a set of open, explicit models of function evolution across nearly 9000 protein families, with feedback and review mechanisms to allow efficient updating, modification and extension of the models as additional scientific knowledge becomes available. Our initial analysis of these models and annotations provides insights into the nature of function evolution and the importance of gene duplication, as well as a quantitative estimate of the contribution of model organism studies to our current understanding of human gene function.

Functional and evolutionary significance of unknown genes from uncultivated taxa

Álvaro Rodríguez del Río¹, Joaquín Giner-Lamia¹, Carlos P. Cantalapiedra¹, Jorge Botas¹, Ziqi Deng¹, Ana Hernández-Plaza¹, Lucas Paoli², Thomas S.B. Schmidt³, Shinichi Sunagawa², Peer Bork³, Luis Pedro Coelho⁸, and Jaime Huerta-Cepas¹

1. Centro de Biotecnología y Genómica de Plantas - CBGP, (UPM-INIA/CSIC), Madrid, Spain
2. Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, Switzerland
3. Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
4. Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China

Most microbes on our planet remain uncultured and poorly studied. Recent efforts to catalog their genetic diversity have revealed that a significant fraction of the observed microbial genes are functional and evolutionary untraceable, lacking homologs in reference databases. Despite their potential biological value, these apparently unrelated orphan genes from uncultivated taxa have been routinely discarded in metagenomics surveys. As a result, the unique genetic repertoire of uncultivated microbial lineages has not yet been incorporated into reference databases of protein domains, orthologous groups or microbial gene families. Thus, despite the astonishing amount of data generated by metagenomics sequencing, their study still rely on reference resources that are heavily biased towards the genetic pool of fully sequenced microorganisms, leaving a gap in our knowledge base and impeding our ability to investigate the true diversity of microbially encoded genes on Earth.

In this talk, I will present our most recent results on the discovery of thousands of novel orthologous groups out of massive metagenomics data, their evolutionary significance and putative functional role. In addition, I will discuss the most important challenges found at the technical and methodological level for de novo orthology delineation, functional prediction and data visualization on very large datasets.

In our work, we analyzed a global multi-habitat dataset covering 151,697 medium and high-quality metagenome assembled genomes (MAGs), 5,969 single-amplified genomes (SAGs), and 19,642 reference genomes, identifying 413,335 highly curated novel protein families under strong purifying selection out of previously considered orphan genes. These new protein families, representing a three-fold increase over the total number of prokaryotic orthologous groups described to date, spread out across the prokaryote phylogeny, can span multiple habitats, and are notably overrepresented in recently discovered taxa. By genomic context analysis, we pinpointed thousands of unknown protein families to phylogenetically conserved operons linked to energy production, xenobiotic metabolism and microbial resistance. Most remarkably, we found 980 previously neglected protein families that can accurately distinguish entire uncultivated phyla, classes, and orders, likely representing synapomorphic traits that fostered their divergence.

The systematic curation and evolutionary analysis of the unique genetic repertoire of uncultivated taxa opens new avenues for understanding the biology and ecological roles of poorly explored lineages at a global scale.

Recent developments in MBGD and its application to genomic functional inference

Ikuo Uchiyama¹, Motohiro Mihara², Dynacom Co. Ltd.
Hiroyo Nishide¹, Hirokazu Chiba³, Masahiko Takayanagi⁴, and Hideto Takami^{5,6}

¹National Institute for Basic Biology

²Dynacom Co. Ltd.

³Database Center for Life Science

⁴Web Brain Co. Ltd.

⁵Atmosphere and Ocean Research Institute

⁶The University of Tokyo

The microbial genome database for comparative analysis (MBGD) is a comprehensive ortholog database for microbial genome comparison. MBGD now contains 15,397 microbial genomes belonging to 1,444 genera and 4,747 species, and comprehensive ortholog analysis among them using a hierarchical ortholog construction method resulted in more than 1 million ortholog groups. We are also making several improvements to effectively utilize this large-scale orthologous data. Especially, we are focusing on utilizing the database to analyze user genome data including metagenome assembled genomes. For analyzing the user's genome data, we have added a rapid ortholog assignment mode in the MyMBGD mode based on MMSeqs profile search against the standard ortholog table. On the basis of this orthology assignment, the user can evaluate the metabolic potential of each query genome using the Genomapple software (formerly MAPLE; Takami et al. 2016) to calculate the module completion ratio (MCR) for each KEGG Module entry. The MCRs of the KEGG Modules in the query genome can be compared with those of other genomes to characterize the unique metabolic functions in the query genome. Moreover, the users can check which genes are missing in the genome when the MCR is not 100%. The phylogenetic profile (PP) method is another approach to utilize the large-scale orthology information to infer gene functions. We implemented a species-aware method to evaluate PP similarities (Ruano-Rubio et al. 2009) and create a new interface to search for ortholog groups that have similar phylogenetic profiles to the specified profile, which can be taken from the specific phenotype or habitat. Applying this phylogenetic profile function, we are now developing a tool to searching for candidates of genes that can substitute for the missing genes identified by the Genomapple analysis.

SonicParanoid enhanced by machine learning allows fast de novo orthology inference of huge MAG datasets

Salvatore Cosentino and Wataru Iwasaki

Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo

Accurate inference of orthologous genes constitutes a prerequisite for genomic and evolutionary studies. SonicParanoid is one of the fastest methods for orthology inference and comparably accurate to well-established methods despite being orders of magnitude faster. Nevertheless, its scalability is hampered by the lengthy all-vs-all alignments, and sequence-similarity search alone is not enough to predict very distant orthologs. In this work we try to tackle these two limitations using machine learning.

We substantially reduced the all-versus-all alignment execution time using an AdaBoost model which exploits the properties of the Bidirectional-Best-Hit and the factors affecting the computational time in local sequence alignment. Evaluation based on multiple datasets showed reductions in execution time up to 50% without negative effects on the accuracy of the orthology inference.

To address the second limitation we trained a doc2vec model with domain-architectures extracted from the input proteins, and we used it to infer orthologs based on domain-architecture similarities, resulting in an increase of one-quarter in the number of predicted orthologs. The time required to perform the domain-based orthology inference grows linearly to the number of input proteomes, making it highly scalable.

Lastly, we evaluated the scalability using a dataset of 2000 MAGs. SonicParanoid was able to infer the orthologs in about 38 hours using 128 CPUs, where other well-established orthology inference tools were still running after two weeks.

The way we reduced all-vs-all execution time could be used by other graph-based methods, while the domain-based approach could in the future, thanks to its scalability, eliminate the need for sequence-based homology searches in orthology inference.

Documentation is available at: <http://iwasakilab.k.u-tokyo.ac.jp/sonicparanoid>

PyPI: <https://pypi.org/project/sonicparanoid>

Eliminating the bottleneck of orthology inference with OMAMer unleashes the full potential of comparative genomics

Victor Rossier^{1, 2, 3}, Alex Warwick Vesztröcy^{1, 3}, Marc Robinson-Rechavi^{2, 3}, and Christophe Dessimoz^{1, 3}

¹ Department of Computational Biology

² Department of Ecology and Evolution

³ University of Lausanne

Comparative genomics is a powerful approach to study evolution and discover the genetic basis of phenotypes. At the core of this approach lies the ability to identify comparable genes across species: the orthologs. However, methods to infer orthologs struggle to cope with the deluge of NGS data. Recently, fast methods that place protein sequences into groups of orthologs, or on gene trees, have gained popularity (Cantalapiedra et al. 2021; Tang et al. 2019; Emms and Kelly 2022). However, their reliance on sequence alignments inherently limits their scalability. To overcome this challenge, we introduced OMAMer, an alignment-free method which can process an entire human proteome within a few minutes on a laptop. In this highlight talk, I will outline the method, and present two unpublished applications of OMAMer. In the first, we investigated the genetic basis of convergent venom evolution by reconstructing protein repertoires of 68 venomous and closely related non-venomous animal species, and identifying gene expansions associated with the emergence of venom. In the second application, OMAMer was used to scale-up the inference of orthologs and in-paralogs for 363 bird genomes recently released by the B10K initiative (Feng et al. 2020). With this dense species sampling, we were able to characterise the role of gene duplications and losses for more than 10 convergent adaptations in birds, such as diving in penguins and auks, or loss of flights in ostriches and kiwis. We therefore believe that methods like OMAMer will help pave the way toward biologically informative "Big data" comparative genomics.

BIOQA: toward a representative benchmark dataset of biological questions/answers involving orthology, gene expression, and complementary omics data

Borbala Banfalvi¹, Petros Liakopoulos², Xinyi Wang¹, Christophe Dessimoz², [Sina Majidian](#)²,
and Ana Claudia Sima²

¹ University College London

² University of Lausanne

What would it take for Siri or Alexa to assist us in retrieving relevant orthology, gene expression, and other kinds of biological data? Question Answering (“QA”) is the subfield in information science which seeks to build systems to answer questions posed by humans in natural language. QA has made rapid progress in recent years thanks to the deep learning revolution. However, such supervised learning methods require large training datasets, and only few QA benchmark sets are available in the context of biological data, let alone orthology data.

To address this issue, we are working toward the creation of a large and representative compendium of questions and answers, whereby questions are identified by surveying the scientific literature, and answers are federated SPARQL queries which can retrieve the relevant data from databases with public SPARQL endpoints.

In this talk, we will present this endeavour and distil what we have learned from reading and analysing over 300 scientific papers which integrate orthology information alongside complementary databases (e.g. gene expression, Gene Ontology annotations, or protein-protein interactions). In particular, we will report on the kinds of questions we could extract from these papers, which constitute a survey of the uses of orthology in the literature.

Ultimately, our benchmark dataset will enable the application, evaluation, and improvement of QA systems in the context of orthology data, and in turn increase the use and impact of orthology databases in scientific research.

AlphaFold 2, ultra-fast structural comparisons and deep learning to distinguish relatives

Nicola Bordin

University College London

Breakthrough methods in protein structure prediction, novel ultra-fast structural aligners and AI are revolutionizing structural biology.

Experimental structures were, until recently, too sparse and available mostly for well-studied organisms. Obtaining accurate models of proteins, annotating and validating their functions and evolutionary relationships are no longer limited to certain species or by time and resources.

Deep-learning methods based on embeddings from protein language models outperform traditional HMM-based tools for detecting extremely remote homologs and orthologs, as well as diverging paralogs, allowing for fast and precise assignments in sequence space.

De-novo structure prediction of these sequences with AlphaFold2 and RoseTTAFold 2 in most cases generate models with qualities comparable to experimental techniques.

The availability of precomputed 3D models for the entirety of UniProt in the AlphaFold Protein Structure Database enabled a new generation of ultra-fast aligners such as Foldseek to traverse the protein structure space in the search for relatives.

Assessing these relationships, building a model and validating its relationships are now achievable in a reasonable timescale, allowing for different approaches previously unattainable, such as multi-domain architecture scanning, domain and chain archaeology and evolution, as well as detecting elusive relatives in different branches of the Tree of Life.

These methods enabled the expansion of homology-based superfamilies, the detection of new folds and architectures, as well as the generation of a new framework for protein structure classification algorithms.

In this talk we'll showcase how these three major breakthroughs complement each other and could radically change the approaches used to search for relatives in sequence and structure space.

DIOPT: an integrative resource of ortholog/paralog prediction

Yanhui Hu, Aram Comjean, Norbert Perrimon, and Stephanie Mohr

Harvard Medical School

Evolutionarily related genes, known as orthologs, tend to encode proteins with similar functions at a biochemical, cellular, and organismal level. Having the ability to quickly predict orthologs across species is arguably one of the most important tools for functional genomic studies. We developed the DRSC Integrative Ortholog Prediction Tool (DIOPT), which combines results from multiple ortholog prediction algorithms. Importantly, DIOPT provides a more sensitive and specific mapping than could be achieved by any given resource. Indeed, we found that the number of resources that predict a given ortholog pair provides a measure of confidence in the results and display the count as part of the DIOPT output results (DIOPT score). After its launch, DIOPT quickly became our most-used online resource. We have continued to update and support DIOPT by adding new algorithms and species, as well as improving DIOPT user-interface based on user feedback. DIOPT-based ortholog mapping is incorporated into other resources including the Alliance for Genome Resources (AGR), human disease centered resources (eg MARRVEL and ModelMatcher), FlyBase and other model organism databases. Customized support is provided for this type of collaboration. For example, to meet the specific requirement of AGR, we have been working with Quest for Ortholog consortium to provide the integrated ortholog and paralog mapping using selected algorithms based on benchmark result. Altogether, DIOPT resource provides an easy-to-use online portal and customized ortholog/paralog mapping for use at external databases and data pipelines.

NcOrtho: Accurate identification of microRNA orthologs

Felix Langschie¹, Matthias Leisegang², Ralf Brandes^{2,3}, and Ingo Ebersberger^{1,4,5}

¹Applied Bioinformatics Group, Inst. of Cell Biology and Neuroscience, Goethe University Frankfurt

²Institute for Cardiovascular Physiology, Goethe University Frankfurt

³German Center of Cardiovascular Research (DZHK), Partner site RheinMain

⁴Senckenberg Biodiversity and Climate Research Centre (BIK-F)

⁵LOEWE Centre for Translational Biodiversity Genomics (TBG)

MicroRNAs (miRNAs) are post-transcriptional regulators that are involved in a broad spectrum of essential biological processes. Despite their functional importance, flexible and accurate frameworks for the detection of miRNA orthologs across species collections are missing. Here, we present ncOrtho, a synteny informed pipeline for training covariance models of miRNAs, which then form the basis of a targeted search for miRNA orthologs. A benchmark of ncOrtho against the manually curated reference database MirGeneDB reveals that ncOrtho identifies orthologs of 556 human miRNAs in 36 vertebrate species with an accuracy of 93% and a sensitivity of 97%. Extending these phylogenetic profiles to 402 vertebrates solidifies the claim that miRNA families are seldomly lost and traces the footprint of whole genome duplications on individual vertebrate lineages. The observation that the genetic diversity of human miRNAs is substantially lower than that of other genomic loci makes them promising novel markers for phylogenomic studies that should be less susceptible to the effects of incomplete lineage sorting (ILS). Although the loss of miRNA families is rare, few clade-specific absence patterns direct the attention to loss-events of otherwise conserved miRNAs. For two exemplary miRNA families that have been lost in most rodents, we investigate the functional implications of miRNA loss by comparing fluctuations of transcript levels brought about by the presence of the respective miRNAs in human and murine stem cells. Additionally, we determine gain and loss of miRNA target sites in ten species to investigate the plasticity of the miRNA regulatory network.

Big catalogs and small genes. Finding structure in prokaryotic genes using metagenomics

Luis Pedro Coelho

Fudan University, Shanghai, China

To investigate how prokaryotic genes are organized at the global scale, we collated >10k high-quality metagenomes and 86k isolate prokaryotic genomes and, using a consistent pipeline, built a global catalogue of prokaryotic genes. We named this catalogue GMGCv1 for Global Microbial Gene Catalogue, version 1. GMGCv1 contains 300 million unigenes (95% nucleotide clustering, corresponding to a species-level cutoff), which we clustered into 30 million protein families (any statistically-significant homology). Most unigenes are members of a small fraction of very large protein families (0.6% of families already contain half of all unigenes), which is consistent with a model where diversification is driven by (nearly) neutral evolution rather than adaptation. Most genes are rare (with the median number of observations being 10, out of >10k) and habitat-specific, although there is gene sharing between similar habitats (e.g., the gut microbiota of different mammals).

More recently, we have expanded this work to also include small open reading frames (we currently consider all open reading frames ≥ 10 amino acids). This poses additional bioinformatics challenges. For example, unlike in the case for full-length genes, we are not able to reliably cluster more than half of all sequences. In fact, when considering small sequences, even 50% amino acid identity over the whole sequence may not be statistically-significant, so that only high-identity matches can be used to infer homology.

Classifying Viral Proteins into Strict Ortholog Groups Using Domain-architecture Aware Inference of Orthologs

Christian M. Zmasek¹ and Richard H. Scheuermann^{1,2,3}

¹ Department of Informatics, J. Craig Venter Institute, La Jolla, CA 92037, USA

² Department of Pathology, University of California, San Diego, CA 92093, USA

³ Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA 92037, USA

When analyzing protein orthology relationships, multidomain proteins pose a unique problem, in particular, if their individual domains exhibit different evolutionary histories. In order to simplify this issue for the purpose of classifying and annotation proteins, we developed the concept of “Strict Ortholog Groups” (SOGs). A strict ortholog group is a group of proteins from different species which are not only orthologous, that is, they share an evolutionary history, but also contain the exact same domain architecture. For the automated inference of SOGs, we developed a computational approach called Domain-architecture Aware Inference of Orthologs (DAIO) for the analysis of protein orthology by combining phylogenetic and protein domain-architecture information.

We present results from two groups of viruses: the order *Nidovirales*, which includes Coronaviruses, and the family *Herpesviridae*. The main findings for *Nidovirales* are that the genomic evolution of *Coronaviridae* is associated with significant gains and losses of protein domains. The section of the genomes that show the largest divergence in protein domains are found in the proteins encoded in the amino-terminal end of the polyprotein (PP1ab), the spike protein (S), and many accessory proteins. The diversity among the accessory proteins is particularly striking, as each *Coronaviridae* subgenus possess a set of accessory proteins that is almost entirely specific to that subgenus. The only notable exception to this is ORF3b, which is present and orthologous over all Alphacoronaviruses. In contrast, the membrane protein (M), envelope small membrane protein (E), nucleoprotein (N), as well as proteins encoded in the central and carboxy-terminal sections of PP1ab (such as the 3C-like protease, RNA-dependent RNA polymerase, and Helicase) show stable domain architectures across all *Coronaviridae*.

For *Herpesviridae*, the results indicate that while many herpesvirus proteins evolved without any detectable gene duplication or domain rearrangement event, numerous herpesvirus protein families do exhibit relatively complex evolutionary histories. Some of them acquired additional domains during evolution (e.g., DNA polymerase), whereas others show a combination of domain rearrangements and gene duplications (e.g., US22 domain proteins).

The Darwin tree of life project and linking large inventories of genome data to understanding orthologue groups and phenotype evolution

Thomas A. Richards

Department of Biology, Oxford, UK

The Darwin tree of life (DTOL) seeks to generate reference genome quality genome data for all eukaryotic species resident in the UK. This effort includes large-scale sequencing of animals, plants, fungi, 'seaweeds' and protists. The initial effort is guided by taxon lists and is a collaborative effort led by the Wellcome Trust Sanger Institute and involving numerous institutions across the UK. In many cases the available list of taxonomy does not match realistic patterns of natural biodiversity. To account for this problem for the Protists we are sequencing both genus representatives of every protist in UK culture collections and conducting very large-scale single cell genome sequencing of cell sorted environmental samples. I will discuss progress in this project. I will then outline our own efforts to generate annotated orthologues that stretch back to the Last Eukaryotic Common Ancestor, our efforts to link these data with large scale genome sampling (e.g. DTOL) and understanding of phenotype functions in non-standard model organisms.

Abstracts – Posters

| Poster No. | Name | Institute | Country | Poster Titel |
|------------|------------------------|--|----------------|--|
| 01 | Laura Portell Silva | Barcelona Supercomputing Center (BSC) | Spain | OpenEBench Scientific Communities in 2022: More Communities, Better Support, Deeper Interactions |
| 02 | Odile Lecompte | University of Strasbourg/ICUBE | France | OrthoInspector 3.5: improving efficiency and scalability of orthology predictions |
| 03 | Maria Martin | EMBL-EBI | United Kingdom | Improved selection of canonical proteins for reference proteomes |
| 04 | Erik Sonnhammer | Stockholm University | Sweden | InParanoiDB 9: Ortholog groups for proteins and protein domains |
| 05 | Silvia Prieto | University of Lausanne | Switzerland | The effect of gene annotation on orthology |
| 06 | Saioa Manzano Morales | Barcelona Supercomputing Center (BSC) | Spain | A pan-genome view of the Asgard archaea, the closest relatives of eukaryotes |
| 07 | Salvatore Cosentino | The University of Tokyo | Japan | Scalable De Novo Orthology Inference Using Protein Representation Learning |
| 08 | Alex Warwick Vesztröcy | University of Lausanne | Switzerland | Testing the Least Diverged Orthologue Conjecture |
| 09 | Sina Majidian | University of Lausanne | Switzerland | A fast pipeline for species tree inference using placement in Hierarchical Orthologous Groups |
| 10 | Vinh Tran | Goethe University Frankfurt | Germany | Searching for orthologs in un-annotated genome assemblies with fDOG-Assembly |
| 11 | Sakhaa Alsaedi | King Abdullah University of Science and Technology | Saudi Arabia | Automated Diagnostic System for Medical Treatment of Infectious Diseases using Causal transfer learning and biological knowledge graph embedding |

OpenEBench Scientific Communities in 2022: More Communities, Better Support, Deeper Interactions

Laura Portell-Silva, Anna Redondo Guitarte, Lidia Lopez, Josep Ll. Gelpí, Salvador Capella-Gutierrez, and José-Maria Fernández

BSC, INB/ELIXIR-ES

OpenEBench is the ELIXIR benchmarking and technical monitoring platform for bioinformatics software and it has been part of the ELIXIR Tools Platform since its inception. OpenEBench provides scientific communities with an online infrastructure to perform unbiased and objective benchmarking evaluations and to make the results freely available on a public website (<https://openebench.bsc.es/>).

Between 2019 and 2022, the number of communities collaborating with OpenEBench has doubled, from four to eight, and there are more to come. Among the new communities, the most active one is the Continuous Automated Model EvaluatiOn (CAMEO), which runs weekly automated benchmarks on predictions of protein structures. In addition, there are communities from diverse areas such as alternative RNA polyadenylation (APA) and Antimicrobial Resistance detection. Within the existing communities, Quest for Orthologs (QfO) has produced a new benchmarking event fully managed using OpenEBench.

OpenEBench is designed to cater for the needs of very different communities, depending on the level of engagement that they want to have with the platform. This flexibility comes with a high level of complexity, which is why all communities receive support from the onset of the collaboration to the publication of their results. This support includes access to comprehensive documentation (<https://openebench.readthedocs.io/en/0.3.1/>), which has been fully revamped and updated.

OrthoInspector 3.5: improving efficiency and scalability of orthology predictions

Arnaud Kress¹, Yannis Nevers², Latitia Poidevin¹, Dorine Merlat¹, Kirsley Chennen¹, and Odile Lecompte¹

¹ University of Strasbourg/ICUBE

² University of Lausanne/Swiss Institute of Bioinformatics

OrthoInspector (Linard et al., 2011) is an orthology and paralogy relationship prediction program that has demonstrated an excellent balance of specificity and sensitivity in several assessments (Nevers et al, 2022; Altenhoff et al., 2020, 2016). OrthoInspector is also a resource (<https://lbgi.fr/orthoinspector/>) that provides the community with pre-computed orthology relationships between species. The website integrates numerous tools for exploring orthology relationships, including phylogenetic profiling tools. This resource has grown considerably over time, from 59 eukaryotic species in the initial version (Linard et al., 2011), to 1947 species in the 2nd version (Linard et al., 2015) to 4753 species in OrthoInspector v3.0 (Nevers et al., 2018). In the new release 3.5, we have further extended the coverage of the databases. They now include a total of 5247 species (930 eukaryotes, 4132 bacteria, and 185 archaea), representing about 30 million protein sequences. In addition to the 3 domain-specific databases, release 3.5 offers a cross-domain database of 370 species of eukaryotes, bacteria and archaea. These species were chosen to best sample the taxonomic diversity in the 3 domains while respecting quality criteria of the selected proteomes. Faced with an ever-increasing number of available proteomes, the software has been modified to improve its scalability. In this new version, all computations are distributed in a fully independent way, allowing to leverage massively parallel computing resources (computing grid, HPC, etc.). We also focused on a simplified database building procedure. A new text-based output format allows data to be manipulated using only sequential accesses, which considerably improves the speed of database generation. This new process also allows an existing database to be easily extended with new species to better and quicker fit researcher's interests.

Improved selection of canonical proteins for reference proteomes

Giuseppe Insana¹, Martin Maria J.¹, and William R. Pearson²

¹ EMBL-European Bioinformatics Institute

² University of Virginia

The Reference Proteomes dataset seeks to provide complete proteomes for an evolutionarily diverse, less redundant, set of organisms. As higher eukaryotes often encode multiple isoforms of a protein from a single gene, the Reference Proteomes pipeline selects a single representative ('canonical') sequence. UniProt identifies canonical isoforms using a 'Gene-Centric' approach: proteins are grouped by gene-identifier and for each gene a single protein sequence is chosen. For unreviewed

(UniProtKB/TrEMBL) protein sequences (and for some reviewed sequences), the longest sequence in the Gene-Centric group is chosen as canonical. This can create inconsistencies, selecting sequences with dramatically different lengths as canonical for orthologous genes. Biologically, it is unlikely that orthologous mammalian proteins differ greatly in length, but this happens about 10% of the time

for the 8 mammals in the Quest for Orthologs set. The Ortho2tree data pipeline examines Gene-Centric canonical and isoform sequences from sets of orthologous proteins from PantherDB, and suggests replacements for canonicals that have lengths

different from closely related orthologs. For the ~140,000 proteins in ~24,000 Panther orthogroups from the eight mammalian proteomes (human, chimp, gorilla, mouse, rat, dog, cow and opossum), ortho2tree proposed 7782 canonical changes, while confirming 69,226 canonical assignments. Of the 4,588 orthogroups with proposed changes and MANE (Matched Annotation from NCBI and EMBL-EBI) labels, the Ortho2tree proposed accession agreed with MANE 80% of the time. When Ortho2tree supported the current HUMAN canonical, MANE agreed 95% of the time. Ortho2tree can reduce canonical assignment errors among sequences that are more than 50% identical, such as sequences from vertebrates or higher plants.

InParanoiDB 9: Ortholog groups for proteins and protein domains

Erik Sonnhammer and Emma Persson

Stockholm University

The InParanoiDB database <https://inparanoid.sbc.su.se/> [1] is a well known resource that includes ortholog predictions between a wide variety of species. However, as the number of sequenced genomes continues to grow, the need for orthology databases to cover more species has increased dramatically. Seeing that prediction of orthologs on the level of protein domains can provide additional information that can not be inferred from predictions on the full protein sequence level, with Domainoid [2] performing well in the QFO benchmark [3], adding information on domain orthologs could be a valuable contribution to ortholog databases.

We here present InParanoiDB 9, an update to the InParanoid database including domain ortholog predictions from Domainoid and now covering 640 species, utilizing the much faster inParanoid-DIAMOND algorithm [4]. The proteomes originate from the UniProt reference proteomes, and consist of a total of 447 eukaryotes, 158 bacteria and 35 archaea, and include all QFO reference proteomes, as well as all reference proteomes used in the InParanoid database, release 8. In total, InParanoiDB 9 consists of 746,501,557 ortholog pairs, and to better accommodate the massive increase in data, and to support visualization of domain ortholog information, the InParanoid website has been updated to a new underlying framework as well as a new graphical interface.

References

1. Sonnhammer ELL, Östlund G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* 2015;43: D234–9.
2. Persson E, Kaduk M, Forslund SK, Sonnhammer ELL. Domainoid: domain-oriented orthology inference. *BMC Bioinformatics.* 2019;20: 523.
3. Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, et al. The Quest for Orthologs orthology benchmark service in 2022. *Nucleic Acids Res.* 2022. doi:10.1093/nar/gkac330
4. Persson E, Sonnhammer ELL. InParanoid-DIAMOND: faster orthology analysis with the InParanoid algorithm. *Bioinformatics.* 2022;38: 2918–2919.

The effect of gene annotation on orthology

Silvia Prieto¹, Christophe Dessimoz¹, and Natasha Glover²

¹ University of Lausanne

² Swiss Institute of Bioinformatics

Computational gene annotation, i.e. finding the genes present in a genome, remains a challenging task. The quality of gene models and gene repertoire lags behind the pace at which available genome assemblies are increasing. Although annotation methods are improving, there is no community standards on their standards and in practice, most published gene annotations result from ad hoc pipelines. As a result, only a few non-model species have complete and accurate gene models. This annotation quality affects downstream analyses, including comparative genomics. Yet there is no focus in understanding the impact of using one or the other annotation for one species on orthology inference, often the first step of these studies. Here, we show that different annotation methods render different orthology results. We ran OMA Standalone to infer orthology on gene models obtained by four frequently used protein-coding gene annotation pipelines or databases (Augustus 3.4 de novo, NCBI Eukaryotic Pipeline, Ensembl and the QfO Reference Proteomes). Preliminary results show important differences among the four annotation methods, namely in the number of orthology pairs and on the accuracy of orthology prediction on a standard benchmark.

A pan-genome view of the Asgard archaea, the closest relatives of eukaryotes

Saioa Manzano-Morales and Toni Gabaldón

Barcelona Supercomputing Center (BSC) / Institute for Research in Biomedicine (IRB)

Asgard archaea were only discovered in the last decade and have since been broadly accepted as the closest relatives of eukaryotes, yielding important insights into the evolutionary origin of eukaryotic cells. Despite recent progress, much of the physiology of Asgard archaea remains unknown. The most widely-studied Asgard clades, Lokiarchaeota and Thorarchaeota, are suggested to encode high ecological adaptability.

The availability of over 200 Asgard group genomes provides an opportunity to analyze the diversity of this clade of prokaryotes from a pangenome perspective. A pangenome, which can be defined as the non-redundant set of all genes (clusters of orthologs) found in all genomes of a taxon, and is potentially altered by both habitat and phylogeny. The main processes shaping pangenomes are gene duplication and loss during vertical inheritance and gene acquisition via horizontal gene transfer (HGT). Therefore, a pangenome analysis, coupled and complemented with phylogenomics, can yield unprecedented insight into the evolutionary forces shaping Asgard genomes.

Here we report a pangenomic analysis of the Asgard group, with focus on the Candidatus Prometheoarchaeum syntrophicum MK-D1 proteome. Such analysis reveals a vast diversity inside this clade, with many clusters of orthologous groups (COGs) belonging to one or few organisms. We further analyze the pangenome architecture of the Asgard group, assessing its openness, as well as characterizing the patchiness of the presence of Eukaryotic Signature Proteins (ESPs) across the organisms of this clade. Additionally, we analyze the Gene Ontology (GO) enrichment of protein functions for the core and accessory gene categories.

Scalable De Novo Orthology Inference Using Protein Representation Learning

Hannah Muelbaier^{1,2}, Salvatore Cosentino³, and Wataru Iwasaki³

¹ Applied Bioinformatics Group, Institute of Cell Biology and Neuroscience, Goethe University

² LOEWE Centre for Translational Biodiversity Genomics, Frankfurt, Germany

³ The University of Tokyo, Japan

Thanks to the steep increase in the number of publicly available genomic datasets, it is now common to include hundreds or thousands of genomes in a single study. This data deluge is challenging orthology inference methods, which may be unable to process datasets of such magnitudes. For example, graph-based de novo orthology inference methods that rely on all-vs-all alignments may require days to weeks to infer orthologous relations even for a few dozens of eukaryotes.

In SonicParanoid2, we proposed a method that adopts machine learning to reduce the runtime of homology searches and uses protein representation learning to increase the breadth of orthology inference. The method infers orthologous relationships using functional domains annotated using the PfamA database, and then filters those predictions based on the orthologs obtained using homology searches. However, SonicParanoid2 still requires all-vs-all alignments due to the low amount of domains that can be annotated using PfamA as training data. The amount of orthologs predicted by the domain-based pipeline is about 1/3 of that inferred by the graph-based method alone.

In this study, we explored alternative resources that can be used to increase the amount of training data for the domain-based orthology pipeline. By using the CDD domain database instead of PfamA and adding regions with a compositional bias to the training data, we were able to increase the number of predictions by 50%. More importantly, we decreased the number of false positives and doubled recall in most of the tests in the QfO benchmark. Lastly, we showed that the domain-based pipeline is highly scalable using a set of 2,000 prokaryotic MAGs. We are now working on a method that infers orthologs solely based on functional domains, which could have accuracies comparable to graph-based methods.

Testing the Least Diverged Orthologue Conjecture

Alex Warwick Vesztröcy

University of Lausanne

The orthologue conjecture - that orthologous genes are functionally more similar than paralogous genes - has been the subject of much debate. However, annotation bias leads to issues when studying the orthologue conjecture in previous studies using gene ontology annotations.

In this study the gene families from the PANTHER database [1] are combined with expression data from the Bgee database [2]. Using expertly curated ancestral anatomical entity similarity annotations [3], this enables the reconstruction of ancestral gene expression profiles. Then, with the application of a simple evolutionary model to allow for different rates of gene family evolution, it is possible to compare sequence evolution with changes in gene expression profile.

This enables the investigation of a specific case of the orthologue conjecture: that after gene duplication the least evolutionary diverged copy maintains the ancestral function, whilst the other copy is no longer under selective pressure to maintain function and is free to diverge. We call this the least diverged orthologue conjecture.

[1] Mi, Huaiyu, et al. "PANTHER version 16: a revised family classification, treebased classification tool, enhancer regions and extensive API." *Nucleic acids research* 49.D1 (2021): D394-D403.

[2] Bastian, Frederic B., et al. "The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals." *Nucleic Acids Research* 49.D1 (2021): D831-D847.

[3] "Anatomical Similarity Annotations." BgeeDB, github.com/BgeeDB/anatomicalsimilarity-annotations. Accessed 30 June 2022.

A fast pipeline for species tree inference using placement in Hierarchical Orthologous Groups

Sina Majidian¹, Adrian Altenhoff², and Christophe Dessimoz¹

¹University of Lausanne

²ETH Zurich

Species tree reconstruction usually requires many orthologous marker loci. Inferring those is computationally demanding and requires complicated pipelines to extract the single copy ortholog groups (OGs). Alternatively, precomputed markers such as BUSCO can be used, but they are not available for all clades. Here, we propose a fast pipeline for tree inference for arbitrary species sets. We exploit the OMamer software to place proteins of species of interest onto a database of hierarchical orthologous groups from OMA. Then, we map each protein to an OG if possible. Next, we select the most informative OGs, compute MSAs and infer the species tree from the super-matrix. As a proof of concept, we ran our pipeline on a dataset of 363 bird proteomes. Within 50 CPU hours, we computed the super-matrix of orthologous characters. IQ-Tree needs 11h on 48 CPUs to infer the tree, which is in good agreement with the NCBI taxonomy.

Searching for orthologs in un-annotated genome assemblies with fDOG-Assembly

Hannah Muelbaier^{1,2}, Vinh Tran¹, and Ingo Ebersberger^{1,2,3}

¹ Applied Bioinformatics Group, Inst. of Cell Biology and Neuroscience, Goethe University Frankfurt

² LOEWE Centre for Translational Biodiversity Genomics, Frankfurt, Germany

³ Senckenberg Biodiversity and Climate Research Centre (BIK-F)

The identification of orthologs in the genomes of newly sequenced species is a relevant step for their integration into a broad range of evolutionary, comparative and functional studies. Numerous approaches varying in computational complexity, sensitivity and specificity have been developed for this purpose. However, one dependency is common to all tools: they require comprehensively annotated gene sets as input where any overlooked gene will result in a missed ortholog. Here, we present fDOG – Assembly, a targeted profile-based ortholog search tool that can identify orthologs in unannotated genome assemblies. Using aligned pre-computed core orthologs as a start, the algorithm generates a consensus sequence that serves as query for a tblastn search. Hit regions are scanned for the presence of a gene using Augustus guided by a block profile or MetaEuk, a reference-based approach to identify protein coding genes. In case a gene is annotated, the encoded protein is tested for orthology and, upon success, will be added to the core orthologous group. An assessment of domain architecture similarity to a reference protein in the core set is optional. We benchmarked fDOG – Assembly, which revealed a performance that is comparable to the ortholog search in fully annotated gene sets. We envision that fDOG – Assembly will be helpful for closing gaps in phylogenetic profiles due to annotation artefacts, but even more for studies requiring the identification of candidate genes in genomes irrespective of their annotation status.

Automated Diagnostic System for Medical Treatment of Infectious Diseases using Causal transfer learning and biological knowledge graph embedding

Sakhaa Alsaedi, Katsuhiko Mineta, Xin Gao, and Takashi Gojobori

King Abdullah University of Science and Technology

Understanding the molecular pathways between host and pathogen is an essential step to tackle infectious diseases. In precision medicine, molecular diagnostics open a new horizon to clinical practices by assisting physicians in understanding the situation of infected patients before the emergence of symptoms and complications. Moreover, leveraging the advantage of deep learning algorithms to assist medical practitioners in clinical decision-making and diagnosis is critical for patient treatment decisions and outcomes. Current automated diagnostic systems, however, solely employ associative deep learning algorithms to identify diseases that are significantly related a patient's symptoms. Such approaches do not consider the genetic risk factors that may cause difficulties. In addition, these risk factors might also be connected to other complicated illnesses, impacting the patient's conditions. Understanding how different viral strains impact individual patients, specifically how they interact with distinct human host cells and immunological responses, is a critical step in developing effective treatment programs. Since the onset of the COVID-19 disease, host genetic variants have been identified to play an important role in the display of varying degrees of sickness severity among different individuals. Hence, it is critical that investigations are conducted with this condition as the initial case study. Thus, we develop DeepCARES, a causal risk estimation score prediction model that prioritizes multi-organs dysfunction and predicts the severity score of each organ dysfunction during infection by integrating multi-omics data with genetic risk factors to provide prior knowledge graphs that can be used to infer the roles of causal risk factors in inducing severe outcomes during infection. Preliminary results indicate that our model outperforms baseline approaches on electronic medical records and synthetic data. The projected scores provide clinicians with a better understanding of COVID-19, and with such knowledge give more appropriate treatment plans to lessen the severity and complications of COVID-19.

Information for Presenters

Talks

Duration

Invited Talks: Please plan your presentation for 30 mins + 5 mins Q&A

Contributed Talks: Please plan your presentation for 20 mins + 5 mins Q&A

AV equipment

The hotel will equip the meeting room with the following equipment

- Projector with HDMI connector
- Laptop with operating system Windows 10
- PowerPoint
- HDMI adaptors for USB-C, mini-display port, display port

Before your presentation

Please make sure upload your presentation to the conference laptop before the session starts. If you have to use your own laptop, make sure to bring an HDMI adapter with you and test the setup before the session start

Posters

The poster boards are adapted to standard A0 poster size / portrait (A0 is 84.1 x 118.9cm ; or 33.1 x 46.8 inches). The maximum size of the poster must be 90 cm wide and 120 cm high.

List of Participants

| Name | Talk No. | Poster No. |
|---|-------------|------------|
| Alsaedi, Sakhaa KAUST Saudi Arabia Sakhaa.Alsaedi@kaust.edu.sa | | P11 (p.36) |
| Altenhoff, Adrian SIB Swiss Institute of Bioinformatics Switzerland adrian.altenhoff@sib.swiss | | P09 (p.34) |
| Avignoli, Alessio CRG Spain alessio.vignoli@crg.eu | | |
| Bordin, Nicola University College London United Kingdom n.bordin@ucl.ac.uk | T14 (p. 19) | |
| Capella-Gutierrez, Salvador Spanish National Bioinformatics Institute (INB) Coordination Node Spain salvador.capella@bsc.es | | P01 (p.26) |
| Chakraborty, Abhisek Indian Institute of Science Education and Research, Bhopal India chakrabortyabhi2013@gmail.com | | |
| Coelho, Luis Pedro Fudan University China luis@luispedro.org | T17 (p.22) | |
| Cosentino, Salvatore The University of Tokyo Japan salvocos@edu.k.u-tokyo.ac.jp | T11 (p.16) | P07 (p.32) |

| Name | Talk No. | Poster No. |
|---|---|------------------------|
| Dessimoz, Christophe University of Lausanne and Swiss Institute of Bioinformatics Switzerland christophe.dessimoz@unil.ch | T01 (p.6), T02 (p.7), T06 (p.11), T12 (p.17), T13 (p.18) | P05 (p.30), P09 (p.34) |
| Durand, Dannie Carnegie Mellon University USA durand@cmu.edu | T05 (p.10) | |
| Ebersberger, Ingo Goethe University Frankfurt Germany ebersberger@bio.uni-frankfurt.de | T16 (p.21) | P10 (p.35) |
| Forslund, Sofia Kirke MDC/Charité Germany Sofia.Forslund@mdc-berlin.de | | |
| Fuentes, Diego IRB Barcelona - Institute for Research in Biomedicine Spain diego.fuentes@irbbarcelona.org | T03 (p.8) | |
| Gabaldón, Toni Barcelona Supercomputing Centre Spain toni.gabaldon.bcn@gmail.com | T03 (p.8) | P06 (p.31) |
| Glover, Natasha Swiss Institute of Bioinformatics Switzerland natasha.glover@sib.swiss | T01 (p.6) | P05 (p.30) |
| Hadziahmetovic, Armin Ludwig-Maximilians-Universität München Germany hadziahmetovic@bio.ifl.lmu.de | | |
| Hernandez Plaza, Ana CSIC Spain ana.hernandez.plaza@gmail.com | | |

| Name | Talk No. | Poster No. |
|--|------------|------------|
| Hiller, Michael Senckenberg Frankfurt Germany michael.hiller@senckenberg.de | | |
| Hu, Claire Harvard Medical School USA claire_hu@genetics.med.harvard.edu | T15 (p.20) | |
| Huerta Cepas, Jaime CBGP (UPM-INIA/CSIC) Spain jhcepas@gmail.com | T09 (p.14) | |
| Langschied, Felix Goethe University Frankfurt Germany langschied@bio.uni-frankfurt.de | T16 (p.21) | |
| Lecompte, Odile University of Strasbourg/ICUBE France odile.lecompte@unistra.fr | | P02 (p.27) |
| Majidian, Sina University of Lausanne Switzerland sina.majidian@unil.ch | T13 (p.18) | P09 (p.34) |
| Manzano Morales, Saioa BSC (Barcelona Supercomputing Center) Spain saioa.manzano@bsc.es | | P06 (p.31) |
| Martin, Maria EMBL-European Bioinformatics Institute United Kingdom martin@ebi.ac.uk | | P03 (p.28) |
| Moi, David University of Lausanne DBC Switzerland dmoi@unil.ch | T06 (p.11) | |

| Name | Talk No. | Poster No. |
|---|-----------------------|------------|
| Nevers, Yannis University of Lausanne Switzerland yannis.nevers@unil.ch | T02 (p.7) | P02 (p.27) |
| Palero, Ferran University of Valencia Spain Ferran.Palero@uv.es | | |
| Pilizota, Ivana EMBL-European Bioinformatics Institute United Kingdom ivana@ebi.ac.uk | T04 (p.9) | |
| Portell Silva, Laura Barcelona Supercomputing Center (BSC) Spain laura.portell@bsc.es | | P01 (p.26) |
| Prieto, Silvia University of Lausanne Switzerland silviabaprieto@gmail.com | | P05 (p.30) |
| Richards, Thomas University Oxford United Kingdom thomas.richards@biology.ox.ac.uk | T19 (p.24) | |
| Rossier, Victor University of Lausanne Switzerland victor.rossier@unil.ch | T02 (p.7), T12 (p.17) | |
| Sonnhammer, Erik Stockholm University Sweden erik.sonnhammer@scilifelab.se | | P04 (p.29) |
| Szklarczyk, Damian UZH / SIB Swiss Institute of Bioinformatics Switzerland damian.szklarczyk@sib.swiss | T07 (p.12) | |

| Name | Talk No. | Poster No. |
|---|------------|------------|
| Thomas, Paul University of Southern California USA pdthomas@usc.edu | T08 (p.13) | |
| Tran, Vinh Goethe University Frankfurt Germany tran@bio.uni-frankfurt.de | | P10 (p.35) |
| Uchiyama, Ikuo National Institute for Basic Biology Japan uchiyama@nibb.ac.jp | T10 (p.15) | |
| Warwick Vesztrocy, Alex University of Lausanne Switzerland alex.warwickvesztrocy@unil.ch | | P08 (p.33) |
| Zmasek, Christian J Craig Venter Institute USA czmasek@jcvl.org | T18 (p.23) | |

Conference Notes

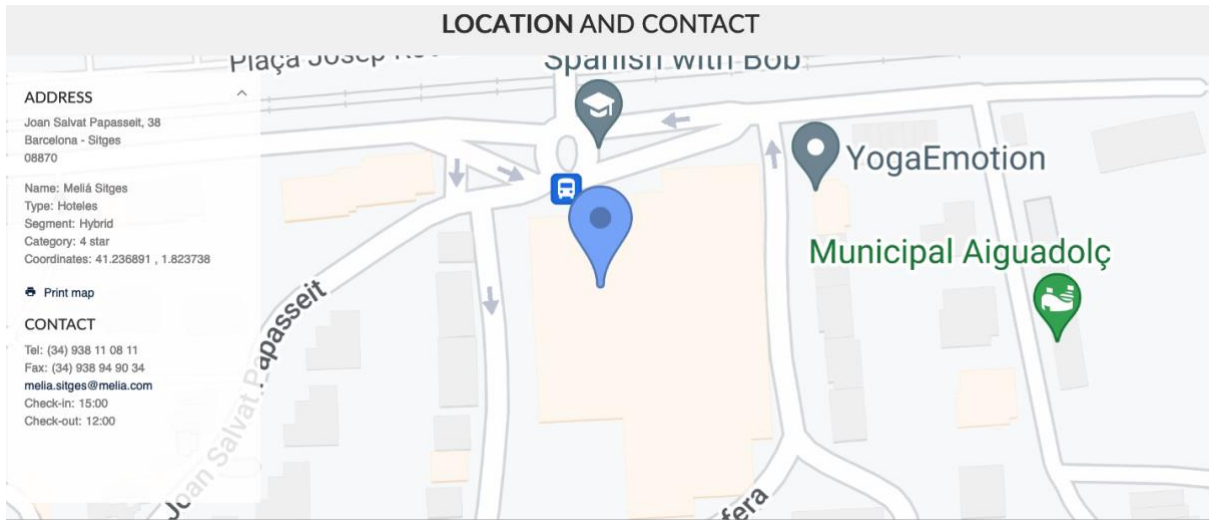
Like with all QFO meetings, we also have this year a Google Doc, and everybody is invited to contribute by writing down ideas, things that have been discussed, and the like. You can access the Google Doc via the following URL: <https://docs.google.com/document/d/1UrSitqVHHWCmui4BD-fbf1YTv7-v0XlxnQALTFpV6V4/edit?usp=sharing>. Alternatively, scan the QR code below.



Addresses and Maps

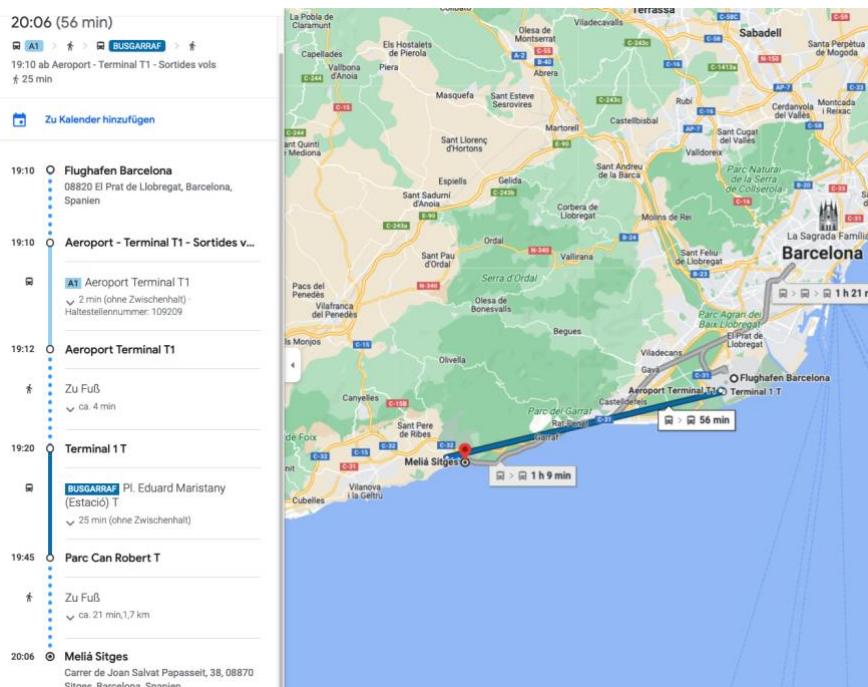
Venue:

Llevant 1, Hotel Meliá, Joan Salvat Papasselt 38, Barcelona - Sitges, Spain
(<https://www.melia.com/en/hotels/spain/sitges/melia-sitges/index.htm>)



Getting to Sitges:

Getting to Sitges from the airport of Barcelona (BCN) is pretty easy, even if you are relying on public transportation. There is a bus running from Terminal 1T directly to Sitges (see below for an example connection in the evening). From the bus stop in Sitges (Parc Can Robert T), the hotel is within walking distance (if your luggage is not too heavy). For more details on the travel please see the [ECCB2022 Travel web site](#).








Stay Safe

Dear participants,

COVID was and still is a considerable threat to human health. We still felt it necessary that the community meets in person, and it is for this reason that we have organized QFO-7 face-to-face. In order to avoid that our meeting becomes a COVID spreading event, it would be great if you obey the following guidelines on a voluntary basis:



-  test yourself for a SARS-CoV-2 infection regularly
-  wear face masks (either FFP2 or medical) in the lecture hall
-  wash your hands regularly
-  avoid hand shakes
-  after more than two years living with the pandemic, you all know what else you should and should not do...

In case you feel symptoms that could indicate a SARS-CoV-2 infection, please inform the [organizing team](#) immediately and we would kindly ask you to not attend the meeting until a negative test result is obtained.

Please see also the [Stay Safe pages of the Hotel Melia](#) for further information

If you have any questions, please contact the [QFO-7 organizing](#) team.