Bioinformatics Group - Thesis projects

Last updated: April 30th, 2021

Reconstructing symbiont strains from long-read metagenomics

Supervisor	Anne Kupczok
Type	Data analysis, Methodology
Requirements	Advanced Bioinformatics
Skills	Genome assembly, Metagenomics, Comparative genomics
Timestamp	March 2021

Description

In the dark part of the deep ocean, chemosynthetic bacteria can exist by gaining energy from the oxidation of inorganic compounds instead of by photosynthesis. These bacteria are the basis for a rich ecosystem that also includes diverse invertebrates that live in a nutritional symbiosis with these bacteria. For example, *Bathymodiolus* mussels harbor chemosynthetic sulfur-oxidizing (SOX) and methane-oxidizing (MOX) bacteria in a special organ, the gill. Although most *Bathymodiolus* species harbor only a single 16S phylotype for each symbiont, metagenomic analyses of multiple *Bathymodiolus* species showed that different SOX and MOX strains can be present within an individual mussel (Ansorge et al. 2019; Romero Picazo et al. 2019). Strains are different variants of a species and might harbor important functional differences (Rossum et al. 2020). Notably, co-occurring SOX strains can differ in the content of genes involved in energy and nutrient utilization and viral defence mechanisms (Ansorge et al. 2019). SOX genomes were also found to contain high numbers of mobile genetic elements such as transposases, integrases, restriction-modification systems, and toxin-related genes, where the latter are also linked to mobile genetic elements (Sayavedra et al. 2015).

The presence of different strains and of mobile genetic elements resulted in highly fragmented assemblies for short-read metagenomes, where gene presences could not be linked over long genomic distances. Here we use the PacBio sequencing technology which delivers long reads to reconstruct genome-wide SOX and MOX strains including the accessory gene content and mobile genetic elements. Thereby recent methods for the assembly of PacBio metagenome data will be explored (such as Kolmogorov et al. 2020) and an approach to extract strains from metagenomes will be developed. The gene content of the reconstructed strains can then be compared.

This project is a collaboration with the group of Prof. Nicole Dubilier, Max Planck Institute for Marine Microbiology Bremen, Germany.

References

Ansorge R, Romano S, Sayavedra L, Porras MÁG, Kupczok A, Tegetmeyer HE, Dubilier N, Petersen J. 2019. Functional diversity enables multiple symbiont strains to coexist in deep-sea mussels. Nature Microbiology 4:2487–2497.

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, et al. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nature Methods 17: 1103–1110.

Romero Picazo D, Dagan T, Ansorge R, Petersen JM, Dubilier N, Kupczok A. 2019. Horizontally transmitted symbiont populations in deep-sea mussels are genetically isolated. ISME J 13:2954–2968.

Rossum TV, Ferretti P, Maistrenko OM, Bork P. 2020. Diversity within species: interpreting strains in microbiomes. Nature Reviews Microbiology 18:491–506.

Sayavedra L, Kleiner M, Ponnudurai R, Wetzel S, Pelletier E, Barbe V, Satoh N, Shoguchi E, Fink D, Breusing C, et al. 2015. Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. eLife 4:e07966.

Annotation and analysis of the blue shark (*Prionace glauca*) genome

Supervisor	Judith Risse, Sandra Smit
Type	Data analysis
Requirements	Advanced Bioinformatics, Genomics
Skills	Genome analysis, genomics, Python
Timestamp	April 2021

Description

Cartilaginous fishes (Chondrichthyes) are divided into two subclasses, elasmobranchs (Elasmobranchii, including sharks, rays and skates) and chimaeras (Holocephali), and evolved independently from the rest of jawed vertebrates around 450 million years ago. For elasmobranchs few reliable genome-wide sequence resources are available. Most cartilaginous fishes contain large genome sizes and massive chromosome karyotypes (2n = 66–104). This suggests high genome stability, potentially resulting in low vulnerability to aged-related diseases and abnormalities. Of particular interest are shark immune or olfactory systems, as well as their exceptional wound-healing capabilities. Their genomes may help to reveal the evolutionary evidence to explain the efficiency of such systems, but also may provide genes of interest with biotechnological potential.

The blue shark (*Prionace glauca,* genome size ~ 4.2 Gbp, 2n = 86) is a cosmopolitan species, and most frequently caught pelagic shark species in fishing activities, therefore classified as “Near Threatened” by several IUCN regional assessments. Their ability to populate almost all oceans of the globe, along with their high ecological, economic, and evolutionary significance, makes it a particularly relevant species for genome sequencing. As part of this project the genome of a single female was sequenced on 8 PacBio HiFi SMRT cells (~45x coverage) and will be assembled at the Leiden Centre for Applied Bioscience.

In this project, you will annotate the assembled genome using PacBio IsoSeq data (various tissues) to complete the blue shark reference genome. The core tasks will be: 1) researching approaches to genome annotation using only PacBio HiFi data, 2) structural annotation of the blue shark genome using the selected tools, 3) functional annotation of the transcriptome using e.g. InterProScan and diamond blast. Optional tasks may include the investigation of genes and gene families specific to blue shark and genes specific to shark ‘powers’, or the investigation of tissue- and haplotype-specific gene-expression patterns making use of the full length, high accuracy transcript data. In any case, you will make use of high-quality data produced with the extremely accurate PacBio HiFi method.

This project is a collaboration with Dick Roelofs (Keygene), Ken Kraaijeveld (Leiden), Tiago Simões & Sara Novais (MARE, Portugal)

References

Alves et al. Science of the Total Environment (2016) 563-564. https://doi.org/10.1016/j.scitotenv.2016.04.085
Brůna et al. NAR Genomics and Bioinformatics (2021) 3(1). https://doi.org/10.1093/nargab/lqaa108
Yandell & Ence. Nature Reviews Genetics (2012) 13(5). https://doi.org/10.1038/nrg3174

The evolution of auxiliary metabolic genes in bacteriophages

Supervisors	Anne Kupczok, Marnix Medema
Type	Data analysis
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Comparative genomics, databases, phylogenetics, python, statistics
Timestamp	September 2020

Description

Bacteriophages (short: phages) are viruses that infect bacteria. Some phage genomes carry auxiliary metabolic genes (AMGs), which are metabolic genes that are derived from bacterial genes. AMGs were found to alter the bacterial metabolism during viral infection resulting in increased viral proliferation (Warwick-Dugdale et al. 2019). Phages with AMGs are particularly well studied in the marine environment, where phages are known to contribute substantially to bacterial mortality and where they play a central role in biogeochemical cycles (Breitbart et al. 2018).

Phages need to package their genome into a protein capsid for transmission, which limits the total genome size of a particular phage. Due to this restriction on genome length, AMGs are often shorter, more compact versions of the bacterial gene. In addition, it has been observed that phage metabolic genes form a separate clade which is distinct from the bacterial genes. Thus, AMGs might have evolved into distinct phage-adapted versions, which also have a deviating functionality (e.g., Rihtman et al. 2019). On the other hand, some AMGs might be transferred back to bacterial genomes (e.g., Lindell et al. 2004). These cases provide interesting evolutionary scenarios, where the evolvability of bacteria is increased due to the acquisition of gene versions that evolved in phages.

This project will assess whether AMGs generally evolve into distinct shorter versions of the bacterial gene and whether the transfer of metabolic genes from phages to bacteria is a prevalent phenomenon. To this end, publicly available genomes of phages and bacteria will be scanned for metabolic genes (Shaffer et al. 2020). First, the hypothesis that phage metabolic genes are shorter will be tested on this data set. Second, phylogenies will be constructed to test whether bacterial and phage genes form distinct clades. These phylogenies will then be analyzed to identify instances where metabolic genes that evolved in a phage have been transferred back to bacterial genomes.

References

Breitbart M, Bonnain C, Malki K, Sawaya NA. 2018. Phage puppet masters of the marine microbial realm. *Nature Microbiology* 3:754–766.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. *PNAS* 101:11013–11018.

Rihtman B, Bowman‐Grahl S, Millard A, Corrigan RM, Clokie MRJ, Scanlan DJ. 2019. Cyanophage MazG is a pyrophosphohydrolase but unable to hydrolyse magic spot nucleotides. *Environmental Microbiology Reports* 11:448–455.

Shaffer M, Borton MA, McGivern BB, Zayed AA, Rosa SLL, Solden LM, Liu P, Narrowe AB, Rodríguez-Ramos J, Bolduc B, et al. 2020. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *bioRxiv*:2020.06.29.177501.

Warwick-Dugdale J, Buchholz HH, Allen MJ, Temperton B. 2019. Host-hijacking and planktonic piracy: how phages command the microbial high seas. *Virology Journal* 16:15.

BiG-SCAPE 2.0 – updates to a successful software for large-scale genome mining

Supervisor	Jorge Navarro, Jérôme Collemare, Marnix Medema
Type	Data analysis, Genome mining
Requirements	Advanced bioinformatics
Skills	Programming (Python), Genomics
Timestamp	April 2021

Description

Microbes and plants are able to synthesize secondary metabolites (SMs) that allow them to thrive in their environment by offering means of communication with other organisms, defense against competitors or environmental conditions, acquisition of additional resources or facilitating host colonization. The genes that make up the biosynthetic pathways for these metabolites are co-regulated and, in microbes, often co-localized in the same genomic loci, generally referred to as “biosynthetic gene clusters” (BGCs).

The study of BGCs is of great relevance not only as a starting point to elucidate the bioactivities of the SMs whose production they encode, but also to predict the biosynthetic capacity of newly-sequenced genomes, to detect the presence of known harmful toxins and to characterize novel pathways with potential to be of relevance to humankind, such as those involved in producing disease-treating drugs, pigments and crop protection agents.

Identification of regions containing putative BGCs is straightforward nowadays thanks to software such as antiSMASH (1), which uses the principle of co-localization and knowledge about core biosynthetic genes—those that synthesize the scaffold of the metabolite. However, as genome sequencing is being scaled up, large groups of BGCs need to be compared across genomes, or within large sets of metagenome-assembled genomes. Such large-scale comparisons can help researchers discover BGCs that are similar to characterized ones, avoiding time-consuming experiments (dereplication) or, on the contrary, to find interesting new versions of known BGCs. Such analyses can also uncover large groups of similar BGCs (gene cluster families, GCFs) yet to be characterized, which can be prioritized for further studies.

One tool that uses sequence similarity networks to compare large datasets of BGCs is the “biosynthetic gene similarity clustering and prospecting engine”, or BiG-SCAPE (2), which uses conserved protein regions (domains) within predicted BGCs to carry out three different similarity calculations that are combined into a final distance value for each pair of BGCs. By applying a clustering algorithm to each BGC subnetwork, BiG-SCAPE then groups BGCs into GCFs.

In this project, you will work on bringing BiG-SCAPE to the next level by restructuring its code and implementing new features to make it faster, less resource-intensive, more accurate, and more useful to researchers in the secondary metabolism field.

References

(1) https://academic.oup.com/nar/article/47/W1/W81/5481154

(2) https://pubmed.ncbi.nlm.nih.gov/31768033/

Gene content variation mediated by recombination in bacteriophage evolution

Supervisors	Anne Kupczok, Dick de Ridder
Type	Data analysis
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Comparative genomics, databases, alignments, python, statistics
Timestamp	September 2020

Description

Bacteriophages (short: phages) are viruses that infect bacteria, i.e. they only reproduce during bacterial infection (Dion et al. 2020). Their mode of evolution is characterized by high mutation rates and frequent recombination and horizontal gene transfer (e.g., Kupczok et al. 2018). Different molecular mechanisms can result in phage recombination. Homologous recombination utilizes the host recombination machinery. Relaxed homologous recombination is facilitated by phage-encoded recombination proteins and allows for recombination at sites of limited homology (de Paepe et al. 2014). Finally, illegitimate recombination takes place between non-homologous regions. Notably, all these mechanisms require co-infection of the same bacterial cell by two phages. It is debated, however, whether (relaxed) homologous recombination (de Paepe et al. 2014) or illegitimate recombination (Hatfull 2015) is the most prevalent mechanism.

Many instances of gene content variation between closely related phages have been observed. This project will investigate whether they originated by illegitimate or (relaxed) homologous recombination. In the case of (relaxed) homologous recombination, it is expected to find traces of homology with an elevated polymorphism density in the stretches surrounding the variable genes.

In this project, an approach will be implemented to identify whether (relaxed) homologous recombination could have resulted in the observed gene content variation. To this end, variable regions between pairs of related bacteriophages will be identified and the surrounding regions will be analyzed for polymorphisms. All available genomes of bacteriophages and archaeal viruses will be analyzed and it will be investigated whether there are differences in the prevalence of (relaxed) homologous recombination between different taxonomic groups of viruses or between different host species.

References

Dion MB, Oechslin F, Moineau S. 2020. Phage diversity, genomics and phylogeny. Nat Rev Microbiol 18:125–138.

Hatfull GF. 2015. Dark matter of the biosphere: The amazing world of bacteriophage diversity. Journal of Virology:JVI.01340–15.

Kupczok A, Neve H, Huang KD, Hoeppner MP, Heller KJ, Franz CMAP, Dagan T. 2018. Rates of mutation and recombination in Siphoviridae phage genome evolution over three decades. Molecular Biology and Evolution 35:1147–1159.

De Paepe M, Hutinet G, Son O, Amarir-Bouhram J, Schbath S, Petit M-A. 2014. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. PLOS Genetics 10:e1004181.

Inferring viral recombination from metagenomes

Supervisors	Anne Kupczok, Dick de Ridder
Type	Data analysis, Methodology
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Metagenomics, population genetics, databases, python
Timestamp	September 2020

Description

Bacteriophages (short: phages) are viruses that infect bacteria, i.e. they only reproduce during bacterial infection (Dion et al. 2020). Their mode of evolution is characterized by high mutation rates and frequent recombination and horizontal gene transfer (e.g., Kupczok et al. 2018). However, the prevalence of recombination has rarely been quantified for phages.

Our view of bacteriophage diversity has recently been expanded by viral metagenomics, i.e., the community sequencing of viruses in an environment. Those data sets also provide a snapshot of the population diversity at the timepoint of sequencing. This population diversity encompasses all the sampled variants within a species, in particular, variation in gene content and single-nucleotide variants (SNVs) in shared regions. The correlation between SNVs provides a signal for the strength of recombination; this signal is, for example, used by the mcorr method (Lin and Kussell 2019). Phage genomes, however, are substantially shorter than bacterial genomes. In addition, phage-encoded recombination proteins allow for recombination at sites of limited homology (de Paepe et al. 2014); thus, the pattern of recombination might be different in phages. It is thus unclear, whether the mcorr approach results in as clear correlation profiles in phages as for bacteria.

This project will assess the applicability of the mcorr approach to viral metagenomes. To this end, public viral metagenomes will be selected that cover phage reference genomes and these genomes will be analyzed with mcorr. Then, the project can take two different directions. If found that mcorr is applicable to phages, different phages and phages from different environments will be compared for their inferred recombination rates. On the other hand, if mcorr is not applicable to phages, alternative methods will be investigated and an approach that takes the specificities of phage recombination into account will be designed.

References

Dion MB, Oechslin F, Moineau S. 2020. Phage diversity, genomics and phylogeny. Nat Rev Microbiol 18:125–138.

Kupczok A, Neve H, Huang KD, Hoeppner MP, Heller KJ, Franz CMAP, Dagan T. 2018. Rates of mutation and recombination in Siphoviridae phage genome evolution over three decades. Molecular Biology and Evolution 35:1147–1159.

Lin M, Kussell E. 2019. Inferring bacterial recombination rates from large-scale sequencing datasets. Nature Methods 16:199–204.

De Paepe M, Hutinet G, Son O, Amarir-Bouhram J, Schbath S, Petit M-A. 2014. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. PLOS Genetics 10:e1004181.

Machine learning to predict enzyme regioselectivity

Supervisor	Marnix Medema, Aalt-Jan van Dijk
Type	Sequence analysis, Cheminformatics, Machine Learning
Requirements	Advanced bioinformatics, Machine Learning
Skills	Programming, Machine Learning, Chemistry
Timestamp	March 2021

Description

Plants and microbes produce a wide range of specialized metabolites that provide them with ecologically and physiologically specialized functions. These metabolites are often synthesized by the action of complex enzymatic pathways, in which one class of enzyme constructs a core scaffold of the molecule and other enzymes further modify this scaffold through group transfer or redox reactions. Predicting where on the target molecule such tailoring enzymes act to make these modifications (their ‘regioselectivity’) is challenging.

Recent developments in machine learning methodology allow the prediction of enzyme functions from large-scale data, thereby enabling effective charting of the massive diversity of enzymes in the biosphere (1). Furthermore, integrative approaches are being developed that combine crystallographic data, genomic context and cheminformatics to map enzymes to proposed catalytic activities and even pathways (2).

In this project, you will make use of the latest technologies to develop a (prototype for) a new algorithm that will use chemical fingerprinting of regioselectivities using a combination of both 2D and 3D molecular features to predict where on a molecule an enzyme will act. Taking methyltransferases as a case study, you will first assess the level of phylogenetic conservation of regioselectivities based on experimental data (3) and identify functionally related groupings across protein sequence diversity, before using a combination of protein sequence and 3D structural features to train machine-learning classifiers for regioselectivity prediction.

The resulting methods will provide a foundation for future studies to systematically chart the functional diversity of enzyme families and predict metabolite structures from genome sequence data by combining predictions of multiple enzymes in a pathway.

References

1. Calhoun et al. (2018) Prediction of enzymatic pathways by integrative pathway mapping. *eLife* 7, e31097.
2. Durairaj et al. (2021) Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Comp. Biol.* 17, e100819.
3. Kautsar et al. (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* 48, D454–D458.

Strategic genome mining for novel antimicrobial compounds

Supervisor	Mohammad Alanjary, Marnix Medema
Type	Data analysis, Genome mining
Requirements	Advanced bioinformatics
Skills	Programming, Genomics, Biology
Timestamp	January 2021

Description

The fortunate discovery of broad-spectrum antimicrobial compounds, such as penicillin, has been a crucial component to healthcare not only for thwarting infectious diseases but also ensuring recovery after life-saving procedures. Unfortunately, microbes resistant to many known compounds, with some showing pan-resistance to all known treatments (1), are on the rise. This threat of antimicrobial resistance (AMR) is further compounded by a waning development of new drugs to market, most of which were known nearly 50 years ago. Despite new discoveries using synthetic and traditional pipelines, none have shown to be promising in the treatment of 'critical priority' Gram-negative pathogens (2). Today, there is an opportunity to re-invigorate this pipeline by leveraging genome mining methods to discover natural compounds that have a low chance of re-discovery (3,4). In this project you will help develop prioritization schemes and search for biosynthetic gene clusters (BGCs) that are likely to offer broad-spectrum activity. By leveraging gene cluster networking methods (5) to group BGCs into families, and prioritizing gene cluster families based on several criteria, you will help generate leads for novel antimicrobial compounds.

Criteria will include 1) Taxonomic 'dispersion' patterns of BGCs (part of the 'accessory genome' but shared between distantly related species); 2) Chemical structure predictions indicating potential to penetrate Gram-negative cell envelope; 3) Presence of associated active export or self-resistance markers; 4) Functional replacement (by another antibiotic BGC) in closely related species or potential synergistic association with other BGCs (e.g., supercluster configurations); 5) Ecological or environmental metadata associated with a need to produce antibiotics.

Leads will be further analysed and pursued in collaboration with the lab of Kim Lewis, Northeastern University, Boston (USA).

References

1. Göttig,S., Gruber,T.M., Higgins,P.G., Wachsmuth,M., Seifert,H. and Kempf,V.A.J. (2014) Detection of pan drug-resistant Acinetobacter baumannii in Germany. J. Antimicrob. Chemother., 69, 2578–2579. http://doi.org/10.1093/jac/dku170.
2. Lewis,K. (2020) The Science of Antibiotic Discovery. Cell, 181, 29–45. http://doi.org/10.1016/j.cell.2020.02.056.
3. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes-a review. Nat. Prod. Rep., 33. http://doi.org/10.1039/c6np00025h.
4. Ziemert, N., Weber, T. & Medema, M. H. Genome Mining Approaches to Bacterial Natural Product Discovery. Comprehensive Natural Products III (Elsevier Inc., 2020)
5. Navarro-Muñoz et al. (2020) A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology* 16: 60-68.

Chemical structure encoder for identifying monomer composition in polyketides and non-ribosomal peptides

Supervisors:	Mohammad Alanjary, Barbara Terlouw, Marnix Medema
Type:	Software engineering, genome analysis
Requirements:	Programming in Python, Advanced Bioinformatics
Skills:	Programming, genome analysis
Timestamp:	April 8, 2021

Natural products, compounds produced by a variety of organisms, have been a major source of life-saving drugs and antibiotics [1]. Two major enzyme classes that form these molecules, non-ribosomal peptide synthetases (NRPS) and polyketide synthases (PKS), have a remarkable biosynthesis route utilizing a multi-modular architecture that forms a factory-line of enzymatic domains to assemble the product [2]. These have been shown to be amenable to modification through a variety of approaches, such as module replacement, and can potentially be leveraged to increase production of their respective products, or create new derivatives [3,4]. Likewise, polyketide synthases (PKS) offer similar opportunities for modification and operate in a predictable modular fashion [5]. To identify similar monomer sequences that lead to predictable motifs and to enable targeted synthesis and engineering of new molecules, it is necessary to deconstruct these products to their likely starting monomer sequences.

In this project, you will take part in efforts to create a system to encode monomer sequences from PKS/NRPS products and provide methods to organize and search for shared or known monomer sequences in a generated database. These efforts can also help accelerate re-engineering of these biosynthetic routes by identifying likely starting points for a particular chemical structure. This will entail expanding on previous work that includes input of chemical structures though a web interface and cataloging currently known products for natural product discovery.

Key references

1. Newman,D.J. and Cragg,G.M. (2020) *J. Nat. Prod.*, 83, 770–803.
2. Süssmuth,R.D. and Mainz,A. (2017) *Angew. Chemie Int. Ed.,* 56, 3770–3821.
3. Alanjary,M., Cano-Prieto,C., Gross,H. and Medema,M.H. (2019) *Nat. Prod. Rep.*, 36, 1249–1261.
4. Bozhüyük,K.A.J., Linck,A.,et al. (2019) *Nat. Chem.*, 10.1038/s41557-019-0276-z.
5. Helfrich E.J.N. et al. (2021) *Nat. Commun.* 12, 1422.

Mining microbiomes for novel peptidic natural products

Supervisor	Marnix Medema
Type	Data analysis, Genome mining
Requirements	Advanced bioinformatics
Skills	Programming, Genomics, Biology
Timestamp	March 2021

Description

Microbial natural products constitute a wide variety of chemical compounds, many which can have antibiotic, antiviral, or anticancer properties that make them interesting for clinical purposes. Natural product classes include polyketides (PKs), nonribosomal peptides (NRPs), and ribosomally synthesized and post-translationally modified peptides (RiPPs). While variants of biosynthetic gene clusters (BGCs) for known classes of natural products are easy to identify in genome sequences, BGCs for new compound classes escape attention. In particular, evidence is accumulating that for RiPPs, subclasses known thus far may only represent the tip of an iceberg (1). We recently introduced decRiPPter (Data-driven Exploratory Class-independent RiPP TrackER), a RiPP genome mining algorithm aimed at the discovery of novel RiPP classes (2). DecRiPPter combines a Support Vector Machine (SVM) that identifies candidate RiPP precursors with pan-genomic analyses to identify which of these are encoded within operon-like structures that are part of the accessory genome of a genus. Subsequently, it prioritizes such regions based on the presence of new enzymology and based on patterns of gene cluster and precursor peptide conservation across species.

While decRiPPter has already been used to successfully identify multiple dozens of new candidate RiPP classes, the pipeline has only been systematically applied to *Streptomyces* bacteria. Here, you will use the tool to analyse a wide range of genera that are important members of the human and/or plant microbiota, in order to identify new types of biosynthetic pathways that might be important in mediating host-microbiome interactions and conferring microbiome-associated phenotypes. You will then systematically chart the diversity of detected gene clusters using sequence similarity networking (3) and map metagenomic and metatranscriptomic reads to these gene clusters to assess their abundance and expression across healthy and disease-associated samples using state-of-the-art tools we recently developed for this (4). Leads will be further analysed and pursued in collaboration with the lab of Mohamed Donia, Princeton University (USA).

References

1. Kloosterman A.M., Medema M.H., van Wezel G.P. (2021) Omics-based strategies to discover novel classes of RiPP natural products. *Curr. Opin. Biotechnol.* 69, 60-67.
2. Kloosterman, A. M. et al. Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. *PLoS Biol.* 18, e3001026 (2020).
3. Navarro-Muñoz et al. (2020) A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology* 16: 60-68.
4. Pascal Andreu et al. (2021) BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. *BioRxiv*, doi: 10.1101/2020.12.14.422671.

A novel metabolite annotation approach combining LC-MS and LC-MS/MS data

Supervisors	Justin van der Hooft and Ric de Vos (BU Bioscience)
Type	Metabolomics Data Analysis / Assembling molecular fragments / Databases
Requirements	Progr. in Python, Advanced Bioinformatics
Skills	Programming, Data analysis, Statistics
Timestamp	24 April 2020

Description

Metabolomics is nowadays a world-wide used, deep-chemical phenotyping tool. For example, It is enabling scientists to determine the effects of natural and induced genetic variation. Liquid chromatography with a high-resolution mass detector is the favourite platform for multi-parallel detection of non-volatile compounds. With current methodologies, only a few percent of metabolite features can confidently be identified in natural extracts due to their enormous structural diversity. The main reason for the relatively low annotation level in untargeted LC-MS data is that although the elemental formula of a detected compound can be frequently deduced from its accurate mass. Additional compound characterization based on molecular fragmentation can reveal more structural information. Such datasets are either concurrently generated in the ion-source upon LC-MS analysis [1] or deliberately generated in LC-MSMS mode [2], as well as tools for high mass resolution-spectral library matching are needed to enable a better and more automated annotation of as many detected compounds as possible. In this project you will use dedicated data sets specifically generated at the plant metabolomics group of WR-Bioscience: analysed in 2 full scan LC-MS modes, and 2 LC-MSMS fragmentation modes, i.e. the commonly used data-dependent (DD)-MSMS mode and the relatively new All Ion Fragmentation (AIF)-MSMS mode. In LC-AIF MSMS mode all compound ions entering the mass spectrometer are simultaneously fragmented resulting in a mix of fragments from several co-eluting compounds. While AIF-MSMS data are more challenging to interpret, since the precursor (parent) ion is by definition undefined, the method is clearly more sensitive and thus potentially more powerful in comparison to classical DD-MSMS. To tackle this, you will assemble molecular ion masses from the full scan LC-MS data set with their specific fragment masses from the LC-AIF MSMS data set, based on their correspondence in both LC-retention time and abundance pattern across a series of contrasting samples, a possibility enabled by the MSClust tool [3]. This new, MSClust-based molecular ion-AIF MSMS data clustering approach will be compared with classical DD-MSMS data for metabolite fragment coverage, as well as for their performance in automated compound annotation by matching the obtained mass spectra information with metabolite mass spectral libraries and Mass2Motifs substructure classification using MS2LDA [2]. Your project will thus lead to a novel and potentially powerful approach for high-throughput annotation of metabolites detected by untargeted LCMS approaches, which is key to all metabolomics research both within WUR and overall.

Key references

[1] Xue et al. 2020, Anal Chem. https://dx.doi.org/10.1021/acs.analchem.0c00409 [2] van der Hooft et al. 2016, Proceedings of the National Academy of Sciences, 113, 13738-13743. [3] Tikunov et al. 2012, Metabolomics 8, 714-718.

Latent space models to link metabolite structures and MS spectra

Supervisors	Justin van der Hooft, Aalt-Jan van Dijk
Type	Machine learning
Requirements	Progr. in Python, Advanced Bioinformatics, Machine Learning
Skills	Programming, data analysis
Timestamp	March 2020

Description

Metabolites produced by microbes, fungi, and plants play a lot of important functions such as energy transport or signalling, and are used for various applications like antibiotics and drugs. Understanding biochemical characteristics of metabolites is an essential part of metabolomics to enlarge fundamental biological knowledge and to further develop applications in areas such as biotechnology or biomedicine. However, metabolite identification remains a challenging task in metabolomics with complex mixtures that typically contain a huge number of unknown metabolites. In the standard approach based on mass spectrometry fragmentation workflows (LC-MS/MS), the key problems are that the resulting MS/MS spectra cannot easily be connected to known metabolites and unknown metabolite structures are still difficult to elucidate. Given the increased availability of datasets of spectra with known associated metabolite structures, a promising direction is to use machine learning to aid in this process. Recently, an approach has been proposed which learns a mapping between metabolite structures and spectra, via molecular vectors [Nguyen 2019]. These molecular vectors are generated by a so-called message passing neural network (MPNN), which is a deep learning approach specifically aimed at learning representations of (molecular) graphs. Parameters of the model are learnt by maximizing the correlation between given spectra and molecular vectors. Importantly, in this approach, only metabolite-spectrum pairs can be used, i.e., spectra without known metabolite assignment or metabolites without measured spectrum are ignored. In this project, we will investigate the use of these ‘unlabelled’ data sources. One possible approach is to use the above-mentioned MPNN method, and include metabolites without measured spectra with an additional term in the loss function. Key in this case would be the construction of this additional term in the loss function, which would capture chemical similarities between metabolites and/or other additional constraints one has on the representation of metabolites. Alternatively, unsupervised approaches have been developed to model molecular graphs, in particular, an approach called graphNVP [Madhawa 2019]. This learns a latent representation of molecules, without taking spectra into account at all. We could explore the use of graphNVP as a starting point, and then use metabolite-spectrum pairs to optimize the correlation between a representation of spectra and the latent representation of molecules. Finally, latent representations of spectra can also be made, for example by applying natural language processing (NLP) algorithms and correlating the two latent spaces using known metabolite – spectra links as anchors could be an interesting route. This project will investigate the use of novel machine learning approaches to facilitate metabolite identification from LC-MS/MS data. The outcome can be incorporated in widely used metabolomics platforms such as GNPS [Wang 2016].

Nguyen 2019, Bioinformatics, 35, 2019, i164–i172. Madhawa 2019, https://arxiv.org/abs/1905.11600
Wang 2016, Nature Biotechnoloy, 34, 828–837.

MS2ChemClass: automated mass spectral-based chemical class annotation

Supervisors	Joris Louwen, Florian Huber, Aalt-Jan van Dijk, Justin van der Hooft
Type	Machine learning
Requirements	Progr. in Python, Advanced Bioinformatics, Machine Learning
Skills	Machine Learning, Programming, Data Analysis
Timestamp	April 2021

Description

Metabolites produced by microbes, fungi, and plants play many important functions such as energy transport or signalling, and are used for various applications like antibiotics and drugs. Understanding biochemical characteristics of metabolites is an essential part of metabolomics to enlarge fundamental biological knowledge and to further develop applications in many areas. However, metabolite identification remains a challenging task in metabolomics with complex mixtures that typically contain a huge number of unknown metabolites. In the standard approach based on mass spectrometry fragmentation workflows (LC-MS/MS), the key problems are that the resulting MS/MS spectra cannot easily be connected to known metabolites and unknown metabolite structures are still difficult to elucidate. Therefore, there is an increased interest in annotating metabolite features at the level of substructures (partial structure) or chemical compound classes, i.e., a group of molecules sharing chemistry and/or biosynthesis pathways. Chemical class ontologies for metabolomics (ChemOnt) (Djoumbou Feunang 2016) and natural products (NPClassifier) (Kim 2020) were recently coined.

Given the increased availability of datasets of spectra with known associated metabolite structures, a promising direction is to use machine learning to aid in this process. For example, metabolite features in metabolomics datasets annotated with chemical compound classes would allow for a comprehensive chemical overview to effectively direct further analysis efforts into the relevant type of molecules, as for example shown by MolNetEnhancer. It would also help to connect metabolomics output to that of genomics analysis, since genome mining can also predict the chemical compound classes of the - in the genome - encoded molecules. The current state of the art is CANOPUS that works with computed fragmentation trees to predict molecular fingerprints before assigning chemical compound classes to MS/MS spectra of metabolite features. This works very well for metabolites below 500 Da, but takes considerable more time for larger-sized metabolites as typically faced in natural product research.

In this project, you will work with ~100,000 curated reference spectra of >15,000 unique metabolites from GNPS public spectral libraries [Wang 2016] for which structure information is available and chemical compound class information can thus be obtained. Based on available unsupervised (Huber 2021) and supervised mass spectral embeddings, you will test which classification algorithm, for example based on K-Nearest Neighbours, can quickly and reliably assign which chemical compound classes to MS/MS spectra and design an optimal workflow for quick classification of MS/MS spectra from large-sized metabolites. The resulting algorithm will become part of the matchms (Huber 2020) and Spec2Vec (Huber 2021) codebase and could be used by platforms like GNPS and will be integrated into NPLinker to facilitate integrated genome-metabolome analyses.

Djoumbou Feunang 2016, Journal of Cheminformatics, 8, 61; Kim 2020, ChemRXiv; Ernst 2019, Metabolites, 9(7), 144; Dührkop 2020, Nature Biotechnology; Wang 2016, Nature Biotechnoloy, 34, 828–837; Huber 2020, Journal of the Open Source Software.

Elemental Formula and Mass Difference-enhanced Substructure Discovery in Metabolomics Profiles

Supervisors	Justin van der Hooft, Kai Dührkop (Jena, Germany), Joris Louwen, Simon Rogers (Glasgow, UK).
Type	Computational Metabolomics, Machine learning
Requirements	Progr. in Python, Advanced Bioinformatics, Machine Learning
Skills	Programming, Data Analysis
Timestamp	7 December 2020

Description

Metabolites produced by microbes, fungi, and plants play many important functions such as energy transport or signalling, and are used for various applications like antibiotics and drugs. Understanding biochemical characteristics of metabolites is an essential part of metabolomics to enlarge fundamental biological knowledge and to further develop applications in areas such as biotechnology or biomedicine. However, metabolite identification remains a challenging task in metabolomics when faced with complex natural mixtures that typically contain a huge number of unknown metabolites. In the standard approach based on mass spectrometry fragmentation workflows (LC-MS/MS), the key problems are that the resulting MS/MS spectra cannot easily be connected to known metabolites and the elucidation of unknown metabolites.

Computational metabolomics workflows have been developed to overcome the above limitations by linking structure or substructure information to metabolomics profiles. For substructure discovery, co-occurrence patterns in mass fragments and neutral losses (difference between precursor m/z and mass fragment) are sought. Mass spectra also contain many so-called mass differences, i.e., the difference between mass fragments - and the presence of some of those may contain relevant structural information for substructure annotation as well. For example, in peptidic spectra they may relate to the presence of particular amino acid residues in the fragmented metabolite. Here, you will explore the synergies between two popular computational metabolomics tools, SIRIUS-CSI:FingerID fragmentation tree-based annotation [Dührkop, 2019] and MS2LDA substructure discovery [Van der Hooft, 2016], as a basis to annotate elemental formulas in spectra and select informative mass differences and use both as basis for substructure discovery. One main problem resides in the large amount of mass differences that can be determined in a typical MS/MS spectrum of which only a small subset contains relevant information not captured by mass fragments and neutral losses alone. It is this subset that can be used as diagnostic mass features as part of substructure patterns in metabolomics profiles. Using fragmentation trees of public reference spectra and datasets and the corresponding masses and elemental formula assignments, you will build up a mass difference library by using SIRIUS annotations of mass differences and then feed them into the MS2LDA algorithm to discover mass difference-enhanced Mass2Motifs. With help of state-of-the-art tools such as QemisTree [Tripathi, 2020] and CANOPUS [Dührkop, 2020], you will annotate a number of mass spectral patterns. The project will result in a means to select informative mass difference to enhance substructure discovery, a list of relevant mass differences for the studied datasets, and an annotated motif set for MotifDB (www.ms2lda.org/motifdb).

Dührkop et al., 2019, Nature Methods, 16, 299–302. Van der Hooft et al., 2016, PNAS, 113(48), 13738-13743. Tripathi et al., 2020, Nature Chemical Biology. Dührkop et al., 2020, Nature Biotechnology.

Tracing Drug and Food Substructures & their Biotransformations in Urine

Supervisors	Justin van der Hooft, Madeleine Ernst (SSI, Denmark), Joris Louwen, Florian Huber (eScienceCenter).
Type	Computational Metabolomics, Machine learning
Requirements	Progr. in Python, Advanced Bioinformatics, Machine Learning
Skills	Programming, data analysis
Timestamp	7 December 2020

Description

Despite containing >95% of water, urine is a complex metabolite mixture containing traces of the food and medicines we take in. There is increasing interest to develop unbiased screenings to map someone's dietary and medicinal "footprints" in order to link this to different phenotypes such as risk to the onset of a particular disease. Urine analysis is complex and typically done using a wide-screening metabolomics approach thereby collecting spectra for hundreds to thousands of metabolites. However, metabolite identification remains a challenging task in metabolomics with complex mixtures that typically contain a huge number of unknown metabolites. In the standard approach based on mass spectrometry fragmentation workflows (LC-MS/MS), the key problems are that the resulting MS/MS spectra cannot easily be connected to known metabolites and unknown metabolite structures are still difficult to elucidate. Here, you will have the chance to work on a unique urine metabolomics dataset that was obtained following the untargeted metabolomics workflow as set out in Van der Hooft et al., 2016a, for 25 urine samples, but now applied to 100 urine samples of which anonymized metadata is available including their drug use, to develop a substructure and biotransformation-inspired approach to gain insight in the chemical complexity and find drug and food metabolites and their biotransformations. In there, 125 drug metabolites derived from 39 different drugs were previously annotated. You will start with unsupervised MS2LDA substructure discovery [Van der Hooft, 2016b], and network analysis using molecular networking [Wang 2016], combined in MolNetEnhancer [Ernst et al., 2019] and mass spectral embedding-based approaches [Huber, 2020]. To map relevant biotransformations in the urine metabolomics data, you could use various routes such as a combination of Ion Identity Networking [Schmid, 2020] and Paired Mass Differences [Yu, 2020] or perform more supervised approaches by linking spectral features to the available metadata. This will also aid in prioritizing molecular families and substructure motifs that are related to drug and food intake. Using spectral libraries and public data available from sources like GNPS-MassIVE, you will create a curated list of relevant mass differences that indicate microbial or human biotransformations, annotate series of food and drug related metabolites in the urine metabolomics profiles and create a public library thereof (also including the 125 previously annotated drug metabolites), and build drug- and food-related public motif sets for MotifDB (www.ms2lda.org/motifdb). This project will aid in starting to build the necessary computational metabolomics infrastructure to start using a wide-screen metabolomics approach as a diagnostic toolkit in the clinic.

Van der Hooft et al., 2016a, Metabolomics, 12, 125. Van der Hooft et al., 2016b, PNAS, 113(48), 13738-13743. Wang et al., 2016, Nature Biotechnology, 34, 828–837. Ernst et al., 2019, Metabolites, 9(7), 144. Huber et al., 2020, bioRxiv, 10.1101/2020.08.11.245928v2. Schmid et al., 2020, bioRxiv, 10.1101/2020.05.11.088948v1. Yu and Petrick, 2020, Communications in Chemistry, 3, no. 157.

Toward “Natural Products aware” Chemical Fingerprints

Supervisors	Justin van der Hooft and Maria Sorokina (University of Jena, Germany)
Type	Cheminformatics, Computational metabolomics
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Basic Chemistry and Biology, Programming
Timestamp	7 December 2020

Description

Mammals and plants are externally and internally colonized by diverse microbial communities [1, 2]. Microorganisms have evolved to adapt well to the different environments they are facing; in particular, bacteria can produce an arsenal of structurally diverse specialized molecules, also referred to as natural products [3]. Natural products (NPs) inspire the pharmaceutical industry and research due to their exceptional structural characteristics from which they derive their activities. NPs also present an extraordinary structural diversity, sometimes difficult to discover and rank. General interest in NPs is surging as it is regarded as a promising source of bioactive molecules. Therefore, initiatives to capture natural products through chemical fingerprints have also started [4]; however, these do not yet fully appreciate and cover the diversity of natural products and are not easy to use. Furthermore, an NP-likeness score [5] allows to estimate how likely a molecule is to be a natural product from the purely structural point of view, but this score can be considerably improved, for example through more accurate “NP-aware” chemical fingerprints. The main goals of this project are to improve the NP-likeness score to prioritise natural product structure discovery and to define a general NP chemical fingerprint that works for all NP classes. During the project, you will: i) establish an advanced NP-likeness score [5], based on the co-occurrence of the molecular substructures and features in the natural products, ii) train the new score on a high-quality natural product set of over 200 000 molecules (OpenNP database/LOTUS) and establish the most frequently co-occurring substructures in natural products and run comparative statistics between the different fingerprints (e.g. circular, PubChem, extended fingerprints, NPRules [6]) and on the various subsets from COCONUT [7] (https://coconut.naturalproducts.net/). Furthermore, you may investigate the role of sugar moieties in the natural product fingerprint and add extensions to the fingerprint for sugars and their types. This project will contribute toward a better prioritization of natural products from a large set of structures and could eventually be used to accelerate the linking of mass spectral and genomic features to natural products.

Key references:

1. Mendes, R. and J.M. Raaijmakers,. The Isme Journal, 2015; 2. Hacquard, S., et al., Cell Host & Microbe,2015; 17(5): p. 603-616; 3. Donia, M.S. and M.A. Fischbach, Science, 2015. 349(6246); 4. Seo et al., J. of Cheminformatics, 2020, no 6;, 5. Sorokina and Steinbeck, J. of Cheminformatics, 2019, no 55; 6. Sam Stokman, Thesis WUR, 2019, https://library.wur.nl/WebQuery/groenekennis/2249753; 7. Sorokina and Steinbeck, J. of Cheminformatics, 2020, no 20.

Local assembly of QTL regions in plants

Supervisors	Sara Diaz, Dick de Ridder
Type	Quantitative trait loci analysis, data analysis
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Genomics, programming
Timestamp	March 2020

Description

Plant breeding relies on identifying, characterizing and using Quantitative Trait Locus (QTLs). QTLs are genomic regions from one specific plant that associate with a desired trait. More than 76K QTLs are described in literature for crops. Many of those are only described by a few genetic markers and (almost) no information on the plant material source.

In this project, we aim to develop methods to locally assemble QTL regions in crops in order to analyze them for candidate gene prioritization. We will make use of available data from specific crops (10x genomics, PacBio, Nanopore) and adjust local assembly methods to these long read technologies. We will then compare the QTLs in the source vs the reference genome, both in euchromatic regions and complex regions such as resistance gene clusters.

References

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2949-4
https://www.ncbi.nlm.nih.gov/pubmed/25039268

Using machine learning to predict underlying factors for plant meiotic recombination rate variation

Supervisors	Aalt-Jan van Dijk
Type	Algorithm development, Data analysis
Requirements	Machine Learning; Programming in Python; Adv. bioinf., Adv. statistics;
Skills	Programming, statistics
Timestamp	April 8, 2021

Description

Meiotic recombination plays a key role in ensuring accurate segregation of homologous chromosomes by the formation of at least one crossover (CO). Plant breeding exploits the genetic variation resulting from COs to introduce favorable genes from wild varieties. Using current DNA sequencing technology, COs can be studied at unprecedented levels of throughput and resolution. However, so far, we lack a clear understanding of what determines variation in CO occurrence. We recently demonstrated an initial Random Forest model for crossovers observed based on crosses in a few different plant species [1]. This model predicts regions which are likely to be targeted by recombination, using features such as DNA sequence and predicted DNA shape. In this project, you will add additional features, including epigenetics-based (e.g. histone or DNA methylation). Moreover, historical recombination rate datasets obtained from natural populations will be analyzed with machine learning as well, albeit that here regression and not classification has to be applied.

From the machine learning models, features will be found which are important to predict recombination rate variation. Comparison of these features for natural populations vs. data obtained from crosses will help understanding similarities and differences between these populations. In case the data and models indicate sufficient similarity between crossovers from breeding crosses and from natural populations, we will develop a combined model for these two types of datasets, potentially leading to improved prediction accuracy. Such more advanced modelling approach will combine regression and classification. Our Random Forest based approach will be extended to have a classification and regression part which both use the same partition of feature space. Alternatively, logistic regression could be used for the classification part and linear regression for the regression part; regularization could be used to encourage the regression and the classification part of the model to select similar features. Finally, deep learning might be explored as an alternative modelling strategy, taking inspiration from recently developed related approaches [2].

References

[1] Demirci, S., S.A. Peters, D. de Ridder, and A.D.J. van Dijk, DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. Plant J, 2018.

[2] Adrion, J.R., Galloway, J.G., and A.D. Kern, Predicting the landscape of recombination using deep learning, Mol. Biol and Evol., 2020, 37, 1790-1808.

Explaining functional properties of non-domain protein sequence regions

Supervisors	Aalt-Jan van Dijk, Richard Immink
Type	Data analysis
Requirements	Advanced bioinformatics
Skills	Programming, data analysis
Timestamp	April 8, 2021

Description

Functional properties of protein sequences are often linked to specific domains. Such domains are recognizable in various related proteins, and are captured e.g. by PFAM using an HMM description. However, in addition to recognizable domains, many proteins have relatively long stretches of non-domain sequence regions in-between. A given non-domain sequence region in a protein has by definition less similarity to non-domain regions in other proteins, compared to similarity between domains. Hence, predicting functional sub-regions or residues based on e.g. a multiple sequence analysis is difficult or even impossible. However, in some cases, clear evidence has accumulated that indicates that such non-domain regions, have important functional roles. We recently performed an analysis on MADS domain proteins, important plant transcription factors involved in regulating e.g. flowering time and flower formation. For these proteins, it is known that in addition to two recognizable domain regions, there are also two non-domain regions. We assembled a large set of >10000 of MADS domain protein sequences from ~1000 different plant species, and mined these for sequence motifs. This resulted in a large set of motifs, for which we then analyzed evolutionary properties such as conservation and clade-specificity. In addition, we found that some of these motifs displayed clear overlap with sites in these proteins found experimentally to contribute to function.

In this project, we will finalize this analysis on MADS domain proteins and apply a similar analysis to the PEBP family of proteins. The famous Florigen protein, acting as a hormone like substance to initiate flowering, belongs to this highly conserved family. PEBP proteins can form complexes with transcription factors, but also with sugar transporters, making them versatile and multitasking proteins. Nevertheless, we are far from understanding the molecular characteristics essential for these various important functions and we believe that a thorough investigation of possible functionality of non-domain regions could aid in unlocking these unsolved mysteries.

One additional analysis step which seems promising but is yet untested on plant proteins is to look at specific features using the approach described in [1].

[1] https://elifesciences.org/articles/46883

Matrix factorization for gene function prediction

Supervisors	Aalt-Jan van Dijk, Dick de Ridder
Type	Algorithm development, Data analysis
Requirements	Programming in Python, Adv. bioinf., Adv. statistics; Par. estimation / Modern stat. for the life sciences / Machine Learning
Skills	Programming, statistics
Timestamp	September 4, 2019

Description

Current experiment-based Gene Ontology annotations are far from complete, and various data sources are available that can be used to infer novel gene annotations for genes with so far unknown function. Matrix Factorization (MF) is a promising method to integrate data sources. MF is widely used in various application areas, e.g. in recommender systems (deployed when items such as movies are to be recommended to users). We recently experimented using the approach presented by Zitnik (2015), a penalized low-rank matrix factorization, to predict gene function for A. thaliana. In this project we will extend this in the following directions:

1. Using additional data: so far, expression, protein domains, and KEGG were used. We will investigate which additional data might be predictive. This will include data from Genome Wide Association Studies and from transcription factor – target interactions. These datasources have so far been underexplored for gene function prediction but contain valuable information. A subsequent step could also be to investigate using data from multiple species simultaneously.
2. Model development: one key issue that became clear during the initial investigations was that the dimension of the learned representations strongly influences performance. We will investigate how to set values for this parameter, using a measure for complexity of the learned model vs how well the model fits the data. Secondly, the objective function contains a term which describes how well the factorized representation resembles the input data, and a term for gene-gene similarity. We will investigate how to scale these two terms. A third step which deserves attention is how the learned model is used to deliver gene function prediction for novel genes. Finally, we will consider the relationships inferred by the model between e.g. genes and pathways; it might be possible to use prior knowledge on such relationships as constraints, which might then indirectly lead to improved gene function predictions.
3. Interpreting the learned models is a so far underexplored part of MF. Similar to how PCA loadings can be used to learn about underlying structure in data, we envision to use the learned latent representations in MF to understand in what ways expression relates to gene function.

Depending on your background and interest, we could focus on either of these directions. MF is an active area of research, and it would also be possible to explore additional approaches, including recent deep learning based methods (https://bitbucket.org/cdal/dcmf/src/master/)

References

Zitnik and Zupan, IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 2015, 41.

The origins of promiscuity in fragrance-producing enzymes

Supervisors	Aalt-Jan van Dijk
Type	Data analysis
Requirements	Advanced Bioinformatics
Skills	Data analysis, Programming
Timestamp	April 8, 2021

Description

Terpene synthases (TPSs) are a large specialized class of enzymes responsible for the production of terpenes, key-compounds contributing to the diverse fragrances of plants. There are more than 25,000 different plant terpenes all deriving from the same 5-carbon precursor units, coupled together linearly and then cyclized, rearranged, and modified. TPS enzymes, which catalyse these reactions, are very diverse in terms of sequence [1] but share a common structural fold across plants, animals, bacteria, and fungi, pointing to their common evolutionary origin.

Though many TPSs are highly-specific and produce only a single terpene product, many others are more promiscuous and can produce a number of products. The record being an enzyme synthesizing an astounding 52 different terpenes [2]. In addition, due to duplication and sub/neofunctionalization events, a single plant species can have anywhere between 20 and 150 TPS genes with varying product profiles [3]. These two factors result in plants producing mixtures of terpenes, with varied roles in pollinator attraction and herbivore repulsion, that give each species its unique scent.

This project will concentrate on unravelling the origins of terpene enzyme promiscuity arising from plants. Depending on your interests you may choose to pursue one of two different lines of inquiry:
1) A phylogenetic analysis will help reveal evolutionary pressures on specific residues determining the number and type of terpenes produced by an enzyme.
2) Comparative phylogenomics of TPS genes across multiple plant genomes will establish the role of genomic environments in the number and type of terpenes produced by a species.

References

[1] Durairaj, J., Di Girolamo, A., Bouwmeester, H. J., de Ridder, D., Beekwilder, J., & van Dijk, A. D. (2018). An analysis of characterized plant sesquiterpene synthases. Phytochemistry.
[2] Degenhardt, J., Köllner, T. G., & Gershenzon, J. (2009). Monoterpene and sesquiterpene synthases and the origin of terpene skeletal diversity in plants. Phytochemistry, 70(15-16), 1621-1637.
[3] Chen, F., Tholl, D., Bohlmann, J., & Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. The Plant Journal, 66(1), 212-229.

Protein-protein interaction prediction using co-evolution

Supervisors	Aalt-Jan van Dijk
Type	Data analysis
Requirements	Advanced Bioinformatics, Machine Learning, Algorithms in Bioinformatics
Skills	Genomics, Programming
Timestamp	April 8, 2021

Description

We recently developed a method to predict protein-protein contacts through correlated mutation analysis on a pair of multiple sequence alignments of homologs of interacting proteins (Ouroboros; Correa Marrero, 2018). Our method iterates between (i) weighting proteins according to how likely they are to interact based on the correlated mutations signal, and (ii) predicting correlated mutations based on the weighted sequence alignment. This approach accurately discriminates between protein interaction versus non-interaction and simultaneously improves the prediction of intermolecular contact residues compared to a naive application of correlated mutation analysis. We aim to widen the scope of the algorithm by i) investigating its applicability at much broader scale; ii) analyze the impact of various sources of noise or bias, e.g. from phylogeny; and iii) develop strategies to improve the algorithm. The first step will be performed by analyzing data from two recent publications (Perlaza-Jiménez, 2018; Cong 2019) in which a genome-wide scan for correlated mutations predicted putative protein-protein interactions (in in Arabidopsis thaliana and in E.coli, resp.). Homologs for these proteins will be obtained across the plant kingdom and Ouroboros' performance on these data will be investigated. Potential biases that will be investigated include the effect of biased sampling of sequences across the species phylogeny. In addition, for interactions where the two proteins are related to each other (members of the same protein family) previous preliminary analyses indicated specific problems with our approach, which so far does not yet take into account similarity between the interacting proteins. Solutions to this problem will be further investigated, which will involve modifications of the current algorithm. Another modification might be that currently, if a protein in one of the two multiple alignments occurs multiple times (with different putative interaction partner), this is not taken into account properly. The probability of a given position in such protein should be a (joint) function of all putative interacting partners in the other alignment. This can be implemented in a straightforward way using a weighted sum of the one-hot-encoded features for the different putative interaction partners.

References

1. Correa Marrero et al, *Bioinformatics, Volume 35, Issue 12, June 2019, Pages 2036–2042*. "Improved inference of intermolecular contacts through protein–protein interaction prediction using coevolutionary analysis"
2. Perlaza-Jiménez and Walther, 2018, Nucleic Acids Research, Volume 46, Issue 16, 19 September 2018, Pages 8114–8132, A genome-wide scan for correlated mutations detects macromolecular and chromatin interactions in *Arabidopsis thaliana*
3. Cong, Anishchenko, Ovchinnikov and Baker, 2019, Science 365, 185-189. Protein interaction networks revealed by proteome coevolution.

A phylogenetic framework for linking genes to molecules in large-scale genomic/metabolomic datasets

Supervisors	Marnix Medema, Justin van der Hooft
Type	Genome analysis, Computational genomics, Phylogenetics
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Genomics, programming, chemical biology
Timestamp	August 29, 2019

Description

Specialized metabolites produced by microbes, fungi, and plants are used for various applications like antibiotics and anticancer drugs. The genes encoding the enzyme ensembles that produce those specialized molecules (a.k.a. natural products) are often physically clustered together in biosynthetic gene clusters (BGCs). Algorithms have been developed that can predict the presence of such BGCs in whole genome sequences [1]. These tools yield large numbers of putative BGCs with some known but predominantly yet unknown molecular products - of which the structural prediction remains difficult. Experimental validation of links between BGCs and the products which they encode has also led to the discovery of several subclusters: modules of co-evolving genes that are responsible for the production of specific molecular substructures across otherwise structurally diverse molecules [2].

Several methods have been devised that make use of large genomic and metabolomic datasets to identify, by means of correlation, which BGCs are responsible for the production of which molecules, either by correlating absence/presence patterns of full BGCs across strains with those of full molecules [3, 4] or by correlating absence/presence patterns of subclusters with substructures. However, these approaches suffer from the problem that many shared absence/presence patterns are due to the phylogenetic relatedness of the underlying strains (co-inheritance) rather than independent parallel acquisition of the same BGC. Hence, a species-phylogenetic framework is required to correct for (or at least interpret/adjust) such correlation metrics.

The main goal of this project is to develop an improved, phylogenetically aware, way of identifying how ‘special’ the co-occurrence of molecules and BGCs across strains is, in order to judge their statistical significance. Given the large datasets that are already available to apply this on, the expectation is that this will make it possible to link BGCs to molecules in high throughput and generate a concrete reduced list of candidates coming from correlation analyses that can be tested in the laboratory by collaborators. Thus, this will accelerate the discovery of natural products such as antibiotics and anticancer agents.

Key references

1. Blin et al., Nucleic Acids Research, 2017. 45(W1): p. W36-W41.
2. Del Carratore et al., Communications Biology, 2019, 2: 83.
3. Doroghazi et al., Nature Chemical Biology, 2014, 10: 963-968.
4. Navarro-Muñoz et al., Nature Chemical Biology, in press.

Pangenomic QTL analysis

Supervisor	Sandra Smit, Aalt-Jan van Dijk
Type	Algorithm development
Requirements	Advanced Bioinformatics, Algorithms in Bioinformatics
Skills	Programming, Genomics, Algorithm development
Timestamp	May 25, 2020

Description

For a trait of interest, such as flowering time or yield, researchers or plant breeders often want to be able to indicate the most likely causal gene(s). Therefore they often analyze Quantitative Trait Loci (QTLs), which describe associations between genome regions and traits. However these QTL regions on a genome may contain hundreds of genes. Comparing genomes of parental species used to generate the QTL data can be helpful in order to indicate likely causal genes [1]. Further strategies for gene prioritization are under development [2,3,4], e.g. using enrichment of certain gene annotations in these regions, or using machine learning.

In this project, the idea is to add further power to these prioritization methods by applying them in the context of multiple annotated genomes. This is in particular relevant because QTL data by design is obtained from individuals with different genomes. We will use so called “pangenomes” for this. A pangenome is a data structure, which contains multiple annotated genomes and facilitates comparative analyses across these genomes [5]. The data structure is stored in a Neo4j graph database, which can incorporate heterogeneous data types, e.g. QTL annotations. Recently, we implemented QTL-functionalities in a pangenome datastructure, which now allows to annotate QTL reginos in a pangenome and to extract homologous regions and annotations from multiple assemblies. The functionality was tested on small sets of Arabidopsis and rice QTL data.

To follow up on this, the following activities will be performed in this project:

1. Incorporate GO terms in the pangenome to facilitate GO-centric navigation.
2. Develop a method to incorporate genes and annotations from multiple genomes in gene prioritization
3. Apply the pangenome-QTL-prioritization approach to available QTL datasets in rice, Arabidopsis and tomato.

References

[1] Lim et al. (2014) Quantitative Trait Locus Mapping and Candidate Gene Analysis for Plant Architecture Traits Using Whole Genome Re-Sequencing in Rice. Mol Cells. 37(2): 149–160

[2] Kourmpetis et al. (2010) Bayesian Markov Random Field analysis for protein function prediction based on network data. *PLoS One* 5(2):e9293.

[3] Bargsten et al. (2014) *BMC Plant Biology* 14:330.

[4] Lin et al. (2019) *G3: Genes|Genomes|Genetics 3129-3138.*

[5] Sheikhizadeh S, Schranz ME, Akdel M, de Ridder D, Smit S (2016) PanTools: representation, storage and exploration of pan-genomic data. Bioinformatics 32(17):i487-i49.

Integrative QTL analysis

Supervisors	Harm Nijveen, Dick de Ridder
Type	Algorithm development
Requirements	Adv. Bioinformatics, Adv. Statistics / Modern Statistics for the Life Sciences
Skills	Genomics, Programming, Statistics, Machine learning
Timestamp	September 17, 2019

Description

In a recent national study on maternal effects on seed quality, 165 homozygous recombinant lines of *Arabidopsis thaliana* grouped in a number of different growth conditions were genotyped based on 1059 markers and transcript levels were measured. These lines were also extensively phenotyped, with the goal of performing generalized genetical genomics [1] – correlating genotype with phenotype (expression) under a range of conditions. Levels of a number of primary metabolites were measured as well.

In this project, the goal is to develop methods to learn which genes influence which genotype, extending the QTL approach by incorporating expression and metabolic pathway information [2]. Prior knowledge on metabolic regulation and the relation between condition and metabolic activation can be used to refine the search and zoom in on possible mechanistic explanations of the observed phenotypes. The desired outcome is a method to optimally combine genetical genomics data with prior knowledge.

References

[1] Y. Li *et al*. (2008) Generalizing genetical genomics: getting added value from environmental perturbation. *Trends Genetics* 24(10):518-24.

[2] R.C. Jansen *et al.* (2009) Defining gene and QTL networks. *Current Opinion in Plant Biology* 2009, 12:1–6.

[3] Nijveen, H. *et al.* (2017) AraQTL – workbench and archive for systems genetics in *Arabidopsis thaliana*. Plant J, 89: 1225–1235. doi:10.1111/tpj.13457

Novel enzymes for fragrance and flavour

Supervisors	Aalt-Jan van Dijk, Marnix Medema
Type	Algorithm development
Requirements	Programming in Python, Machine Learning
Skills	Programming, Statistics, Chemical biology, Data analysis
Timestamp	September 4, 2019

Description

Ingredients from plants used in the flavour and fragrance industry are increasingly produced by microbial production platforms. Terpenes (a class of >10,000 natural compounds) are prime examples of such plant flavour compounds, used in a wide range of products. Microbial platforms for plant compounds often work by expression of the plant biochemical pathway in the microorganism, upon which the microorganism will produce the plant metabolite. The production of terpenes is largely mediated by a single class of enzymes, the terpene synthases. These synthases are often limiting for production, and differ greatly in their efficiency. To improve microbial production platforms, it is imperative to identify superior plant terpene synthases. In this project, you will apply machine learning, to recognize synthases with a specific product, e.g. valencene or patchoulol, among thousands of uncharacterized terpene synthases in daily expanding plant genomics data. We will focus on one particular class of terpene synthases, producing sesquiterpenes. Sesquiterpenes can be categorized based on their cyclization pattern. The terpene synthase reaction can comprise one or a few of a set of 13 reactions, all catalyzed by a single enzyme. A particular combination of these reactions results in a specific cyclization pattern of the final product.

Machine learning includes methods that convert data (here: terpene synthase sequences) into numerical representations (“features”) and find patterns in these features that distinguish various “classes” (here: types of terpene products produced). Training data consisting of sequences with known product specificity is available, and additional cases will be obtained from literature and databases. Features can be derived by e.g. counting sub-strings in a sequence, or by analysing conservation of amino acids in an alignment. By associating particular patterns in these features with product specificity, the algorithm can learn how to recognize different classes of proteins. Specifically, support vector machines (SVMs) will be trained to predict absence or presence of each of the 13 different reactions as labels, each time for the entire set of enzymes with known product specificity. After training and estimation of classification performance using cross-validation, the SVMs will be applied to predict functionality for all available sesquiterpene synthase sequences. This will allow to prioritize sequences as candidate enzymes for particular compounds of interest, and will allow targeted exploration of the rich biosynthetic diversity encoded in plant genomes.

References

1. Medema & Osbourn (2016) Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. Natural Product Rep. 33: 951.
2. Röttig M, Rausch C, Kohlbacher O (2010) Combining Structure and Sequence Information Allows Automated Prediction of Substrate Specificities within Enzyme Families. PLoS Comput Biol 6(1): e1000636. doi: 10.1371/journal.pcbi.1000636

Connecting transcription factor - DNA interaction to flowering time regulation

Supervisors	Aalt-Jan van Dijk
Type	Data analysis
Requirements	Machine Learning, Programming in Python, Advanced Bioinformatics
Skills	Programming, statistics
Timestamp	April 8, 2021

Description

Regulation of flowering time requires highly specific regulation of gene expression. In *Arabidopsis thaliana* some of the most prominent transcription factors involved in this regulation belong to the MADS-box transcription factor family. The canonical DNA binding motif for this transcription factor family is the CArG-box, which has the consensus $CC(A/T)_6GG$. The availability of genome-wide binding maps for transcription factors based on ChIP-seq allows to investigate binding patterns of MADS-box proteins in a lot of detail. Recently, we re-analyzed eight ChIP-seq datasets of MADS-box proteins. The preferred DNA binding motif of each protein was found to be a CArG-box with the 3' extension 5'-NAA-3' [1]. Furthermore, motifs of other transcription factors were found in the binding sites of the MADS-box transcription factors, suggesting that interaction of MADS-box proteins with other transcription factors is important for target gene regulation.

In this project we will connect knowledge on MADS binding to putative phenotypic effects of this binding (i.e. changes in flowering time). We will do so by integrating phenotypic information available from for the sequenced Arabidopsis 1001 genome project [2] with the above mentioned ChIP-seq data. We will focus on a set of ~200 genes known to be involved in later stages of flowering time regulation, and analyze sequence variation in the ChIP-seq peaks for the MADS domain transcription factors in promoter regions of these ~200 genes. This data will be used in a machine learning approach (classification or regression, depending on how the flowering time data will be encoded) which will learn to predict flowering time based on the absence or presence of specific sequence variation in the binding sites of the MADS domain proteins. Importantly, this will allow to make mechanistic predictions on how MADS transcription factor binding is involved in flowering time regulation, and how evolutionary changes in this binding result in changes in flowering time.

References

[1] https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-018-1348-8

[2] https://arapheno.1001genomes.org/phenotype/262/

Finding genes related to regeneration

Supervisors	Harm Nijveen, Renze Heidstra
Type	Data analysis
Requirements	Advanced Bioinformatics, Adv. statistics/Modern Statistics for the Life Sciences
Skills	Genomics, Programming, Statistics, Machine learning
Timestamp	October 17, 2019

Description

Introduction of transgenes in plants has been a force in molecular biology and biotechnology for decades. *Agrobacterium tumefaciens* is generally used as a vehicle to introduce genetic material in the plant cell. Unfortunately, not all plants (particularly agronomically important crops) have the ability to regenerate a complete plant from a single (transgenic) cell.

Therefore, identification and knowledge on the molecular factors involved in the regeneration process is required[1]. To gain more information on the regeneration process, an RNA-sequencing experiment was conducted using *Arabidopsis thaliana* regenerating tissue at multiple time-points up to the fully regenerated shoot.

In this project the goal is to uncover candidates whose function is currently not associated with the regeneration process, as well as characterize the differential expression of genes suspected to be involved in regeneration through time. This requires the analysis of the RNAseq dataset individually, comparing expression data from different time-points, and to existing datasets. Genes/transcripts can be clustered based on co-expression [2,3] measures to find genes that show expression patterns similar to regeneration related genes. Clustered genes can then be analysed for enriched Gene Ontology annotations or common transcriptional regulators[4] to learn more about the underlying regulatory mechanisms.

References

1. Radhakrishnan D, Kareem A, Durgaprasad K, Sreeraj E, Sugimoto K, Prasad K: Shoot regeneration: a journey from acquisition of competence to completion. *Current Opinion in Plant Biology* 2018, 41:23-31.
2. Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008, 9:559.
3. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W: Learning from Co-expression Networks: Possibilities and Challenges. *Frontiers in Plant Science* 2016, 7:444.
4. Kulkarni SR, Vaneechoutte D, Van de Velde J, Vandepoele K: TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Research* 2018, 46:e31-e31.

Prediction in pangenome graphs

Supervisors	Dick de Ridder
Type	Algorithm development
Requirements	Machine learning, Algorithms in Bioinformatics
Skills	Programming, Machine learning
Timestamp	April 9, 2021

Description

In recent years, the number of genomes has grown rapidly. Many species are no longer represented by a single reference genome but by numerous related genomes. To capitalize on the genomic diversity in large collections of genomes, we need to transition from a reference-centric approach to a pangenome approach. Computational pangenomics is currently a hot topic and a challenging field of research [1]. We have developed a pangenome solution called PanTools which compresses multiple annotated sequences into a single graph representation, constructed, stored, and annotated in a Neo4j graph database [2].

There are numerous applications for such a pangenome representation at different levels, containing sequences, variants, genes, expression levels, orthologous groups, functional annotations, phenotypes etc. of a set of genomes. Whereas traditional studies mostly relate variation at one individual level to a phenotype, for example SNPs in GWAS or expression differences in transcriptomics studies, a particularly intriguing application of a pangenome would be to combine various levels of annotation to predict such a phenotype. This could help learn whether nonlinear combinations of annotations predict a phenotype; for example, whether a certain SNP in a gene with a certain function is predictive of the phenotype only in genomes in which it is expressed. A long term goal of such an approach would be to use machine learning to infer novel relations and add these as computationally derived edges to the graph.

In this project, we will further develop a prototype machine learning pipeline based on PanTools. A basic framework is available to extract sets of relevant features from a pangenome graph, to predict phenotypes and functions in a machine learning setting. Neo4j offers Cypher, a query language to formulate complex graph structure and property queries; extracted features are then converted to training/test sets for a machine learning application in Python or R. This approach has been developed and validated on a set of bacterial genomes. Open questions are what gene function prediction performance is possible on larger, less well annotated genomes (plants), what computational bottlenecks arise in queries (and whether the underlying datastructure limits this), how predicted annotations should best be added to the pangenome graph and whether prediction “on the fly” when querying can be implemented.

References

[1] Computational Pan-Genomics Consortium (2018) Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics* 19(1):118-135.

[2] Sheikhizadeh S, Schranz ME, Akdel M, de Ridder D, Smit S (2016) PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics* 32(17):i487-i494.

The INSIDE study: can probiotics improve quality of life of ileostoma patients?

Supervisors	Edoardo Zaccaria, Peter van Baarlen (HMI), Dick de Ridder (BIF)
Type	Data analysis
Requirements	Machine Learning, Advanced Statistics
Skills	Python/R
Timestamp	April 9, 2021

16 ileostoma volunteers were recruited to participate in a three-arm cross-over, placebo-controlled intervention study. The interventions are two probiotic products and a placebo product, administered in random order. The intervention periods lasted 2 weeks, with 2 weeks of wash-out between each intervention. The study was preceded and ended with 2 weeks on run-in/run-out periods. At the first and last day of the intervention period the volunteers were invited at the hospital, for a total of 6 visit/participant. Two sampling schemes have been adopted:

- 27 small size ileostomy effluent have been collected per each participant. 3 during the run in, 6 during each intervention, 4 during the wash-out and 2 during the run-out period. Samples were prepared for microbiota composition analysis (16S amplicon sequencing).
- During 6 visits the following samples have been taken from each individual:
    1. urine for metabolites analysis
    2. blood samples for PBMCs transcriptome analysis, biomarkers profiling, metabolome analysis and immunoresponsiviness (cytokine profiling after *ex vivo* stimulus, 6 stimuli + 1 control)
    3. ileostomy samples for metatranscriptome analysis and SCFA composition

The following datasets are available for analysis:

- Microbial 16S amplicon sequencing
- Metabolites in blood
- Metabolites in urine
- Analytes/biomarkers in blood
- Immunoresponsiveness of subjects
- Transcriptomes of PBMCs
- Metatranscriptome

The goal of this project will be to integrate and mine this data using unsupervised and supervised machine learning techniques. Particular interest will be in visual exploration (PCA, MDS, t-SNE, clustering etc.) and in nonlinear feature extraction using classification algorithms.

[1] M. Kleerebezem, S. Binda, P.A. Bron, G. Gross, C. Hill, J.E. van Hylckama Vlieg, S. Lebeer, R. Satokari and A.C. Ouwehand. Understanding mode of action can drive the translational pipeline towards more reliable health benefits for probiotics. *Curr Opin Biotechnol.* 56:55-60, 2018.

High throughput single cell RNA sequencing to study plant root development

Supervisors	Rens Holmer, Harm Nijveen, Dick de Ridder
Type	Data analysis
Requirements	Advanced Bioinformatics, Machine Learning
Skills	Programming, RNA-seq analysis
Timestamp	October 17, 2019

Description

Measuring RNA abundance is one of the most important tools for studying transcriptional regulation in eukaryotic development. However, since the fundamental unit of transcriptional regulation is a single cell, bulk RNA sequencing methods are of limited use. For this precise reason, developmental research in human and animal systems has shifted towards measuring RNA abundance in single cells. Technological advancements in recent years make it possible to sequence full transcriptomes for thousands to millions of cells in high throughput, uncovering many previously unknown aspects of transcriptional regulation and cell type differentiation. However, whereas isolating individual cells from animal tissues is relatively straight forward, this is not the case for plant tissues, mainly due to the rigid plant cell wall.

There are two main approaches to isolate RNA from individual cells in plants. The first approach is to dissolve the plant cell wall with enzymes, producing protoplasts. This is a technically challenging and labor intensive process that only works for certain tissues. However, several studies have shown that this is at least feasible in *Arabidopsis thaliana* roots. The second approach is to only isolate nuclei instead of full cells, and to sequence the RNA present in these nuclei. This should be a straightforward process that could work for any plant tissue. We have recently generated such single nucleus RNA-seq datasets from several *Medicago truncatula* tissues to investigate the feasibility of this approach.

The goal of this project is to investigate the feasibility of single nucleus RNA sequencing in plants. To this end you will work with newly generated data from *Medicago truncatula* and publicly available data from *Arabidopsis thaliana* [1–3]. Important questions are if it is possible to identify cell types, and specific marker genes for these cell types. A strong emphasis will lie on identifying whether existing methods developed in animal systems are applicable for plant datasets. Single cell RNA sequencing data is both high dimensional and sparse in nature, so you will need some affinity with data normalization and dimensionality reduction techniques .[4]

References

1. Shulse, C. N. *et al.* High-throughput single-cell transcriptome profiling of plant cell types. *bioRxiv* 402966 (2018). doi:10.1101/402966
2. Ryu, K. H., Huang, L., Kang, H. M. & Schiefelbein, J. Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells. *Plant Physiol.* **179**, 1444–1456 (2019).
3. Zhang, T.-Q., Xu, Z.-G., Shang, G.-D. & Wang, J.-W. A Single-Cell RNA Sequencing Profiles the Developmental Landscape of Arabidopsis Root. *Mol. Plant* **12**, 648–660 (2019).
4. Laehnemann, D. *et al. 12 Grand Challenges in Single-Cell Data Science*. (PeerJ Preprints, 2019). doi:10.7287/peerj.preprints.27885v3

Timing of transcriptional regulation during early root nodule organogenesis

Supervisors	Rens Holmer, Harm Nijveen
Type	Data analysis
Required	Advanced Bioinformatics, Machine Learning
Skills	Programming, RNA-seq analysis
Timestamp	October 17, 2019

Description

Root nodules are specialized plant organs housing symbiotic rhizobium bacteria, most commonly found in legumes [1]. Once inside the nodule, the rhizobium bacteria start fixing environmental nitrogen, effectively providing fertilizer to the plant. To better understand how a plant can intracellularly house symbiotic bacteria, we are interested in how a novel organ can be formed upon perception of the bacteria. Previous work in the laboratory of molecular biology has focused on describing the exact timing of morphological changes during nodule organogenesis in the model legume *Medicago truncatula* [2]. Specifically, 24hours after perception of the bacteria cells in the pericycle cell layer start dividing, initiating the first visible steps of organogenesis.

We are now looking to get a better idea of the early transcriptional changes preceding the first cell divisions leading to a functioning nodule. For this, we have generated time series RNA-seq data of the first eight hours after bacterial contact. Additionally, to get a more detailed view into the spatial element of transcriptional regulation, we have subdivided the root into five sections of 5mm. Combined with control and treatment groups, the dataset consists of 82 transcriptome sequencing samples.

There are several approaches that can be explored during this thesis. The main challenges lie in developing a strategy to statistically test for differential expression between treated and untreated roots in a time series experiment [3–5]. Another option is to identify groups of genes that have a similar expression pattern in all samples, i.e. to look for coexpression [6]. A key part of the project will be studying expression profiles of specific genes for which we know they are crucial for forming a nodule, but don't know how they are transcriptionally regulated. Finally, we are currently measuring plant hormone concentrations in the same samples for which we sequenced RNA, opening up some possibilities for data integration approaches.

References

1. Oldroyd, G. E. D. Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants. *Nat. Rev. Microbiol.* **11**, 252 (2013).
2. Xiao, T. T. *et al.* Fate map of Medicago truncatula root nodules. *Development* **141**, 3517–3528 (2014).
3. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687 (2017).
4. Stegle, O. *et al.* A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *J. Comput. Biol.* **17**, 355–367 (2010).
5. Heinonen, M. *et al.* Detecting time periods of differential gene expression using Gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics* **31**, 728–735 (2015).
6. Serin, E. A. R., Nijveen, H., Hilhorst, H. W. M. & Ligterink W. Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.* (2016).

Protein interactions involved in SOBIR1/BAK1-mediated plant immunity

Supervisors	Wen Huang, Matthieu Joosten (Phytopath.), Aalt-Jan van Dijk (Bioinf.)
Type	Data analysis
Requirements	Advanced Bioinformatics, Machine Learning
Skills	Programming, data analysis
Timestamp	December 2020

Description

Plants are challenged by a plethora of agents causing biotic stress. Their first layer of defense is mediated by pattern recognition receptors (PRRs) that localize on the plasma membrane (PM). So far, all known plant PRRs that carry an extracellular receptor domain consisting of leucine-rich repeats (LRRs) are either receptor-like kinases (RLKs) or receptor-like proteins (RLPs). They share the same overall structure; however, in contrast to RLKs, RLPs lack a cytoplasmic domain for downstream signaling [1]. RLPs, such as the tomato PRR Cf-4 [2] that mediates resistance against strains of the pathogenic extracellular fungus *Cladosporium fulvum* secreting the matching avirulence factor Avr4, constitutively interact with the RLK SUPPRESSOR OF BIR1-1 (SOBIR1). SOBIR1 is essential for Cf-4 accumulation and function [3], and upon recognition of Avr4 by Cf-4, the RLK BRI1-ASSOCIATED KINASE 1 (BAK1), which is a regulatory co-receptor involved in development and defense, is recruited by the activated Cf-4/SOBIR1 complex [4]. BAK1 recruitment to the activated RLP/SOBIR1 complex appears to be a general process. It has been proposed that subsequent trans-phosphorylation events between the kinase domains of SOBIR1 and BAK1 eventually initiate downstream defense signaling [5]. However, little is known about how the RLP/SOBIR1/BAK1 complex functions at the level of complex formation and downstream signal initiation.

In this project, you will contribute to the elucidation of PRRs and receptor-like cytoplasmic kinases that are essential in SOBIR1/BAK1-mediated immunity in tomato and other crop plants. In particular, you will further develop and apply a coevolution-based interaction prediction approach [6]. This method uses sequence alignments of putatively interacting proteins and finds patterns of intermolecular correlated mutations. Using those patterns, the method simultaneously predicts protein interactions and intermolecular contact residues. Coevolution analysis itself does not require training data on which proteins are interacting or which residues are involved. However, by combining the method with such information, we will improve its prediction performance. Structural information on RLPs/RLKs will be used to predict which residues are potentially involved in the interaction, using e.g. interface prediction or protein docking [7]. Available protein interaction data include an all vs. all screen of the extracellular network of Arabidopsis LRR-RLKs [8], containing over 500 interactions. In this project, you will develop a machine learning (classification) approach, integrating co-evolution predictions with the available knowledge, in order to predict PRR/RLCK interactions and the residues involved.

References

[1] Couto D and Zipfel C. *Nat Rev Immunol* **16**:537-552 (2016).

[2] Thomas CM et al., *Plant Cell* **9**:2209-2224 (1997).

[3] Liebrand TWH et al., *Proc Natl Acad Sci USA* **110**:10010-10015 (2013).

[4] Postma J et al., *New Phytol* **210**:627-642 (2016).

[5] Van der Burgh AM et al., *Mol Plant Pathol* **20**:410-422 (2019).

[6] Correa Marrero et al., *Bioinformatics* **35**:2036–2042 (2019).

[7] Xue et al., *FEBS Letters* **23**: 3516 (2015).

[8] Smakowska-Luzan et al., *Nature* **553**:342-346 (2018).

Using protein structures and interactions to screen for mechanosensors

Supervisors	Eva Deinum (Biometris), Aalt-Jan van Dijk
Type	Data analysis
Requirements	Programming in Python, Advanced bioinformatics
Skills	Programming, statistics, structural biology
Timestamp	April 8, 2021

Description

Plant cells have an anisotropic structure, so that they expand more in one direction than another. This makes plants grow straight. The fibres of the cell wall, cellulose microfibrils, are deposited along cortical microtubules (MTs). This way, the cell wall inherits (a history of) the orientation of the MTs (Gutierrez). The cortical microtubules themselves can align along the stresses in the cell wall, thus assuring that the new cell wall material meets the local mechanical requirements (Colin 2020).

How this happens at the molecular level, no one knows. Two things are known: 1) there are many microtubule associated proteins (MAPs) that affect MT stability, either directly, or by recruiting other proteins; 2) the microtubule severing enzyme katanin is important for this coalignment with cell wall stress. Katanin can fully cut MTs, which can amplify or suppress MTs (Deinum 2017), but also take small "bites" that are subsequently "healed" with GTP-tubulin, leaving behind temporary “GTP-islands” that could increase MT stability. Perhaps, some MAPs can recognize slight deformations in the microtubule lattice, resulting in a differential regulation of microtubule stability depending on their orientation with respect to these stresses. If this difference is large enough, that could introduce an orientation bias (Saltini 2020).

In this project, your **aim** is: investigation of the microtubule binding domains of MAPs to identify potential molecular mediators that make microtubules coalign with wall stresses.

Approach:

1) Collect relevant plant protein domains and sequences, available information on the effect of mutations, available homologous structures, and available interaction data.
2) Use template-based protein structure modelling for the domains/proteins of interest for which a homologous structure is available. For cases for which no suitable template exists, we will explore the use of template free modelling.
3) Generate microtubule structures with e.g. seam defects and GTP/GDP differences.
4) Apply data-driven protein-protein docking (van Zundert 2016) using input from steps 1-3 in order to predict binding conformations and interfaces.
5) Look for and further analyze "weirdos": e.g., MAPs that are lattice binding, but fit better on GTP-tubulin; MAPs that have a preference for a slightly different spacing; affinity for the "seam" of the microtubule; or fitting other structural defects in the microtubule lattice.
6) If time allows, we will also explore the use of a sequence-based protein interaction prediction approach which makes use of co-evolution (Correa Marrero 2019).

References

Colin et al., PNAS, (2020) doi: 10.1073/pnas.2008895117
Correa Marrero et al., Bioinformatics, 35, 2036-2042 (2019).
Deinum et al., PNAS, 114, 6942-6947 (2017).
Gutierrez et al., Nat Cell Biol, 11, 797-806 (2009):
Saltini: and Mulder, Phys Rev E, 101, 052405 (2020)
G.C.P van Zundert et al., J. Mol. Biol., 428, 720-725 (2016).

Improving deep learning based genomic prediction models by incorporating prior knowledge

Supervisors	Farooq Mohammed, Aalt-Jan van Dijk, Dick de Ridder
Type	Algorithm development
Requirements	Advanced bioinformatics, Machine Learning/Deep Learning
Skills	Genomics, Programming, Statistics, Machine Learning
Timestamp	April 8, 2021

Description

Genomic prediction involves prediction of trait values, e.g. yield, based on genotype data. Single nucleotide polymorphisms (SNPs) are usually used to characterise genotypes. Conventionally, linear mixed effect models are widely used [1] but more recently deep learning (DL) techniques such as Convolutional Neural Network (CNN) [2] start to be applied. Deep learning has been effectively applied in various imaging applications; but its performance in genomic prediction is reported to be inferior to linear models and other machine learning methods like random forest and gradient boosting [3]. This is likely because of the limited number of phenotyped individuals typically available, compared to thousands of SNPs. Pook et al. [4] proposed a local CNN (LCNN) approach that defines convolutional layer filters for contiguous genomic regions. The authors reported some improvement in the prediction performance but could not improve much compared to existing methods. Incorporation of prior knowledge has been successfully employed to regularize ML methods to reduce dimensionality and increase model sparsity and generalization. Motivated by this fact, procedures have been developed where prior biological knowledge is employed to conduct feature selection prior to prediction task [5].

In this project, our **aim** is to develop a new method for genomic prediction based on LCNN with improved prediction performance, making use of publicly available prior biological knowledge.

To do so, you will upgrade the existing LCNN method [4] to apply knowledge guided weight optimization in the convolution layer. We can achieve this by using publicly available gene-gene interaction information from the STRING database and adding a regularization term to the objective function [6] to encourage the model to give similar weights to interacting genes. This will help the convolutional layer to learn sparse feature representations which it feeds to its subsequent fully connected layers; thereby, reducing the overfitting problem [7] and improving prediction performance.

References

[1] Meuwissen, T.H.E., B. Hayes, and M. Goddard, *Prediction of total genetic value using genome-wide dense marker maps.* Genetics, 2001. **157**(4): p. 1819-1829.

[2] Montesinos-López, O.A., et al., *A review of deep learning applications for genomic selection.* BMC Genomics, 2021. **22**(1): p. 19.

[3] Abdollahi-Arpanahi, R., D. Gianola, and F. Peñagaricano, *Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes.* Genetics Selection Evolution, 2020. **52**(1): p. 1-15.

[4] Pook, T., et al., *Using local convolutional neural networks for genomic prediction.* bioRxiv, 2020.

[5] Ma, J.Z., et al., *Using deep learning to model the hierarchical structure and function of a cell.* Nature Methods, 2018. **15**(4): p. 290-+.

[6] Andel, M. and F. Masri, *Sparse Omics-network Regularization to Increase Interpretability and Performance of SVM-based Predictive Models.* IEEE Explore, 2015.

[7] Kong, Y. and T. Yu, *A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification.* Scientific Reports, 2018. **8**(1): p. 16477.

Plant-specific basecalling of Oxford Nanopore data

Supervisors	Dick de Ridder, Richard Finkers
Type	Algorithm development
Requirements	Programming in Python, Machine Learning, Deep Learning
Skills	Programming, machine learning
Timestamp	April 9, 2021

Description

Nanopore technology provides a novel approach to DNA sequencing that yields long, label-free reads [1]. The first commercial implementation of this approach, the MinION, has shown promise in various sequencing applications: it produces (very) long reads, historically at a very high per-base error rate but recently of increasing quality. The raw output signal of a nanopore sequencer consists of traces of measured changes in current over time, which need to be translated into sequences (i.e. basecalled). Originally this step used hidden Markov models (HMMs), but over the past years a number of approaches based on deep learning have been proposed (e.g. [2-4]). Reviews have shown that the training set plays a crucial role: basecallers trained on taxon-specific data perform better than generic basecallers [2,5], likely due to differences in methylation.

In this project, the goal is to benchmark basecallers specifically on plant data and to explore the consequences of taxon-specific training. Specifically, the question is (1) what basecaller best lends itself to taxon-specific training, (2) what accuracy improvements can be obtained and (3) how accuracy improvement relates to the taxonomic level, i.e. what the optimal aggregation level is. To this end, a number of state-of-the-art basecallers will be selected and implemented to run on our GPU server. Experience in a high-level deep learning framework (PyTorch or TensorFlow) is recommended.

References

[1] De Lannoy C, de Ridder D, Risse J. The long reads ahead: de novo genome assembly using the MinION. *F1000 Research* 6:1083, 2017.

[2] Vereecke N, Bokma J, Haesebrouck F, Nauwynck H, Boyen F, Pardon B, Theuns S. High quality genome assemblies of *Mycoplasma bovis* using a taxon-specific Bonito basecaller for MinION and Flongle long-read nanopore sequencing. *BMC Bioinformatics* 21:517, 2020.

[3] Zhang Y, Akdemir A, Tremmel G, Imoto S, Miyano S, Shibuya T, Yamaguchi R. Nanopore basecalling from a perspective of instance segmentation. *BMC Bioinformatics* 21:136, 2020.

[4] Teng H, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *Gigascience* 7:1–9, 2019.

[5] Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* 20:129, 2019.

Pangenomics in tetraploids

Supervisor	Sandra Smit, Richard Finkers, Dick de Ridder
Type	Data analysis
Requirements	Advanced Bioinformatics, Programming in Python
Skills	Comparative genomics, databases
Timestamp	April 9, 2021

Comparison of genomes of multiple related individuals has led to the realization that there is a high degree of genomic variation (e.g., genes and regions present in one genome and not in the other) and that a single reference genome does not represent the genomic diversity of a species. This led to the concept of pangenomes (Bayer et al. 2020), data structures that aim to represent the full genomic diversity within a species. Pangenomes and pangenome tools have been mostly developed for diploid species, with some examples in allopolyploid crops (Montenegro et al. 2017); limited application has been observed in autopolyploid crops. In contrast to allopolyploid crops (where the genome basically behaves like a diploid), in autopolyploids, all copies of a chromosome can pair with all copies during meiosis, resulting in a much more complex variation landscape. The high similarity between the multiple haplotypes present combined with an often fragmented view of the genome (lack of long-range contiguity information) make the analysis of genomic variation in autopolyploid species hard.

In this project, the goal is to load recently available data of phased autotetraploid genomes into the PanTools pangenome platform (Sheikhizadeh et al. 2016) and query these graphs for core genes (genes present in all four alleles), shell genes (genes present in at least one allele, in all cultivars) or dispensable genes (genes present only in some cultivars). The aim is to test whether existing PanTools functionality suffices to interrogate pangenomes of autopolyploid species (both theoretically and in practice) and, if not, to improve or extend it. The goal is to obtain insight into current limitations of PanTools for (auto)polyploid crop species and recommendations for future improvement.

- Bayer, P. E., A. A. Golicz, A. Scheben, J. Batley, and D. Edwards, 2020 Plant pan-genomes are the new reference. Nat. Plants 6:.
- Montenegro, J. D., A. A. Golicz, P. E. Bayer, B. Hurgobin, H. T. Lee et al., 2017 The pangenome of hexaploid bread wheat. Plant J. 90: 1007–1013.
- Sheikhizadeh, S., M. E. Schranz, M. Akdel, D. de Ridder, and S. Smit, 2016 PanTools: representation, storage and exploration of pan-genomic data. Bioinformatics 32: i487–i493.

Sequencing-based high density marker development

Supervisor	Dick de Ridder, Richard Finkers
Type	Workflow development
Requirements	Advanced Bioinformatics, Machine Learning
Skills	Comparative genomics, programming
Timestamp	April 9, 2021

Next generation sequencing technologies, such as Illumina, have become cheap enough that large experimental plant breeding populations can now be screened via whole genome sequencing (WGS) approaches. As breeding populations have a degree of structure (e.g., the number of different alleles expected in the population is limited), individuals do not have to be sequenced at large depth, as intermediate positions are linked on the same chromosome and can be imputed. However, short read lengths hamper the detection of which markers share a genome and strategies using reference genomes can introduce noise (e.g., because of repetitive sequences in the genome). However, presence/absence signals of markers in the individuals can be used to cluster alleles into their respective chromosomes, constrained by knowledge on the parental genomes and the possible forms of recombination.

In this project, we will assess strategies to efficiently analyze WGS data of an autotetraploid breeding population. Such populations are particularly challenging as variants can occur on one or more chromosomes. We will investigate different strategies to call variants (e.g., GATK vs. K-mer-based approaches), and (advanced) clustering techniques to assign groups of variants to their respective homologous chromosomes. The optimal strategy should be implemented into an easy-to-use workflow to be used for, for example, interactive scaffolding of assemblies into pseudomolecules or performing genetic analyses such as QTL mapping.

Fold2vec: protein structure embedding for deep learning

Supervisor	Aalt-Jan van Dijk, Dick de Ridder
Type	Method development
Requirements	Advanced Bioinformatics, Deep Learning
Skills	machine learning, deep learning
Timestamp	April 8, 2021

The structure of proteins to a large extent determines their function and activity in the cell. Yet the use of protein structures in computational biological research has thus far been underdeveloped compared to the use of sequences, due to a lack of experimental structures and the challenges fof working with 3D data. Recently however, the number and quality of available protein structures has started to increase dramatically. This leads to a major need for novel methods to work with protein structures, particularly in deep learning, which has recently had a large impact in bioinformatics. An important aspect of the success of deep learning is the ability to work with pre-trained embeddings to tackle new problems, avoiding duplication of effort and wasteful computation. Such approaches have already been applied to DNA and protein sequences, allowing to capitalize on the large amounts of unlabelled data available.In this project, we will explore the development of a generic protein structure embedding method in the context of deep learning-based protein function prediction.

To do so, we will build on a framework recently developed in our group, in which shape-mers are used as descriptors of local structure. These shape-mers generalize the ubiquitous use of *k*-mers to represent protein sequences [1]. To represent protein structures based on these shape-mers, simple count vectors can serve as a baseline “bag of words”-like embedding in which no additional sequential or structural information is retained. To better capture such information, a fully connected autoencoder can be used. Finally, to take structural and sequential context of shape-mers into account, we will explore NLP-based approaches, ideally context-sensitive methods, e.g. those based on Transformers [2] such as BERT [3]. To train an embedding, a neural network will be required to reconstruct a representation after a pass through a low-dimensional bottleneck (autoencoder) or to predict masked information given a context (NLP-based approaches).

[1] Durairaj, Janani, Mehmet Akdel, Dick de Ridder, and Aalt D.J. van Dijk. 2020. “Geometricus represents protein structures as shape-mers derived from moment invariants”. *Bioinformatics,* accepted for publication; preprint available on *bioRxiv*, DOI 10.1101/2020.09.07.285569.

[2] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need”. *In: NIPS 2017, Proc. 31st Int. Conf. on Neural Inf. Proc. Syst.,* 6000–6010.

[3] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-training of deep bidirectional transformers for language understanding”. arXiv:1810.04805.

Deep learning on protein-DNA sequence pairs using the encoder-decoder

Supervisor	Aalt-Jan van Dijk, Dick de Ridder
Type	Method development
Requirements	Advanced Bioinformatics, Machine Learning, Deep Learning
Skills	machine learning, deep learning
Timestamp	April 8, 2021

Deep learning methods such as recurrent neural network-based sequence2sequence [1], or more recently attention-based Transformer [2], have led to major improvements in the field of natural language processing, e.g. in language translation or methodology for interpreting or generating natural language. A common characteristic of several of these approaches is that they adhere to the so-called encoder-decoder architecture. In this architecture, the encoder “encodes” a sequence, transmits information about it to the decoder; whereas the decoder, using the information it obtains from the encoder, predicts an output sequence. Recently, similar architectures have started to be applied to biological sequences. This involves for example the prediction of optimal codons for a protein sequence of interest, using a model which uses an LSTM recurrent neural network as part of an encoder-decoder [3].

In this project you will explore the use of an encoder-decoder architecture on pairs of protein-DNA sequences. This could in particular involve proteins and their promoter sequences. An encoder-decoder system should then learn to predict the best matching protein sequence for a given promoter, or vice versa, the best matching promoter for a given protein sequence. A key step in building this predictor will be to decide on what representation or embedding should be used to represent the input protein and promoter sequences. For proteins, we will explore the use of recently proposed Transformer-based embeddings [4]. For promoter sequences, a convolutional neural network-based approach might be a good option, in order to capture the presence of various short-sequence motif based biological signals present. Various CNN-based approaches have previously been developed for promoter sequences, e.g. [5], and we will explore their use in this project.

[1] Sutskever et al., Sequence to sequence learning with neural networks.

[2] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need”. *In: NIPS 2017, Proc. 31st Int. Conf. on Neural Inf. Proc. Syst.,* 6000–6010.

[3] Yang et al., “Generative models for codon prediction and optimization”; https://mlcb.github.io/mlcb2019_proceedings/papers/paper_29.pdf

[4] Rives et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, https://doi-org.ezproxy.library.wur.nl/10.1073/pnas.2016239118

[5] Kelley et al., “Sequential regulatory activity prediction across chromosomes with convolutional neural networks. “, https://genome-cshlp-org.ezproxy.library.wur.nl/content/early/2018/03/27/gr.227819.117.full.pdf

Deep learning with DNABERT to predict genomic properties

Supervisor	Aalt-Jan van Dijk, Dick de Ridder
Type	Method development
Requirements	Advanced Bioinformatics, Machine Learning, Deep Learning
Skills	machine learning, deep learning
Timestamp	April 8, 2021

The analysis of sequence data using neural networks has been revolutionized by the development of the attention mechanism, in particular in the context of the Transformer architecture [1]. A very exciting trend has been the development of Transformer-based models that are pre-trained on large amounts of data, which subsequently with relatively minor effort can be fine tuned to perform a range of specific tasks. A particular example of this paradigm is given by BERT [3] which has had a major impact in natural language tasks. Very recently, inspired by BERT, similar approaches have been developed on genome sequence data, in particular with the DNABERT approach [4]. This model showed good performance in various specific prediction tasks related to genome annotation (e.g. transcription factor binding site prediction), after minor fine-tuning for those tasks. In this project you will explore how to exploit the DNABERT model to predict additional genomic properties, in particular quantitative ones. This could include the prediction of gene expression, but also e.g. various epigenetic modifications, or evolutionary signals such as conservation or recombination. The available DNABERT model has been developed for the human genome. Our final goal is to apply a similar approach in plants, and in this project you will investigate if it is possible to indeed develop a similar approach on plant genomes. In case this turns out to be too computationally intensive, we will instead focus on the human genome.

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need”. *In: NIPS 2017, Proc. 31st Int. Conf. on Neural Inf. Proc. Syst.,* 6000–6010.

[2] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-training of deep bidirectional transformers for language understanding”. arXiv:1810.04805.

[3] Ji et al, DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, https://academic.oup.com/bioinformatics/advance-article-abstract/doi/10.1093/bioinformatics/btab083/6128680