Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Javier del Campo
Email	jdelcampo@ibe.upf-csic.es
Institution	Institut de Biologia Evolutiva (CSIC-UPF)
Website	delcampolab.com
Group	del Campo Lab. Microbial Ecology and Evolution

Project

Computational genomics

Project Title:

A Ribosomal Operon Reference Database

Keywords:

rrn, metabarcoding, eDNA, microbiome, biomonitoring

Summary:

The laboratory The del Campo Lab is based at the Institut de Biologia Evolutiva (UPF-CSIC) in Barcelona. The research at the del Campo Lab is focused on the study of host-associated microbes and the effect of global warming on the microbiomes of benthic and planktonic marine animals. We have a wet and dry lab, to perform experiments and bioinformatics analysis, enabling the broadest possible goals. The ongoing climate change and its effects on the environment, such as rising sea temperature, has strong impacts on free-living marine microbial communities. However, the effects of global warming have not been properly studied on host-associated microbiomes. Microbiomes (both prokaryotic and eukaryotic) associated with host organisms have a strong influence on host evolution, physiology, and ecological functions. We study how environmental changes resulting from global warming affect the composition and function of the microbiomes in key members of the marine fauna and consequently how these changes affect the hosts. Currently, our study focuses on these impacts on corals, teleost fish, and zooplankton. To tackle this novel research topic, we use a combination of molecular biology, ecophysiology, and bioinformatics. The proposed project Metabarcoding has been for many years a useful approach to study the diversity and distribution of micro and macroorganisms across environments. Furthermore, metabarcoding is currently being implemented successfully as a biomonitoring tool. This methodology is used for diagnosis of microbial pathogens, to study the health of lakes, rivers and beaches, to track the presence of invasive or endangered species, etc. However, the current metabarcoding methodologies present certain limitations, being the most significant the lack of phylogenetic resolution. The most popular metabarcoding approach is the use of short read barcodes generated using Illumina. These fragments, that are commonly not longer than 400 bp, despite providing very useful information cannot reach the level of detail that would allow us to infer from them species or strain identities (the latest in the case of microbes). We propose the use of the whole rRNA operon (rrn) as a barcode for life. Many fragments of the rrn such as the 16S and 23S in bacteria, the 18S, ITS1, ITS2 and 28S in eukaryotes, or fragments of them, are commonly used as barcodes. By using the rrn we are using a barcode that is many times longer than the current barcodes and that also includes many of them. So, it does not have only the advantage of providing more phylogenetic resolution but also allows to bring previous information generated using other rrn derived barcodes under the same phylogenetic and taxonomic framework. In order to stablish the rrn as a barcode the first thing we need to generate is a reference database. As we are just starting to generate now the first rrn amplicons using third generation sequencings (Nanopore, PacBio) we still do not have access to this type of data to generate such a reference database. However, genomes and metagenomes can be sources of rrn that can be used as references after placing them in a phylogenetic tree in order to assign them an identity. We propose to use extracted rrn from publicly available genomes and metagenomes and build a phylogenetically aware reference database using R and MySQL.

References:

Guillou, L. et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41, D597-604 del Campo, J. et al. (2018) EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. PLOS Biol. 16, e2005849 Jamy, M. et al. (2020) Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve

environmental diversity Molecular Ecology Resources 20, 429–443

Expected skills::

R, HMMER, Python, MySQL, phylogeny

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Luis A Pérez Jurado & Carlos Ruiz Arenas
Email	luis.perez@upf.edu / carlos.ruiza@upf.edu
Institution	UPF
Website	
Group	Genetics Unit

Project

Computational genomics

Project Title:

Combination of common and rare genetic variants to improve the diagnosis of complex disease

Keywords:

Human genetics Polygenic risk score Pathogenic variants Complex diseases

Summary:

Complex diseases, such as obesity or autism, are caused in most cases by a combination of environmental and genetic factors. Genetic factors can be classified as pathogenic or disease susceptibility variants, depending on the strength of their association with the phenotype. On one hand, pathogenic variants are genetic variants enough to cause the disease with high penetrance. These variants are normally ultra rare (<0.1%), preventing the application of association studies and usually requiring family studies and/or other functional studies for their validation. On the other hand, disease susceptibility variants increase the risk to suffer the disease but they are not enough by themselves to cause it, requiring additive effects of other concurrent genetic or environmental factors. These susceptibility variants tend to be relatively common in the population (from 0.1% to 50%) and association studies, comparing their frequency in cases and controls, can be used to define and quantify their relation to disease. The results of these association studies can be collapsed in Polygenic Risk Scores (PRS). PRS are a measure of the risk alleles for a disease carried by an individual. Thanks to the availability of big public datasets, PRS have improved their performance and can identify individuals with high susceptibility to disease (Khera et al 2018). In recent years, both approaches have been independently applied to study complex diseases. For instance, the genetic heritability of autism has been estimated to be 3-10% due to de novo rare variants, 3-10% to inherited rare variants and around 50% due to common variants (Alonso-Gonzalez et al, 2018). Despite this success, a significant proportion of heritability is still missing. We hypothesize that missing heritability is mainly due to rare variants with low penetrance, i.e. variants that are only pathogenic in a specific genetic background (Fahed et al, 2020). In this project, we propose to combine the analysis of common and ultra rare variants to improve our understanding of some common diseases. We propose three main tasks: - Compare PRS between controls and cases with and without high penetrant variants - Prioritize variants considering PRS - Propose new candidate genes We will use autism as an example of a complex disease and we will apply our methods to data from public repositories (dbGAP, SFARI, UK Biobank) and internal data. The applicant who will work in this project will learn to perform variant calling, prioritize genetic variants, define and compute polygenic risk scores (PRS), work with public data, develop analysis pipelines and work with software containers.

References:

Alonso-Gonzalez A, Rodriguez-Fontenla C, Carracedo A. De novo Mutations (DNMs) in Autism Spectrum Disorder (ASD): Pathway and Network Analysis. Front Genet. 2018 Sep 21;9:406. doi: 10.3389/fgene.2018.00406. PMID: 30298087; PMCID: PMC6160549. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL, Lai C, Brockman D, Philippakis A, Ellinor PT, Cassa CA, Lebo M, Ng K, Lander ES, Zhou AY, Kathiresan S, Khera AV. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun. 2020 Aug 20;11(1):3635. doi: 10.1038/s41467-020-17374-3. PMID: 32820175; PMCID: PMC7441381. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018 Sep;50(9):1219-1224. doi: 10.1038/s41588-018-0183-z. Epub 2018 Aug 13. PMID: 30104762; PMCID: PMC6128408. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, Samocha KE, Goldstein JI, Okbay A, Bybjerg-Grauholm J, Werge T, Hougaard DM, Taylor J; iPSYCH-Broad Autism Group; Psychiatric Genomics Consortium Autism Group, Skuse D, Devlin B, Anney R, Sanders SJ, Bishop S, Mortensen PB, Børglum AD, Smith GD, Daly MJ, Robinson EB. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet. 2017 Jul;49(7):978-985. doi: 10.1038/ng.3863. Epub 2017 May 15. PMID: 28504703; PMCID: PMC5552240.

Expected skills::

Good level of bash and R scripting and a good background in human genetics.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Comments:

Lab experiments to confirm the project results might be considered.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Lorenzo Pasquali
Email	lorenzo.pasquali@upf.edu
Institution	UPF
Website	https://www.endoregulatorygenomics.org/
Group	Endocrine Regulatory Genomics

Project

Computational genomics

Project Title:

Genetics and regulatory genomics of glucose metabolism diseases

Keywords:

Regulatory genomics, pancreatic islets, diabetes, chromatin, regulatory functions

Summary:

In the present project we will characterize the dynamics of tissue-specific cis-regulatory networks in tissues central to the glucose metabolism. The project will include the analysis and integration of chromatin data such as open chromatin profiles (ATAC-seq), histone modifications (ChIP-seq), 3D chromatin structure (4C-seq/Hiseq) and transcriptomic maps (RNA-seq), with the aim of identifying unexplored paths in the context of the molecular mechanisms that maintain tissue-specific functions and cell fate.

References:

Ramos-Rodríguez et al. DOI: 10.1038/s41588-019-0524-6 Eizirik el al. doi: 10.1038/s41574-020-0355-7

Expected skills::

High motivation, team work, knowledge of R, experience with Unix operating systems, basic knowledge of regulatory genomics, expertise in statistical analysis.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	David Comas
Email	david.comas@upf.edu
Institution	UPF
Website	https://www.biologiaevolutiva.org/dcomas/
Group	Human Genome Diversity

Project

Computational genomics

Project Title:

Analysis of the human genome diversity: unravelling demographic and genomic processes

Keywords:

Genome diversity, human populations, demography, adaptation

Summary:

The interests of our research are focused on the human genome diversity analysis in order to infer the (genomic and population) processes responsible for this diversity and try to establish the (population and epidemiological) consequences of the human genetic variability. Thus, our main research lines are focused on aspects of human genome diversity, population genetics, genome variation and disease susceptibility, and genome evolution and disease. 1. Population processes: Concerning population processes that have modeled the human genetic diversity, we have focused our research on the use of molecular tools to reconstruct the human population history through the phylogeny of genetic markers. Our interest has been focused on the genetic consequences at population level of human migrations and admixtures. The use of well-established phylogenies in the mitochondrial and Y-chromosome human genomes allowed us to unravel the population history of several populations. Nonetheless, we have recently used whole genome variation in the autosomes in order to establish the structure of human populations. 2. Genomic processes: Concerning genomic processes that have modeled the human genetic diversity, our research has have been focused on the relationship between human diversity and complex traits, including complex diseases. The genetic analysis in human populations of genes of biomedical interest might shed light on the evolution of these genes. In this context, we have focused our research in the analysis of genes that have been previously associated to complex diseases, such as psychiatric and immunological diseases. The analysis of these genes has allowed us to conclude that some of the failures in replicating genetic associations are due to extreme genetic differences between populations. In addition, we are also interested in other complex traits, such as height, not directly related to disease.

References:

1. Lorente-Galdos B, Lao O, Serra-Vidal G, Santpere G, Kuderna LFK, Arauna LR, Fadhlaoui-Zid K, Pimenoff VN, Soodyall H, Zalloua P, Marques-Bonet T, Comas D (2019) Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biology 20:77. 2. Font-Porterias, Arauna LR, Poveda A, Bianco E, Rebato E, Prata MJ, Calafell F, Comas D (2019) European Roma groups show complex West Eurasian admixture footprints and a common South Asian genetic origin. PLoS Genetics 15(9): e1008417. 3. Serra-Vidal G, Lucas-Sanchez M, Fadhlaoui-Zid K, Bekada A, Zalloua P, Comas D (2019) Heterogeneity in Palaeolithic population continuity and Neolithic expansion in North Africa. Current Biology 29:3953-3959. 4. Castro e Silva MA, Nunes K, Lemes RB, Mas-Sandoval A, Amorim CEG, Krieger JE, Mill JG, Salzano MS, Bortolini MC, da Costa Pereira A, Comas D, Hünemeier T (2020) Genomic insight into the origins and dispersal of the Brazilian coastal natives. Proceedings of the National Academy of Sciences USA 117 (5) 2372-2377. 5. Bianco E, Laval G, Font-Porterias N, García-Fernández C, Dobon B, Sabido-Vera R, Sukarova Stefanovska E, Kučinskas V, Makukh H, Pamjav H, Quintana-Murci L, Netea MG, Bertranpetit J, Calafell F, Comas D (2020) Recent Common Origin, Reduced Population Size, and Marked Admixture Have Shaped European Roma Genomes. Molecular Biology and Evolution 37(11):3175-3187.

Expected skills::

Computational skills to manage and analyze genotype and DNA sequence data from whole genomes

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Robert Castelo & Irene Madrigal
Email	robert.castelo@upf.edu
Institution	Universitat Pompeu Fabra / Hospital Clínic
Website	https://functionalgenomics.upf.edu
Group	Functional Genomics / Molecular Genetics

Project

Computational genomics

Project Title:

Deciphering genetics of hereditary hemorrhagic telangiectasia

Keywords:

WES, variant calling, variant filtering and interpretation

Summary:

Hereditary hemorrhagic telangiectasia (HTT) is an autosomal dominant vascular dysplasia leading to epistaxis, telangiectasia and visceral arteriovenous malformations. Pathogenic variants in ENG and ACVRL1 are the main genetic cause responsible of the disease. Historically, genetic testing for HTT consisted of the analysis of ENG and ACVRL1. Nowadays whole exome sequencing (WES) has been introduced as diagnostic tool in patients with this disease. WES allowed the identification of several pathogenic genetic variants; nevertheless the proportion of unresolved exomes is much higher than expected. Particularly in HTT, in which the clinical phenotype is very specific, WES did not reveal, in the studied genes (ACVRL1, ENG, EPHB4, GDF2, RASA1 and SMAD4) , any pathogenic variant in 75% of patients. We assume the existence of other responsible genes o genetic mechanisms in HTT. The main objective of the project is to develop new algorithms for WES analysis in order to detect new candidate genes for HTT. This project will be jointly supervised with Dr. Irene Madrigal from the Molecular Genetics department at the Hospital Clínic de Barcelona, who is in charge for finding a genetic diagnosis for these patients.

Expected skills::

basic knowledge of human genetics, programming and analysis of next generation sequencing data

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Javier del Campo
Email	jdelcampo@ibe.upf-csic.es
Institution	Institut de Biologia Evolutiva (CSIC-UPF)
Website	delcampolab.com
Group	del Campo Lab. Microbial Ecology and Evolution.

Project

Computational genomics

Project Title:

The Microeukaryotic Virome

Keywords:

Virome, Giant Viruses, Microeukaryotes, Viral Endogenization

Summary:

The laboratory The del Campo Lab is based at the Institut de Biologia Evolutiva (UPF-CSIC) in Barcelona. The research at the del Campo Lab is focused on the study of host-associated microbes and the effect of global warming on the microbiomes of benthic and planktonic marine animals. We have a wet and dry lab, to perform experiments and bioinformatics analysis, enabling the broadest possible goals. The ongoing climate change and its effects on the environment, such as rising sea temperature, has strong impacts on free-living marine microbial communities. However, the effects of global warming have not been properly studied on host-associated microbiomes. Microbiomes (both prokaryotic and eukaryotic) associated with host organisms have a strong influence on host evolution, physiology, and ecological functions. We study how environmental changes resulting from global warming affect the composition and function of the microbiomes in key members of the marine fauna and consequently how these changes affect the hosts. Currently, our study focuses on these impacts on corals, teleost fish, and zooplankton. To tackle this novel research topic, we use a combination of molecular biology, ecophysiology, and bioinformatics. The proposed project Virus have been reported as a significant component of the nuclear genomes of different microeukaryotes from algae to heterotrophic protists. These viruses have been proved to be relevant for different aspects of microeukaryotic biology, shaping the genome of their algal host or protecting the host from other viral infections. The number of viruses described from microeukaryotes is relatively low compared to those infecting bacteria or macroorganisms. The microeukaryotic virome is a source of novel viral diversity, particularly of giant viruses. The aim of this project in collaboration with Professor Richard A. White from the University of North Caroline Charlotte is to characterize the viral landscape of the unicellular eukaryotes. Initially we will build a comprehensive database of microeukaryotic genomes and transcriptomes. Using this dataset, we will proceed to extract the viral signal from the different organisms' genome and proceed to their characterization using phylogenetic trees. We expect in this project to unveil a significant amount of viral diversity. As a byproduct of it we will also generate a comprehensive microeukaryotic genomic database.

References:

Fischer, M. G. et al. (2016) Host genome integration and giant virus-induced reactivation of the virophage mavirus. Nature 540, 288–291 Moniruzzaman, M. et al. (2020) Widespread endogenization of giant viruses shapes genomes of green algae. Nature 588, 141-145

Expected skills::

R, Phyton, Genome Analysis, Phylogenies, Database Management

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Javier del Campo

Email	jdelcampo@ibe.upf-csic.es
Institution	Institut de Biologia Evolutiva (CSIC-UPF)
Website	delcampolab.com
Group	del Campo Lab. Microbial Ecology and Evolution

Project

Computational genomics

Project Title:

The genomic mechanisms of ichthyocarbonates precipitaion

Keywords:

fish, climate change, carbon cycle, genome, microbiome,

Summary:

The laboratory The del Campo Lab is based at the Institut de Biologia Evolutiva (UPF-CSIC) in Barcelona. The research at the del Campo Lab is focused on the study of host-associated microbes and the effect of global warming on the microbiomes of benthic and planktonic marine animals. We have a wet and dry lab, to perform experiments and bioinformatics analysis, enabling the broadest possible goals. The ongoing climate change and its effects on the environment, such as rising sea temperature, has strong impacts on free-living marine microbial communities. However, the effects of global warming have not been properly studied on host-associated microbiomes. Microbiomes (both prokaryotic and eukaryotic) associated with host organisms have a strong influence on host evolution, physiology, and ecological functions. We study how environmental changes resulting from global warming affect the composition and function of the microbiomes in key members of the marine fauna and consequently how these changes affect the hosts. Currently, our study focuses on these impacts on corals, teleost fish, and zooplankton. To tackle this novel research topic, we use a combination of molecular biology, ecophysiology, and bioinformatics. The project Calcium carbonate released by teleost fish in marine environments (AKA ichthyocarbonates) represents one of the main carbon sinks in the open ocean, so a mitigator of climate change. The ichthyocarbonate pellets released by fish have an impact on the global carbon cycles and based on the most recent predictions of temperature increase and acidification as a result of climate change its importance will increase in the future. The formation of ichthyocarbonates in the gut of the teleost fish is also a key mechanism for the fish survival because allows them to maintain their osmotic balance. However, despite its physiological importance and its role as an alternative carbon sequestration method, little is known about the genomic mechanisms involved in the precipitation of ichthyocarbonates. Using genomics and transcriptomics data from the Gulf Toadfish (Opsanus beta), a model organism for the study of osmoregulation, such genes involved in the calcium carbonate precipitation have not been found neither in any other fish genome as far as we know. Classically it has been thought that the responsible for the precipitation of calcium carbonate was the fish, but recently microorganisms have been reported on the surface of the ichthyocarbonates opening the door to the possibility that the fish microbiota might be playing a role in this process. So, it is possible that the genes directly involved in the precipitation of ichthyocarbonates are present in the microbiome. The aim of this project is to characterize the complete mechanism of ichthyocarbonates precipitation targeting at the same time the piscine host and its microbiome. We will compile a set of reference genomes of teleost fish and re-analyze them using alternatives approaches that would allow us to obtain a better assembly to minimize the loss of information and to assemble the genomes of the most abundant associated microbes using binning strategies on the “contaminant” fraction of the raw genomic data. We hope that using this strategy will allow us to reconstruct the complete carbonate precipitation pathway.

References:

Wilson, R. W. et al. 2009. Contribution of Fish to the Marine Inorganic Carbon Cycle Science 323, 359–362.

Expected skills::

R, Python, Genome Assembly and Annotation, Phylogeny, Binnig Strategies, Database Management

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Lorenzo Pasquali
Email	lorenzo.pasquali@upf.edu
Institution	UPF
Website	https://www.endoregulatorygenomics.org/
Group	Endocrine Regulatory Genomics

Project

Computational genomics

Project Title:

Genetics and regulatory genomics of glucose metabolism diseases

Keywords:

Regulatory genomics, pancreatic islets, diabetes, chromatin, regulatory functions

Summary:

In the present project we will characterize the dynamics of tissue-specific cis-regulatory networks in tissues central to the glucose metabolism. The project will include the analysis and integration of chromatin data such as open chromatin profiles (ATAC-seq), histone modifications (ChIP-seq), 3D chromatin structure (4C-seq/Hiseq) and transcriptomic maps (RNA-seq), with the aim of identifying unexplored paths in the context of the molecular mechanisms that maintain tissue-specific functions and cell fate.

References:

Ramos-Rodríguez et al. DOI: 10.1038/s41588-019-0524-6 Eizirik el al. doi: 10.1038/s41574-020-0355-7

Expected skills::

High motivation, team work, knowledge of R, experience with Unix operating systems, basic knowledge of regulatory genomics, expertise in statistical analysis.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Amelie Baud
Email	abaud@ebi.ac.uk
Institution	CRG
Website	https://www.crg.eu/en/programmes-groups/baud-lab
Group	Amelie Baud

Project

Computational genomics

Project Title:

Dissecting the genetic basis of handling-induced micturition in BXD recombinant inbred mice

Keywords:

Genotype to phenotype path; Complex traits genetics; Systems genetics; Animal models

Summary:

We have observed significant and strong differences in handling-induced micturition/urination between two inbred strains of mice, C57BL/6J and DBA2/J. We have collected phenotype data (micturition) on a large number of recombinant inbred strains derived from C57BL/6J and DBA2/J (BXD recombinant inbred mice), and a wealth of additional phenotypes as well as sequence data are available for these mice (http://www.genenetwork.org/). The project aims at dissecting the genetic basis of this phenotype, namely quantifying the proportion of phenotypic variation explained by genetics (heritability), mapping the underlying genomic loci (quantitative trait loci), and identifying phenotypes that are genetically correlated with micturition, in order to better understand what this phenotype represents (e.g. Is it a response to stress? Does it instead reflect morphological differences in the urinary system of the mice?).

References:

http://www.genenetwork.org/ (database and analysis toolkit to study BXD recombinant inbred mice); https://doi.org/10.1016/j.cels.2020.12.002; DOI 10.1007/978-1-4939-6427-7_4

Expected skills::

Experience programming in R would be a plus

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Donate Weghorn
Email	dweghorn@crg.eu
Institution	CRG
Website	http://weghornlab.net/
Group	Evolutionary Processes Modeling

Project

Computational genomics

Project Title:

Selection on cancer genomes exerted by the immune system

Keywords:

cancer genomics, immune evasion

Summary:

Cancer is a genetic disease, caused by DNA mutations that accumulate in cells of the human body over the course of time. One of the most important lines of defense against cancer is the immune system. Consequently, detectable cancer tumors must have been able to evade the body's immune surveillance. We expect this feature of successful tumors to leave a footprint of selection in the cancer genome. The aim of this project is to investigate differences in selection between cancer tumors that evolved under different strengths of the immune response. To this end, we will use somatic mutations detected in over 10,000 tumors as well as the tumors' gene expression data. The project has a bioinformatics, a statistical data analysis, and a population genetics component. The student will learn all the corresponding techniques and tools regarding data analysis, partly in collaboration with other lab members.

References:

https://www.nature.com/articles/ng.3987 https://www.nature.com/articles/s41588-020-0687-1

Expected skills::

Programming, logical-analytical thinking

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Josefa Gonzalez
Email	josefa.gonzalez@ibe.upf-csic.es
Institution	IBE (CSIC-UPF)
Website	gonzalezlab.eu
Group	Evolutionary and Functional Genomics

Project

Computational genomics

Project Title:

Discovering new targets for malaria vector control strategies in urban settings

Keywords:

malaria, Structural variants, adaptation, urbanization, transposable elements

Summary:

Malaria is a deadly disease that kills ~400.000 people per year mostly in Africa, but also in other worldwide regions. Urban environments were until recently considered to be unfit for Anopheles larvae development. However, during the last decades the two major African malaria vectors, Anopheles gambiae and An. coluzzii, have rapidly adapted to polluted habitats threatening current vector-control strategies. While genomic approaches have already been applied to develop vector-control strategies, so far they have focused on single nucleotide changes in coding regions applied to traits previously known to be relevant for the mosquito vector capacity, such as insecticide resistance. This project puts forward a new strategy based on the emergent field of urban adaptation to identify new genetic and epigenetic targets for malaria vector control. The project aims are (i) identifying all the genetic variants present in Anopheles genomes including SNPs, transposable elements, and copy number variants; (ii) identifying signatures of selection at the DNA level to pinpoint the most relevant genes for urban adaptation; and (iii) identifying the environmental factors more relevant for adaptation to urban environments. This project goes beyond the state-of-the-art by combining two emerging fields of research, urban adaptation and the functional role of structural variants, to tackle a relevant societal challenge that not only affects African countries, as re-emergence of malaria associated with climate change and increased human mobility is already being recorded in non-African countries.

Expected skills::

NGS data processing

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

An interview to further discuss the project is required before acceptance to the lab

---

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Arnau Sebe-Pedros

Email	arnau.sebe@crg.es
Institution	CRG
Website	https://www.sebepedroslab.org/
Group	Single-cell genomics and evolution

Project

Computational genomics

Project Title:

Investigating animal cell type diversity, evolution and regulation using single cell genomics and epigenomics approaches

Keywords:

Evolutionary biology; Single-cell genomics; Genome regulation; Animal phylogenetics; Comparative genomics

Summary:

Projects and specific tasks We are looking for students to join our team to work on a computational project involving integrative analysis of high-throughput single-cell genomics and chromatin data in different animals and unicellular relatives of animals. You will analyse single-cell datasets from different species and perform comparative genomics analyses. The goal is to reconstruct the evolutionary origin and diversification of animal cell types. We also have a second position to work on the development of a phylogenetics pipeline to infer genome-wide gene orthologies. You will learn about phylogenetics methods, protein alignment tools, and gene family evolution. The goal is to set-up a robust orthology framework to integrate single-cell atlases from diverse organisms; as well as to focus on the evolution of particular multi-gene families that are important for animal multicellularity and cell type differentiation (e.g. transcription factors). About the group Our group studies genome regulation from an evolutionary systems perspective. In particular, we are interested in deciphering the evolutionary dynamics of animal cell type programs and in reconstructing the emergence of genome regulatory mechanisms linked to cell type differentiation (from transcription factor binding through chromatin states to the physical architecture of the genome). To this end, we apply advanced single-cell genomics and chromatin experimental methods to molecularly dissect cell types and epigenomic landscapes in phylogenetically diverse organisms. We also develop computational tools to integrate these diverse data sources into models of cell type gene regulatory networks and we use phylogenetic methods to comparatively analyze these models. Our recent work has provided the first whole-organism cell type atlases in different species and mapped key regulatory genome features underlying these cellular programs (see Sebé-Pedrós 2018, Cell, Sebé-Pedrós 2018 NEE, Sebé-Pedrós 2016 Cell). By sampling additional species and chromatin features at single-cell resolution, we now aim at dissecting the evolution of cell types and their underlying gene regulatory networks.

References:

Check our website: https://www.sebepedroslab.org/

Expected skills::

We are seeking for creative and highly motivated students with an interest in evolutionary biology, genome regulation and/or comparative genomics. We are preferentially looking for dry/computational candidates, but there is also a possibility to work on dry+wetlab projects. Basic bioinformatics skills (command-line terminal, R/python scripting) are highly desirable, while ability to work in collaborative projects is a must. Possibility to continue with PhD after the master.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Roderic
Email	roderic.guigo@crg.cat
Institution	CRG
Website	https://genome.crg.cat/
Group	Bioinformatics and Genomics

Project

Computational genomics

Project Title:

: Efficient gene annotation across the entire phylogenetic spectrum

Keywords:

Bioinformatics, gene finding, transcriptomics,

Summary:

Understanding Earth's biodiversity and responsibly administrating its resources is among the top scientific and social challenges of this century. The Earth BioGenome Project (EBP) aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of 10 years (http://www.pnas.org/content/115/17/4325). The outcomes of the EBP will inform a broad range of major issues facing humankind, such as the impact of climate change on biodiversity, the conservation of endangered species and ecosystems, and the preservation and enhancement of ecosystem services. It will contribute to our understanding of biology, ecology and evolution, and will facilitate advances in agriculture, medicine and in the industries based on life: it will, among others, help to discover new medicinal resources for human health, enhance control of pandemics, to identify new genetic variants for improving agriculture, and to discover novel biomaterials and new energy sources, among others. The value of the genome sequence depends largely on the precised identification genes. The aim of the research project is to develop a gene annotation pipeline that produces high quality gene annotations that can be efficiently scaled to more than one million species. Our group has a long-standing interest in gene annotation. Roderic Guigo developed one of the first computational methods to predict genes in genomic sequences (geneid, Guigó et al, 1992), which has been widely used to annotate genomes during the past years. On the other hand, we are part of GENCODE, which aims to produce the reference annotation of the human genome. Within GENCODE we have developed experimental protocols to efficiently produced full-lengh RNA sequences. Our pipeline will be based on identifying the genes that can be precisely predicted computationally in a given species, subtract them from RNA samples, and produced high quality RNA sequences for the genes that are more difficult to annotate. The master student will work specifically on the identification of selenoprotein genes

Expected skills::

Good programming skills python, C, or similar. Good unerstandgin of molecular biology concets

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Andrés Ozaita
Email	andres.ozaita@upf.edu
Institution	UPF
Website	https://www.upf.edu/neurophar
Group	Laboratory of Neuropharmacology

Project

Computational systems biology

Project Title:

Intellectual disability can be caused by genetic mutations

Keywords:

intellectual disability, treatment, synaptic proteome, synaptic transcriptome, splicing

Summary:

Synaptic Intellectual disability may derive from specific genetic alterations, as found in neurodevelopmental disorders such as fragile X syndrome (FXS) and Down syndrome (DS), both disorders associated to relevant alterations in synaptic plasticity. Mouse models of these disorders mimicking the genetic alterations found in humans have demonstrated relevant tools to understand the physiopathology of the disorders and to test pharmacological approaches that may improve cognitive performance. In the lab we have described an approach to improve cognitive performance in models of FXS and DS, but the impact of these treatments in the biology of the synapse has not been addressed. We are now investigating, using high throughput proteomic and transcriptomic analysis of sorted synaptic contacts, the characteristics of pathological synapses, and the effects that pharmacological treatments have in improving synaptic plasticity in both models of intellectual disability.landmarks of intellectual disability

References:

Navarro-Romero A, Vázquez-Oliver A, Gomis-González M, Garzón-Montesinos C, Falcón-Moya R, Pastor A, Martín-García E, Pizarro N, Busquets-Garcia A, Revest JM, Piazza PV, Bosch F, Dierssen M, de la Torre R, Rodríguez-Moreno A, Maldonado R, Ozaita A. Cannabinoid type-1 receptor blockade restores neurological phenotypes in two models for Down syndrome. Neurobiol Dis. 2019 May;125:92-106. doi: 10.1016/j.nbd.2019.01.014. Epub 2019 Jan 25. PMID: 30685352. Busquets-Garcia A, Gomis-González M, Guegan T, Agustín-Pavón C, Pastor A, Mato S, Pérez-Samartín A, Matute C, de la Torre R, Dierssen M, Maldonado R, Ozaita A. Targeting the endocannabinoid system in the treatment of fragile X syndrome. Nat Med. 2013 May;19(5):603-7. doi: 10.1038/nm.3127. Epub 2013 Mar 31. PMID: 23542787.

Expected skills::

Bioinformatics

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Eva Maria Novoa
Email	eva.novoa@crg.eu
Institution	CRG
Website	https://www.crg.eu/en/programmes-groups/novoa-lab
Group	Epitranscriptomics and RNA Dynamics

Project

Computational systems biology

Project Title:

PREDICTING AND MONITORING CANCER USING MACHINE LEARNING ON NANOPORE SEQUENCING DATA

Keywords:

nanopore sequencing, RNA modifications, small RNA, machine learning, deep learning, cancer, prognosis, sample classification

Summary:

Dysregulation of small RNA abundances and their RNA modifications is a well-known feature in cancer cells, which leads to enhanced expression of specific oncogenic transcripts and proteins [1,2]. Despite the well-established association between small RNA dysregulation and cancer progression and malignancy, small RNA abundances and modifications are still not being used as screening, diagnostic or prognostic markers for cancer detection or progression, mainly due to the lack of a simple, unbiased and cost-effective method to quantify small RNA abundances and their modifications. Our laboratory has pioneered the use of direct RNA sequencing for the detection and quantification of RNA abundances and their modifications, including both development of improved library preparation protocols as well as the development of novel algorithms to predict and quantify RNA modifications [3-5]. Here we propose to use native RNA nanopore sequencing technology to predict the malignancy of biological samples in a high-throughput, rapid, multiplexed and cost-effective manner. Specifically, the candidate MSc student will benefit from a recently developed method in our lab to sequence small RNAs using nanopore sequencing. The candidate will then develop and apply deep learning algorithms to classify small RNA profiles into "normal", "tumoral" and "metastasic". Once the classification model is benchmarked and validated using cell lines, the methodology will then be applied to patient-derived samples.

References:

1. Begik O, Lucas MC, Ramirez JM, Liu H, Mattick JS and Novoa EM#. Integrative analyses of the RNA modification machinery reveal tissue- and cancer-specific signatures. Genome Biology 2020, 21:97. doi: 10.1186/s13059-020-02009-z 2. Gingold et al., A Dual Program for Translation Regulation in cellular profileration and differentiation. Cell 2014, 158(6):1281-1292. 3. Liu H*, Begik O*, Lucas MC, Ramirez JM, Mason CE, Wiener D, Schwartz S, Mattick JS, Smith MA and Novoa EM#. Accurate detection of m6A RNA modifications in native RNA sequences. Nature Comm 2019, 10:4079. doi:10.1038/s41467-019-11713-9 4. Smith MA*, Ersavas T*, Ferguson JM*, Liu J, Lucas MC, Begik O, Bojarski L, Barton K and Novoa EM#. Molecular barcoding of native RNAs using nanopore sequencing and deep learning. Genome Research 2020 30(9): 1345-1353 5. Begik O*, Lucas MC*, Ramirez JM, Milenkovic I, Cruciani C, Vieira HGS, Medina R, Liu H, Sas-Chen A, Mattick JS, Schwartz S and Novoa EM#. Quantitative profiling of native RNA modifications and their dynamics using nanopore sequencing. bioRxiv 2021, 189969 (accepted in Nature Biotechnology)

Expected skills::

python (required), R (required), prior experience with machine learning is a plus but not required, familiarity with third-generation sequencing (e.g. nanopore) is a plus but not required

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

Option for funding, as well as option for PhD continuity.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Pia Cosma
Email	pia.cosma@crg.es
Institution	CRG
Website	http://www.crg.eu/en/maria_pia_cosma
Group	Reprogramming and Regeneration

Project

Computational systems biology

Project Title:

Identification of master regulators of reprogramming

Keywords:

retina, reprogramming, regeneration, master regulators, gene networks

Summary:

We use gene regulatory network to identify master regulators of reprogramming and pluripotency. We are now investigating master regulators that can be enhances to induce the regeneration of the retina in mammals.

Expected skills::

bioinformatics, math lab,

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Master project 2021-2022

Personal Information

Supervisor	Jana Selent
Email	jana.selent@upf.edu
Institution	IMIM-UPF
Website	www.jana-selent.org
Group	GPCR Drug Discovery Group

Project

Structural bioinformatics

Project Title:

Unraveling signaling bias at G protein-coupled receptors (GPCRs)

Keywords:

G protein-coupled receptors, molecular dynamics, data analysis, drug design

Summary:

G protein-coupled receptors (GPCRs) are the most abundant class of receptors in the human organism. They are present in almost every type of cell, and govern almost every process in the human body (i.e. cognitive and inflammatory processes or control of the cardiovascular system). Owing to their ubiquity, they are targets of more than 30% of current drugs, and every day new GPCRs are revealed to be pharmacological targets for existing diseases. GPCRs can initiate signalling through binding with several intracellular partners, which in turn elicit distinct signalling cascades. It is established that in physiological conditions, a receptor binds to each of the partners in equal proportion. Interestingly some drugs act by altering the GPCR structure so that it preferentially binds to one specific partner - a phenomenon known as signaling bias. Such ligands, named biased ligands, offer great promise, as they enable to modify pathways associated with symptoms while not modifying other pathways - which could cause side effects. However, the molecular requirements for a molecule to act as a biased agonist within a receptor are still poorly understood. Molecular dynamics (MD) is a novel and sophisticated technique that enables to simulate protein behaviour in a physiological environment. The Master student will apply this approach to study time-resolved molecular mechanisms underlying signaling bias induced by small drug-like molecules. For this, the student will be trained on setting up simulated systems, running production simulations as well as the application of a wide range of analysis tools that allow capturing subtle structural events related to signaling bias. Structural insights will be exploited for the design of a novel class of GPCR modulators with a tailored signaling profile. We expect that the results of the analysis will be published in a high impact journal, and the expertise acquired by the student will make her/him a valuable asset for pharma companies in future. We are looking for a highly motivated and skilled student with exceptional academic records that allows pursuing a PhD afterwards.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods 2020, DOI: 10.1038/s41592-020-0884-y

Expected skills::

Experience in structural biology, programming in python/bash, molecular dynamics engines (GROMACS, NAMD, etc.), analysis tools/packages (VMD, Chimera, MDtraj…) and high level of English, oral and written.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Baldo Oliva
Email	baldo.oliva@upf.edu
Institution	UPF
Website	sbi.upf.edu
Group	SBI

Project

Structural bioinformatics

Project Title:

Protein Folding ab initio, based on contact maps and supersecondary structures using distance-restraints derived from multiple sequence alignments.

Keywords:

ab initio fold prediction; protein design; threading

Summary:

Background We work on the modeling of macro-complex regulatory structures formed by transcription factors (TF), co-factors, and DNA. One of the main contingencies to structurally model the macro-molecular complex formed by TFs and co-factors either in the enhancer or promoter binding sites is the completion of the structures. Modelling the structure of the DNA-binding domain of a TF bound to DNA can be achieved by homology modelling. However, the domains of TF that bind co-factors or other elements of the macro-molecular complex often involve unstructured or non-structurally solved regions. To solve this contingency we propose two approaches: 1) we will use docking of binary interactions if the structure of the unbound proteins is known; and 2) if the structure of one or both partners of a binary interaction we will apply threading and ab initio approaches to generate a structural model. Both approaches can be addressed by standard methods(1,2), but here we propose to update two innovations produced in our group during the previous grant. First, we developed iFrag to predict binding interfaces of proteins using only the sequences of the partners(3); second, we developed RADI (https://www.biorxiv.org/content/10.1101/406603v1), a program to predict residue-residue interactions based on amino-acid covariations in the sequence using Direct Coupling Analysis. We plan to combine RADI and iFrag to improve the prediction of binding and apply the approach on directed docking, modelling the structure of binary interactions when the structures of both partners are known or modelled. Furthermore, we will use RADI, combined with local super-secondary structures, named sMotifs, as classified in ArchDB(4), to model ab initio the structure of proteins that cannot be modelled by homology modelling. Approach We have already developed a pilot version, using restraints obtained with RADI, the prediction of secondary structure, and short fragment sMotifs templates from ArchDB to apply MODELLER(5) and construct an ab initio model structure of a query sequence. However, we need to update this version to be applicable on a largest extend of proteins. First, we need to update the database of sMotifs. Second, many proteins can share similar local conformations

but still they may not be classified. Therefore, we will extend the number of short-fragment templates by searching with psi-blast and selecting all short fragments with known structure that can potentially be used as templates. One of the problems already detected in the pilot is that more than one template can be assigned to the same region of the query sequence. We selected the longest fragment in the original version of the pilot, but the most appropriate approach is to model all combinations of fragments to produce several models after the application of restraints. Then, because neither the restraints of RADI are exact, nor the short-fragment templates correct, we will have several but different solutions of the conformation, so we will cluster the solutions, score and rank the conformers by statistical potentials using SPserver (6) and select the common restraints and common short fragment templates. The scores of SPserver can be improved by training a combination of energies that benefits from a large knowledge of structures and the use of several potentials. Also, we need a machine-learning method to optimize the selection of structures. A potential approach is to split the alignment in sections of 100 column-positions by sliding windows and construct the final model from the overlaps. The structure of the model will be iteratively repeated and improved until all convenient restraints and templates are congruent with the model.

References:

1. Garcia-Garcia, J., Bonet, J., Guney, E., Fornes, O., Planas, J. and Oliva, B. (2012) Networks of ProteinProtein Interactions: From Uncertainty to Molecular Details. Mol Inform, 31, 342-362. 2. Mirela-Bota, P., Aguirre-Plans, J., Meseguer, A., Galletti, C., Segura, J., Planas-Iglesias, J., Garcia-Garcia, J., Guney, E., Oliva, B. and Fernandez-Fuentes, N. (2020) Galaxy InteractoMIX: An Integrated Computational Platform for the Study of Protein-Protein Interaction Data. J Mol Biol. 3. Garcia-Garcia, J., Valls-Comamala, V., Guney, E., Andreu, D., Munoz, F.J., Fernandez-Fuentes, N. and Oliva, B. (2017) iFrag: A Protein-Protein Interface Prediction Server Based on Sequence Fragments. J Mol Biol, 429, 382-389. 4. Bonet, J., Planas-Iglesias, J., Garcia-Garcia, J., Marin-Lopez, M.A., Fernandez-Fuentes, N. and Oliva, B. (2014) ArchDB 2014: structural classification of loops in proteins. Nucleic Acids Res, 42, D315-319. 5. Webb, B. and Sali, A. (2016) Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics, 54, 5 6 1-5 6 37. 6. Aguirre-Plans, J., Meseguer, A., Molina-Fernández, R., Marin-Lopez, M.A., Jumde, G., Casanova, K., Bonet, J., Fornes, O., Fernandez-Fuentes, N. and Oliva, B. (2020) SPServer: Split-Statistical Potentials for the analysis of protein structures and protein-protein interactions. BMC Bioinformatics.

Expected skills::

Python programming

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Jana Selent
Email	jana.selent@upf.edu
Institution	IMIM-UPF
Website	www.jana-selent.org
Group	GPCR Drug Discovery Lab

Project

Structural bioinformatics

Project Title:

Why is water essential for GPCR functionality?

Keywords:

G protein-coupled receptors, molecular dynamics, data analysis, drug design

Summary:

G protein-coupled receptors (GPCRs) are the most abundant class of receptors in the human organism. They are present in almost every type of cell, and govern almost every process in the human body (i.e. cognitive and inflammatory processes or control of the cardiovascular system). Owing to their ubiquity, they are targets of more than 30% of current drugs, and every day new GPCRs are revealed to be pharmacological targets for existing diseases. As for the whole life, water is critical for GPCR functionality at a nanoscopic scale. It forms important intermolecular networks that mediate the signaling response of the receptor. Our group owns an extraordinary dataset with unprecedented information about the implication of water molecules in GPCR structural dynamics and the binding of small drug-like molecules. We believe that stabilized or disrupted intermolecular water signatures drive the functional consequences of a drug-like molecule. In this project, the Master student will seek for the underlying molecular mechanism of the water-mediated functional responses. For this, she/he will (i) develop an analysis pipeline to extract relevant information from our unique dataset and (ii) setup control simulations to validate obtained conclusions. We expect that the results of the analysis will be published in a high impact journal, and the expertise acquired by the student will make her/him a valuable asset for pharma companies in future. We are looking for a highly motivated and skilled student with exceptional academic records that allows pursuing a PhD afterwards.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods 2020, DOI: 10.1038/s41592-020-0884-y

Expected skills::

Experience in structural biology, programming in python/bash, molecular dynamics engines (GROMACS, NAMD, etc.), analysis tools/packages (VMD, Chimera, MDtraj…) and high level of English, oral and written.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Jana Selent
Email	jana.selent@upf.edu
Institution	IMIM-UPF
Website	www.jana-selent.org
Group	GPCR Drug Discovery Lab

Project

Structural bioinformatics

Project Title:

Simulation meets experiment: refinement of cryo-EM GPCR structures using molecular dynamics

Keywords:

G protein-coupled receptors, molecular dynamics, data analysis, drug design

Summary:

G-protein coupled receptors (GPCRs) are the most abundant class of receptors in the human organism. They are present in almost every type of cell, and govern almost every process in the human body (i.e. cognitive and inflammatory processes or control of the cardiovascular system). Owing to their ubiquity, they are targets of more than 30% of current drugs, and every day new GPCRs are revealed to be pharmacological targets for existing diseases. Recent advances in cryo-electron microscopy (cryo-EM) and image classification provide insights into ensembles of low- to high-resolution that describe differently populated conformational states of proteins. However, methods for deriving accurate atomistic models from cryo-EM density maps lag behind this resolution revolution. The increasing amount of molecular detail requires the development of new methodologies and software to accurately and timely interpret experimental densities. Molecular dynamics (MD)-based refinement methods have grown into a valuable approach to tackle this challenge. In this project, the Master student will develop an MD-based pipeline that can be applied to GPCRs. This represents an important milestone for the scientific community as it can provide novel structural insights into this important drug targeting class. For this, the student will learn how to setup simulations and how to use correlation-driven MD for the refinement of atomistic models into cryo-electron microscopy maps. We expect that the results of the analysis will be published in a high impact journal, and the expertise acquired by the student will make her/him a valuable asset for pharma companies in future. We are looking for a highly motivated and skilled student with exceptional academic records that allows pursuing a PhD afterwards.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods 2020, DOI: 10.1038/s41592-020-0884-y Igaev et al. Automated cryo-EM structure refinement using correlation-driven molecular dynamics, Elife 2019, DOI: 10.7554/eLife.43542

Expected skills::

Experience in structural biology, programming in python/bash, molecular dynamics engines (GROMACS, NAMD, etc.), analysis tools/packages (VMD, Chimera, MDtraj…) and high level of English, oral and written.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Baldo Oliva
Email	baldo.oliva@upf.edu
Institution	UPF
Website	http://sbi.upf.edu
Group	SBI

Project

Structural bioinformatics

Project Title:

Proposal to score protein structures using distance-restraints derived from multiple sequence alignments and statistic potentials.

Keywords:

Protein structure quality

Summary:

We work on the modeling of macro-complex regulatory structures formed by transcription factors (TF), co-factors, and DNA. One of the main contingencies to structurally model the macro-molecular complex formed by TFs and co-factors either in the enhancer or promoter binding sites is the completion of the structures. Modelling the structure of the DNA-binding domain of a TF bound to DNA can be achieved by homology modelling. However, the domains of TF that bind co-factors or other elements of the macro-molecular complex often involve unstructured or non-structurally solved regions. To solve this contingency we propose two approaches: 1) we will use docking of binary interactions if the structure of the unbound proteins is known; and 2) if the structure of one or both partners of a binary interaction we will apply threading and ab initio approaches to generate a structural model. Both approaches can be addressed by standard methods(1,2), but here we propose to update two innovations produced in our group during the previous grant. First, we developed iFrag to predict binding interfaces of proteins using only the sequences of the partners(3); second, we developed RADI (https://www.biorxiv.org/content/10.1101/406603v1), a program to predict residue-residue interactions based on amino-acid covariations in the sequence using Direct Coupling Analysis. We have used RADI, combined with local super-secondary structures, named sMotifs, as classified in ArchDB(4), to model ab initio the structure of proteins that cannot be modelled by homology modelling. We have already developed a pilot version, using restraints obtained with RADI, the prediction of secondary structure, and short fragment sMotifs templates from ArchDB to apply MODELLER(5) and constructed an ab initio model structure of a query sequence. However, we need to update this version to select the correct folds, because neither the restraints of RADI are exact, nor the short-fragment templates correct, and as a consequence, we obtain several but different solutions of the conformation. Therefore, we must cluster the solutions, score and rank the conformers and select the best conditions of the models. We have developed a program of statistical potentials, the SPserver (6) to select the correct folds. The scores of SPserver can be improved by training a combination of energies that benefits from a large knowledge of structures and the use of several potentials. The program can also benefit from the results of RADI to improve the learning. This machine-learning method will be useful to optimize the selection of structures. Proposed approach. 1) Define a method to compare as residue-profile the structures of model decoys with the experimental structure. This can be obtained by the superposition of the complete structure, using the RMSD fluctuation of CA atoms, the local matrix of distances around residues, or the local GDT score. Local distance matrices have the benefit to compare straightforward with potentials under a distance cut-off threshold. The information will be transformed into a score with values between 0 and 1, being 1 the best match (for example, the ratio of identic contacts within a radius that are the same between the decoy and the experimental structure). 2) Construct several ML methods of python “sklearn” library (SVM, NN, LogisticProgression, RF, etc.) using the statistic potential profiles per residue as inputs, the

matrix of interacting potentials with the residues within a 3D-ball and the RADI direct and mutual information. The output will be a normalized score between 0 and 1 of the local quality. Additional approaches, using PyTorch, Keras and Tensorflow will also be considered. 3) The final score of the model will be obtained as the average of the total of scores along the sequence.

References:

1. Garcia-Garcia, J., Bonet, J., Guney, E., Fornes, O., Planas, J. and Oliva, B. (2012) Networks of ProteinProtein Interactions: From Uncertainty to Molecular Details. Mol Inform, 31, 342-362. 2. Mirela-Bota, P., Aguirre-Plans, J., Meseguer, A., Galletti, C., Segura, J., Planas-Iglesias, J., Garcia-Garcia, J., Guney, E., Oliva, B. and Fernandez-Fuentes, N. (2020) Galaxy InteractoMIX: An Integrated Computational Platform for the Study of Protein-Protein Interaction Data. J Mol Biol. 3. Garcia-Garcia, J., Valls-Comamala, V., Guney, E., Andreu, D., Munoz, F.J., Fernandez-Fuentes, N. and Oliva, B. (2017) iFrag: A Protein-Protein Interface Prediction Server Based on Sequence Fragments. J Mol Biol, 429, 382-389. 4. Bonet, J., Planas-Iglesias, J., Garcia-Garcia, J., Marin-Lopez, M.A., Fernandez-Fuentes, N. and Oliva, B. (2014) ArchDB 2014: structural classification of loops in proteins. Nucleic Acids Res, 42, D315-319. 5. Webb, B. and Sali, A. (2016) Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics, 54, 5 6 1-5 6 37. 6. Aguirre-Plans, J., Meseguer, A., Molina-Fernández, R., Marin-Lopez, M.A., Jumde, G., Casanova, K., Bonet, J., Fornes, O., Fernandez-Fuentes, N. and Oliva, B. (2020) SPServer: Split-Statistical Potentials for the analysis of protein structures and protein-protein interactions. BMC Bioinformatics.

Expected skills::

Python programming.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Baldo Oliva
Email	baldo.oliva@upf.edu
Institution	UPF
Website	http://sbi.upf.edu
Group	SBI

Project

Structural bioinformatics

Project Title:

TF-DNA Binding strength affected by methylations

Keywords:

enhancer methylation

Summary:

Changes in DNA methylation are involved in development, disease, and the response to environmental conditions. Methylation of DNA is thought to regulate transcription both directly and indirectly. CpG methylation can directly repress transcription by preventing binding of some transcription factors (TFs) to their recognition motifs(17). For further insights, Lea et al. developed mSTARR-seq(18), a method that assesses the causal effects of DNA methylation on regulatory activity at genomic high-throughput level. Our objective is to predict the changes of TF binding caused by methylation. First we will build a database of methylated DNA binding with known TF binding. In a first approach the database will be extracted from experimental data of Yin et al. (17) and Lea et al. (18), indicating the loss or gain of TF binding. In a second approach, we will infer the effect from the comparison of bound TF binding sites with and without methylations. We will use the dataset of UniBind(13) to select the binding sites confirmed bound by TFs or the predictions of Viestra et al. (15). Then, we will select the tracks from UCSC Genome Browser with assays of DNA methylation (i.e. Methyl-RBBS) specific for tissue. We will compare the percentage that cytosines are methylated in the binding site with respect to any other location in the genome (this can be further refined by comparing with cytosines in the same TAD region). We will use the hypergeometric distribution to compare the ratio of methylation versus the expected ratio according to the length of binding recognition (as derived by the ChIP-Seq experiment). We will split the results in three categories: 1) If the ratio of methylation is lower than expected, then the methylation of cytosines reduces the TF binding. 2) On the contrary, if the methylation in the binding regions is higher than expected, the methylation is required for TF binding. 3) Otherwise, the methylation has no effect on TF binding. With the new bindings (case 2) we will generate statistical potentials specific of methylated cytosines by modelling the structure of TF-DNA binding, introducing a new symbol for methylated cytosines and including them in the general statistical potentials. We will calculate statistical potentials specific of the family of each TF including the new symbol for methyl-cytosine as in Meseguer et al. (19). The effect of disruption (case 1) will be used to generate statistical potentials specific of disruption. As before, the structure of TF-DNA binding will be modelled and the frequencies of the interactions between amino-acids and nucleotides will be obtained from the models. However, these potentials will be used to determine the potential of disruption, as these are the models of interactions lost after methylation. As before, a general potential will be derived with all TFs and their disrupted DNA binding sites and another set of potentials, specific for each TF family will be constructed. Finally, we will test the capacity of predicting TF disruptions after cytosine methylation or TF-DNA new bindings and specific PWMs for methyl cytosines. Two tests will be used for validation. First, using a 5-fold protocol with partially hidden data; and second, by training the method with one set of methylation (i.e. using the experiments of Yin et al. (17) and Lea et al. (18)) and testing the potentials in a different set (i.e. using data of ENCODE and removing redundancies with the training).

References:

1. Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME Suite. Nucleic Acids Res, 43, W39-49. 2. Fuxman Bass, J.I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., Yi, S., Hill, D.E., Vidal, M. and Walhout, A.J. (2015) Human gene-centered transcription factor networks for enhancers and disease variants. Cell, 161, 661-673. 3. Fuxman Bass, J.I., Pons, C., Kozlowski, L., Reece-Hoyes, J.S., Shrestha, S., Holdorf, A.D., Mori, A., Myers, C.L. and Walhout, A.J. (2016) A gene-centered C. elegans protein-DNA interaction network provides a framework for functional predictions. Mol Syst Biol, 12, 884. 4. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327-339. 5. Vockley, C.M., Guo, C., Majoros, W.H., Nodzenski, M., Scholtens, D.M., Hayes, M.G., Lowe, W.L., Jr. and Reddy, T.E. (2015) Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res, 25, 1206-1214. 6. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S. and Kellis, M. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat Biotechnol, 34, 1180-1190. 7. Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D. and Shendure, J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol, 27, 1173-1175. 8. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B. et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol, 30, 271-277. 9. Fornes, O., Gheorghe, M., Richmond, P.A., Arenillas, D.J., Wasserman, W.W. and Mathelier, A. (2018) MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. Sci Data, 5, 180141. 10. Kumar, S., Ambrosini, G. and Bucher, P. (2017) SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res, 45, D139-D144. 11. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A. et al. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583, 699-710. 12. Wang, J., Zhuang, J., Iyer, S., Lin, X.Y., Greven, M.C., Kim, B.H., Moore, J., Pierce, B.G., Dong, X., Virgil, D. et al. (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res, 41, D171-176. 13. Gheorghe, M., Sandve, G.K., Khan, A., Cheneby, J., Ballester, B. and Mathelier, A. (2019) A map of direct TF-DNA interactions in the human genome. Nucleic Acids Res, 47, e21. 14. Mathelier, A., Shi, W. and Wasserman, W.W. (2015) Identification of altered cis-regulatory elements in human disease. Trends Genet, 31, 67-76. 15. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E. et al. (2020) Global reference mapping of human transcription factor footprints. Nature, 583, 729-736. 16. Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A. et al. (2020) Index and biological spectrum of human DNase I hypersensitive sites. Nature, 584, 244-251. 17. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F. et al. (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science, 356. 18. Lea, A.J., Vockley, C.M., Johnston, R.A., Del Carpio, C.A., Barreiro, L.B., Reddy, T.E. and Tung, J. (2018) Genome-wide quantification of the effects of DNA methylation on human gene regulation. Elife, 7. 19. Meseguer, A., Årman, F., Fornes, O., Molina-Fernández, R., Bonet, J., Fernandez-Fuentes, N. and Oliva, B. (2020) On the prediction of DNA-binding preferences of C2H2-ZF domains using structural models: application on human CTCF. NAR Genomics and Bioinformatics, 2.

Expected skills::

Python programming

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Baldo Oliva
Email	baldo.oliva@upf.edu
Institution	UPF
Website	http://sbi.upf.edu
Group	SBI

Project

Structural bioinformatics

Project Title:

TF-DNA Binding strength affected by mutations on DNA binding site

Keywords:

Single nucleotide variants; Cis-regulation; Transcription factors;

Summary:

The principal reason to understand changes on the binding strength of a TF for a specific DNA binding sequence is to study cis-regulation and mutations that will affect it. Therefore, we focus on the prediction of the effect of one or more single nucleotide substitutions disrupting the recognition of the specific TF. The approach can be used to predict causal regulatory haplotypes that likely contribute to human phenotypes and to functionally fine map causal regulatory variants in regions of high linkage disequilibrium identified by expression quantitative trait loci (eQTL) analyses. We will predict binding strength to classify strong and weak changes and deciding if they imply the loss of the interaction. We will apply structural modelling with several templates, testing all potential PWMs, and the analysis of various statistical potentials on the TF-DNA interaction, comparing the native binding site with all other DNA variants. In order to train an AI/ML model, we will use the information from experimental Protein Binding Micro-arrays (PBM) of each TF. We will use as inputs the profiles of statisticalenergies and also the profiles of the PWM search along the DNA sequence, using the scores of FIMO (1) and the enrichment achieved with different structural models of the same TF. In order to learn and test with these scores, using different TFs and DNA binding lengths, the scores of the profiles will be normalized. The normalization will help us to better characterize the magnitude of the change produced by single nucleotide substitutions of the DNA binding sequence. The use of PBM will help us to handle an overwhelming amount of information, as all combinations of DNA sequences (formed by 8 or 12-mer nucleotides) are experimentally tested. We will train on PBM and the test will be performed on PBM

(using a 10-fold approach), yeast-one-hybrid experiments on the specificity-changes of mutant DNA sequences(2,3), and other high-throughput experiments available such as SELEX(4), STARR-seq (5) or Sharpr-MPRA(6) (a modification of the MPRA(7,8) protocol that was developed to unveil at genome scale the effect of SNVs). Additional training and testing sets will be extracted from the datasets MANTA2 (9) and SNP2TFBS (10), composed by binding-site predictions in the human genome with the potential impact on TF binding for all possible SNVs. The impact of SNVs in MANTA2 is assessed by means of PWM scores computed on the alternate alleles. Similarly, we will use the theoretical PWMs plus all the energy profiles obtained by scanning the DNA sequence with the collection of TF-DNA structural models. Furthermore, the approach will be iteratively improved to be applicable on direct TF-DNA interactions in the human genome, using ChiP-seq data from recently expanded ENCODE encyclopaedia (11), updated versions of FactorBook (12), the dataset of UniBind (13) and datasets of altered cis-regulatory elements (14) and the location of DNase I hypersensitive regions of the genome with human genetic variation within transcription factor footprints(15,16)

References:

1. Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME Suite. Nucleic Acids Res, 43, W39-49. 2. Fuxman Bass, J.I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., Yi, S., Hill, D.E., Vidal, M. and Walhout, A.J. (2015) Human gene-centered transcription factor networks for enhancers and disease variants. Cell, 161, 661-673. 3. Fuxman Bass, J.I., Pons, C., Kozlowski, L., Reece-Hoyes, J.S., Shrestha, S., Holdorf, A.D., Mori, A., Myers, C.L. and Walhout, A.J. (2016) A gene-centered C. elegans protein-DNA interaction network provides a framework for functional predictions. Mol Syst Biol, 12, 884. 4. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327-339. 5. Vockley, C.M., Guo, C., Majoros, W.H., Nodzenski, M., Scholtens, D.M., Hayes, M.G., Lowe, W.L., Jr. and Reddy, T.E. (2015) Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res, 25, 1206-1214. 6. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S. and Kellis, M. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat Biotechnol, 34, 1180-1190. 7. Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D. and Shendure, J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol, 27, 1173-1175. 8. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B. et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol, 30, 271-277. 9. Fornes, O., Gheorghe, M., Richmond, P.A., Arenillas, D.J., Wasserman, W.W. and Mathelier, A. (2018) MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. Sci Data, 5, 180141. 10. Kumar, S., Ambrosini, G. and Bucher, P. (2017) SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res, 45, D139-D144. 11. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A. et al. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583, 699-710. 12. Wang, J., Zhuang, J., Iyer, S., Lin, X.Y., Greven, M.C., Kim, B.H., Moore, J., Pierce, B.G., Dong, X., Virgil, D. et al. (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res, 41, D171-176. 13. Gheorghe, M., Sandve, G.K., Khan, A., Cheneby, J., Ballester, B. and Mathelier, A. (2019) A map of direct TF-DNA interactions in the human genome. Nucleic Acids Res, 47, e21. 14. Mathelier, A., Shi, W. and Wasserman, W.W. (2015) Identification of altered cis-regulatory elements in human disease. Trends Genet, 31, 67-76. 15. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E. et al. (2020) Global reference mapping of human transcription factor footprints. Nature, 583, 729-736. 16. Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A. et al. (2020) Index and biological spectrum of human DNase I hypersensitive sites. Nature, 584, 244-251. 17. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F. et al. (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science, 356. 18. Lea, A.J., Vockley, C.M., Johnston, R.A., Del Carpio, C.A., Barreiro, L.B., Reddy, T.E. and Tung, J. (2018) Genome-wide quantification of the effects of DNA methylation on human gene regulation. Elife, 7. 19. Meseguer, A., Årman, F., Fornes, O., Molina-Fernández, R., Bonet, J., Fernandez-Fuentes, N. and Oliva, B. (2020) On the prediction of DNA-binding preferences of C2H2-ZF domains using structural models: application on human CTCF. NAR Genomics and Bioinformatics, 2.

Expected skills::

Python programming

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Jana Selent
Email	jana.selent@upf.edu
Institution	IMIM-UPF
Website	www.jana-selent.org
Group	GPCR Drug Discovery Lab

Project

Structural bioinformatics

Project Title:

Detection of novel druggable binding sites at G protein-coupled receptors

Keywords:

G protein-coupled receptors, molecular dynamics, data analysis, drug design

Summary:

G protein-coupled receptors (GPCRs) are the most abundant class of receptors in the human organism. They are present in almost every type of cell, and govern almost every process in the human body (i.e. cognitive and inflammatory processes or control of the cardiovascular system). Owing to their ubiquity, they are targets of more than 30% of current drugs, and every day new GPCRs are revealed to be pharmacological targets for existing diseases. Molecular dynamics (MD) is a sophisticated technique that enables to simulate protein behaviour in a physiological environment. In this project the Master student will develop a simulation-based pipeline that allows detecting druggable binding sites including cryptic pockets ("transient binding pockets"). This pipeline involves (i) the setup of simulation systems including small chemical fragments to probe the entire protein surface for druggable binding sites, (ii) running production runs, (iii) automated detection of binding sites and (iii) intuitive visualization. The pipeline will be implemented into our GPCRmd server and detected sites will be exploited for the discovery of new molecular GPCR modulators. We expect that the results of the analysis will be published in a high impact journal, and the expertise acquired by the student will make her/him a valuable asset for pharma companies in future. We are looking for a highly motivated and skilled student with exceptional academic records that allows pursuing a PhD afterwards.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods 2020, DOI: 10.1038/s41592-020-0884-y

Expected skills::

Experience in structural biology, programming in python/bash, molecular dynamics engines (GROMACS, NAMD, etc.), analysis tools/packages (VMD, Chimera, MDtraj…) and high level of English, oral and written.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Javier del Campo
Email	jdelcampo@ibe.upf-csic.es
Institution	Institut de Biologia evolutiva (CSIC-UPF)
Website	delcampolab.com
Group	del Campo Lab. Microbial Ecology and Evolution

Project

Web development & bioinformatic tools

Project Title:

Shiny Tree

Keywords:

reference databases, shiny, eDNA, microbiome

Summary:

The laboratory The del Campo Lab is based at the Institut de Biologia Evolutiva (UPF-CSIC) in Barcelona. The research at the del Campo Lab is focused on the study of host-associated microbes and the effect of global warming on the microbiomes of benthic and planktonic marine animals. We have a wet and dry lab, to perform experiments and bioinformatics analysis, enabling the broadest possible goals. The ongoing climate change and its effects on the environment, such as rising sea temperature, has strong impacts on free-living marine microbial communities. However, the effects of global warming have not been properly studied on host-associated microbiomes. Microbiomes (both prokaryotic and eukaryotic) associated with host organisms have a strong influence on host evolution, physiology, and ecological functions. We study how environmental changes resulting from global warming affect the composition and function of the microbiomes in key members of the marine fauna and consequently how these changes affect the hosts. Currently, our study focuses on these impacts on corals, teleost fish, and zooplankton. To tackle this novel research topic, we use a combination of molecular biology, ecophysiology, and bioinformatics. The proposed project Phylogenetic trees are essential to interpret evolutionary relationships and have become crucial in order to curate taxonomically the references databases used for microbiome and eDNA (biomonitoring) analysis. However, not all the researchers that keep the have the neeeded taxonomic know-how to properly curate the aforementioned databases are familiar with the use of the terminal or the use of phylogenetic tree building software. The aim of the proposed project is to build a Shiny app to facilitate tree building, edition and manipulation to non-experts in a user-friendly manner. Using established tools such as fasttree, RAxML or ggtree we will build a Shiny app that will be integrated into a reference database curation pipeline that we are designing in collaboration with The Carpentries (https://carpentries.org/) as part of the EukRef project (https://pr2-database.org/eukref/about/).

References:

del Campo, J. et al. (2018) EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. PLOS Biol. 16, e2005849 Yu, G. et al. (2017) Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol. Evol. 8, 28–36

Expected skills::

R, Shiny

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed
