Master project 2020-2021

Personal Information

Supervisor	Roberto Malinverni and Marcus Buschbeck
Email	rmalinverni@carrerasresearch.org
Institution	Josep Carreras Leukaemia Research Institute (IJC)
Website	http://www.carrerasresearch.org/en/Chromatin_Metabolism_and_Cell_Fate
Group	Chromatin, metabolism and cell fate

Project

Computational genomics

Project Title:

Role of macroH2A histone variant in three-dimensional genomics context

Keywords:

Chromatin, macroH2A, HiC, Histone, R, NGS

Summary:

Chromosome conformation capture (C-)techniques allow to assess the nuclear architecture and distribution of chromatin in unprecedented level and have boosted the growth of nuclear organization field during the recent years (1). A large number of different variants of C-techniques including Hi-C and , Hi-ChIP have become routine in basic research, this has led to the creation of a massive amount of data stored in different public databases. Our laboratory has particular interest for years in the study of a particular histone variant called macroH2A (2). Histone variants replace canonical histones in a sub-fraction of the core structural units of chromatin, the nucleosomes. Recently, we demonstrated surprising impact of macroH2A on nuclear organization and heterochromatin architecture (3). The proposal of this master project is to investigate the role of macroH2A and other heterochromatin regulators, through the integration of the data created by our laboratory (Hi-C, HiChIP, ChipSeq, RNAseq) with those present in public databases. Specifically, we will address the following questions: 1. To evaluate association of macroH2A with respect to self-interacting genomic regions such as topological associated domains (TADs) and genome compartments. 2. To modify a framework of our previously created tool (regioneR (4)) to allow its application in a three-dimensional genomics context. 3. To create, pipelines and bioinformatics tools to query and visualize such a complex mass of data. Technically we will mainly use resources in R (Bioconductor) and Python, in Linux environment. High Performance Computer calculation will be carried out at CSUC (www.csuc.cat).

References:

1 - Grob S., Cavalli G. (2018) Technical Review: A Hitchhiker’s Guide to Chromosome Conformation Capture. In: Bemer M., Baroux C. (eds) Plant Chromatin Dynamics. Methods in Molecular Biology, vol 1675. Humana Press, New York, NY. 2 - Post-Translational Modifications of H2A Histone Variants and Their Role in Cancer. Corujo D, Buschbeck M. Cancers (Basel). 2018 Feb 27;10(3). 3 - MacroH2A histone variants maintain nuclear organization and heterochromatin architecture. Douet J, Corujo D, Malinverni R, Renauld J, Sansoni V, Posavec Marjanović M, Cantariño N, Valero V, Mongelard F, Bouvet P, Imhof A, Thiry M, Buschbeck M. J Cell Sci. 2017 May . 4 - regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. Bioinformatics. 2016 Jan 15.

Expected skills::

Experience in programming languages (preferably R). Basic knowledge of NGS data and Linux operating system. Enthusiasm to answer biological questions.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

This project will be co-supervised by Roberto Malinverni and Marcus Buschbeck

---

Universitat Pompeu Fabra Barcelona
Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Toni Giorgino
Email	toni.giorgino@cnr.it
Institution	Consiglio Nazionale delle Ricerche
Website	www.giorginolab.it
Group	Istituto di Biofisica

Project

Computational genomics

Project Title:

Machine learning-based assessment of SARS-CoV-2 genome variability

Keywords:

Coronavirus, machine learning, deep neural networks, genome

Summary:

We shall exploit the genetic-epidemiological evidence collected during the present SARS-CoV-2 outbreak to characterise the genomic regions with high mutation potential that could play a role in future outbreaks and in acquisition of drug resistance. Statistical learning and artificial-intelligence methods will be used to produce mutation models; the selected hot-spots will be cross-referenced in order to build early lead strategies of use in future outbreaks. The work is essentially computational. The student may work, either locally or remotely, with the computational group at the Institute of Biophysics of the Italian National Research Council, located at the University of Milan (Italy). Further collaborations are possible.

References:

* Smith M, Smith JC. Repurposing Therapeutics for the Wuhan Coronavirus nCov-2019: Supercomputer-Based Docking to the Viral S Protein and Human ACE2 Interface. 2020 Feb 20 (Chemrxiv) * https://viralzone.expasy.org/8996 * www.giorginolab.it * https://users.unimi.it/biolstru/molbd3-lab.html

Expected skills::

The project is heavily computationally focused. A good grasp of Python and an interest in machine learning are essential.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

Also: structural bioinformatics. Contact: toni.giorgino@cnr.it

---

Master project 2020-2021

Personal Information

Supervisor	Toni Gabaldón
Email	toni.gabaldon.bcn@gmail.com
Institution	Barcelona Supercomputing Centre

Website	http://cgenomics.org
Group	Comparative Genomics

Project

Computational genomics

Project Title:

Evolution of hybrid genomes

Keywords:

Hybridization, genome evolution, phylogenomics, pathogens

Summary:

Evolution of eukaryotic species and their genomes has been traditionally understood as a vertical process in which genetic material is transmitted from parents to offspring along a lineage, and in which genetic exchange is restricted within species boundaries. However, mounting evidence coming from comparative genomic studies indicates that this paradigm is often violated. Horizontal gene transfer and mating between diverged lineages blur species boundaries and complicates the reconstruction of evolutionary histories of species and their genomes. Non-vertical evolution might be more restricted in eukaryotes as compared to prokaryotes, yet it is not negligible and can be common in certain groups. Recognition of such processes brings about the need to incorporate this complexity in our tools and models, as well as to conceptually re-frame eukaryotic diversity and evolution. In this project you will work on several hybrid genomes, including those of some pathogenic species, using comparative genomics and populations genomics tools.

References:

https://www.ncbi.nlm.nih.gov/pubmed/28681409

Expected skills::

Python, Phylogenetics, Variant calling analysis,

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Comments:

Alternative projects, within the scope of interests of the group (see publications and webpage) can be discussed.

Master project 2020-2021

Personal Information

Supervisor	Julio Rozas
Email	jrozas@ub.edi
Institution	Universitat de Barcelona
Website	http://www.ub.edu/molevol/EGB/
Group	Evolutionary Genomics & Bioinformatics

Project

Computational genomics

Project Title:

Comparative and evolutionary analysis of repetitive elements in spider genomes

Keywords:

Comparative genomics; Transposable elements; Repetitive elements; phylogenomics; Adaptive genomics; genome annotation

Summary:

Understanding the origin, amplification and functional role of repetitive sequences in eucaryotic genomes is a central question in Evolutionary Biology. Despite that modern high-throughput sequencing (HTS) technologies are currently accessible for many labs, the accurate identification and annotation of gene family is one of the major challenges in the field. This scenario will change in the near future thanks to the irruption of the so called third-generation sequencing technologies (i.e., long-read sequencing). In this sense, our research group is generating new high quality genomic data from a group of Canary Island endemic spiders (Chelicerata) using long-read sequencing technologies but also chromosome-scale assembly techniques, such as Hi-C and Chicago libraries. The objective of this project is to study the abundance, distribution and evolution of repetitive elements in chelicerates and, by extension, in arthropods, including transposable elements (TEs) and other types of repetitive sequences. A large body of evidence suggest that these elements have structural functional significance. TEs, for instance, can generate variability by movement and insertion, are responsible of defining centromeric regions, or can activate gene expression under stress conditions. Our bioinformatic study in a comparative context will enable understanding of the nature and behavior of this important genomic components. The student will participate in the identification, annotation and/or analysis of repetitive elements in complete genomes of several spiders (and chelicerates) species. For that, he/she will use high quality genome sequences (data generated by our group based on third generation sequencing technologies, and sequences already available in databases), bioinformatics tools (software and scripts to manipulate and visualizte sequences and genomic annotations, to identify repetitive elements, to conduct evolutionary genetics analyses). The basic work-flow will consist in the identification and primary annotation of repeats, the determination of families, types and classes, the estimation of gene turnover rates, or the characterization of the distribution of these repetitive sequences across chromosomes or with respect to other genomic elements, such as protein-coding genes . Many of these analyses will be carried out in our high performance computer cluster.

References:

References from our researhch group • Frías-López, C., Sánchez-Herrero, J. F., Guirao-Rico, S., Mora, E., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2016.DOMINO: Development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms. Bioinformatics 32: 3753-3759. doi:10.1093/bioinformatics/btw534. • Rendón-

Anaya, M. et al. 2019.The Avocado Genome Informs Deep Angiosperm Phylogeny, Highlights Introgressive Hybridization, and Reveals Pathogen-Influenced Gene Space Adaptation. Proc. Natl. Acad. Sci. USA. 116: 17081-17089. doi: 10.1101/654285. • Sánchez-Herrero, J. F., Frías-López, C., Escuer, P., Hinojosa-Alvarez, S., Arnedo, M. A., Sánchez-Gracia, A., Rozas, J. 2019.The draft genome sequence of the spider Dysdera silvatica (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. GigaScience 8: 1-9. doi: 10.1093/gigascience/giz099. • Vizueta, J., Macías-Hernández, N., Arnedo, M. A., Rozas, J. Sánchez-Gracia, A. 2019.Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation. Mol. Ecol. 28: 4028-4045. doi: 10.1111/mec.15199. • Vizueta, J., Sánchez-Gracia, A., Rozas, J. 2019.BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. bioRxiv XX:. doi: 10.1101/593889. • Vizueta, J., Rozas, J., Sánchez-Gracia, A. 2018.Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertoires across Chelicerates Genome Biol. Evol. 10: 1221-1236. doi:10.1093/gbe/evy081. Research Group References: (http://www.ub.edu/molevol/julio/SelPublications.html)

Expected skills::

Basic knowledge on NGS data handling and analysis, especially in genome assembly and annotation, notions of comparative genomics and transcriptomics approaches and phylogenetic methods, and experience with Linux operating systems and some of the high level programing languages commonly used in bioinformatics (Perl, Python, R).

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 2020-2021

Personal Information

Supervisor	Anthony Mathelier
Email	anthony.mathelier@ncmm.uio.no
Institution	Centre for Molecular Medicine Norway, University of Oslo
Website	https://mathelierlab.com/
Group	Computational Biology & Gene Regulation

Project

Computational genomics

Project Title:

A pan-cancer computational study of the interplay between transcription factor binding, DNA methylation, and enhancer activity

Keywords:

DNA methylation, Transcription Factor, Enhancer, Quantitative Trait Loci, Machine Learning

Summary:

Methylation of DNA is a prominent DNA modification linked to gene expression alteration in cancers [1,2]. While DNA methyltransferase (DNMT) enzymes de-methylate DNA, Ten-Eleven Translocation (TET) proteins are involved in demethylation. As DNMTs and TETs do not bind DNA in a sequence-specific manner, how these proteins are recruited to their specific sites of action is still an open question. Further, our understanding of the cascading effect of these aberrant DNA methylation patterns on gene expression deregulation is still limited. With a better characterization of the cascading effects of DNA methylation in cancer patients, we could reveal key regulatory networks critical for an improved molecular understanding of the diseases, as we recently showed for breast cancer [3]. In this project, the Master student will perform pan-cancer computational analyses to study the interplay between TF binding, DNA methylation, and enhancer activity. Specifically, the project will aim at (1) unravelling the interplay between DNA methylation and TF-binding in cancer types with limited statistical power and (2) unravelling the interplay between DNA methylation and enhancer activities. 1. We recently computed expression-methylation quantitative trait loci (emQTL) between TF expression and methylation at high-confidence TF-DNA interaction information stored in our UniBind database [4]. emQTL highlighted an interplay between DNA 5mC marks and TF-binding and showed that the binding of key pioneer TFs at their binding sites are likely to trigger local DNA demethylation that could lead to carcinogenesis (unpublished). Unfortunately, the small sample size for some cancer types prohibited the identification of the TFs involved, due to reduced statistical power. The student will use Generative Adversarial Networks (or alike machine-learning approaches) to simulate synthetic data for both methylation and gene expression from available patient data. This approach will alleviate the statistical power bottleneck currently observed. The generated data will be used to perform emQTL analyses and highlight key TFs modulating DNA methylation landscape in these cancer genomes. 2. In the second part of the project, the emQTL framework will be extended to investigate the relationship between DNA methylation and enhancer activity. Specifically, we will use RNA-seq data mapped at enhancers annotated by the FANTOM5 consortium with DNA methylation information from both normal and cancer tissues. The results will be used to investigate how the interplay between DNA methylation, TF binding, and enhancer activity mimics cell fate transition. Indeed, recent reports found that, during cell fate transition, pioneer TFs prime inaccessible enhancers, leading to increased chromatin accessibility and loss of DNA methylation [5]. This project will equip the student with computational biology skills employed in studying gene regulation and cancer genomics. She/he will build computational workflow using Snakemake and scripts in Python, R, and bash. The master student will be introduced to and learn to handle large public cancer genomics data (from ICGC, TCGA, and BASIS) and gene regulation resources (e.g. UniBind, JASPAR).

References:

1. Suzuki T, Maeda S, Furuhata E, Shimizu Y, Nishimura H, Kishima M, et al. A screening system to identify transcription factors that induce binding site-directed DNA demethylation. Epigenetics Chromatin 2017;10:60. 2. Suzuki T, Shimizu Y, Furuhata E, Maeda S, Kishima M, Nishimura H, et al. RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells. Blood Adv 2017 3. Fleischer T, Tekpli X, Mathelier A, Wang S, Nebdal D, Dhakal HP, et al. DNA methylation at enhancers identifies distinct breast cancer lineages. Nat. Commun. 2017 4. Gheorghe M, Sandve GK, Khan A, Chèneby J, Ballester B, Mathelier A. A map of direct TF–DNA interactions in the human genome. Nucleic Acids Res. 2019 5. Barnett KR, Decato BE, Scott TJ, Hansen TJ, Chen B, Attalla J, et al. ATAC-Me Captures Prolonged DNA Methylation of Dynamic Chromatin Accessibility Loci during Cell Fate Transitions. Mol. Cell 2020

Expected skills::

Proficiency in Python, R, and/or bash, previous experience in genomics data analysis, team spirit, English proficiency, Exposure to gene regulation and cancer biology will be a plus but not a strict requirement

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Anthony Mathelier
Email	anthony.mathelier@ncmm.uio.no
Institution	Centre for Molecular Medicine Norway, University of Oslo
Website	https://mathelierlab.com/
Group	Computation Biology & Gene Regulation

Project

Computational genomics

Project Title:

Exploring the links between DNA methylation, transcription factor binding, and alternative splicing

Keywords:

Alternative splicing, DNA methylation, transcription factor, cancer genomics, gene regulation

Summary:

RNA splicing is a process involved in mRNA maturation that involves removal of introns from the pre-mRNA. The machinery engaged in this process, the spliceosome, recognizes conserved nucleotide sequences in the introns in order to promote their excision from the pre-mRNA. Alternative splicing is a process that allows the cells to selectively control which parts of a pre-mRNA will be represented in the mature mRNA to be translated into protein. Despite current knowledge, the mechanism of alternative splicing is still not fully understood. The CCCTC-binding factor (CTCF) is a transcription factor (TF) that has recently been shown to be relevant in this process as mutations in its binding site are linked to specific exon inclusion or exclusion [1]. As binding of CTCF to the DNA is altered by DNA methylation [2], the study of the impact of DNA methylation at CTCF binding site on alternative splicing could shed light on aberrant splicing patterns observed in cancers. In this project the Master student will explore the interplay between DNA methylation, transcription factor binding and aberrant alternative splicing in cancers. Specifically, we plan to: (1) detect differentially used exons in cohorts of tumor and normal samples obtained from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC); (2) identify binding sites for CTCF and other TFs in the vicinity of the differentially used exons; and (3) characterize the effects of somatic mutations and DNA methylation at these binding sites on aberrant alternative splicing. Some details are provided below. (1) The student will use the DEXSeq [3] Bioconductor package on the RNA-seq data from normal and cancer samples to detect differentially used exons. This analysis will be performed on cohorts of samples from TCGA and/or ICGC for which RNA-seq, DNA methylation, and somatic mutations are available. (2) In the second step of the project, the student will use our UniBind database of high confident direct TF-DNA interactions [4] and TF binding analyses to highlight binding sites for CTCF and TFs that could be associated with alternative splicing. A strategy similar to what was used by Ruiz-Velasco et al. [1] for CTCF will be implemented. (3) In the final step of the project, the candidate will combine information from (1) and (2) to overlay somatic mutation and DNA methylation at TF binding sites with the observed alternative splicing events in cancer patients. This project will consolidate the student's knowledge in computational biology for the analysis of genomics data with a focus on gene expression regulation and cancer. Moreover, the student will get familiar with large cancer genomics public data sets available at TCGA and/or ICGC. She/he will learn how to develop computational workflow to analyze large-scale data, such as differential exon usage and differential methylation analyses. The student will also be exposed to different programming languages such as R, Python, and Bash.

References:

1. Ruiz-Velasco M, Kumar M, Lai MC, Bhat P, Solis-Pinson AB, Reyes A, et al. CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. Cell Syst. 2017;5: 628–637.e6. 2. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011;479: 74–79. 3. Reyes A, Anders S, Huber W. Inferring differential exon usage in RNA-Seq data with the DEXSeq package. 2013. Available: https://bioconductor.riken.jp/packages/3.5/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.pdf 4. Gheorghe M, Sandve GK, Khan A, Chèneby J, Ballester B, Mathelier A. A map of direct TF–DNA interactions in the human genome. Nucleic Acids Res. 2018;47: e21–e21.

Expected skills::

Proficiency in Python, R, and/or bash; Previous experience in genomics data analysis; Team spirit; English proficiency

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 2020-2021

Personal Information

Supervisor	Renee Beekman in collaboration with Francois Serra and Alfonso Valencia
Email	renee.beekman@crg.eu; francois.serra@bsc.es; alfonso.valencia@bsc.es
Institution	Centre for Genomic Regulation (CRG) in collaboration with the Barcelona Supercomputing Center (BSC)
Website	https://www.crg.eu/en/programmes-groups/beekman-lab; https://www.bsc.es/discover-bsc/organisation/scientific-structure/computational-biology
Group	Single Cell Epigenomics and Cancer Development (CRG) in collaboration with Computational Biology Life Sciences (BSC)

Computational genomics

Project Title:

Unveiling differences between the 3D chromatin structure of homologous chromosomes in the context of translocations in tumour cells

Keywords:

3D Chromatin Structure, Chromosome/Allele-specific Read Mapping, De Novo Genome Assembly, Normal Cells and Tumour Cells, Lymphoma.

Summary:

Each mammalian cell has two copies of each chromosome. For the vast majority of computational analyses, the two copies of each chromosome are combined. However, it is generally known that many biological processes can show differences between the two copies of the chromosomes. A specific gene can for example be expressed from only one of the two chromosomes (allele-specific expression). Or, genetic mutations in tumour samples occur on only one of the two chromosomes (heterozygous mutations). In our group, we aim to distinguish information of the different copies of the chromosomes to better understand the development of cancer. We will do this in the context of lymphomas, which are tumours that originate from normal immune cells. While the chromosomes are considered linear structures, they are actually folded at the three-dimensional (3D) level in a highly organised way (Dekker et al. Nat Rev Genet. 2013). This organisation is needed for the chromosomes to regulate gene expression. Importantly, in lymphoma cells the 3D chromatin structure is altered in comparison to normal cells (Vilarrasa-Blasi & Soler-Vila et al. BioRxiv 2019). In our group, we aim to study the 3D chromatin structure in lymphoma cells in comparison to normal cells. More specifically, we study how genetic translocations (=a piece of one chromosome fuses to another chromosome) in lymphoma cells affect the 3D chromatin structure. In this project, we will use Hi-C data generated to study the 3D chromatin structure in normal cells (Rao et al. Cell 2014) and lymphoma cells (Vilarrasa-Blasi & Soler-Vila et al. BioRxiv 2019 and unpublished data). Hi-C is a molecular technique coupled to next generation sequencing that allows to reconstruct the 3D folding of the genome in the nucleus (Lieberman-Aiden et al. Science 2009). First, we will computationally separate the two homologous copies of chromosome 14. We focus on this chromosome as in lymphomas one of the copies of chromosome 14 is affected by a genetic translocation we aim to study. More specifically, we will use the variation in the genetic code between the two copies of chromosome 14 to distinguish them. To that end, we will use the genomic sequencing data of this chromosome in these samples and perform a de novo assembly to create a reference sequence for the two copies separately. Next, we will use this reference sequence to map the Hi-C reads representing the 3D chromatin structure to one or the other chromosome. Finally, we will reconstruct the 3D-chromatin structure using these separated sets of reads in TADbit (Serra et al. PLoS Comput Biol. 2017). From these reconstructed copies of chromosome 14 we will analyse the differences in the 3D chromatin structure in order to understand the effect of this genetic translocation specific to lymphoma on the 3D chromatin landscape surrounding it. What will you learn: • Computational biology: basics on network analysis; collaborative software development using GIT; to design and use of computational pipelines for high performance computing (in the 30th most powerful supercomputer in the world). • Structural Genomics: to process and analyse data from Chromosome Conformation Capture techniques (mostly Hi-C and Capture-C). • Genomics and Epigenomics: to explore available data at the interface of genomics and epigenomics; to understand the basics of gene regulation mechanisms and to postulate hypotheses about deregulation of these mechanisms in cancer and test them by analysing the data. • Tumour Biology: to understand the genetic and epigenetic mechanisms underlying the development of lymphomas. • Scientific Dissemination: to present in lab meetings and to write a research article resulting from your work.

References:

Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013 Jun;14(6):390-403. doi: 10.1038/nrg3454. Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M. V, Ragoczy, T., Telling, A., et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 2009 Jul; 326, 289–93. doi: 10.1126/science.1181369 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec 18;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021. Serra F, Baù D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol. 2017 Jul 19;13(7):e1005665. doi: 10.1371/journal.pcbi.1005665. Vilarrasa-Blasi R, Soler-Vila P, Verdaguer-Dot N, Russiñol N, Di Stefano M, Chapaprieta V, Clot G, Farabella I, Cuscó P, Agirre X, Prosper F, Beekman R, Beà S, Colomer D, Stunnenberg HG, Gut I, Campo E, Marti-Renom MA, Martin-Subero JI. Dynamics of genome architecture and chromatin function during human B cell differentiation and neoplastic transformation. bioRxiv 764910

Expected skills::

A strong background in UNIX command line tools as well as in python or R programming, in combination with creative thinking and enthusiasm to work in a multi-disciplinary team with wet lab and bioinformatic experience.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

Our lab in the CRG can be divided into two branches: a wet lab branch and a bioinformatic branch. A key aspect of our group is that these two branches are intermingled, whereby the different team members can interact on a day-to-day basis and during weekly lab meetings. On top of that, most team members will have shared wet lab-bioinformatic projects. Moreover, this project in particular will be conducted in collaboration with the lab of Alfonso Valencia in the Barcelona Supercomputing Center, giving a significant boost in computational power and bioinformatics expertise. We strongly believe that both wet lab and bioinformatic analyses and especially the interaction between these two fields are critical to better understand biological phenomena.

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Laura Isús
Email	laura.isus@genomcore.com
Institution	Made of Genes (Genomcore)
Website	https://genomcore.com, https://madeofgenes.com
Group	Bioinformatics Unit

Project

Computational genomics

Project Title:

Design and development of bioinformatic tools for precision medicine

Keywords:

Precision medicine, Computational Genomics, data integration, Report automatization

Summary:

Genomcore/Made of Genes (https://genomcore.com, https://madeofgenes.com) is a company founded in 2015 with the objective to allow the effective implementation of precision medicine in healthcare. We developed a unique B2B technological framework designed to manage large volumes of personal, health-related, highly sensitive biomedical and health data aimed to diagnosis laboratories and healthcare providers. We also feature a packetized B2C/B2B2C personalized healthcare solution that combines genomic and metabolic analysis in a single test. Our innovative solutions are recognized worldwide through different international awards, such as MIT Technology Review Innovators Under 35, Dubai Future Accelerators or Seal of Excellence of the European Commission. If you want to join a unique fast-growing, high-potential, trend-making company, this is your chance. We are looking for a talented and motivated Bioinformatics Student to collaborate with our bioinformatics team in the research, design and development of new bioinformatic tools for precision medicine. The project will entail processing, quality and annotation pipelines for omics datasets. Data visualization and report generation for prevention, diagnostic or treatment recommendations. The project will allow the student to participate and learn from a real setting and a selection of activities aimed to complete the researchers' career development. The position will be located in our Esplugues de Llobregat (Barcelona) offices.

Expected skills::

1) Experience working in Linux environments (Unix tools, Bash scripting, SSH, Unix filesystem...). 2) Experience in scripting language (Python is preferred). 3) Knowledge of general genetics and genetic inheritance. 4) Academic training in both Computer Sciences and Life Sciences (ie, Degree + Master in course). 5) Knowledge of tools for manipulating NGS data (BWA, Samtools, GATK, etc). 6) Experience using public databases (ClinVar, dbSNP, Reactome, OMIM, GO, PharmGKB, etc) will be valued. 7) Fluency in spoken and written English. 8) Fluency in Spanish or Catalan is a plus. Knowledge of other languages is also valued.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

Gross academic aid of 300€ / month

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Ichiro Hiratani
Email	ichiro.hiratani@riken.jp
Institution	RIKEN Center for Biosystems Dynamics Research (RIKEN BDR)
Website	https://www.bdr.riken.jp/en/research/labs/hiratani-i/index.html

Group	Laboratory for Developmental Epigenetics

Project

Computational genomics

Project Title:

Integrative nucleome analysis of genome-wide scRepli-seq and Hi-C datasets to explore the 3D genome architecture

Keywords:

3D genome organization, 4D nucleome, Hi-C, scRepli-seq, NGS

Summary:

We welcome students with bioinformatics skills who have a keen interest in 3D genome architecture (4D nucleome) through integrative analysis of genome-wide NGS datasets derived from single-cell Repli-seq (scRepli-seq) and Hi-C experiments. We are looking for curator-type bioinformatics students who are responsible for all the work related to data resources and data integration, described as the "second category" bioinformatician in the following link. https://bitesizebio.com/38236/how-to-become-a-bioinformatician/

Expected skills::

(1) Programming skills for analyzing genomic data (unix/python/R/Perl), (2) Statistics background, (3) Basic molecular biology background

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Rosa Trobajo & David Mann
Email	rosa.trobajo@irta.cat
Institution	IRTA
Website	http://www.irta.cat/ca/grup/aigues-marines-i-continentals/
Group	Aigües Marines i Continentals

Project

Computational genomics

Project Title:

Functional evaluation of HTS reads using protein sequence data

Keywords:

HTS, protein sequence, environmental DNA, microalgae

Summary:

High Throughput Sequencing (HTS) is currently being developed as a substitute for traditional methods in biomonitoring aquatic ecosystems (e.g. for the European Union Water Framework Directive). For example, methods involving counting cells of microscopic algae are being replaced by metabarcoding: a short region of the gene (rbcL) coding for the large subunit of RuBisCO (ribulose-1,5-bisphosphate carboxylase/oxygenase, the key enzyme of photosynthesis) is amplified from environmental samples and sequenced by Illumina; the reads are processed through a bioinformatics pipeline, where they are filtered to remove sequences containing errors, and then identified to species by reference to a database of known ‘barcodes’ from Sanger sequencing. Although these pipelines work adequately, some of the methods used to reject ‘faulty’ sequences are crude. For example, sequences that occur only rarely in a dataset are often rejected because of the error rate in Illumina sequencing, which leads to incorrect nucleotides occurring ±anywhere along a sequence. A threshold is therefore set, that a particular sequence must be observed twice or more times before it is accepted as real, on the basis that any particular error during sequencing is unlikely to occur many times repeatedly. The proportion of reads rejected on this basis can be of the order of 50% and probably involves many type II errors in the quantification of diversity. The suggested project would involve developing a method that is able to assess HTS reads, even when the sequences have never been encountered before (and are therefore not in the reference database), by taking account of the fact that RuBisCO (like all proteins) has a function and that function can only be performed if the protein folds in the correct way. Hence certain changes in the DNA coding for RuBisCO are evolutionarily ‘easy’ (because they have no effect on protein function, e.g. many codon 3rd position changes), whereas others are strongly or fully constrained. The project aims to determine the likelihood of different changes at a particular DNA site by evaluating variation among known rbcL gene sequences in the group of organisms being studied and also considering RuBisCO structure (which is well-known), and to use this information to develop an ‘intelligent filter’ for metabarcoding pipelines. In this, reads would be evaluated on the basis of whether they code for biologically plausible peptides, rather than solely on the basis of their frequency. Though applied to rbcL, the approach developed could be applicable with appropriate modification to any coding sequence used for metabarcoding (e.g. the CO1 gene used to barcode animals).

Expected skills::

Bioinformatics pipeline development, protein structure prediction, programming, sequence alignment

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Comments:

This project would require a combination of skills from different areas of specialization in the syllabus, mainly from computational genomics and structural bioinformatics, and include the need for some programming.

Master project 2020-2021

Personal Information

Supervisor	Marta Melé
Email	marta.mele@bsc.es
Institution	Barcelona Supercomputing Center
Website	https://www.bsc.es/discover-bsc/organisation/scientific-structure/transcriptomics-and-functional-genomics-lab-tfgl
Group	Trancritomics and functional Genomics

Project

Computational genomics

Project Title:

Understanding individual variation in splicing in human populations

Keywords:

Transcriptomics, differential gene expression, human populations, splicing, ribosome profiling, posttranscriptional processing, RNA binding proteins.

Summary:

The candidate will join Marta Melé's Transcriptomics and Functional Genomics lab in the Life Sciences Department at the Barcelona Supercomputing Center. The lab is interested in understanding how individual variation in gene expression can explain phenotypic differences between individuals both in the context of health and disease. To address this question, we use large-scale transcriptomic analysis and latest single-cell sequencing technologies combined with methods development to study gene expression, splicing and cell type composition variation across human tissues and phenotypes. In this project, we will perform a large-scale analysis of splicing variation between individuals with different phenotypes and from different ethnic groups. In previous studies, we observed that variation in splicing may play in contrast a comparatively greater role in defining individual phenotypes than variation in gene expression. contributes more to individual variation than to changes in gene expression (Melé et al. Science 2015). Moreover, we observed an enrichment of specific genes showing large splicing variation between individuals that was

especially strong for ribosomal proteins and that will be explored further. Ultimately, in this project we will explore in depth what is the role of splicing in defining why human individuals are different from one another. What you will learn: Development of computational pipelines to analyze and interpret large omics datasets such as RNA-Seq, single-cell RNA-seq, ribosome profiling, and CLiP-seq). Working in a high performance computing (HPC) environment. Effective communication of research findings, scientific writing, critical thinking.

References:

Melé, M. et al. The human transcriptome across tissues and individuals. Science (80-.). 348, 660–665 (2015).

Expected skills::

Availability to start in July 2020 is preferred Strong programming skills in bash, python, R, perl, or similar, Some experience working in HPC clusters Some experience with Next Generation Sequencing data analysis Excellent communication skills in spoken and written English Capacity to contribute to research projects with novel research ideas and analysis Capacity to work as a team in a highly collaborative and diverse environment

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Master project 2020-2021

Personal Information

Supervisor	Yasushi Okada
Email	y.okada@riken.jp
Institution	RIKEN, BDR
Website	https://www.bdr.riken.jp/en/research/labs/okada-y/index.html
Group	Laboratory for Cell Polarity Regulation

Project

Computational genomics

Project Title:

Decoding genome by imaging

Keywords:

protein engineering, image processing, machine learning, NGS data analysis

Summary:

We are developing technologies to estimate the epigenetic state of the cell from the super-resolution live cell imaging. The project includes the following four sub-projects, and the intern student can choose one according to his/her interest. 1) Development of the prove for visualization, which includes structure-based designing of the mutant protein probes, imaging of the designed probes, and analysis of the binding sites in the genome by genome-wide sequencing. 2) Development of the program for automation of the microscope system, which includes automatic search for the cell by deep learning. 3) Image processing, which includes denoising, regularization, and quantification, through the combination of the traditional algorithms and machine learning. 4) Development of the computational models to link the image data to the sequence data.

Expected skills::

no

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Manuel Irimia
Email	mirimia@gmail.com
Institution	Centre for Genomic Regulation

Website	https://www.crg.eu/manuel_irimia
Group	Transcriptomics of vertebrate development and evolution

Project

Computational genomics

Project Title:

The role of alternative splicing in the evolution of animal tissue diversity

Keywords:

Alternative splicing, microexons, exon-intron evolution, tissue-specific regulation, parallel evolution.

Summary:

Why evolution of alternative splicing? Alternative splicing (AS) is a molecular process allowing multiple transcripts to arise from the same gene. The power of AS in expanding the functional potential of a gene is well exemplified by the axon guidance receptor Dscam in Drosophila melanogaster. Dscam produces ~38000 distinct transcripts, which uniquely fine-tune the function of the gene in different neurons. AS contributes to functional diversity not only within cell populations but also between cell types/tissues. Our lab is currently investigating the role of AS in the evolution of animal tissue diversity: here we propose a parallel analysis of AS evolution in vertebrates and insects, two monophyletic clades in the bilaterian tree. Set up: we inferred exon orthology groups in a set of 20 bilaterian species (8 vertebrates, 8 insects and two pairs of relative outgroups) and we assembled a comprehensive RNA-seq dataset covering 8 homologous tissues in all species. We used the RNA-seq data to identify AS exons within each species and tissue. Experimental design: the project will be divided into three main parts: 1) We will investigate evolutionary patterns involving the entire tissue AS landscapes. Preliminary results show that neural and muscle AS networks seem to be well conserved in vertebrates but not in insects, suggesting different rewiring rates between the two clades. 2) We will focus on the exons specifically spliced within each tissue. Many tissue-specific exons have acquired tissue-specific regulation millions of years after their birth. An exciting perspective is the identification of a causal relationship between changes in exon regulation and simultaneous phenotypic innovation/adaptations. 3) We will explore the regulatory mechanisms underlying the rise of tissue-specific AS. The master project will be developed as part of this bigger project on exon evolution. The student will become familiar with the principles of alternative splicing and gene regulation, while getting hands-on experience with genome annotations, RNA-seq data analysis, comparative transcriptomics, and network reconstruction.

References:

- Torres-Méndez, A., Bonnal, S., Marquez, Y., Roth, J., Iglesias, M., Permanyer, J., Almudí, I., O’Hanlon, D., Guitart, T., Soller, M., Gingras, A.-C., Gebauer, F., Rentzsch, F., Blencowe, B.J.B., Valcárcel, J., Irimia, M. (2019). A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons. Nature Ecol Evol, 3:691-701. - Marletaz, F., Firbas, P., Maeso, I., Tena, J.J., Bogdanovic, O., Perry, M., Wyatt, C.D.R., [+50 authors], Holland, P.W.H., Escriva, H., Gomez-Skarmeta, J.L., Irimia, M. (2018). Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature, 564:64-70. - 6) Burguera, D., Marquez, Y., Racioppi, C., Permanyer, J., Torres-Mendez, T., Esposito, R., Albuixech, B., Fanlo, L., D'Agostino, Y., Gohr, A., Navas-Perez, E., Riesgo, A., Cuomo, C., Benvenuto, G., Christiaen, L.A., Martí, E., D'Aniello, S., Spagnuolo, A., Ristoratore, F., Arnone, M.I., Garcia-Fernàndez, J., Irimia, M. (2017). Evolutionary recruitment of flexible Esrp-dependent splicing programs into diverse embryonic morphogenetic processes. Nat Commun, 8:1799.

Expected skills::

Ideally, experience on RNA-seq analyses and/or comparative genomics. Interest on genome evolution.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Mar Albà
Email	mar.alba@upf.edu
Institution	IMIM-UPF
Website	evolutionarygenomics.imim.es
Group	Evolutionary Genomics, Research Programme on Biomedical Informatics

Project

Computational genomics

Project Title:

Gene expression dysregulation in cancer and neoantigen formation

Keywords:

Transcriptomics; cancer; small ORFs; neoantigens; immunotherapy.

Summary:

In cancer, genomic structural rearrangements and mutations result in the expression of many novel transcripts that are not expressed in normal conditions. Recent studies suggest some of these transcripts translate peptides that can be presented by MHC molecules and be an important source of neoantigens. Such neoantigens are non-self proteins and could thus trigger a potent immune response and be very relevant for immunotherapy approaches to fight against cancer. However, the lack of studies measuring novel transcriptional events in cancer prevents us from fully understanding the contribution of these neoantigens. The aim of the project will be to perform transcriptome assembly directly from RNA-Seq data using large publicly available cancer cell datasets. In the group we have previously employed massive transcriptomics data to identify recently originated transcripts in human and mouse and predict any encoded protein products (Ruiz-Orera et al., 2015; Ruiz-Orera et al., 2018). Here we will use similar techniques to identify novel, non-annotated, transcripts in cancer cell RNA-Seq data and to characterize putative neoantigens.

References:

Ruiz-Orera, J., Hernández-Rodríguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., Albà, M.M. (2015). Origins of de novo genes in human and chimpanzee. Plos Genetics, 11(12): e1005721. Ruiz-Orera, J., Grau-Verdaguer, P.,

Villanueva-Cañas, J-L., Messeguer, X., Albà, M.M. (2018). Translation of neutrally evolving peptides provides a basis for de novo gene evolution. Nature Ecology and Evolution, 2:890–896.

Expected skills::

Interest in computational genomics and transcriptomics; knowledge of a programming language; knowledge of R; good command of English.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Miquel Angel Pujana
Email	mapujana@iconcologia.net
Institution	Catalan Institute of Oncology IDIBELL
Website	http://ico.gencat.cat/en/recerca/Programa-ProCURE/index.html
Group	Cancer Resistance & Bioinformatics

Project

Computational genomics

Project Title:

Discovery of “Dr Jekyll & Mr Hyde” genes

Keywords:

Cancer, genetics, outcome, tumor suppressor, oncogene

Summary:

Integration of genomic and clinical information using the Cox proportional hazard model is commonly used to identify biological factors that influence cancer outcome. Typically, this is applied to analyze the connection between gene expression and cancer progression, therapeutic response or patient survival. This approach has generated hundreds of biomarkers, of which several are nowadays applied in the clinic. However, some genes might not show a single facet during the course of the disease: they can act as tumor suppressors or as oncogenes depending on other variables (so called “Dr Jekyll and Mr Hyde” genes). These genes, their features and impact on cancer outcome remain completely unknown. Objectives In this project, we aim to identify this type of genes by pan-cancer interrogation of gene expression and clinical outcomes. This proposal will be integrated into experimental assays performed at the recipient group.

References:

An Integrated TCGA Pan-Cancer Clinical Data Resource to drive high quality survival outcome analytics. Cell. 173, 2: p400-416.e11, 10.1016/j.cell.2018.02.052 (2018). Kourou et al., Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13, 8–17 (2015).

Expected skills::

Candidate(s) are expected to be proficient in programming in R and to have strong background on statistics.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Oriol Dols Icardo and Jordi Clarimón
Email	odols@santpau.cat and jclarimon@santpau.cat
Institution	Sant Pau Biomedical Research Institute
Website	http://santpaumemoryunit.com/

Group Genetics of Neurodegenerative Diseases Unit

Project

Computational genomics

Project Title:

Deep transcriptome characterization of the frontal cortex of frontotemporal lobar degeneration patients

Keywords:

Neurodegenerative disease; RNA sequencing; Transcriptome; Human brain; RNA alterations

Summary:

Frontotemporal lobar degeneration (FTLD) is a neuropathological term for a group of neurodegenerative dementias, mainly characterized by the aberrant deposition of TDP-43 (FTLD-TDP) or tau (FTLD-tau) proteins in the frontal and temporal lobes. Dysfunction of the RNA metabolism has proven to be one of the major pathological hallmarks of FTLD. In order to investigate RNA alterations in FTLD human brains, we have performed high-throughput RNA sequencing (encompassing total and small RNA) to deeply characterize the transcriptome of the frontal cortex of 12 FTLD-tau, 20 FTLD-TDP and 10 healthy controls. In this project, bioinformatics tools will be applied in order to disentangle gene and isoform differential expression, gene co-expression networks and alternative splicing events associated with FTLD which will be integrated with small RNA sequencing data from the same individuals. Finally, cell-type deconvolution algorithms using human single-nucleus RNA sequencing data will be applied to disentangle cellular heterogeneity in FTLD. Since this approach has not yet been performed in this neurodegenerative disorder, outcomes from this study will have a very high potential to be published in specialized journals.

Expected skills::

Linux/Ubuntu, R and python

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Personal Information

Supervisor	Mar Albà
Email	mar.alba@upf.edu
Institution	IMIM-UPF
Website	evolutionarygenomics.imim.es
Group	Evolutionary Genomics, Research Programme on Biomedical Informatics

Project

Computational genomics

Project Title:

Building transcriptomes with Nanopore sequencing data

Keywords:

Transcriptomics; transcript discovery; long reads; Nanopore; gene expression.

Summary:

Long read sequencing techniques, such as Oxford Nanopore Technologies (ONT), have great potential to sequence complex transcriptomes and to discover new transcripts beyond the annotated ones. While methods that work with Illumina short reads are quite mature, the development of software to work with Nanopore RNA sequencing reads is a very active area of research. We and co-workers have recently developed a method to build transcriptomes from RNA-derived Nanopore reads (cDNA and direct RNA) that does not require a reference genome and that could be used to investigate highly rearranged genomes (such as those in cancer cells) or species that currently lack a sequenced genome (de la Rubia et al., 2020). The aim of the project will be to compare methods based on Nanopore or Illumina reads for building eukaryotic transcriptomes in the absence of a reference genome and to identify novel, non-annotated, transcripts in species that already have a genome and reference annotations. We would like to determine when it is more convenient to use one sequencing technology over the other one, and if the combination of the two tecnologies – using Illumina reads to correct errors in Nanopore reads – is a real advantage. For this we will use already available datasets for yeast, human and mous especies, as well as datasets that are currently being generated in the group.

References:

de la Rubia, I., Indi, J.A., Carbonell, S., Lagarde, J., Albà, M.M., Eyras, E. (2020). Reference-free reconstruction and quantification of transcriptomes from long-read sequencing. bioRxiv, https://doi.org/10.1101/2020.02.08.939942

Expected skills::

Interest in computational genomics and transcriptomics; knowledge of a programming language; knowledge of R; good command of English.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Master project 2020-2021

Personal Information

Supervisor	François Serra
Email	francois.serra@bsc.es
Institution	BSC - Barceona Supercomputing Center
Website	https://www.bsc.es/
Group	Computational Biology

Project

Computational genomics

Project Title:

A dynamic epigenetic network

Keywords:

epigenetics; networks; bioinformatics

Summary:

The laboratory of Alfonso Valencia specializes in many areas of computational biology focusing on the integration of data and development of computational frameworks that could help bridge the gap between fundamental research and personalized medicine. One of the areas of expertise of the lab is epigenomics, producing pioneer studies in the field. The project will be mainly supervised by François Serra an expert in epigenetics and chromatin conformation (coauthor of the works mentioned below on leukemia and cell differentiation), with experience in mentoring master and PhD students. The project we propose here is based on the work of Vera Pancaldi that built an epigenetic network based on the 3D conformation of the chromatin (Pancaldi et al. 2016). We aim to reproduce this work using more genomic interaction data (Davies et al. 2017) and to study it in the dynamic models of leukemia and cell differentiation (Beekman et al. 2018) or cell dedifferentiation (Stadhouders et al. 2018). Concretely the student will work on the development of a computational pipeline to discover interactions between DNA binding proteins and epigenetic marks. Finally, the data will be represented as a network of interactions to be analyzed at different time stages. We expect to be able to understand the functional association between the different actors in the epigenetic landscape and to understand the dynamics behind its remodeling upon disease or in development. What you will learn: - Computational biology: basics on network analysis; collaborative software development using GIT; design and use of computational pipelines for high performance computing (in the 30th most powerful supercomputer in the world). - Epigenomics: explore available data in the interface between genomics and epigenomics, postulate hypotheses about the mechanisms of gene regulation, and analyze the results. - Scientific Dissemination: to present in lab meetings and to write a research article resulting from your work.

References:

Beekman, R., Chapaprieta, V., Russiñol, N., Vilarrasa-Blasi, R., Verdaguer-Dot, N., Martens, J.H.A., Duran-Ferrer, M., Kulis, M., Serra, F., Javierre, B.M., Wingett, S.W., Clot, G., Queirós, A.C., Castellano, G., Blanc, J., Gut, M., Merkel, A., Heath, S., Vlasova, A., Ullrich, S. and Martin-Subero, J.I. 2018. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nature Medicine 24(6), pp. 868–880. Davies, J.O.J., Oudelaar, A.M., Higgs, D.R. and Hughes, J.R. 2017. How best to identify chromosomal interactions: a comparison of approaches. Nature Methods 14(2), pp. 125–134. Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B.M., Juan, D., Fraser, P., Spivakov, M., Valencia, A. and Rico, D. 2016. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biology 17(1), p. 152. Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., Hecht, J., Filion, G.J., Beato, M., Marti-Renom, M.A. and Graf, T. 2018. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. Nature Genetics 50(2), pp. 238–249.

Expected skills::

1- Critical thinking and creativity 2- Good statistical and programming skills (R/Bioconductor or Python) 3 - Basic knowledge of molecular biology 4- Ability to access and evaluate scientific literature

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

This project is a follow up of the work currently being done by a UPF master student. The new student will benefit from the results and analysis generated until now. It is also wide enough, biologically and methodologically, to leave room for the student to decide in which direction she/he would prefer to go in either way. The project will be supervised by François Serra and co-supervised by Alfonso Valencia.

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Ramiro Logares
Email	ramiro.logares@icm.csic.es
Institution	ICM - CSIC
Website	http://www.log-lab.barcelona

Group	Ecology of Marine Microbes & log-lab

Project

Computational genomics

Project Title:

Population dynamics and evolution of the ocean microbiome

Keywords:

microbiome, ocean, metagenomics, evolution, ecology

Summary:

The global ocean and the tiny organisms it contains are crucial for global ecosystem function. Microbial phytoplankton in the ocean fix as much carbon from the atmosphere as land plants, and other heterotrophic microbes guarantee that most of the fixed carbon is circulated through food webs. The genomic machinery that marine microbes use for performing a myriad of metabolic processes remained unknown until ca. 15 years ago, when large-scale DNA sequencing projects became feasible. With the advent of high-throughput DNA sequencing, we started unveiling the ocean microbiome at unprecedented levels of detail. During the last 5 years, very large genomic datasets have been extracted from the global ocean microbiome. In particular, the global expeditions TARA-Oceans (https://oceans.taraexpeditions.org) and Malaspina (http://www.expedicionmalaspina.es) have produced a goldmine of genomic data that we are continuously explored. This data is the best representation we have of the diversity and function of marine microbes, and considers mostly metagenomes and metatranscriptomes (ca. 30 Terabytes of compressed DNA data). My group at the ICM-CSIC (log-lab http://www.log-lab.barcelona at the EMM https://emm.icm.csic.es) is involved in both global marine expeditions. The proposed project aims at interrogating these datasets in order to 1) determine the population variation of selected microbes (using mutations; a.k.a. SNPs or Single Nucleotide Polymorphisms) in the global ocean and 2) find out whether some of the previous variation is due to evolutionary processes that occurred relatively recently in geological time. For investigating the above, we will build metagenome-assembled genomes (MAGs) and then map metagenomic or metatranscriptomic reads from the global ocean to a number of selected MAGs. Afterwards, we will perform a SNP calling analysis, aiming to determine fine-grained genomic variation. The analysis of these SNPs is what will indicate how much variation is present in the selected microbial populations and whether part of this variation has emerged through adaptive evolution. Most analyses for this work will be performed at our marine bioinformatics platform Marbits https://marbits.icm.csic.es

References:

Sunagawa, S., et al., Structure and function of the global ocean microbiome. Science, 2015. 348(6237): p. 1261359. Carradec, Q., et al., A global ocean atlas of eukaryotic genes. Nat Commun, 2018. 9(1): p. 373. Logares, R., et al., Disentangling the mechanisms shaping the surface ocean microbiota. 2020. Microbiome. In press. https://www.researchsquare.com/article/rs-7862/v2 Falkowski, P., The power of plankton. Nature, 2012. 483(7387): p. S17-20. Alberti, A., et al., Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci Data, 2017. 4: p. 170093. de Vargas, C., et al., Eukaryotic plankton diversity in the sunlit ocean. Science, 2015. 348(6237): p. 1261605.

Expected skills::

Proficiency with bash and R. Familiar with python.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

motivation, interest to learn new bioinformatics techniques and to work in clusters

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Shigehiro Kuraku
Email	shigehiro.kuraku@riken.jp
Institution	RIKEN BDR
Website	https://www.bdr.riken.jp/en/research/labs/kuraku-s/
Group	Laboratory for Phyloinformatics

Project

Computational genomics

Project Title:

Elucidating the rules of genomic scaling: how does the size of DNA regions influence their physiological output?

Keywords:

genomic scaling, rate of living, c-value paradox, longevity

Summary:

Genome sizes exhibit a remarkable variation even among vertebrate animals. This project is assumed to be conducted only with computational solutions and aims at understanding the effect of variable physical spacing between exons and genes in animal genomes, by investigating which portion of the genomes are susceptible to the variation, with an emphasis on genes responsible for physiological controls.

References:

Hara et al. Nat Ecol Evol, 2018 2:1761- (https://www.nature.com/articles/s41559-018-0673-5) and Kowalczyk et al., eLife 2020 9:e51089 (https://elifesciences.org/articles/51089)

Expected skills::

Basic skills of programming, basic knowledge of molecular biology

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Tomas Marques-Bonet
Email	tomas.marques@upf.edu
Institution	UPF
Website	http://biologiaevolutiva.org/tmarques
Group	Comparative Genomics

Project

Computational genomics

Project Title:

Population genomics and conservation for whole genomes of 200 species of primates

Keywords:

Whole genomes DNA sequencing, Population genetics, variant calling, admixture

Summary:

Genomic diversity is at the core of many evolutionary inferences. The finer study of primates, our closest relatives, is relevant for several reasons. They are the only living organisms with whom we share a higher proportion of genetic material as we have a shared evolutionary history over time. Thus, studying the genetics of the primates is a necessary endeavour to define the similarities among primates, the uniqueness of humans, and to strengthen the foundations of primate management and conservation. The latter of which should be an international effort, as these species should be considered a treasure of humanity. In the recent years, we have shown that it is possible to study full genome information from apes (Prado-Martinez et al. Nature 2013, Xue et al. Science 2015; deManuel et al. Science 2016; Nater et al. Current Biology 2017). Considering the population decline that all primates are experiencing, it is time-sensitive to act rapidly and generate the global dataset of variation for all primates. We have generated high quality full genome information for a large panel of primates all over the world. By using samples from the wild, we will further elucidate the role of demography, admixture and selection on genome diversity. In so doing, fundamental insights will be gained into the study of primates with multiple ramifications to biology.

References:

Prado-Martinez et al. Nature 2013, Xue et al. Science 2015; deManuel et al. Science 2016; Nater et al. Current Biology 2017

Expected skills::

Programming, population genetics.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Sarah Djebali
Email	sarah.djebali@inserm.fr
Institution	IRSD, INSERM U1220
Website	www.en.irsd.fr
Group	Genetic and regulation of iron metabolism

Project

Computational genomics

Project Title:

Bioinformatics methods for the identification of enhancer/gene relationships in vertebrate genomes

Keywords:

enhancer/gene regulatory relationships; high-through functional sequencing data; chromatin structure; program evaluation; machine learning

Summary:

For many complex genetic diseases, the majority of the identified variants are located outside protein-coding genes [1], making it difficult to understand their function. And when variants are located far away from any gene, they are usually assumed to act on the nearest gene, which can often prove totally wrong [2]. The regulatory element that can explain this long distance action of the variant on the gene is the enhancer. Enhancers are genomic regions on which transcription factors bind, and which activate the expression of one or several genes by being brought close to (in 3D) the uptream regulatory elements (promoters) of those genes. Enhancers can therefore be far away from the genes they activate on the 1D genome, but being close to them in the 3D space of the nucleus. Today the best approaches to identify enhancer/gene relationships in the genomes are genetic screening [3] and targeted chromatin structure (3D), such as polymerase II ChIA-PET [4] or promoter capture HiC [5]. The problem is that the first one can only targets a handful of genes and the second one is very difficult and costly to generate. For this reason and because many international consortia such as ENCODE, FANTOM or Epigenome Roadmap have recently produced and made publicly available large quantities of functional 1D data (such as RNA-seq, ATAC-seq, histone marks or methylation data), the favoured approach is the integration of high-throuput functional 1D data. Although many programs exist to identify enhancer/gene relationships from functional 1D data [6,7,8], there is no consensus about what the best approach is. Here we would like to fill in this gap by assessing the different existing methods on reference sets and proposing a new method that uses a minimal amount of different data. The student will therefore have to: - Make a complete state-of-the-art of the existing 1D methods - Plan the evaluation * Define reference sets * Define criteria to include programs in the evaluation * Define the input data to use for each program to evaluate - Make the programs to evaluate work on small and real evaluation datasets - Determine the best approach and propose a new one that uses as few different input data as possible

References:

[1] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009 Jun 9;106(23):9362-7. [2] Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, Wei Y. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature genetics. 2017 Nov;49(11):1602. [3] Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA, Nguyen TH. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature Genetics. 2019 Dec;51(12):1664-9. [4] Zhang J, Poh HM, Peh SQ, Sia YY, Li G, Mulawadi FH, Goh Y, Fullwood MJ, Sung WK, Ruan X, Ruan Y. ChIA-PET analysis of transcriptional chromatin interactions. Methods. 2012 Nov 1;58(3):289-99. [5] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature genetics. 2015 Jun;47(6):598. [6] He B, Chen C, Teng L, Tan K. Global view of enhancer–promoter interactome in human cells. Proceedings of the National Academy of Sciences. 2014 May 27;111(21):E2191-9. [7] Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MT, Cheng C, Fan X, Gerstein M, Cheng AS. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nature genetics. 2017 Oct;49(10):1428. [8] Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic acids research. 2019 Jun 4;47(10):e60-.

Expected skills::

Linux command line; Programming skills (bash, awk, python, ...); having already manipulated high-throughput (functional) sequencing data; know the basics of statistics and the R language; know how to run jobs on a cluster; understand written English well

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

PhD funding is not available yet but several options can be envisionned. This question also depends on the success of applications for funds that will be done during the fall.

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Nuria Lopez-Bigas
Email	nuria.lopez@irbbarcelona.org
Institution	IRB Barcelona
Website	bbglab.irbbarcelona.org
Group	Biomedical Genomics

Project

Computational genomics

Project Title:

Understanding cancer biology

Keywords:

Cancer drivers, selective advantage, mutational processes, tumorigenesis

Summary:

A tumor has between hundreds and thousands of mutations and only a few are directly involved in tumorigenesis, frequently called driver mutations. These mutations affect genes which when mutated confer the cell with a growth advantage with respect to its neighbors. Our lab has developed methods to identify these driver genes, and has analyzed tens of thousands of tumors, producing a catalog of the genes underlying tumorigenesis in the most frequent cancer types. Currently, we are interested in cataloguing the downstream effect that mutations affecting these driver genes have in different tumor types. While many mutations in driver genes are capable of driving tumorigenesis, some are not, and the range of driver mutations of a cancer gene varies between tumor types. Understanding the functional effect of driver mutations thus constitutes a key goal of cancer genomics research.

References:

Tamborero et al, 2018. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Medicine. 10:25 Pich et al, 2018. Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. Cell doi:10.1016/j.cell.2018.10.004 Sabarinathan et al., 2016. Nucleotide excision repair is impaired by binding of transcription factors to DNA. Nature 532, 264-267 Mularoni et al, 2016. OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biology. 17: 128 Rubio-Perez et al, 2015. In silico prescription of anti-cancer drugs to cohorts of 28 tumor types reveals novel targeting opportunities. Cancer Cell. 27(3):382-396 Gonzalez-Perez et al, 2013. IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods. doi:10.1038/nmeth.2642

Expected skills::

Basic programming, data analysis and statistics skills. Willing to learn

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 2020-2021

Personal Information

Supervisor	Chaysavanh Manichanh
Email	cmanicha@gmail.com
Institution	Vall d'Hebron Research Institute
Website	https://sites.google.com/site/manichanhlab/ and http://www.vhir.org/portal1/
Group	Microbiome Lab

Project

Computational genomics

Project Title:

Development of bioinformatics and statistical tools to integrate meta-omics data to decipher the human microbiome

Keywords:

Human Microbiome; Metagenomics; Metatranscriptomics; Metabolomics; Composition and functions

Summary:

Meta-omics approaches have been intensively used over the last 20 years to study the composition and functions of the human microbiome (the other Human Genome) in health and disease conditions. The aim of the present work is to develop and/or implement bioinformatics tools to analyze and integrate meta-omics data. • You will work in the dry-lab conducting bioinformatics and biostatistical research. You will be integrated in a young and collaborative environment: medical doctors, nutritionist, molecular biologists, bioinformaticians, statistician. • You will learn from your colleagues, and take responsibility, in writing your conclusions into academic papers, which eventually will be published in High Impact Journals. We want to help you build solid foundations on the research method, so you will be assisted by more experienced colleagues.

References:

https://sites.google.com/site/manichanhlab/our-publications

Expected skills::

Fluent in English (most of our team are foreigners, thus English is our language); Theoretical and practical knowledge of classical statistical inference and Machine Learning; Strong coding experience

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Comments:

We are looking for a motivated student who is seeking to pursue his/her career in research. The candidate will be remunerated 1000 euros/month (gross salary) during his master internship and will be offered the possibility to apply for a PhD fellowship (INPhINIT “la Caixa”, FPU, AGAUR, VHIR...).

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Rory Johnson

Email	rory.johnson@ucd.ie
Institution	University College Dublin
Website	https://www.gold-lab.org/
Group	Laboratory for Genomics of Long Noncoding RNAs and Disease (GOLD Lab)

Project

Computational genomics

Project Title:

CRISPR, Cancer, Noncoding RNAs

Keywords:

CRISPR, cancer, lncRNA

Summary:

One of the biggest biological surprises of the last decade has been the discovery of a completely new class of genes in the human genome – long non-coding RNAs (lncRNAs). These RNA transcripts are not translated into protein, but instead seem to function as regulatory molecules that control the expression of other genes. As part of the international ENCODE consortium, our group has helped catalogue >10.000 of these genes, 99% of which remain completely uncharacterised. LncRNAs represent an extremely promising source of new drug targets. The objective of our lab is to develop a new generation of anti-cancer therapies based on designed lncRNA inhibitors. We identify lncRNA targets via in-house developed, interdisciplinary strategies combining bioinformatics with CRISPR-Cas9 genome-engineering tools. We can offer a variety of tailor-made projects to interested students. These may be, for example, integrative data analysis to make testable predictions about lncRNA functionality, or creation of pipelines for identification of cancer-causing lncRNAs from our CRISPR screens. Previous MSc students have gone on to publish first-author papers based on their MSc thesis, and have successful scientific careers with us and other groups (eg from UPF: Carlevaro-Fita, Pulido-Quetglas, Mas-Ponte, Lanzos – see Pubmed). Students in our lab get exposed to latest bioinformatic and experimental practices on a daily basis. They get closely mentored and have numerous opportunities to present their work internally. Our lab is involved in several international collaborations and consortia, including the International Cancer Genome Consortium (https://www.nature.com/collections/afdejfafdb) and we were recently awarded a prestigious Future Research Leaders grant from the President of Ireland (https://www.sfi.ie/research-news/news/president-higgins-honours/). If you are motivated to work at the forefront in computational cancer genomics and genome-engineering in a fun, supportive and motivated team, then contact us for more information! See also: gold-lab.org https://twitter.com/GOLDLab_Bern https://people.ucd.ie/rory.johnson

References:

Selected recent papers: Rheinbay E... PCAWG Consortium (including Johnson R, Carlevaro-Fita J, Lanzos A) Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020 Feb;578(7793):102-111. Bergadà-Pijuan J, Pulido-Quetglas C, Vancura A, Johnson R#. CASPR, an analysis pipeline for single and paired guide RNA CRISPR screens, reveals optimal target selection for long noncoding RNAs. Bioinformatics. 2019 (In Press) Carlevaro-Fita J, Polidori T, Das M, Navarro C, Zoller TI, Johnson R#. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. Genome Research 2019 Feb;29(2):208-222 Joana Carlevaro-Fita, Rory Johnson#. Global Positioning System: Understanding long noncoding RNAs through subcellular localisation. Molecular Cell 2019 Mar 7;73(5):869-883 Roberta Esposito, Núria Bosch, Andrés Lanzós, Taisia Polidori, Carlos Pulido-Quetglas, Rory Johnson#. Hacking the cancer genome: Profiling therapeutically-actionable long noncoding RNAs using CRISPR-Cas9 screening. Cancer Cell 2019 Apr 15;35(4):545-557. Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R#, Johnson R#. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nature Genetics 2017 Dec;49(12):1731-1740 Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R#. Towards a complete map of the human long non-coding RNA transcriptome. Nature Reviews Genetics 2018 Sep;19(9):535-548.

Expected skills::

Unix, R, python

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Miquel Angel Pujana
Email	mapujana@iconcologia.net
Institution	Catalan Institute of Oncology IDIBELL
Website	http://ico.gencat.cat/en/recerca/Programa-ProCURE/index.html
Group	Cancer Resistance Research & Bioinformatics

Project

Computational genomics

Project Title:

Discovering unexpected cancer-protective effects of common medications

Keywords:

Cancer, therapy, common medication, genetics, GWAS, epidemiology

Summary:

Development of a new cancer-target drug costs hundreds of millions of EUR and on average 10 years of experimental work before approval. However, overall success rate is less than 10%. Thus, drug repurposing is received much attention and new indications of existing drugs are accounting for 20% of new products. Systematic analyses of thousands of developed/approved drug or compounds have found many with previously unrecognized anti-cancer activity. While these evidence mainly derive from in vitro cellular assays, large-scale studies of population-based health care records integrated into genetic information are currently possible. Preliminary data from our group has discovered that certain drugs used for non-cancer common conditions have large protective effects regarding cancer progression and metastasis. In this project, we aim to estimate the beneficial effects of common medications on cancer patient survival by integrating and modeling epidemiological and health care data from two European populations. The effects will be further deciphered at the germline genetic level by meta-analyses of GWASs. This proposal is integrated into experimental assays also performed at the recipient group.

References:

• Bycroft et al., The UK Biobank resource with deep phenotyping and genomic data, Nature 562, 203–209(2018). • Bolivar et al., SIDIAP Database: Electronic Clinical Records in Primary Care as a Source of Information for Epidemiologic Research, Med Clin. 138(14):617-21 (2012). • Pantziarka et al., Hard Drug Repurposing for Precision Oncology: The Missing Link? Front Pharmacol. 9: 637 (2018). • Corsello et al., Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. Nat Cancer doi:10.1038/s43018-019-0018-6 (2020).

Expected skills::

Candidate(s) are expected to be proficient in programming in R and to have strong background on statistics.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat
Pompeu Fabra
Barcelona

Master in
Bioinformatics for
Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Lluís Ribas de Pouplana
Email	lluis.ribas@irbbarcelona.org
Institution	IRB Barcelona
Website	www.irbbarcelona.org
Group	Gene Translation Laboratory

Project

Computational genomics

Project Title:

Identification of domain-specific adaptations for protein synthesis

Keywords:

Protein synthesis, proteome diversity, transfer RNA, modified bases, evolution

Summary:

We have developed tools to detect and analyze protein sequences that are impossible to synthesize for some organisms. Those species that can make this proteins do so thanks to special adaptations ('upgrades') of the protein synthesis machinery. Now we want to further develop and use these tools to map the global proteome landscape and identify all possible protein sequences possibly unique to some group of species thanks to the existance of 'upgrades'. We offer one or two payed (modestly) positions to carry out these analyses. We look for candidates interested in evolution, biochemistry, and confident in the use of R.

References:

References: 1. The mitochondrial tRNA conundrum. (2020) Ribas de Pouplana L. Nat Rev Mol Cell Biol. doi: 10.1038/s41580-020-0220-5. 2. Differential expression of human tRNA genes drives the abundance of tRNA-derived fragments. (2019) Torres AG, Reina O, Stephan-Otto Attolini C, Ribas de Pouplana L. Proc Natl Acad Sci U S A. 116(17):8451-8456. 3. The Expansion of Inosine at the Wobble Position of tRNAs, and Its Role in the Evolution of Proteomes. (2019) Rafels-Ybern À, Torres AG, Camacho N, Herencia-Ropero A, Roura Frigolé H, Wulff TF, Raboteg M, Bordons A, Grau-Bove X, Ruiz-Trillo I, Ribas de Pouplana L. Mol Biol Evol. 36(4):650-662. 4. Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla. (2018) Rafels-Ybern À, Torres AG, Grau-Bove X, Ruiz-Trillo I, Ribas de Pouplana L. RNA Biol. 2018;15(4-5):500-507.

Expected skills::

R. Desirable but not essential: Python, and experience in phylogenetic analysis.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Mario Cáceres
Email	mcaceres@icrea.cat
Institution	Institut de Biotecnologia i de Biomedicina (IBB), Unversitat Autònoma de Barcelona (UAB)
Website	https://invfest.uab.cat/
Group	Comparative and Functional Genomics Group

Project

Computational genomics

Project Title:

Functional and evolutionary impact of polymorphic inversions in the human genome

Keywords:

Expected skills depend on the actual line of research chosen, but should include scripting/programming skills (python, bash, R and/or perl) and experience in genomic variants and functional analysis. Knowledge of MySQL and PHP would also be helpful for working with the InvFEST database.se.

Summary:

The master student will integrate in a young, interdisciplinary and highly-dynamic group. In particular, the proposed tasks span a diverse range of themes focused in the functional and evolutionary impact of inversions, which are a little studied class of genomic variants, and the project could vary according to the interest and background of the candidate. 1. Bioinformatic analysis of the functional consequences of inversions and their association with phenotypic traits and disease susceptibility through imputation of inversion genotypes in large-scale datasets, in which the effect of these changes has been typically missed. 2. Development of new functionalities and visualization tools for our human polymorphic inversion data base InvFEST (http://invfestdb.uab.cat/), the world reference of human inversions. 3. Comparative study of known human inversion regions in other mammal species genomes to determine if there are inversion recurrence hotspots conserved over long evolutionary distances that might indicate a potential functional role.

References:

M. Puig et al. Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR. Genome Research (in press) (2020). C. Giner-Delgado et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. Nature Communications 10: 4222 (2019). D. Vicente-Salvador et al. Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. Human Molecular Genetics 26:567-581 (2017). M. Puig et al. Functional impact and evolution of a novel human polymorphic inversion that disrupts a gene and creates a fusion transcript. PLoS Genetics 11(10): e1005495. doi:10.1371/journal.pgen.1005495 (2015). A. Martínez-Fundichely et al. InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Research 42 (D1): D1027-D1032 (2014).

Expected skills::

Expected skills depend on the actual line of research chosen, but should include perl, python and bash programming and experience in working with DNA sequence data and functional analysis. Knowledge of MySQL and PHP would also be helpful for working with the InvFEST database.

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Comments:

Depending on the degree of experience of the candidate and the task performed it is possible to obtain financial support for the master practice, Also, at the end of the master there is the possibility to apply for a PhD fellowship.

Universitat
Pompeu Fabra
Barcelona

Master in
Bioinformatics for
Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Albert Jordan
Email	ajvbmc@ibmb.csic.es
Institution	Institute of Molecular Biology of Barcelona (IBMB-CSIC)
Website	http://www.ibmb.csic.es/groups/chromatin-regulation-of-human-and-viral-gene-expression
Group	Chromatin regulation of human and viral gene expression

Project

Computational genomics

Project Title:

Occupancy of histone H1 variants genome-wide and consequences of altering H1 levels on human chromatin organization.

Keywords:

Chromatin, histones, genomics, 3D nuclear structure, ChIP-seq

Summary:

We focus our research on the control of gene expression in human cells by chromatin organization, components and modifications. We investigate the role and specificity of histone H1 variants in chromatin organization and gene expression control. By RNA interference of the different human H1 variants we have found that they have different involvement in cellular processes such as cell cycle progression and gene expression. We have also described a differential role of H1 variants in pluripotency and differentiation. Currently, we are investigating the occupancy of H1 variants genome-wide by ChIP-seq (NGS) and the consequences of altering H1 levels on chromatin organization (ATAC-seq, DNA methylation, hiC, etc), with an extensive use of Genomics and

Bioinformatics. Additionally, we are performing proteomics of H1 variant specific protein complexes in chromatin and nucleoplasm.

References:

▲ Izquierdo-Bouldstridge A*, Bustillos A*, Bonet-Costa C, Aribau P, Garcia D, Dabad M, Esteve-Codina A, Pascual L, Peiro S, Esteller M, Murtha M, Millán-Ariño Ll, Jordan A (2017) Histone H1 depletion triggers an interferon response in cancer cells via activation of heterochromatic repeats. Nucleic Acids Research 45(20): 11622-42. ▲ Millán-Ariño Ll, Izquierdo-Bouldstridge A, Jordan A (2016) Specificities and genomic distribution of somatic mammalian histone H1 subtypes. BBA Gene Regulatory Mechanisms 1859(3): 510-19. ▲ Mayor R*, Izquierdo-Bouldstridge A*, Millán-Ariño Ll, Bustillos A, Sampaio C, Luque N, Jordan A (2015) Genome distribution of replication-independent histone H1 variants shows H1.0 associated with nucleolar domains and H1X associated with RNA polymerase II-enriched regions. Journal of Biological Chemistry 290(12):7474-91. ▲ Millán-Ariño Ll, Islam A, Izquierdo-Bouldstridge A, Mayor R, Terme JM, Luque N, Sancho M, López-Bigas N, Jordan A (2014) Mapping of six somatic linker histone H1 variants in human breast cancer cells uncovers specific features of H1.2. Nucleic Acids Research. doi: 10.1093/nar/gku079 ▲ Terme JM*, Sesé B*, Millán-Ariño L, Mayor R, Izpisua-Belmonte JC, Barrero MJ, Jordan A (2011) Histone H1 variants are differentially expressed and incorporated into chromatin during differentiation and reprogramming to pluripotency. Journal of Biological Chemistry 286(41):35347-57 ▲ Sancho M, Diani E, Beato M, Jordan A (2008) Depletion of human histone H1 variants uncovers specific roles in gene expression and cell growth. PLOS Genetics- Oct;4(10):e1000227.

Expected skills::

Strong motivation for research. Background or interest in Biologgy/Biomedicine and Epigenetics. The student will work in analyzing high-throughput genomic data such as ChIP-seq, RNA-seq, ATAC-seq and hi-C. To do so, experience in handling aligners, peak calling softwares, differential gene expression analysis and statistics tests will be an advantage. In addition, programming skills in R, Python and/or Perl are also necessary.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Robert Castelo
Email	robert.castelo@upf.edu
Institution	Universitat Pompeu Fabra
Website	https://functionalgenomics.upf.edu

Group	Functional Genomics

Project

Computational genomics

Project Title:

Functional genomics

Keywords:

genetics, genomics, statistics, bioconductor

Summary:

The research in the functional genomics group is geared towards the development of computational methods and pipelines to address questions of biological and clinical relevance. Depending on the profile of the candidate, different types of master projects are possible, ranging from software engineering and development in R/Bioconductor, development of new methods for the analysis of high-throughput genomics data, to the analysis of specific datasets to answer particular biological and clinical questions. Some of our contributions in all these aspects can be found in the list of references.

References:

1. Costa et al. Genome-wide postnatal changes in immunity following fetal inflammatory response. medRxiv, 19000109, 2020. 2. Roverato and Castelo. Path weights in concentration graphs. Biometrika, in press (arXiv:1907.05781) 3. Puigdevall et al. Genetic linkage analysis of a large family identifies FIGN as a candidate modulator of reduced penetrance in heritable pulmonary arterial hypertension. Journal of Medical Genetics, 56:481-490, 2019. 4. Puigdevall and Castelo. GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor. Bioinformatics, 18:3208-3210, 2018. 5. Roverato and Castelo. The networked partial correlation and its application to the analysis of genetic interactions. Journal of the Royal Statistical Society Series C -Applied Statistics, 66:647-665, 2017. 6. Costa and Castelo. Umbilical cord gene expression reveals the molecular architecture of the fetal inflammatory response in extremely preterm newborns. Pediatric Research, 79:473-481, 2016. 7. Baumstark et al. The propagation of perturbations in rewired bacterial gene networks. Nature Communications, 6:10105, 2015. 8. Tur et al. Mapping eQTL networks with mixed graphical Markov models. Genetics, 198(4):1377-1383, 2014. 9. Hänzelmann et al. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics, 14:7, 2013.

Expected skills::

Programming, scripting, minimum understanding of statistics.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Master project 2020-2021

Personal Information

Supervisor	Josep F Abril
Email	jabril@ub.edu
Institution	Department of Genetics, Microbiology & Statistics
Website	https://compgen.bio.ub.edu/
Group	Computational Genomics Lab @ UB

Project

Computational genomics

Project Title:

Refactoring Viral Metagenomic Pipelines

Keywords:

metagenomics, sequence analysis, kmer frequencies, taxonomy annotation pipelines

Summary:

In collaboration with the VirCont research lab, we have already developed a number of analysis procedures for the characterization of viral species found from high-throughput sequencing experiments of complex samples, as well as the diversity parameters from environmental samples. We want to integrate those into a semi-automatic/fully-automatic pipeline to perform the whole process, from data-gathering to the generation of summary reports. We will try to extend the current analyses with k-mer based approaches, as well as more efficient ways to assign species by fast homology-based approaches.

References:

Natalia Timoneda PhD Thesis: https://compgen.bio.ub.edu/dl2650

Expected skills::

Student should master Unix/bash, python/perl/C, R, SQL, web apps (HTML/CSS/shiny/django)..

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

If applicant is interested in doing a PhD, there is the possibility to apply for a Generalitat FI or Ministerio FPU PhD grants on the next announcement.

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Arnau Sebe-Pedros
Email	arnau.sebe@crg.es
Institution	CRG
Website	https://www.crg.eu/en/programmes-groups/sebe-pedros-lab
Group	Single-cell genomics and evolution

Project

Computational genomics

Project Title:

Evolutionary modeling of cell type gene regulatory networks using single cell genomics data

Keywords:

evolution; gene regulation; cell types; transcription factors; chromatin accessibility

Summary:

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 analyse 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. 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. We are seeking highly motivated master students to join our team and work on inferring and comparing cell type gene regulatory networks (GRNs) across species. Methodologically, this project will involve the integrative computational analysis high-throughput single-cell genomics and chromatin data in different systems.

References:

https://www.ncbi.nlm.nih.gov/pubmed/29856957 https://www.ncbi.nlm.nih.gov/pubmed/29942020
https://www.ncbi.nlm.nih.gov/pubmed/27114036

Expected skills::

R and Python programming; experience working on a computing cluster; good understanding of functional genomics methods and experience analyzing genomic data.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Biola M. Javierre
Email	bmjavierre@carrerasresearch.org
Institution	Josep Carreras Leukaemia Research Institute (IJC)
Website	http:www.carrerasresearch.org/
Group	3D Chromatin Organization

Project

Computational genomics

Project Title:

Dissecting The Role Of Spatial-Temporal Genome Architecture In Pediatric Acute Lymphoblastic Leukemia

Keywords:

leukemia, ¨omics¨ data, 3D genome architecture

Summary:

Most of mutations and epimutations associated with complex diseases lie in non-coding regions, frequently at regulatory regions, and potentially exert their functions by altering the regulation of the target genes. The vast majority of regulatory elements that regulate each gene in each cell type are uncharted, constituting a major missing link in understanding genome control. We previously developed a new method called Promoter Capture Hi-C (PCHi-C), which allows the pioneer genome-wide systematic identification of the long-range regulatory elements that control more than 20000 genes. Using this method, we connected for the first-time non-coding autoimmune disease variants to putative target promoters prioritizing thousands of disease-candidate genes and implicating disease pathways, quarters of which not previously implicated (Cell 2016). Based on preliminary data recently generated, we hypothese that the novel description of the regulatory elements that control each gene along human B lymphpoiesis could allow to understand the contributions of mutations and epimutations in B cell cancer development and to discover new genes potentially implicated in malignant transformation. First, we are developing a novel experimental and computational methodology to genome-wide detect distal interacting regions of the genome for all genes in rare cell types with an improved resolution. Second, using this new methodology and other omics such as CHiP-seq, RAN-seq and ATAC-seq, we will unravel the dynamic rewiring of promoter interactomes along B cell differentiation. Third, we will link non-coding mutations and epimutations to their putative target genes, describing potential novel genes and gene pathways associated with B cell pediatric acute lymphoblastic leukemia. In summary, this interdisciplinary project will provide unprecedented insights into our understanding of how cells decide their identity with an impact on regenerative medicine, autoimmunity, immunodeficiency and B cell malignancies. SPECIFIC AIMS. - Mapping, filtering, interaction peak calling and analysis of the new method inspired on PCHiC data (HICUP and CHiCAGO pipelines). - Mapping, filtering, calling and analysis of CHIP-seq, ATAC-seq and RNA-seq - Analysis of GWAS data, WGS and WBGS data. - Integration of non-coding mutations and epimutations (Differentially Methylated Regions) with the previously omics data to define new genes and gene pathways associated with pediatric acute lymphoblastic leukemia.

References:

-Javierre, B.M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 167, 1369-1384.e19 (2016). This study focused on 17 human primary hematopoietic cell types, demonstrates that promoter interactions are highly cell-type- and lineage-specific and that they allow the association of non-coding mutations with potential target genes. - Cairns, J. Freire-Pritchett, P., Wingett, S.W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.M. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 17, 127 (2016). This paper describes the algorithms for detecting significant interactions from capture Hi-C. - Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B.M. et al. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biol. 17, 152 (2016) This manuscript summarizes the computational method used to calculate the enrichment of specific epigenomic features in the chromatin fragments constituting the nodes of the network. - Azagra A, Marina-Zarate E, Ramiro AR, Javierre BM #, Parra M # (#Corresponding author). From Loops to Looks: Transcription Factors and Chromatin Organization Shaping Terminal B Cell Differentiation. Trends Immunol (2020) This review summarizes the role of genome architecture in B cell differentiation and biology - Watt S, Vasquez L, Walter K, Mann AL, Kundu K, Chen L, Yan Y, Ecker S, Burden F, Farrow S, Farr B, Iotchkova V, Elding H, Mead D, Tardaguila M, Ponstingl H, Richardson D, Datta A, Flicek P, Clarke L, Downes K, Pastinen T, Fraser, P, Frontini M, Javierre BM #, Spivakov M#, Soranzo N# (#Corresponding author). Variation in PU.1 binding and chromatin looping at neutrophil enhancers influences autoimmune disease susceptibility. Nat Commun. (Under review) bioRxiv 620260; doi: https://doi.org/10.1101/620260 This manuscript describes the interplay between SNPs, transcription factor binding, gene expression, histone modifications, 3D chromatin organization and disease.

Expected skills::

High level of motivation and interest, Proficiency in at least one scripting or programming language, Proficiency in scripting environments for statistics and data analysis, Competitive CV, High level of collaborative and communicative skills, Good level of English speaking and writing skills.

Possibility of funding::

No

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Andreu Paytuvi
Email	apaytuvi@sequentiabiotech.com
Institution	Sequentia Biotech
Website	http://www.sequentiabiotech.com
Group	Sequentia Biotech

Project

Computational genomics

Project Title:

Analysing human microbiomes: towards personalized medicine

Keywords:

microbiome nutrition health-care

Summary:

The study of the microbiomes present in the human body is of fundamental importance as it is highly relevant for clinical applications. For instance, dysbiosis in distinct communities has been related to some diseases. Traditionally, studying these populations required the isolation and culture of each individual microorganism, which is a significant limitation considering that small portion prokaryotes are culturable. However, using sequencing technologies allows the study of these populations in a high-throughput manner. These technologies have been essential for the development of metagenomics, which is defined as the culture-independent genomic analysis of all the microorganisms in an environmental niche. Human microbiomes are taxonomically different whether they come from the gut, skin, vagina or from the mouth. For instance, the genus Bacteroides is very abundant in the gut while it is Lactobacillus in the vagina. Changes in the normal microbiota composition (dysbiosis) have been linked to some diseases such as diabetes (gut), obesity (gut), autism (gut), fertility (vagina), acne (skin), Parkinson (gut), among others. About 16,000 samples from the American Gut Project (AGP) have been already analysed with Gaia to obtain the taxonomic profile of the samples. Metadata for these 16,000 samples is available. With this taxonomic matrix and the available metadata, the student will develop methods, especially related to machine-learning, that will help doctors to diagnose. Therefore, the final aim of the project is the development of models to classify a sample (e.g. from a patient) to specific groups (e.g. potentially diabetic, Parkinson-like profile, etc.).

Expected skills::

The student considering this thesis proposal must have a strong Bash (command-line) and Python/R knowledge.

Possibility of funding::

No

Possible continuity with PhD: :

No

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Josep Vilardell
Email	josep.vilardell@ibmb.csic.es
Institution	institute of Molecular Biology of Barcelona
Website	www.ibmb.csic.es/vilardell
Group	Mechanisms of pre-mRNA splicing

Project

Computational genomics

Project Title:

Impact of the spliceosome on protein synthesis

Keywords:

splicing, spliceosome, ribosome, mRNA. differential expression

Summary:

The information content of genomes can be greatly expanded by pre-mRNA splicing. Virtually all human pre-mRNAs need to be spliced to become mRNAs. Furthermore, most pre-mRNAs can be spliced into different mRNAs by alternative splicing. Therefore, it is hardly surprising that perturbations in splicing are linked to disease. However, we know little on how the splicing of particular RNAs may be affected, and even less on how a number of splicing changes are coordinated during development or disease. To start addressing this question, we are analyzing WGS and RNASeq data from a number of cancer datasets. Although we are interested in all events of regulated splicing, we pay special attention to those related to the biosynthesis and function of the ribosome. A cycling cell depends on a suitable set of ribosomes to provide the necessary amount of structural and functional proteins before mitosis; paradoxically, making this machinery requires most of the cell's energy (as an illustrative example, a growing HeLa cell is making 1.6 x105 ribosomal proteins per minute). Thus, we expect that fast-growing cell, subjected to a strong selection (such as a tumor cell), will tweak this process to get any advantage. However, the analysis of the transcriptome of ribosomal proteins presents specific challenges because (a), it includes the mRNAs that are most abundant in the cell, but the amounts of each one are variable (while the ribosome has one copy of each protein); (b), the corresponding pre-mRNAs undergo little alternative splicing; and (c), the majority of human pseudogenes come from them, which introduces ambiguity when mapping reads to the genome. Our initial results suggest that processing of this set of transcripts is altered in cancer in unexpected ways, and we plan on strengthening our conclusions by expanding our analyses. In this context there are many opportunities for those with a strong motivation to document genomic strategies that control the tran-scriptome of specific gene families, like those related to the ribosome or the spliceo-some. The tasks involve quality analysis of raw RNASeq data, mapping using standard tools (for example, Hisat, STAR, and those related to direct sequencing of RNA), statistical analysis (Ballgown, Salmon, Vast-tools, DexSeq, or others), and modeling. Subject to progress, we would explore the use of transcriptomics data as a disease prognosis tool; namely, is a distinct distribution of transcripts indicative of a particular disease out-come?

References:

* Hussain, S. (2018) “Native RNA- Sequencing Throws its Hat into the Transcriptomics Ring” TiBS 1434. https://
/doi.org/10.1016/j.tibs.2018.02.007 * Guimaraes, JC. and Zavolan, M. (2016) “Patterns of ribosomal protein expression specify normal and malignant human cells” Genome Biol. 17:236-248 * Gupta, V. and J. R. Warner (2014). "Ribosome-omics of the human ribosome." RNA 20: 1004-1013. * Bitton, D. A., et al. (2014). "LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq." Genome Res 24(7): 1169-1179. * Acuna, L. I. and A. R. Kornblihtt (2014). "Long range chromatin organization: a new layer in splic-ing regulation?" Transcription 5. * Kawashima T et al (2014) Widespread use of non-productive alternative splice sites in Saccharo-myces cerevisiae. PLoS Genet. 2014 Apr 10;10(4):e1004249. * Zhang, J. and J. L. Manley (2013). "Misregulation of pre-mRNA alternative splicing in cancer." Cancer Discov 3(11): 1228-1237. * Fu, R. H., et al. (2013). "Aberrant alternative splicing events in Parkinson's disease." Cell Trans-plant 22(4): 653-661. * Plass, M., et al. (2012). "RNA secondary structure mediates alternative 3'ss selection in Saccha-romyces cerevisiae." RNA 18(6): 1103-1115

Expected skills::

knowledge of R is highly desirable

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Comments:

We are a wet lab but with knowledge of Bioinformatics and several questions to be approached using Bioinformatics but for which we have many molecular data. This is therefore an excellent setting for any knowledgeable, independent, ambitious, and highly motivated Bioinformatics student.

Master project 2020-2021

Personal Information

Supervisor	David Comas
Email	david.comas@upf.edu
Institution	Universitat Pompeu Fabra (UPF)
Website	http://www.biologiaevolutiva.org/dcomas/
Group	Human Genome Diversity Group

Project

Computational genomics

Project Title:

Human genome diversity: demography and adaptation

Keywords:

Human genome, genome diversity, demography, adaptation

Summary:

Our research is focused on the understanding of the current genomic diversity in human populations in order to establish the mechanisms, causes and consequences of this genetic variation. We are mainly focused on trying to disentangle two types of processes: - Demographic processes. Population history, such as migrations, expansions, bottlenecks and admixtures have modelled the extant genome diversity of humans. Using several genetic markers, such as mitochondrial or Y-chromosome lineages as well as high-throughput SNP coverage of several human populations, we have addressed some demographic questions, from regional aspects such as the genetic impact of the Bantu expansion in Central Africa to more global issues such as the colonization of whole continents. - Selective processes. The human genome as also been modelled by selective processes as a result of adaptations during the species history. We have analyzed parts of our genome in order to detect genetic signals yield by selective processes, such as adaptation to different environments and its relationship with human diseases. Disentangle both types of processes is not an easy task and our research deals with the analysis of the diversity of the human genome at a population level in order to detect demographic and selective processes.

Expected skills::

Basic bioinformatic skills in genome data analysis

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Tanya Vavouri
Email	tvavouri@carrerasresearch.org
Institution	Josep Carreras Leukaemia Research Institute (IJC)
Website	http://www.carrerasresearch.org/en/regulatory-genomics
Group	Regulatory Genomics

Project

Computational genomics

Project Title:

Evolution and biogenesis of mammalian PIWI-interacting RNAs (piRNAs).

Keywords:

genomic repeats, small non-coding RNAs, gene regulation

Summary:

Transposons and other repeats substantially contribute to genetic diversity in a species, to spontaneous mutations and regulatory innovations. PIWI-interacting RNAs (piRNAs) bound to PIWI proteins repress transposon activity in the germline. Repression of transposons is essential for normal progression of mammalian spermatogenesis. Transposons are highly enriched among piRNA producing loci and are transcriptionally and post-transcriptionally repressed by piRNAs. Nearly half of all piRNA-producing loci are protein-coding genes but, to date, it remains unknown why/how certain protein-coding genes are targeted for piRNA production during gametogenesis. The dynamic landscape of mammalian transposon insertions in genes and the strong association between piRNAs and transposons raise the question whether transposon insertions in genes have triggered piRNA production from these genes. The goal of this project is to use bioinformatics tools and available data (both from the lab and from other publications) to understand the effect of transposon insertions on gene function in the mammalian male germline. The specific objectives are to understand the extent and genetic causes of inter-individual variation in piRNA expression in mouse and to gain mechanistic insight into piRNA production from protein-coding genes.

References:

Ozata, D.M., Gainetdinov, I., Zoch, A. et al. PIWI-interacting RNAs: small RNAs with big functions. Nat Rev Genet 20, 89–108 (2019). https://doi.org/10.1038/s41576-018-0073-3

Expected skills::

R, scripting in bash, perl/python

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Eulàlia de Nadal
Email	eulalia.nadal@irbbarcelona.org
Institution	IRB Barcelona
Website	https://www.irbbarcelona.org/en/research/cell-signaling
Group	Cell Signaling

Project

Computational genomics

Project Title:

Decoding transcriptional heterogeneity one cell at a time (Deletome-seq)

Keywords:

single-cell RNA-seq, transcriptional heterogeneity, stress-responses, SAPK

Summary:

Single cell RNA-seq (scRNA-seq) has become the method of choice to dissect complex samples. These studies have provided striking insights such as the identification of novel cell types, but they also unveiled an unexpectedly high degree of transcriptional heterogeneity. The molecular mechanisms underlying this variability are not understood. Yet, cell-to-cell heterogeneity provides a mechanism to alter cell fate and cell identity. Currently, it remains a challenge to understand which mechanisms regulate transcriptional heterogeneity and their consequences. Here we propose to combine single cell transcriptomics with functional genome-wide genetic screen to identify the principles underlying transcriptional heterogeneity.

References:

- Nadal-Ribelles M&, Islam S&, Wei W&, Latorre P&, Nguyen M, de Nadal E, Posas F, Steinmetz LM. Sensitive high-throughput single-cell RNA-seq reveals within-clonal transcript correlations in yeast populations. Nat Microbiol. 4:683-692 (2019). - Nadal-Ribelles M, Islam S, Wei W, Latorre P, Nguyen M, de Nadal E*, Posas F*, Steinmetz LM*. Yeast Single-cell RNA-seq, Cell by Cell and Step by Step. Bio-Protocol Bio-protocol 9: e3359 (2019). - de Nadal E*, Posas F*. Osmostress-induced gene expression - a model to understand how stress-activated protein kinases (SAPKs) regulate transcription. FEBS J. 282: 3275-85 (2015). - de Nadal E, Ammerer G, Posas F. Controlling gene expression in response to stress. Nat Rev Genet. 12: 833-45. (2011).

Expected skills::

Biology, biochemistry or related fields

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Master project 2020-2021

Personal Information

Supervisor	Anna Bigas
Email	yguillen@imim.es
Institution	IMIM
Website	https://www.imim.es/programesrecerca/cancer/en_annabigas.html
Group	Cancer and Stem Cells

Computational genomics

Project Title:

Comparative genomics and transcriptomics in stem cells and cancer

Keywords:

Transcriptomics, Genomics, leukemia, public repositories, cancer genetics

Summary:

Comparative genomics has become an essential tool for understanding genetic changes among different organisms, tissues and diseases. Alongside the latest development of powerful sequencing techniques and complex computational algorithms, researchers have been able to identify genetic drivers of cancer, as well as to provide insights into the biological pathways involved in carcinogenesis. Comparative genomics is thus considered a key element in cancer sciences, and bioinformatics is nowadays implemented in every multidisciplinary research team. The Stem Cells and Cancer group led by Anna Bigas is focused on the study of the molecular mechanisms involved in hematopoietic stem cells generation and hematologic malignancies, specially T-cell acute lymphoblastic leukemia (T-ALL). Moreover, it is implicated in the exploration of biological processes underlying colorectal cancer as it works in close cooperation with Colorectal Cancer group led by Lluis Espinosa. The expertise of Yolanda Guillén, a bioinformatics-trained biotechnologist, is crucial to perform the computational part of these projects and to understand the biological relevance of the results. We are glad to host an enthusiastic and motivated student willing to participate in: • The implementation of different bioinformatics pipelines in ongoing projects. We do use multiple approaches to analyze transcriptional (RNA-Seq and microarrays), genomics (ChIP-Seq) and epigenomics (ATAC-Seq) data. The student will learn how to prepare, run and interpret the results, from the raw sequencing data to the final output. Importantly, we do have access to a supervised computational cluster, which will make the student possible to understand how to work in such computational environment. • The exploration of transcriptional changes in T-ALL. Our main objective is to collect transcriptional data, mainly RNA-Seq, from public resources in order to screen for genetic expression changes in T-ALL. We are not only interested in identifying genes differentially expressed in T-ALL compared to normal cells, but to detect isoform switching patterns in cancer.

Expected skills::

Bash and R basic programming

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Master project 2020-2021

Personal Information

Supervisor	Rosa Fernández
Email	rosa.fernandez@ibe.upf-csic.es
Institution	Institute of Evolutionary Biology
Website	rmfernandezgarcia0.wixsite.com/metazomics
Group	Metazoa Phylogenomics Lab

Project

Computational genomics

Project Title:

Understanding how animals protect their DNA against UV light damage through the lens of comparative genomics

Keywords:

Comparative genomics; Phylogenomics; UV light-induced DNA damage repair; Nonmodel Organisms; Animals

Summary:

Almost 500 millions of years ago, several animal lineages conquered terrestrial environments from marine ones. One of the main challenges they needed to overcome was the protection of their DNA against UV light-induced damage, a threat that did not exist under water. Interestingly, we know almost nothing about how non-vertebrate animals repair their DNA after UV light-induced DNA damage. How do they do it? Do the different animals that conquered land (including nematodes, arthropods, earthworms or planarians, among other creatures) use the same mechanisms or different ones? This project aims at shedding light into the genomic underpinnings of UV light-induced DNA damage repair in non-model organisms through a comparative genomics spyglass.

Expected skills::

Python and/or perl programming. Knowledge on phylogenetics and comparative genomics desirable but not essential.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Ilario De Toma
Email	ilario.detoma@gmail.com
Institution	CRG
Website	
Group	Mara Dierssen

Project

Computational genomics

Project Title:

EGCG induced change in the phospho-proteome of DYRK1A overexpressing cells

Keywords:

Down syndrome; Proteomics; Time-course; EGCG; Hippocampus

Summary:

DYRK1A is a gene triplicated in Down syndrome that regulates the phosphorylation of several targets. Mice overexpressing DYRK1A show cognitive alterations. Interestingly, EGCG, the main polyphenol extracted from green tea, it is a DYRK1A inhibitor and ameliorates the cognitive impairment in DYRK1A transgenic mice and other DS mouse models. We performed an iTRAQ experiment on hippocampal primary neuronal cultures, labeling 5 different time points upon EGCG treatment, 5 uM (0, 5’, 15’, 30’, 120’) both in transegenic and wild type cells. This will shed lights in the acute phase action of EGCG actions, with the future goal to improve its efficacy for treatment purposes.

Expected skills::

Using R and RStudio

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Davide Piscia
Email	davide.piscia@cnag.crg.eu
Institution	CNAG-CRG
Website	www.cnag.cat
Group	Data Analysis Team, Bioinformatics Unit

Project

Computational genomics

Project Title:

Deep learning models applied to rare diseases: where do we stand?

Keywords:

deep learning, rare diseases, variant effect prediction, non-coding regions

Summary:

Deep learning models have been used extensively in image recognition and natural language processing, but in the last years they have been also applied to genomics. In the last years some interesting initiatives were started such as kipoi (https://kipoi.org) and selene (https://selene.flatironinstitute.org) whose aim is to facilitate the use of deep learning in biological contexts. One of the objectives of this project is to evaluate and summarize the state of the art of genomics deep learning, especially for variant effects prediction in non-coding regions . Once the most promising models have been selected, the candidate will have to apply it to some of the rare disease whole genome datasets hosted at CNAG-CRG. In this task she/he will have to be able to run the models in the CNAG-CRG HPC cluster (GPUs enabled) and do a first assessment of the model predictions. The long-term goal of this work is to integrate deep learning as a functionality in the RD-Connect GPAP platform (https://platform.rd-connect.eu).

Expected skills::

The candidate is expected to have good computational skills, especially in python.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

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

Project

Computational systems biology

Project Title:

Creating the framework for a multidimensional understanding of signal transduction

Keywords:

signal transduction, multidimensional, molecular, spatial, and temporal level

Summary:

Remarkably, signal transduction systems use a relatively limited repertoire of intracellular signaling components. This small number of effector proteins nevertheless enables a cellular signaling apparatus that is flexible and versatile. Versatility is achieved by modulation at the molecular, spatial, and temporal levels of the macro-molecular interactions at each node in the pathway. In effect, a limited number of nodes, each with several alternative downstream pathways, can give rise to a vast number of distinct signaling pathways. Although versatile and complex, the biological role of signal transduction demands specificity and precision in signaling. In this respect, many questions remain open about the interplay of the molecular, spatial, and temporal levels of the macro-molecular interactions. This knowledge gap is tackled by the European Research Network on Signal Transduction (ERNEST) which counts currently more than 400 researchers with different expertise in the field. We are looking for a motivated student who is interested in supporting the endeavor of ERNEST. The Master student will have the unique opportunity to interact with known researchers across Europe. The student will be in charge of developing a framework for collecting, organizing and visualizing diverse signaling data. He/she should have knowledge in HTLM/CSS, web page design, MySQL 5.5, database design, Python. C/C++ and JavaScript is a plus. An important benefit of this projects is that the master student will be introduced to a wide European network in signal transduction which provides with valuable contacts and diverse opportunities for future job openings.

References:

Sommer et al. The European Research Network on Signal Transduction (ERNEST): Toward a Multidimensional Holistic Understanding of G Protein-Coupled Receptor Signaling (2020) (https://pubs.acs.org/doi/pdf/10.1021/acsptsci.0c00024)

Expected skills::

He/she should have knowledge in HTLM/CSS, web page design, MySQL 5.5, database design, Python. C/C++ and JavaScript is a plus.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Mireya Plass
Email	mplass@idibell.cat
Institution	Idibell
Website	
Group	Gene regulation of Cell Identity

Computational systems biology

Project Title:

Functional characterization of RNA binding proteins in neural differentiation

Keywords:

single-cell trabscriptomics, gene regulation, RNA binding proteins, alternative polyadenylation

Summary:

Background The Plass lab, located at Bellvitge Biomedical Research Institute (IDIBELL), investigates the role of RNA binding proteins (RBPs) in the regulation of cell identity using a combination of computational and high-throughput experimental approaches. In particular, the lab is interested in understanding how RNA processing mechanisms, including splicing and alternative polyadenylation (APA), impact neuronal cell differentiation and how this process is altered in the development of neurodegenerative diseases. We known that around 70% of human genes are regulated by APA. Despite the growing amount of evidence showing that APA changes according to the proliferation and differentiation status of a cell, it is still not clear if APA contributes to this process and which are RBPs involved. Goals The project proposed will focus on developing new computational approaches to identify RBPs involved in the differentiation of neural cell types in single-cell transcriptomics data (scRNA-seq) and characterize their target genes across different cell types. During this project, the student will learn how to analyze scRNA-seq datasets using existing tools and will develop new computational methods to identify regulatory interactions in single-cell datasets. Approach 1.- Characterize cell and tissue specific RBPs Using a collection of published scRNA-seq datasets, the student will characterize the expression of RBPs across human and mouse cell types from different tissues and differentiation states using existing packages for single-cell transcriptomic analyses. She/he will use these data to make an expression atlas of RBPs and identify cell type and tissue specific RBPs, by comparing the expression of the RBPs across datasets. 2.- Identify RBPs target genes across cells Using a combination of published and newly-developed methods, the student will identify RBP-RBP and RBP-target interactions across cell types that could suggest a functional relation. Assuming that we find interactions in which the expression of an RBP affects the expression of a gene, the student will develop a computational method to recover them from single-cell datasets. We will benchmark the method developed by assessing the ability to identify known interactions between RBPs and genes. Next, we will use a computational pipeline developed in the lab (Plass et al. unpublished), that allows quantifying the expression of individual APA isoforms. Using a similar approach as described before, we will now identify robust associations between specific RBP and individual APA isoforms across cells. 3. – Identify RBP – gene/isoform interactions relevant for neural differentiation The student will use a computational lineage-reconstruction to understand the relationships between different neural cell types in a time-dependent manner, i.e. understand which cell types give rise to other cell types. These methods can also be used to order cells according to their differentiation status. Once she/he has obtained a ordering of cells, she/he will develop a new method to identify interactions between an RBP and a target gene or isoform in time, i.e., identify cases in which the expression of an RBP in time n affects the expression of a target gene/isoform in time n + t. In this way, we will be able to identify new correlations that may be related to the cellular differentiation process.

References:

Derti, A. et al. A quantitative atlas of polyadenylation in five mammals. Genome Res 22, 1173–1183 (2012). Ji, Z., Lee, J. Y., Pan, Z., Jiang, B. & Tian, B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A 106, 7028–7033 (2009). Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O. & Lai, E. C. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res 23, 812–825 (2013). Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science (80-) 360, eaaq1723 (2018). Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol 20, 59 (2019).

Expected skills::

We are looking for a master student with experience in high-throughput sequencing data analyses. Candidates are expected to have experience in R and a scripting language such as Python or Perl. Prior knowledge on post-transcriptional regulation or method development will be a plus. Interest in gene regulation and working in a multidisciplinary team will be valued, as interaction with experimental researchers will be required for the success of the project.

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 2020-2021

Personal Information

Supervisor	Arnau Montagud
Email	arnau.montagud@bsc.es
Institution	BSC - Barcelona Supercomputing Center
Website	https://www.bsc.es/
Group	Computational Biology

Project

Computational systems biology

Project Title:

Simulation of drug interactions in multiscale model tailored to prostate cell-lines

Keywords:

multiscale modelling, drug simulations, gastric cancer, parameter databases

Summary:

General context The candidate will join the area of Precision Medicine in Alfonso Valencia's Computational Biology group within the Life Sciences Department at the Barcelona Supercomputing Center. This research line encompasses the development of different strategies and approaches to improved personalized diagnosis of disease, as well as treatment selection for particular patients, based on their individual characteristics. Computational systems biomedicine relies on the development of in silico models to integrate different sources of experimental information and produce patient-specific mechanistic explanations of cellular behaviour used to design new targeted therapies. In the context of cancer, cell signalling as well as metabolic models have been reconstructed for different cancer types and healthy tissues. Simulation of these models using different computational approaches (e.g. Boolean formalism, Constraint-Based Modelling) have supported the development of targeted therapies that attack specific biological pathways in the cell. The candidate will focus on using and further developing a set of tools aimed for the simulation drug inhibitions of different cell lines. These simulations will explore varying concentrations of single drug inhibition and combinations of them. Scientific context Discovery of efficient anti-cancer drug combinations is a major challenge, since experimental testing of all possible combinations is clearly impossible. Recent efforts to computationally predict drug combination responses retain this experimental search space, as model definitions typically rely on extensive drug perturbation data1,2. Relying on background knowledge extracted from literature and databases, patient-specific dynamical models were developed3 previously tailoring a general cancer model4 to breast-cancer patients. In this work, the study of solutions of the Boolean model led to identifications of particularities among patients and their clinical stratifications3. Currently, we have used this same framework to obtain prostate-cell-line-specific dynamical models and are starting to perform drug perturbation studies. Nevertheless, due to the limitations of the simulation tools used5,6, this study neither identifies sets of concentrations where this synergy is maximal nor it considers population-level constraints and behaviours. In present project, the candidate will simulate varying concentration of inhibitors in the different cell lines model using a multiscale modelling framework, PhysiBoSS7, that that combines agent-based8, Boolean5,6 and environmental dynamics9 modelling. The candidate will first gather from databases and literature biophysical information on parameters that allows for the tailoring of the multiscale simulation to each cell-line such as uptake rates, growth rates, etc. Then, the use of scripts already in place (in python, bash, perl, R) and new ones developed by the student will allow exploring different concentrations of drugs to find maximal synergies specific for each cell-line that would help identifying drug responses potentially relevant in the clinic.

References:

1. Flobak, Å. et al. Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling. PLOS Comput. Biol. 11, e1004426 (2015). 2. Flobak, Å., Vazquez, M., Lægreid, A. & Valencia, A. CImbinator: a web-based tool for drug synergy analysis in small- and large-scale datasets. Bioinformatics 33, 2410–2412 (2017). 3. Béal, J., Montagud, A., Traynard, P., Barillot, E. & Calzone, L. Personalization of logical models with multi-omics data allows clinical stratification of patients. Front. Physiol. 9, 1965 (2019). 4. Fumia, H. F. & Martins, M. L. Boolean Network Model for Cancer Pathways: Predicting Carcinogenesis and Targeted Therapy Outcomes. PLoS ONE 8, e69008 (2013). 5. Stoll, G., Viara, E., Barillot, E. & Calzone, L. Continuous time Boolean modeling for biological signaling: application of Gillespie algorithm. BMC Syst. Biol. 6, 116 (2012). 6. Stoll, G. et al. MaBoSS 2.0: an environment for stochastic Boolean modeling. Bioinformatics 33, 2226–2228 (2017). 7. Letort, G. et al. PhysiBoSS: a multi-scale agent-based modelling framework integrating physical dimension and cell signalling. Bioinformatics bty766 (2018) doi:10.1093/bioinformatics/bty766. 8. Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M. & Macklin, P. PhysiCell: An open source physics-based cell simulator for 3-D multicellular systems. PLOS Comput. Biol. 14, e1005991 (2018). 9. Ghaffarizadeh, A., Friedman, S. H. & Macklin, P. BioFVM: an efficient, parallelized diffusive transport solver for 3-D biological simulations. Bioinformatics 32, 1256–1258 (2016).

Expected skills::

Knowledge of molecular and cell biology // Strong interest in the information gathering, analysis, modelling and simulation of biological systems. // Programming skills (python, R, bash and perl for the scripts and software tools are written in C++). // Ability to access and evaluate scientific literature.

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Comments:

The project will be supervised by Arnau Montagud, co-supervised by Miguel Ponce de León and Alfonso Valencia.

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	José Manuel Mas
Email	raquel.valls@anaxomics.com
Institution	Anaxomics Biotech
Website	www.anaxomics.com
Group	Data Science Department

Project

Computational systems biology

Project Title:

Molecular pattern recognition from high-throughput data of patients in a Real World Database

Keywords:

artificial intelligence, high-throughput, real world data, GEO database

Summary:

GEO database contains the description of millions of experiments including high-throughput data. Some of these experiments are based on primary cells and represent a source of Real World Human Data (RWD), being this type of data of special interest for FDA and EMA during drug development process. After the isolation and preparation of GEO database, it is necessary carrying out tasks associated with the validation of our RWD repository. This validation process is based on pathway enrichment analyses and the study of over/under expression of proteins, and they will be done by using artificial intelligence (AI) techniques. The student enrolled in this project will be the responsible to validate a subset of patients associated with certain specific pathologies (pending to decide). The student will select the patients from our RWD repository and will use AI techniques to compare patients' data in front of data from healthy people.

Expected skills::

programming python, c++ or Matlab

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Master project 2020-2021

Personal Information

Supervisor	Katsuyuki Shiroguchi
Email	katsuyuki.shiroguchi@riken.jp
Institution	RIKEN
Website	https://www.bdr.riken.jp/en/research/labs/shiroguchi-k/index.html
Group	Lab for Prediction of Cell Systems Dynamics

Project

Computational systems biology

Project Title:

Studying cell dynamics by combining live imaging and single-cell RNA-seq

Keywords:

Single cell, RNA sequencing, Optical microscope, Technology development, Challenge,

Summary:

As part of the internship, the student will use our developed single-cell picking system, which combines live imaging and single-cell whole gene expression analysis, to study molecular mechanisms of cell dynamics in a cell population, e.g., cell activation, differentiation, or cell-cell interaction, related to the immune system, cancer, or organoids.

Expected skills::

High motivation and good social manners

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

Good opportunity for those who want to try both experiments and computational analyses

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	José Manuel Mas
Email	raquel.valls@anaxomics.com
Institution	Anaxomics Biotech
Website	www.anaxomics.com
Group	Data Science Department

Project

Computational systems biology

Project Title:

Generation of a Real World Data repository from RNASeq analysis

Keywords:

artificial intelligence, high-throughput, real world data, GEO database

Summary:

GEO database contains the description of millions of experiments including high-throughput data. Some of these experiments are based on primary cells and represent a source of Real World Human Data (RWD), being this type of data of special interest for FDA and EMA during drug development process. GEO database includes around 150.000 patients containing RNASeq data. The protein expression pattern from these data is an interesting information. The student enrolled in this project will be responsible to extract the protein expression pattern from the RNAseq datafiles grouping them by their labelled phenotypes. The student will determine the existing relationship in the protein expression pattern between these labelled patients and other patients with the same labels but with different high-throughput data.

Expected skills::

programming python, c++ or Matlab

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 2020-2021

Personal Information

Supervisor	Jordi Garcia Ojalvo
Email	jordi.g.ojalvo@upf.edu
Institution	Universitat Pompeu Fabra
Website	https://www.upf.edu/web/dsb
Group	Dynamical Systems Biology

Project

Computational systems biology

Project Title:

Self-organization and decision making in cells and tissues

Keywords:

Developmental processes, microbiology, autoimmune diseases, single-cell behavior, dynamics

Summary:

We offer research projects devoted to understanding, using mathematical modeling, how cells make decisions and how cellular populations self-organize in time and space. The specific system and biological process to be studied will depend on the interest of the student and the availability of data/questions from our own lab and the labs of our international collaborators, at the time of the project design.

Expected skills::

Programming experience in Python, C or Julia would be useful, but it's not essential.

Possibility of funding::

No

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Carlos Barata Martí
Email	cbmqam@cid.csic.es
Institution	IDAEA-CSIC
Website	www.idaea.csic
Group	Toxicology Group

Project

Computational systems biology

Project Title:

Funcional transcriptomic analyssis in the crustacean Daphnia magna

Keywords:

functional gene annotation, genome, Daphnia, curate

Summary:

Developing a friendly use pipeline for processing RNA seq data in Daphnia magna which involves ensambling reads, maping, quantifying counts and functional annotation and interpretation using Blast. KEGG and other bioinformatic databases. In addition the student will work on the annotation of the probes of a 8 x 60 K Agilent eArray containing the full set of the 41317gene models representing the full transcriptome of Daphnia magna. Three years ago we developed this array and we were able to annotate 50% of its probes. However, the genome of Daphnia is changing continuously and hence it has to be re-annotated using the existing gene Bancs. The idea is to use the Daphnia magna genome (wfleabase) for a primary annotation of probes and then using NCBI Blast tools using translated proteins or genes across taxa (mainly arthropods). The end product is to provide gene names associated to each probe and its homologous in Drosophila, humans and other species. We also intent to annotate the gene codes to perform GERONTOLOGY, KEGG and other functional analyses

References:

Campos B, Fletcher D, Piña B, Tauler R, Barata C. Differential gene transcription across the life cycle in Daphnia magna using a new all genome custom-made microarray. BMC Genomics 2018; 19. Campos B, Garcia-Reyero N, Rivetti C, Escalon L, Habib T, Tauler R, et al. Identification of metabolic pathways in daphnia magna explaining hormetic effects of selective serotonin reuptake inhibitors and 4-nonylphenol using transcriptomic and phenotypic responses. Environmental Science and Technology 2013; 47: 9434-9443. Campos B, Rivetti C, Tauler R, Piña B, Barata C. Tryptophan hydroxylase (TRH) loss of function mutations in Daphnia deregulated growth, energetic, serotoninergic and arachidonic acid metabolic signalling pathways. Scientific Reports 2019; 9. Fuertes I, Campos B, Rivetti C, Pinã B, Barata C. Effects of Single and Combined Low Concentrations of Neuroactive Drugs on Daphnia magna Reproduction and Transcriptomic Responses. Environmental Science and Technology 2019a; 53: 11979-11987. Fuertes I, Jordão R, Piña B, Barata C. Time-dependent transcriptomic responses of Daphnia magna exposed to metabolic disruptors that enhanced storage lipid accumulation. Environmental Pollution 2019b; 249: 99-108. Piña B, Barata C. A genomic and ecotoxicological perspective of DNA array studies in aquatic environmental risk assessment. Aquatic Toxicology 2011; 105: 40-49.

Expected skills::

knowled in r, pyton, automatic functional annotation, gerontaology, KEGG

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

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

Project

Computational systems biology

Project Title:

Understanding the role of RNA folding in neurodegenerative diseases using third-generation sequencing technologies (oxford nanopore)

Keywords:

Oxford Nanopore sequencing; RNA modiications; RNA structure; machine learning;

Summary:

RNAs are not simple intermediary molecules between DNA and protein, but are in fact functional molecules capable of regulating central cellular processes. Because RNA is a single-stranded molecule, it tends to fold back on itself, forming stable secondary and tertiary structures by internal base pairing and other interactions. The function of RNAs can vary depending on the specific folding that the molecule, and therefore accurate RNA structural maps are needed to understand the complexity, function, and regulation of these molecules. Unfortunately, current methods generating RNA structure maps employ second-generation sequencing technologies (e.g. Illumina), which are unable to produce information on highly repetitive regions of the genome. Here we will use Oxford Nanopore Technologies (ONT), capable of producing full-length RNA molecule reads, to produce RNA structure maps for highly repetitive regions, such as those involved in neurodegenerative diseases such as Amyotrophic Lateral Sclerosis (ALS) or Fronto-temporal dementia (FTD).

References:

1. 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 2. Beaudoin JD*, Novoa EM*, Vejnar CE, Yartseva V, Takacs CM, Kellis M and Giraldez AJ. Analyses of mRNA structure dynamics identify the embryonic RNA regulome. Nat Struct Mol Biol 2018, 25, 677-686 3. Smith MA*, Ersavas T*, Ferguson JM*, Liu J, Lucas MC, Begik O, Bojarski L, Barton K and Novoa EM#. Barcoding and demultiplexing Oxford Nanopore native RNA sequencing reads with deep residual learning. bioRxiv 2019, 864322 (under review in Genome Research) 4. Cozzuto L, Liu H, Pryszcz LP, Hermoso Pulido T, Ponomarenko J and Novoa EM#. MasterOfPores: a workflow for the analysis of Oxford Nanopore direct RNA sequencing datasets bioRxiv 2019, 828336 (accepted in Front in Genet)

Expected skills::

Mandatory: Python, R. Desirable: machine learning, handling of third-generation sequencing data

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Master project 2020-2021

Personal Information

Supervisor	Marta Melé
Email	marta.mele.messeguer@gmail.com
Institution	Barcelona Supercomputing Center
Website	https://www.bsc.es/discover-bsc/organisation/scientific-structure/transcriptomics-and-functional-genomics-lab-tfgl
Group	Transcriptomics and Functional Genomics Lab

Project

Computational systems biology

Project Title:

Single-cell transcriptomic meta-analysis of human disease across tissues

Keywords:

single-cell transcriptomics, Cell type deconvolution, disease, aging, smoking, meta-analysis,

Summary:

The candidate will join Marta Melé's Transcriptomics and Functional Genomics lab in the Life Sciences Department at the Barcelona Supercomputing Center. The lab is interested in understanding how individual variation in gene expression can explain phenotypic differences between individuals both in the context of health and disease. To address this question, we use large-scale transcriptomic analysis and latest single-cell sequencing technologies combined with methods development to study gene expression, splicing and cell type composition variation across human tissues and phenotypes. In this project, we will perform a large-scale analysis of single-cell RNA-sequencing datasets across tissues to address how individual variation in gene expression can explain phenotypic differences between individuals. First, we will analyze hundreds of single-cell RNA-sequencing datasets to explore the impact of aging, smoking, gender and certain disease conditions to changes in gene expression and cell type composition in blood. Second, we will use cell type deconvolution methods to map single-cell signatures in expression data across many tissues from individuals with different conditions including diabetes and cardiovascular diseases. Overall, this project will explore in depth what is the role of gene expression and cell type composition in defining why human individuals are different from one another and how this impacts disease progression. What you will learn: Development of computational pipelines to analyze and interpret large datasets specially from single-cell RNA-seq, and bulk RNA-sequencing. Working in a high performance computing (HPC) environment. Interpret multi-omics data, working through scientific collaboration, effective communicating research, writing scientific articles and critical thinking.

References:

Melé, M. et al. The human transcriptome across tissues and individuals. Science (80-.). 348, 660–665 (2015).

Expected skills::

Availability to start in July 2020 is encouraged Strong programming skills in bash, python, R, perl, or similar, some experience working in HPC clusters Some experience with Next Generation Sequencing data analysis Excellent communication skills in spoken and written English Capacity to contribute to research projects with novel research ideas and analysis Capacity to work as a team in a highly collaborative and diverse environment

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Antonio Julià
Email	toni.julia@vhir.org
Institution	Vall Hebron Research Institute
Website	www.urr.cat
Group	Grup de Recerca de Reumatologia

Project

Computational systems biology

Project Title:

Identification of synergistic drug combinations in autoimmune diseases through single-cell analysis

Keywords:

Autoimmune Disease; Combinatorial drug therapy; Single Cell RNA-seq; Network analysis; Personalized medicine

Summary:

Autoimmune diseases (ADs) are chronic inflammatory diseases that are present at a high frequency in our population. They include diseases like rheumatoid arthritis, psoriasis, lupus and inflammatory bowel disease. They cause a significant reduction in the quality of life of many patients and a significant increase in comorbidities. In the last decade there has been a big increase in number of therapies available to treat ADs. However, these therapies only work for a subset of patients and, in many cases, after a period of time their efficacy diminishes. In our group we are convinced that one solution to this major health problem would be the use of drug combinations. By identifying pairs of drugs that synergize their effect we could provide a more powerful therapy. The present master's thesis project is focused in this interesting research problem. To do so, the student will use single cell RNA-seq data on tissue and immune system cell samples, and different statistical and data mining tools to identify the most likely drug combination for a specific autoimmune disease. During this project, the student will learn a very novel type of data, will acquire advanced data analysis skills and interact with several other bioinformatics specialists in the group as well as clinical researchers.

References:

We have been recently granted a 5-year EU project on combinatorial therapies. We are the coordinators of this translational project, which includes single-cell data analysis. http://doctis.eu/

Expected skills::

Programming skills in R and Python Statistical analysis of data Biological knowledge

Possibility of funding::

No

Possible continuity with PhD: :

Yes

Comments:

While we don't provide funding during the Master's thesis, our aim is to integrate the candidate into our team and provide him/her with funding to be able to pursue his/her PhD.

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Alberto Santos Delgado
Email	alb.santosdel@gmail.com
Institution	Oxford Big Data Institute
Website	https://www.bdi.ox.ac.uk/
Group	Multi Omics Analytics

Project

Computational systems biology

Project Title:

Computational Prediction of Host-pathogen Protein-protein Interactions in Human

Keywords:

Microbiology, Protein-protein interactions, Machine learning, Biomedical databases

Summary:

Protein-protein interactions (PPIs) define the complexity of biological processes in healthy and disease states. Intra-species PPIs describe partially this complexity but symbiotic and pathogenic interactions contribute as well by either benefiting or harming the host. The study of inter-species PPIs are of special interest in the case of pathogens since knowledge extracted from these interactions can lead to new therapeutic targets to avoid their negative effects on the host. There exists an extensive collection of intra-species PPIs enabled by the rapid development of high-throughput experimental technologies. However, experimental identification of inter-species interactions is not simple and computational prediction becomes then necessary (Cuesta-Astroz et al 2018). For instance, we proposed a homology-based prediction method to obtain human-parasite PPIs in 15 parasitic species (Cuesta-Astroz and Santos et al 2019). Additionally, this method incorporated biological context relevant in the parasites' life cycles to obtain accurate spatially-resolved interactions. Here, we want to implement a method that benefits from features derived from intra-species interactions to predict inter-species interactions focusing on human-pathogen interactions. The vast amount of intra-species interactions compiled in publicly available databases can be used to train sequence-based classification algorithms that can then be tested in both intra- and inter-species interactions also available in several resources. Furthermore, this method can be improved by annotating the predicted interactions with known biological context such as infection, survival and pathogenic mechanisms. This project will be carried out in close collaboration with Dr. Yesid Cuesta-Astroz from the University of Antioquia and the Colombian Institute of Tropical Medicine (Medellin, Colombia).

References:

Computational and Experimental Approaches to Predict Host-Parasite Protein-Protein Interactions. Cuesta-Astroz Y, Oliveira G Analysis of Predicted Host–Parasite Interactomes Reveals Commonalities and Specificities Related to Parasitic Lifestyle and Tissues Tropism Cuesta-Astroz Y, Santos A, Oliveira G, and Jensen LJ Comparing two deep learning sequence-based models for protein-protein interaction prediction Richoux F, Servantie C, Borès C, and Téletchéa S

Expected skills::

Python, Machine learning, Fast.ai, PyTorch

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Jordi Mestres
Email	jmestres@imim.es
Institution	IMIM Hospital del Mar Institute of Medical Research
Website	http://syspharm.imim.cat/
Group	Systems Pharmacology

Project

Pharmacoinformatics & systems pharmacology

Project Title:

A knowledge-based approach to PROTACs design

Keywords:

PROTACs design

Summary:

PROteolysis TArgeting Chimeras (PROTACs) have emerged as a new revolutionary modality in drug discovery. PROTACs are heterobifunctional molecules comprising of a ligand targeting a protein of interest, a ligand targeting an E3 ligase and a connecting linker. The aim is, instead of inhibiting the target, to induce its proteasomal degradation [1,2]. In spite its wide exploitation in many therapeutic areas, there is still a lack of well-thought knowledge-based strategies to designing PROTACs. Accordingly, the main aim of this project will be to establish the knowledge basis to develop new approaches to PROTACs design.

References:

[1] M. Konstantinidou et al. PROTACs– a game-changing technology. Expert Opinion in Drug Discovery (2019) 14:1255. [2] Sun et al. PROTACs: great opportunities for academia and industry. Signal Transduction and Targeted Therapy (2019) 4:64.

Expected skills::

The ideal candidate should have good scripting/programming skills and a background on chemistry/biology/pharmacology/pharmacy.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Patrick Aloy
Email	patrick.aloy@irbbarcelona.org
Institution	Institute for Research in Biomedicine (IRB Barcelona)
Website	https://sbnb.irbbarcelona.org
Group	Structural Bioinformatics & Network Biology

Project

Pharmacoinformatics & systems pharmacology

Project Title:

Formatting Biological Big Data to Enable Systems Pharmacology

Keywords:

Systems pharmacology, complex diseases, biological and disease signatures

Summary:

The amount and complexity of the biological data generated in the last years, due to the popularization of high-throughput pipelines, is virtually flooding biomedical research. Indeed, the growth of biological databases is steeper than ever before, and the repertoire of possible read-outs spans all levels of biology. However, the nature of biological data is remarkably complex, and dealing with diversity, inconsistency and incompleteness, among other issues, demands heavy specialist processing, and prevents a widespread predictive approach to disease biology. Indeed, this deluge of data has not spurred the development of truly precision therapies, and the inherent limitations of the prevailing reductionist approaches have highlighted the need of moving away from the ‘one disease, one target, one drug’ paradigm and consider the complexity of human pathologies and physiological responses. The current project builds on the hypothesis that the disease-causing perturbations leave detectable traces at different - and variable - levels of biological complexity (i.e. activation/inhibition of signaling pathways, transcriptional changes, etc) that capture both the direct effect of the perturbation and a global reaction of the system. Accordingly, the main aim of the project is to collect genuinely heterogeneous datasets, and offer a generic and intuitive means to bridge the gap between biological big data repositories and state-of-the-art machine-learning tools. Besides, we shall develop a generalized connectivity mapping, as a form of virtual phenotypic screening, to discover novel chemical or genetic modulators able to revert the specific signatures of disease and ‘cancel out’ the phenotypic traits of the disorder. The successful candidate shall be responsible for the implementation of a pipeline to collect and process biological big data, and to encapsulate it in the form of heterogeneous biological embeddings. Overall, we shall develop a novel strategy to integrate the deluge of biological data in a format that is readily suitable for modern machine learning. Additionally, he/she will develop a General Connectivity Mapping (GCMap) strategy to link biological and chemical signatures from the Chemical Checker (htts://chemicalchecker.org), so that the biological context of each small molecule can be incorporated as a descriptor. We shall then explore the added value of these biological descriptors to identify therapeutic opportunities to treat complex diseases.

References:

- Duran-Frigola M, et al. Formatting biological big data for modern machine learning in drug discovery. WIREs Comp Mol Sci (2018), e1408. - Duran-Frigola M, et al. Extending the small molecule similarity principle to all levels of biology. Nat Biotechnol (2020) In press. Available at bioRxiv.

Expected skills::

Highly motivated. Fluency in English. Good programming and scripting skills, with knowledge of Python and databases management (e.g. postgresSQL).

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Patrick Aloy
Email	patrick.aloy@irbbarcelona.org
Institution	Institute for Research in Biomedicine (IRB Barcelona)
Website	https://sbnb.irbbarcelona.org
Group	Structural Bioinformatics & Network Biology

Project

Pharmacoinformatics & systems pharmacology

Project Title:

Generative Models to create Precision Drugs

Keywords:

Systems pharmacology, generative networks, machine learning, precision drugs.

Summary:

Biological data is accumulating at an unprecedented rate, escalating the role of data-driven methods in computational drug discovery. The urge to couple biological data to cutting-edge machine learning has spurred developments in data integration and knowledge representation, especially in the form of heterogeneous, multiplex and semantically-rich biological networks. Today, thanks to the propitious rise in knowledge embedding techniques, these large and complex biological networks can be converted to a vector format that suits the majority of machine learning implementations. Indeed, we have generated biological embeddings (i.e. bioactivity signatures) that capture complex relationships between small molecules and other biological entities such as targets or diseases (Duran-Frigola et al. 2020 Nat Biotechnol). However, only a tiny fraction of the possible chemical space has been so far explored, meaning that most compounds able to modulate biological activities (i.e. drugs) are yet to be discovered. Accordingly, the main objective of this project is to couple our bioactivity signatures to inverse design algorithms to generate new chemical entities with a desired functionality. In particular, we aim at generating new chemical entities (NCEs) to modulate the activity of a specific set of targets, selected from a combination of perturbagen profiles, to revert the pathological state induced by Alzheimer´s disease (AD) and other complex disorders. All in all, the incorporation of machine learning methods to the drug discovery process will trigger the development of thousands of novel compounds, finally enabling precision medicine. The successful candidate shall be responsible for the implementation of ML-based Generative Models (i.e. cVAEs or GANs) to create new small molecules that fulfill the required polypharmacological properties to revert AD pathological signatures.

References:

- Duran-Frigola M, et al. Formatting biological big data for modern machine learning in drug discovery. WIREs Comp Mol Sci (2018), e1408. - Duran-Frigola M, et al. Extending the small molecule similarity principle to all levels of biology. Nat Biotechnol (2020) In press, available at bioRxiv.

Expected skills::

Highly motivated. Fluency in English. Excellent programming and scripting skills, with deep knowledge of Python. Previous experience on the use of machine learning and data science techniques (e.g. TensorFlow/AdaNet) and HPC environments will be an asset.

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Gerard Pujadas
Email	gerard.pujadas@gmail.com
Institution	Universitat Rovira i Virgili
Website	http://www.cheminformatics-nutrition.recerca.urv.cat/en/
Group	Cheminoformatics & Nutrition

Project

Pharmacoinformatics & systems pharmacology

Project Title:

Drug discovery and drug design for COVID-19 treatment

Keywords:

COVID-19, SARS-CoV-2, drug discovery, drug design

Summary:

The project aims to find new drugs that can be used for the treatment of COVID-19 and related patologies. In order to do that, the student will use structural information of theraphеutic targets for COVID-19 treatment to define energy-based pharmacophores that allow him to mine in commercial databases of small molecules for finding those that can bind with high affinity to the target of interest. Then, the most promosing hits of this virtual screening will be used in hit-to-lead computational experiments to find more potents derivatives. The project will involve different drug-discovery or drug design technologies such protein-ligand docking, shape & electrostatic comparisons, FEP+. To achieve this project, the student will be trained in the use of the most common drug discovery & design suites (Schrödinger, OpenEye and Cresset). The result of his/her research project will be the several drugs for COVID-19 treatment.

References:

Understanding the variability of the S1' pocket to improve matrix metalloproteinase inhibitor selectivity profiles. Gimeno A, Beltrán-Debón R, Mulero M, Pujadas G, Garcia-Vallvé S. Drug Discov Today. 2020 Jan;25(1):38-57 Mining large databases to find new leads with low similarity to known actives: application to find new DPP-IV inhibitors. Ojeda-Montes MJ, Casanova-Martí À, Gimeno A, Tomás-Hernández S, Cereto-Massagué A, Wolber G, Beltrán-Debón R, Valls C, Mulero M, Pinent M, Pujadas G, Garcia-Vallvé S.

Future Med Chem. 2019 Jun;11(12):1387-1401. The Light and Dark Sides of Virtual Screening: What Is There to Know? Gimeno A, Ojeda-Montes MJ, Tomás-Hernández S, Cereto-Massagué A, Beltrán-Debón R, Mulero M, Pujadas G, Garcia-Vallvé S. Int J Mol Sci. 2019 Mar 19;20(6). Combined Ligand- and Receptor-Based Virtual Screening Methodology to Identify Structurally Diverse Protein Tyrosine Phosphatase 1B Inhibitors. Gimeno A, Ardid-Ruiz A, Ojeda-Montes MJ, Tomás-Hernández S, Cereto-Massagué A, Beltrán-Debón R, Mulero M, Valls C, Aragonès G, Suárez M, Pujadas G, Garcia-Vallvé S. ChemMedChem. 2018 Sep 19;13(18):1939-1948 Activity and selectivity cliffs for DPP-IV inhibitors: Lessons we can learn from SAR studies and their application to virtual screening. Ojeda-Montes MJ, Gimeno A, Tomas-Hernández S, Cereto-Massagué A, Beltrán-Debón R, Valls C, Mulero M, Pujadas G, Garcia-Vallvé S. Med Res Rev. 2018 Sep;38(6):1874-1915.

Expected skills::

Good skills with Python and shell scripting

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Juan Fernández Recio
Email	juan.fernandezrecio@icvv.es
Institution	ICVV-CSIC
Website	http://www.icvv.es/english/3dbiowine
Group	Structural Bioinformatics for Wine Sciences

Project

Structural bioinformatics

Project Title:

Structural modeling of bitter taste receptors and their interactions

Keywords:

protein-protein interactions, computational docking, drug discovery, molecular modeling, taste receptors

Summary:

The interplay between genetics and environmental factors is critical in major non-communicable diseases (NCDs), which are the leading cause of death globally and a major cause of premature death. As an example, it is known that genetic variation can affect individual food preferences, which has impact on diet and health. Indeed, genetic sensitivity to bitter taste has been associated to different sensitivity to bitterness and/or linked to variable risk of alcohol dependence, obesity, nicotine dependence, longevity, miocardial infarction, or altered thyroid function (Duffy 2004; Mangold et al 2008; Campa et al 2012; Clark et al 2015). At the molecular level, bitter taste perception in humans is mediated by the 25 members of the Taste 2 receptor (TAS2R) gene family (Conte et al 2002). Each member of the TAS2R family can bind a range of compounds with different specificities, enabling the detection of tens of thousands of bitter molecules (Meyerhof et al 2010 Chem Senses). Therefore, knowing the atomic details of their binding capabilities would be important to understand the impact of these genetic variants in diet preferences and disease risk. In this project, we aim to contribute to the structural characterization of taste receptors to understand their functional mechanisms and the impact of genetic variants in health. We will model four bitter taste receptors that host variants associated to disease: TAS2R16, TAS2R38, TAS2R42, TAS2R50. Preliminary results using available models at www.gpcrdb.es show that critical residues for function gather around active site. However, these models still contain some structural errors that we will need to refine by molecular dynamics (MD). Then, we will model by docking the binding of around 100 bitter compounds to all TAS2R models and will compare the results with their known specificities (Meyerhof et al 2010). This will help to refine the modeling pipeline and will provide a theoretical framework for TAS2R binding to bitter compounds. We will also explore potential homomeric interactions of TAS2Rs by protein-protein docking in collaboration with Hugo Gutiérrez de Terán (Uppsala University). Finally, in collaboration with the groups of Masha Niv (Hebrew University of Jerusalem) and M. Purificación Fernández Zurbano (ICVV-UR), we will use our molecular models to test candidate compounds that are related to bitterness in wine, in order to build a functional model of taste perception in humans for the interpretation of genetic data affecting taste and hence diet preferences and health.

References:

Campa D, de Rango F, Carrai M, Crocco P, Montesanto A, Canzian F, Rose G, Rizzato G, Passarino G, Barale R (2012) PloS ONE 7, e45232. Clark AA, Dotson CD, Elson AET, Voigt A, Boehm U, Meyerhof W, Nanette I. Steinle NI, Munger SD (2015) FASEB J. 29, 164-172. Conte C, Ebeling M, Marcuz A, Nef P, Andres-Barquin PJ (2002) Cytogenet Genome Res 98, 45–53. Duffy VB (2004) Appetite 43, 5-9. Mangold JE, Payne TJ, Ma JZ, Chen G, Li MD (2008) J. Med. Genet. 45, 578-582. Meyerhof F, Batram C, Kuhn C, Brockhoff A, Chudoba E, Bufe B, Appendino G, Behrens M (2010) Chem. Senses 35, 157-170.

Expected skills::

Linux, basic programming capabilities, motivation for structural interpretation of molecular mechanisms

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Master project 2020-2021

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:

Deciphering the mechanism of drug action 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). Owning 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. GPCR drugs can either be agonist, antagonist or inverse agonists. They act by binding to the receptor and establishing transient interactions with protein residues that form the binding pocket. Those interactions alter the GPCR structure, leading to a specific downstream signalling response. However, important features responsible for a distinct drug profile (selectivity, signalling outcome, etc.) are still unclear. Uncovering those factors would provide important structural insight for further drug-design endeavours. Currently, there exist multiple GPCR structures bound to various ligands (chemicals binding to GPCRs), however a static look at ligand-receptor interactions doesn't allow to fully rationalize the signalling profile. Molecular dynamics (MD) is a novel and sophisticated technique that enables to simulate protein behaviour in a physiological environment. They offer a unique opportunity to study GPCR-ligand interactions at single atom resolution, providing insights on receptor behaviour. The development of MD techniques has been rewarded with a Nobel Prize in 2013, and the number of papers using MD is growing exponentially. In our group we have carried out a massive MD projects to unravel general principles of GPCR signalling and drug binding. With the aid of an international consortium we have simulated over 90 GPCRs crystallized with diverse ligands, amounting impressive simulation time (www.gpcrmd.org). We are looking for a motivated student that would be interested in participating in the analysis of this data. The students would be involved in analysing the generated MD data. During the project they will learn how to set up, and simulate their own biological systems. They will learn about GPCR biology, as well as about in silico drug design. To analyse the data the students will learn to write in house scripts in tcl, bash and python, as well as use several statistical methods. The student will have the opportunity to collaborate with international experts renowned in the GPCR field (members of the consortium see reference). We expect that the results of the analysis will be published in a high impact journal, and the skills acquired by the student will make him/her a valuable asset for pharma companies. The project can be extended into a PhD thesis.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome (https://www.biorxiv.org/content/10.1101/839597v2.abstract)

Expected skills::

Experience in structural biology, python, and bash. Experience with molecular dynamics simulations is a plus. Good level of English.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Xavier Barril
Email	xbarril@ub.edu
Institution	Universitat de Barcelona
Website	http://www.ub.edu/bl/
Group	Barril's lab

Project

Structural bioinformatics

Project Title:

Ligand optimisation for drug discovery: Use of MDmix for activity cliff prediction

Keywords:

Computer-aided drug design, structure-activity relationships, drug discovery, lead optimisation, binding free energy, molecular dynamics

Summary:

The goal of the Barril's lab is to discover bioactive molecules that bind to unexplored sites of action, exploiting novel mechanisms to achieve a therapeutic effect. To do so, we apply state of the art structure-based drug discovery methods, many of which have been developed in-house. We introduced the use of molecular dynamics with mixed solvents (MDmix) for druggability prediction,[1] as a computational counterpart of binding site detection by solvent screening.[2,3] This strategy turned out to be extremely successful and the method became widely adopted, with different adaptations (see reference [4] for a recent review). Since then, we have explored and extended the applicability of the method, describing its relationship with protein flexibility,[5] demonstrating its performance in mapping binding hot spots on protein surfaces and predicting water displaceability,[6] or as a guide in docking.[7] An open-source software was produced to help other users adopting the technique: http://mdmix.sourceforge.net Some preliminary work indicates that MDmix can also be used in predicting binding free energies of protein-ligand complexes.[7] In this project we will investigate its efficacy in predicting activity cliffs (i.e. pairs of structurally similar compounds presenting large potency difference). In medicinal chemistry, activity cliffs are crucial in systematic structure− activity relationship (SAR) analysis to identify structural modifications that determine SAR characteristics [8]. An analysis of a large set of matched molecular pairs compiled from the literature[4,5] and already available in our lab will be performed, comparing the performance of MDmix with other computational tools. This project is synergistic with other projects in our lab, and will benefit from substantial previous work and of close collaboration with other group members.

References:

J. Seco, F. J. Luque, X. Barril, Binding site detection and druggability index from first principles. J. Med. Chem. 52, 2363–71 (2009). 2. C. Mattos et al., Multiple solvent crystal structures: probing binding sites, plasticity and hydration. J. Mol. Biol. 357, 1471–82 (2006). 3. E. Liepinsh, G. Otting, Organic solvents identify specific ligand binding sites on protein surfaces. Nat. Biotechnol. 15, 264–8 (1997). 4. P. Ghanakota, H. A. Carlson, Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. J. Med. Chem. 59, 10383–10399 (2016). 5. D. Alvarez-Garcia, X. Barril, Relationship between Protein Flexibility and Binding: Lessons for Structure-Based Drug Design. J. Chem. Theory Comput. 10, 2608–14 (2014). 6. D. Alvarez-Garcia, X. Barril, Molecular simulations with solvent competition quantify water displaceability and provide accurate interaction maps of protein binding sites. J. Med. Chem. 57, 8530–9 (2014). 7. J. P. Arcon et al., Molecular Dynamics in Mixed Solvents Reveals Protein-Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions. J. Chem. Inf. Model. 57, 846–863 (2017). 8. A. M. Wassermann, M. Wawer, J. Bajorath, Activity Landscape Representations for Structure−Activity Relationship Analysis. J. Med. Chem. 53, 8209–8223 (2010). 9. Y. Hu, N. Furtmann, M. Gütschow, J. Bajorath, Systematic identification and classification of three-dimensional activity cliffs. J. Chem. Inf. Model. 52, 1490–8 (2012). 10. X. Hu, Y. Hu, M. Vogt, D. Stumpfe, J. Bajorath, MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J. Chem. Inf. Model. 52, 1138–45 (2012).

Expected skills::

molecular dynamics, protein-ligand docking, structure-based drug discovery

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

The group will offer a fellowship ("Beca de col·laboració") through the Fundació Bosch i Gimpera, for a period of 6 to 12 months. The fellowship is legally capped at 617€ per month. (Subject to funds availability).

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Jordi Villà-freixa
Email	jvilla@uic.cat
Institution	Universitat Internacional de Catalunya
Website	http://mon.uvic.cat/cbbl
Group	Computational Biochemistry and Biophysics Lab

Project

Structural bioinformatics

Project Title:

Finding epitope-MHC interactions through deep learning and molecular simulations

Keywords:

Molecular simulations, MHC, Deep Learning

Summary:

Several Machine Learning methods have been established to classify antigen-MHC interactions with somehow good success. However, precise classification of the binding characteristics within those complexes is still elusive if only classification methods are used. Here we will work in a combination of state of the art deep learning tools with structure based molecular simulations.

References:

"Structure Based molecular simulations" Submitted. Martin Floor, Li Keng Jie, Luís Agulló, Jenn K. Hwang, Jordi Villà-Freixa

Expected skills::

Python, Molecular simulations, Machine Learning

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Master project 2020-2021

Personal Information

Supervisor	Pietro Sormanni
Email	ps589@cam.ac.uk
Institution	University of Cambridge
Website	www.ch.cam.ac.uk/chemistry-of-health/index
Group	Chemistry of Health Initiative

Project

Structural bioinformatics

Project Title:

Automated structure- and sequence-based optimisation of antibody developability potential

Keywords:

Antibody design, Biopharmacueticals developability, drug development

Summary:

Owing to their outstanding performances in molecular recognition, antibodies are extensively used in research, diagnostics, and therapeutics, with more than 90 drugs already approved in the market. However, antibody development for therapeutic applications remains a long and costly process, also because therapeutic applications often require these molecules to withstand stresses that are not present in vivo. Antibody developability is defined as the likelihood of an antibody drug candidate with suitable functionality to be developed into a manufacturable, stable, safe, and effective drug that can be formulated to high concentrations while retaining a long shelf-life. In particular, antibody developability is determined by the presence of chemical liabilities, and by key biophysical properties including thermodynamic stability and solubility. Students are invited to work at the University of Cambridge within the Chemistry of Health initiative in the Department of Chemistry to develop a computational method and associated web server for the automated prediction of mutations that improve antibody developability potential. The applicant will work in a highly multidisciplinary research team where computational method development and corresponding experimental validation are carried out side-by-side. The outcome of this research will have an impact in the emerging field of computational antibody design, and it will improve and accelerate the ‘hit-optimisation’ step in biopharmaceutical pipelines.

References:

Sormanni, P., Aprile, F. A. & Vendruscolo, M. Third generation antibody discovery methods: in silico rational design. Chem. Soc. Rev. 47, 9137–9157 (2018)

Expected skills::

Python programming language is required. Beneficial: web server development and some knowledge of structural biology.

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 2020-2021

Personal Information

Supervisor	Oriol Gallego
Email	oriol.gallego@upf.edu
Institution	DCEXS-UPF
Website	www.gallegolab.org
Group	Live-cell structural biology

Project

Structural bioinformatics

Project Title:

Integrative Structural biology of exocytosis

Keywords:

Integrative Modeling Platform, Python, Integrative structural biology, Exocytosis

Summary:

This project aims to push the limits of integrative structural biology to resolve fundamental problems in cell biology. The student is expected to develop computational tools that strengthen a multidisciplinary team of bioinformaticians, physicists and experimentalists and that provide us with unique capabilities to resolve molecular structures. The biological question that we would like to address is exocytosis, a cellular process responsible to deliver biomolecules to the plasma membrane and extracellular space that is conserved in all eukaryotic cells. Exocytosis controls the growth of cell surface and it is directly coupled with the cell cycle and viability. However, the mechanism that regulates exocytosis is a central question in cell biology that could not be answered yet. Decades of research and the latest developments in gene editing, molecular biology and cryoEM have provided fundamental insight about exocytosis, but failed to resolve the molecular details that control this essential process. The complexity of the protein machinery involved and fast cycles of assembly-activity-disassembly have prevented full understanding of exocytosis. Recently, we developed a new method of fluorescent microscopy capable of resolving the 3D architecture of protein assemblies directly in living cells. Using this approach and computational integration of structural data we reconstructed de novo the exocytic machinery at the nanometre scale (Picco et al 2017 Cell). However, high-resolution structures and conformational dynamics necessary to understand the mechanism of exocytosis remain elusive. We offer a position for a Master student to push further integrative structural biology and that, together with our collaborators (D. Davos, CABD, Sevilla; J. Ries, EMBL, Heidelberg), works to develop the computational tools that can overcome current technical limitations. The student will use Python and the Integrative Modeling Platform (IMP, developed in A. Sali's lab at UCSF) to integrate in vitro and in cellulo datasets (i.e. live-cell imaging, cryo-EM, homology modeling, super resolution microscopy…) and to reconstruct the high-resolution structure of the supra-assembly that controls exocytosis. The student will team-up with a PhD student from our lab to explore new strategies involving Monte Carlo sampling methods and coarse-grained modeling among others. Overall, he/she is expected to contribute to a larger project aiming to resolve the mechanism of exocytosis.

References:

Picco, A., Irastorza-Azcarate, I., Specht, T., B.ke, D., Pazos, I., Rivier-Cordey, A-S., Devos*, D.P., Kaksonen*, M., Gallego*†, O., (2017) "The in vivo architecture of the exocyst provides structural basis for exocytosis." Cell 168, 400-412.e18.

Expected skills::

Expertise with Python is required. Knowledge on structural modeling or molecular dynamics will be a plus.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

High motivation for learning, team work and pushing the project forward is a must.

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Victor Guallar
Email	victor.guallar@bsc.es

Institution	Barcelona Supercomputing Cener
Website	https://www.bsc.es/discover-bsc/organisation/scientific-structure/electronic-and-atomic-protein-modeling-eapm
Group	EAPM

Project

Structural bioinformatics

Project Title:

ML victual Screening

Keywords:

Docking, coronavirus, PELE, deep learning

Summary:

We aim at developing a Deep Docking (DD) approach to screen billions of compounds in a fast manner. It will be integrated in a hierarchical pipeline combining commercial docking techniques and our Monte Carlo scheme (PELE). The application field will include real targets a current project to screen for pancoronavirus (polypharmacology) inhibitors.

References:

https://doi.org/10.1002/minf.202000028,

Expected skills::

Previous knowledge on docking and machine learning will be a plius.

Possibility of funding::

Yes

Possible continuity with PhD: :

Yes

Master project 2020-2021

Personal Information

Supervisor	Miguel Angel Santos Santos
Email	msantoss@santpau.cat
Institution	departamento de Neurologia del Hospital Sant Pau
Website	https://santpaumemoryunit.com/
Group	Unidad de Memoria HSP

Project

Structural bioinformatics

Project Title:

Multi-modal neuroimaging biomarker identification for the improvement of diagnosis and disease monitoring in primary progressive aphasia.

Keywords:

primary-progressive-aphasia, neurodegeneration, voxel-based-morphometry, diffusion-tensor-imaging, resting-state-functional-mri

Summary:

The main objective of this project is to study the clinical utility (to improve diagnosis and track disease progression) of various neuroimaging biomarkers. The secondary but equally fascinating objective is to study the neuroanatomical basis of language using primary progressive aphasia as a "lesion-model." We hypothesize that specific neuroimaging biomarkers including structural, diffusion tensor imaging, resting-state functional MRI, and fluorodeoxyglucose positron emission tomography provide measures of brain damage that reflect differentiable pathophysiologic mechanisms. More specifically, we hypothesize diffusion tensor imaging and functional MRI will be able to capture brain damage at early disease stages when volumetric atrophy is not apparent. To this end, we will preprocess and analyze the before-mentioned neuroimaging scans and then apply various statistical methodologies to compare across diagnostic groups and study their association with other biologic and language measures. The data for this study originates from a multitude of past and ongoing projects based at the Hospital Sant Pau Memory Unit.

References:

Classification of primary progressive aphasia and its variants. Neurology. 2011 Mar 15;76(11):1006-14. doi: 10.1212/WNL.0b013e31821103e6. Epub 2011 Feb 16. Gorno-Tempini ML1, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, Ogar JM, Rohrer JD, Black S, Boeve BF, Manes F, Dronkers NF, Vandenberghe R, Rascovsky K, Patterson K, Miller BL, Knopman DS, Hodges JR, Mesulam MM, Grossman M. Features of Patients With Nonfluent/Agrammatic Primary Progressive Aphasia With Underlying Progressive Supranuclear Palsy Pathology or Corticobasal Degeneration. JAMA Neurol. 2016 Jun 1;73(6):733-42. doi: 10.1001/jamaneurol.2016.0412. Santos-Santos MA1, Mandelli ML1, Binney RJ2, Ogar J1, Wilson SM3, Henry ML4, Hubbard HI1, Meese M1, Attygalle S1, Rosenberg L1, Pakvasa M1, Trojanowski JQ5, Grinberg LT6, Rosen H1, Boxer AL1, Miller BL1, Seeley WW6, Gorno-Tempini ML1. Rates of Amyloid Imaging Positivity in Patients With Primary Progressive Aphasia. JAMA Neurol. 2018 Mar 1;75(3):342-352. doi: 10.1001/jamaneurol.2017.4309. Santos-Santos MA1,2,3,4, Rabinovici GD1,5, Iaccarino L1,6, Ayakta N1,5, Tammewar G1,5, Lobach I7, Henry ML8, Hubbard I1, Mandelli ML1, Spinelli E1,6, Miller ZA1, Pressman PS1,9, O'Neil JP10, Ghosh P1, Lazaris A1, Meyer M1, Watson C1, Yoon SJ1,11, Rosen HJ1, Grinberg L1,12, Seeley WW1,12, Miller BL1, Jagust WJ5,10, Gorno-Tempini ML1. Functional Connectivity is Reduced in Early-stage Primary Progressive Aphasia When Atrophy is not Prominent. Alzheimer Dis Assoc Disord. 2017 Apr-Jun;31(2):101-106. doi: 10.1097/WAD.0000000000000193. Bonakdarpour B1, Rogalski EJ, Wang A, Sridhar J, Mesulam MM, Hurley RS. The Sant Pau Initiative on Neurodegeneration (SPIN) cohort: A data set for biomarker discovery and validation in neurodegenerative disorders. Alzheimers Dement (N Y). 2019 Oct 14;5:597-609. doi: 10.1016/j.trci.2019.09.005. eCollection 2019. Alcolea D1,2, Clarimón J1,2, Carmona-Iragui M1,2,3, Illán-Gala I1,2, Morenas-Rodríguez E1,2, Barroeta I1,2, Ribosa-Nogué R1,2, Sala I1,2, Sánchez-Saudinós MB1,2, Videla L1,2,3, Subirana A1,2, Benejam B1,2,3, Valldeneu S1,2, Fernández S1,2,3, Estellés T1,2, Altuna M1,2, Santos-Santos M1,2, García-Losada L1,2, Bejanin A1,2, Pegueroles J1,2, Montal V1,2, Vilaplana E1,2, Belbin O1,2, Dols-Icardo O1,2, Sirisi S1,2, Querol-Vilaseca M1,2, Cervera-Carles L1,2, Muñoz L1,2, Núñez R1,2, Torres S1,2, Camacho MV4, Carrió I4, Giménez S5, Delaby

C1,6, Rojas-Garcia R7,8, Turon-Sans J7,8, Pagonabarraga J2,9, Jiménez A10, Blesa R1,2, Fortea J1,2,3, Lleó A1,2.

Expected skills::

Basic knowledge in statistics, basic hands on experience with programming and imaging analysis software

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 2020-2021

Personal Information

Supervisor	Magdalena Scharf
Email	magdalena.scharf@pharmazie.uni-marburg.de
Institution	Philipps-University Marburg
Website	www.kolblab.org
Group	Peter Kolb

Project

Structural bioinformatics

Project Title:

In silico prediction of novel ligands for a chemokine receptor

Keywords:

Chemoinformatics, homology modelling, docking calculations, GPCRs, computer-aided drug design

Summary:

Your project will evolve around one member of the chemokine receptors. The ultimate goal is to find novel ligands that modulate the activity of this target. Since there are no crystal structures available for the target receptor, the first step is to prepare a three dimensional structure of it by homology modelling. This model will then be used in docking calculations to screen a large library of molecules against it.

Expected skills::

basic chemical knowledge to evaluate protein-ligand interactions

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Juan Cortés
Email	juan.cortes@laas.fr
Institution	LAAS-CNRS
Website	https://www.laas.fr/public/en
Group	Robotics and Interactions

Project

Structural bioinformatics

Project Title:

Studying how protein evolution (re-)shapes local structural preferences

Keywords:

Protein evolution, local structural preferences, structural database, protein flexibility, intrinsically disordered proteins

Summary:

The structural and dynamical properties of a protein are largely determined by its sequence, and strongly influence its function. To maintain protein functions during evolution, these properties must be robust to sequence variations (for example mutations). Nevertheless, structural/dynamical changes may be beneficial, for instance if they improve the way the protein functions or generate functional innovations. Understanding this trade-off between structural stability and malleability in evolution is of great interest for fundamental biology and for protein design. In recent years, we have contributed to better characterize the sequence-structure/dynamics relationship, particularly in the context of highly-flexible protein regions. We have constructed an extensive database of small fragments involving three consecutive residues (called tripeptides) extracted from coil regions in experimentally-determined protein structures. We have shown that this database is useful to accurately sample the conformational variability of protein loops [1] and intrinsically disordered proteins (IDPs) [2]. We have also developed an approach to characterize the structural preferences of each tripeptide sequence, and we have defined metrics to quantify the structural differences between different sequences. The goal of this project is to investigate how mutations occurring during evolution affect protein local structural propensities. To do so, the candidate will exploit our structural database and use the developed approaches and metrics. Starting from a set of currently observed proteins, s/he will infer the evolutionary history relating them and will quantify the correlation between changes in sequence and changes in local structures. We will consider several protein families. The analysis will be particularly focused on proteins in which (local or global) flexibility plays essential roles, such as antibodies, enzymes and IDPs.

References:

[1] Barozet, A., Molloy, K., Vaisset, M., Simeon, T., Cortés, J. (2020) A reinforcement learning approach to enhance protein loop sampling. Bioinformatics, 36(4):1099–1106 [2] Estana, A., Sibille, N., Delaforge, E., Vaisset, M., Cortés, J., Bernado, P. (2019) Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. Structure, 27(2):381-391.E2

Expected skills::

The candidate should have: background in bioinformatics and structural biology, good programming skills, familiarity with Linux, and knowledge about databases. Teamwork skills are also essential for the achievement of the project.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

The project will be co-supervised by Elodie Laine (Sorbonne Université, Paris)

Master project 2020-2021

Personal Information

Supervisor	Juan Cortés
Email	juan.cortes@laas.fr
Institution	LAAS-CNRS
Website	https://www.laas.fr/public/en
Group	Robotics and Interactions

Project

Web development & bioinformatic tools

Project Title:

A web server to generate conformational ensembles of highly-flexible proteins

Keywords:

Web server, Intrinsically Disordered Proteins (IDPs), Sampling, Conformational ensemble models

Summary:

Contrarily to what was thought in past decades, not all proteins fold into a relatively stable functional structure. Many proteins remain highly flexible in solution, possibly forming local transient structural elements. These proteins are usually called Intrinsically Disordered Proteins (IDPs). They play crucial roles in multiple biological processes and are directly involved in several pathologies, including cancer and neurodegeneration. The high flexibility of IDPs has notably hampered their study. Experimental biophysics technics such as Nuclear Magnetic Resonance (NMR) and Small-Angle X-ray Scattering (SAXS) provide information on conformational trends [1]. However, the quantitative interpretation of these experimental data requires the use of computational approaches that account for their ensemble averaging properties. These computational approaches are based on the construction of large conformational ensembles. We have recently developed a new method to model conformational ensembles of IDPs [2], which provides a more accurate representation than existing approaches. This method will be of great inserts for the scientific community working on the understanding of IDPs. The goal of this project is to provided easy access to this method through a web server, as we did a few years ago for another molecular modeling application (http://moma.laas.fr) [3]. By the end of the project, aiming to disseminate our work, we plan to submit an article describing the new web server for publication in a high-impact scientific journal. The student will work in a team involving other students (PhD and master level), researchers and software engineers working on related topics. He/she will take part in the design phase and the full-stack web development (both front-end and back-end). We aim to use the most recent languages and technologies at both levels (in particular, the Django web framework). Particular importance will be given to the ergonomy of the proposed solution.

References:

[1] T.N. Cordeiro, F. Herranz-Trillo, A. Urbanek, A. Estaña, J. Cortés, N. Sibille, P. Bernadó (2017) Small-angle scattering studies of intrinsically disordered proteins and their complexes. Current Opinion in Structural Biology, 42:15-23. [2] A. Estaña, N. Sibille, E. Delaforge, M. Vaisset, J. Cortés, P. Bernadó (2019) Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. Structure, 27(2):381-391 [3] D. Devaurs, L. Bouard, M. Vaisset, C. Zanon, I. Al-Bluwi, R. Iehl, T. Siméon, J. Cortés (2013) MoMA-LigPath: a web server to simulate protein-ligand unbinding. Nucleic Acids Research, 41(W1):W297-W302.

Expected skills::

Good programming skills are mandatory, mainly C++ and Python. Teamwork skills are also very important for the achievement of the project.

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

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

Project

Web development & bioinformatic tools

Project Title:

Developing web-based analysis tools for the study of GPCRs

Keywords:

web-based tools, data-sharing, structural biology, G protein coupled receptors

Summary:

G protein–coupled receptors (GPCRs) are major targets for the pharmaceutical industry (more than 30% of all FDA-approved drugs act on a GPCR) and present an immense potential for future drug development. Although GPCRs have been extensively studied over the past decades, the underlying molecular and structural mechanisms responsible for many critical regulatory processes of this protein superfamily remain elusive. Understanding the dynamics of receptor functionality is currently a major challenge in molecular biophysics and a requirement for the rational design of drugs with improved therapeutic profile. For this reason, a promising approach to elucidate the molecular basis of GPCR functionality are molecular dynamics (MD) simulations, a potent computational technique capable of generating atomic-resolution simulations of the structural motions of a molecular system. Consequently, MD simulations are increasingly being applied to the study of GPCRs, as reflected by the rapid upsurge of publications concerning this topic. In view of the growing importance of MD simulations, the GPCRmd project was created with the purpose to build the GPCRmd database, a database of MD simulations of GPCRs capable to foster data from all-over the world (www.gpcrmd.org). This web-based platform provides visualization and analysis tools specifically designed for the evaluation of structural and dynamic data of GPCR family members. In this Master project, the student will be involved in the design and development of new interactive analysis tools for the study of GPCRs, which will be incorporated in the GPCRmd viewer webpage. Such analysis tools will be focused on the study of the complex signalling network of intra-protein interactions that ultimately determine the response of the GPCR to a given drug. This includes the automatized detection and classification of different types of relevant interactions, comparison of the interaction network of different receptors (phylogenetically related receptors, wild type vs. mutant, ...), comparison of the network at different stages of a given molecular process, etc. Moreover, the obtained analysis tools will be applied to a case-study with the final aim to better understand how the intra-protein interaction network of a receptor of interest is affected by different stimuli or alterations (binding of a given ligand, receptor mutations, ...). The student will learn about GPCR biology, protein dynamics and in silico drug design, as well as web development, biomedical data analysis and biological databases. The internship can be extended into a PhD thesis. We expect that the GPCRmd database will have high impact on GPCR research and the discovery for new drugs. This will be communicated in a relevant publication to which the Master student will contribute.

References:

https://www.biorxiv.org/content/biorxiv/early/2019/12/17/839597.full.pdf

Expected skills::

Experience in structural biology. Python, HTLM/CSS and web page design. Experience with molecular dynamics simulations and JavaScript is a plus. Good level of English (oral and written).

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Master project 2020-2021

Personal Information

Supervisor	Davide Cirillo
Email	davide.cirillo@bsc.es

Institution	BSC - Barcelona Supercomputing Center
Website	https://www.bsc.es/
Group	Computational Biology

Project

Web development & bioinformatic tools

Project Title:

Evolution and crosstalk of biological ontologies

Keywords:

biological ontologies; graph theory; machine learning

Summary:

The research undertaken at the Barcelona Supercomputing Center (BSC) by the Computational Biology group, led by Prof. Alfonso Valencia, covers a wide range of Artificial Intelligence approaches for biomedicine, in particular in the area of biomedical computational graph theory and algorithms with special focus on biological ontologies. Biological ontologies, such as the Human Phenotype Ontology (HPO) (Köhler et al. 2019) and the Gene Ontology (GO) (The Gene Ontology Consortium 2019), are recognized as essential tools in the grand challenge of biomedical data integration and interpretation. In collaboration with the BSC Computer Science Department, we have developed a system for the efficient traversing and exhaustive path enumeration in interconnected biological ontologies (Cirillo et al. 2019) exerting large-scale parallelism and scalability in High-performance computing (HPC). This framework harnesses machine learning to infer a precise mapping between disease-related phenotypic features and distinct molecular processes allowing knowledge discovery. The proposed activity will be centered on the application and extension of this framework to a larger set of biological ontologies with the aim to study aspects such as (1) the dynamics of biological knowledge accumulation across time; (2) the integration and reconciliation of the multiple biological ontologies; (3) the implementation of machine learning approaches for biological knowledge representation and reasoning. The selected candidate will work in a highly sophisticated HPC environment, will have access to systems and computational infrastructures, and will establish collaborations with experts in different areas. What will you learn - Computational biology: biological knowledge representation; resources, formats and tools related to ontologies for use across the biomedical domain; applications and analytical approaches based on ontological information. - Computer Science: basics of High-performance computing; use of BSC supercomputing resources; BSC biology-oriented HPC implementations. - Scientific Dissemination: acquisition of science communication skills through lab meeting presentations and research article writing.

References:

Köhler et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019 Jan 8;47(D1):D1018-D1027. doi: 10.1093/nar/gky1105. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330-D338. doi: 10.1093/nar/gky1055. Cirillo et al. Graph analytics for phenome-genome associations inference. bioarXiv. 2019 Jun 26. doi: https://doi.org/10.1101/682229.

Expected skills::

- Good statistical and programming skills (Python, R/Bioconductor) - Strong interest in the analysis of biological systems - Basic knowledge of bioinformatics and molecular biology - Ability to access and evaluate scientific literature

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

This project will mainly focus on the application and extension of a previously developed tool, which is currently used in the laboratory. The student will be in close contact with collaborators at the BSC Computer Science Department within the groups “Best Practices for Performance and Programmability” led by Javier Teruel Garcia and Marta Garcia Gasulla, and “High

Performance Artificial Intelligence” led by Ulises Cortés. The project will be supervised by Davide Cirillo and co-supervised by Alfonso Valencia.

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Esteban Vegas i Ferran Reverter
Email	evegas@ub.edu; freverte@ub.edu
Institution	University of Barcelona
Website	
Group	Estadística i Bioinformàtica

Project

Web development & bioinformatic tools

Project Title:

Deep Learning based approaches for mining biomedical databases

Keywords:

Deep Learning, Data mining, Databases, Natural language processing

Summary:

This project aims the implementation and development of a tool based on Deep Learning models for the extraction and abstraction of biomedical knowledge using machine learning analysis of the contents in biomedical references databases (PubMed, MedGen, ...). Specific searches for terms related to a target disease will feed deep clustering algorithms to determine a set of disease-related descriptors. Then, recurrent-neural networks must be trained to assign automatically biomedical articles to disease-related descriptors.

References:

COHEN, Aaron M.; HERSH, William R. A survey of current work in biomedical text mining. Briefings in bioinformatics, 2005, vol. 6, no 1, p. 57-71. Gully A Burns, Xiangci Li, Nanyun Peng, Building deep learning models for evidence classification from the open access biomedical literature, Database, Volume 2019, 2019, baz034, https://doi.org/10.1093/database/baz034 Lan, K., Wang, D., Fong, S. et al. A Survey of Data Mining and Deep Learning in Bioinformatics. J Med Syst 42, 139 (2018). https://doi.org/10.1007/s10916-018-1003-9

Expected skills::

Python, Machine Learning, Databases, Data mining.

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Josep F Abril
Email	jabril@ub.edu
Institution	Department of Genetics, Microbiology & Statistics
Website	https://compgen.bio.ub.edu/
Group	Computational Genomics Lab @ UB

Project

Web development & bioinformatic tools

Project Title:

Integration of omics data related to retinal dystrophies

Keywords:

Interaction networks, retinitis pigmentosa, interologs mapping, web interface, graph databases

Summary:

Inherited retinal dystrophies (IRDs) comprise a highly heterogeneous group of disorders caused by over 200 causative genes. The prevalence of IRDs is 1:3000 worldwide, which make these blinding disorders a health relevant target. The implementation of massive sequencing approaches has greatly facilitated genetic testing and, as a result, the number of IRD genes and mutations is constantly increasing. Nonetheless, a substantial number of cases remain to be accurately diagnosed, as the average yield in IRD genetic diagnosis is roughly 50%. Our lab has contributed to the implementation of the RPGeNet (https://compgen.bio.ub.edu/RPGeNet) network and the curation of RNA-seq data to project differential gene-expression over that interaction network. The main goal is to perform computational analyses, integrating further omics data, to explore and pinpoint novel candidate genes and pathways that can be associated to molecular components of the disease. We plan to extend the current interaction network based on human genes/proteins, into further model organisms, such as mouse and zebrafish, in a way that the network of interologs can facilitate the integration of expression data from the three organisms.

References:

RPGeNet2 publication: https://academic.oup.com/database/article/doi/10.1093/database/baz120/5618821

Expected skills::

Student should master Unix/bash, python/C/perl, R, and perhaps some SQL. We will introduce the student into graph databases and django to provide interactive access to the data.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed

Comments:

If applicant is interested in doing a PhD, there is the possibility to apply for a Generalitat FI or Ministerio FPU PhD grants on the next announcement.

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	María José Rementería
Email	maria.rementeria@bsc.es
Institution	BSC - Barcelona Supercomputing Center
Website	https://www.bsc.es/
Group	Social Link Analytics

Project

Web development & bioinformatic tools

Project Title:

Dynamics of spreading false health news on social networks [RRSSalud]

Keywords:

Health, Fake News, Social Media

Summary:

Summary The RRSSalud is a research project that aims at investigating the typology and dynamics of dissemination on social media of fake news in the area of health. By combining quantitative methodologies (statistical analysis and social network analysis) and qualitative techniques (content analysis and focus groups), we explore the attitudes of people regarding the health information they consume. Specifically, we plan to focus on understanding the ability of Internet users to distinguish between false and true content, as well as, the tactics and strategies they use to detect the trustworthiness of news. In addition, topics, morphologies and rhetorical strategies of fake news will be explored. The investigation is coordinated by three teams of researchers from the University of Navarra and the Barcelona Supercomputing Center. As a result, we expect to develop tools, methods, and guidelines that can be employed by public health institutions, media organizations, and the general public to counteract the dissemination of fake news. Introduction In 2017, the Royal Spanish Academy incorporated a new word into the Dictionary of the Spanish Language: post-truth, which is defined as the "deliberate distortion of a reality, which manipulates beliefs and emotions in order to influence public opinion and social attitudes." The incorporation into the language of this neologism is a symbol of a serious current problem: the organized dissemination of false information, mainly through social media, in order to manipulate the public opinion. This phenomenon, popularly known as false news or "fake news" [1], has shown to have a significant impact in multiple areas and situations in recent years. In the political arena, for example, it has been found that it significantly influenced the results of the Brexit referendum and the 2016 presidential elections in the United States [2]. In business, the phenomenon of disseminating hoaxes and biased information has been identified as a way to deliberately discrediting brands and companies [3]. Information on the environment has also been a fertile area of lies and half-truths, especially with regard to the information on climate change [4]. However, the area of health is one of the areas where disinformation can cause profound damages [5], with anti-vax campaigns or health recommendations on epidemic periods to name a few. RRSSalud project focuses on studying fake news in health-care published on social media in Spain. The aim of the project is to explain the relationship between the vulnerability of Internet users to fake news and the dynamics of propagation and repercussion of these contents on social media. By combining experimental research with quantitative and qualitative methods we plan to explore the phenomenon from a holistic and comprehensive perspective. Specifically, the goals of the project are: i) identify the typology of fake news in the area of health and disseminated on social media in Spain; i) assess the vulnerability of Spain's Internet users to fake news; iii) profile population groups in relation to their critical capacity to identify fake news; iv) understand the dynamics dissemination of fake news and propose actions to mitigate its impact; v) identify the subjective aspects that lead to giving credit to fake news and promote its subsequent dissemination by users.

References:

1 Quandt, T., Frischlich, L., Boberg, S., & Schatto - Eckrodt, T. (2019). Fake news. The International Encyclopedia of Journalism Studies, 1-6. 2 Bastos, M. T., & Mercea, D. (2019). The Brexit botnet and user-generated hyperpartisan news. Social Science Computer Review, 37 (1), 38-54; Rose, J. (2017). Brexit, Trump and Post-Truth Politics, Public Integrity, 19 (6), 555-558. 3 Berthon, P. R., & Pitt, L. F. (2018). Brands, truthiness and post-fact: managing brands in a post-rational world. Journal of Macromarketing, 38 (2), 218-227. 4 Kolmes, S.A. (2011). Climate change: a disinformation campaign. Environment: Science and Policy for Sustainable Development, 53 (4), 33-37. 5 Viviani, M., & Pasi, G. (2017). Credibility in social media: opinions, news, and health information — a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7 (5), e1209.

Expected skills::

Experience with software development and at least one of the following: Proficiency in Python programming language / Proficiency in web programming languages and frameworks (e.g., Javascript, HTML, CSS, Django/Flask) // Ideally, some experience with

Python data science tools (e.g., Pandas, Numpy, Jupyter Notebooks, Scikit-learn, Matplotlib/Seaborn) to obtain, curate, clean, analyze, and visualization of information // Ability to work in an interdisciplinary social-tech environment and interact with relevant stakeholders to understand their needs and formulate solutions

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

RRSSalud is one of the five projects funded by BBVA Foundation as part of its program Scientific Research Teams in Economics and Digital Society. The work will be supervised by Nataly Buslón, Jorge Saldivar, and Maria José Rementeria.

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Carles Hernandez-Ferrer and Leslie Matalonga
Email	carles.hernandez@cnag.crg.eu, leslie.matalonga@cnag.crg.eu
Institution	CNAG-CRG
Website	www.cnag.cat
Group	Data Analysis Team - Bioinformatics Unit

Project

Web development & bioinformatic tools

Project Title:

Implementation of an automated genomic (re)analysis support system for rare disease diagnostic.

Keywords:

rare diseases, genome-phenome analysis

Summary:

It is estimated that 350 million people worldwide suffer from one of the approximately 7000 existing rare diseases (RDs). As 80% of RDs are thought to have a genetic origin, particular emphasis has been placed on the rapidly expanding development of genomic technologies. However, the interpretation of the genome is still a real challenge for molecular geneticists, and innovative bioinformatics solutions combining genomic and clinical data are crucial for reaching accurate patient diagnosis. At the CNAG-CRG and in the context of several EU projects (RD-Connect, Solve-RD, EJP-RD) we developed a data sharing and analysis tool, the RD-Connect Genome- Phenome Analysis platform (GPAP: https://platform.rd-connect.eu/), to provide methods and standardised analyses of phenotypic and (gen)omic data in order to facilitate mutation detection within the context of rare diseases. As part of the Solve-RD project: "solving the unsolved rare diseases" (www.solve-rd.eu), rare disease patient data from 19.000 genomic datasets will be reanalysed using the RD-Connect GPAP. To this purpose, high throughput SNV-indel data (re)analysis solutions are being implemented in the system to automatically allow real-time queries to a high number of samples. A first prototype has been built enabling the filtering of thousands of genomic datasets by specific filters including variant pathogenicity and population databases. The master student joining this project will be working on the development of this innovative approach by improving the tool that queries the RD-Connect API to enable more complex queries (involve family members' genotypes, specific regions of the genome, other individuals with phenotypic similarity, etc.). This work will be done in close collaboration with clinicians and researchers of four ERNs (European Reference Networks) involved in the Solve-RD project enabling continuous analysis feedback and molecular diagnosis confirmation.

Expected skills::

Python programming Data base architecture Genetics background Team working skills

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

---

upf. Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Marc Güell
Email	marc.guell@upf.edu
Institution	UPF

Website	https://www.upf.edu/en/web/synbio
Group	Translational Synthetic Biology

Project

Web development & bioinformatic tools

Project Title:

Computational approaches for efficient and safe engineering of human genomes

Keywords:

Gene editing, CRISPR, Synthetic Biology, Gene therapy

Summary:

Our laboratory is focused on applied synthetic biology for therapeutic purposes. We have two lines of research, one in technology development for gene therapy, and one in skin microbiome engineering. Advanced cell and gene therapies are gaining important impact in medicine. We currently have more than 2,500 on-going gene therapy trials on multiple diseases (cancer, genetic disease, infectious disease, etc...). However, multiple concerns have been raised on the safety of current technologies which prevent a wider deployment. Uncontrolled on-target, pro-cancer pathway activation, controversy on off-target, and lack of efficacy still represent a major concern. We are offering a Bioinformatics master thesis position in developing computational approaches to develop more precise technologies for gene editing. AI and genomics provided tools to significantly improve design (Chuai et al, Genome Biology 2018; Doench et al, Nat Biotech 2016; ...). Nevertheless, these approaches remain not predictable enough for therapeutic purposes. We are developing new algorithms which use clinically relevant data to improve prediction significance for therapeutic purpose.

Expected skills::

Basic statistics and programming

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Personal Information

Supervisor	Gianni De Fabritiis
Email	gianni.defabritiis@upf.edu
Institution	UPF
Website	https://www.compscience.org/
Group	Computational Science Laboratory - GRIB

Project

Web development & bioinformatic tools

Project Title:

Abstraction and reasoning challenge: Create an AI capable of solving reasoning tasks it has never seen before

Keywords:

reinforcement learning; machine learning; inductive programming; AI

Summary:

Can a computer learn complex, abstract tasks from just a few examples? Current machine learning techniques are data-hungry and brittle—they can only make sense of patterns they've seen before. Using current methods like reinforcement learning, an algorithm can gain new skills by exposure to large amounts of data, but cognitive abilities that could broadly generalize to many tasks remain elusive. This makes it very challenging to create systems that can handle the variability and unpredictability of the real world, such as domestic robots or self-driving cars. However, alternative approaches, like inductive programming, offer the potential for more human-like abstraction and reasoning. The abstraction and reasoning corpus (ARC) provides a benchmark to measure AI skill-acquisition on unknown tasks, with the constraint that only a handful of demonstrations are shown to learn a complex task (https://www.kaggle.com/c/abstraction-and-reasoning-challenge). This competition was initially created by the creator of the Keras neural networks library and it’s explained in this paper (https://arxiv.org/abs/1911.01547). The idea is to move beyond the competition timeframe to create an AI that can solve reasoning tasks it has never seen before and set up a path toward a PhD in AI. It is expected that novel work in terms of a paper should be produced during this period. For further details, contact Gianni De Fabritiis (gianni.defabritiis@upf.edu). The research period is paid. We are looking for exceptional candidates passionate about AI and with the willingness to go beyond in AI research. The lab is very well equipped.

References:

http://grib.imim.es/publications/index.php?CATEGORY1=14

Expected skills::

Be confident with maths and programming

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Gianni De Fabritiis
Email	i.escolar@acellera.com
Institution	Acellera Labs
Website	https://www.acellera.com/
Group	-

Project

Web development & bioinformatic tools

Project Title:

Machine learning in computational structural biology and drug discovery

Keywords:

machine learning; GPU computing; medicnal chemistry for drug discovery; PlayMolecule

Summary:

This project aims to develop machine learning methods applied to structural biology, drug discovery and computational chemistry. The aim is to go substantially beyond the state-of-the-art in the use of machine learning and GPU computing, exploring supervised, unsupervised and reinforcement learning approaches. We expect the candidate to participate in the development of new learning approaches and applications derived from deep learning applied to medicinal chemistry for drug discovery. By working in this project, the researcher will have access to state of the art computational resources. This project is expected to lead to discoveries that will be publishable in the highest impact scientific journals. Some examples of applications can be seen in PlayMolecule.org, a drug discovery platform used by thousands of scientists worldwide and pharmas and biotech companies. The platform is based on two main pillars, physical-based molecular simulations on GPUs and machine learning/AI, thus contributing to the company mission of accelerating the transition towards computerized drug discovery process. The platform was born in 2017 and serves as a repository of web applications for molecular modelling tools such as ProteinPrepare [Martínez-Rosell2017; doi:10.1021/acs.jcim.7b00190] and pioneering deep learning applications such as Kdeep [Jiménez2018; doi:10.1021/acs.jcim.7b00650] WHO WE ARE: Founded in 2006, Acellera was one of the first companies worldwide to leverage the use of novel accelerator processor technology (GPU) for molecular simulations. Among our clients, we count 10 of the top 50 pharmaceutical companies. We were selected as one of the Top30 AI Drug Discovery companies in 2019. Our software includes PlayMolecule, ACEMD, HTMD, KDEEP, etc. and it's used by hundreds of users both in academia and the private sector. In particular, PlayMolecule.com is the first platform to democratize the use of molecular dynamics and machine learning applications for drug discovery.

Expected skills::

You have VERY good programming skills and a background in either chemistry, biology, computer science or similar Prior knowledge in neural information processing, deep learning frameworks (pyTorch, Tensorflow) is desirable Very good knowledge of Python and good coding practices This is a strongly computational position, so we encourage application of people that love algorithms, computing, programming and likes to apply it

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

What we offer: - Real in-company work experience. - Possibility to participate in the development of a real software product - After graduation, it is possible to stay in the company

---

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Gianni De Fabritiis
Email	g.defabritiis@acellera.com
Institution	Acellera Labs SL
Website	www.acellera.com
Group	-

Project

Web development & bioinformatic tools

Project Title:

Development of frontend capabilities and graphic user interface (GUI) for PlayMolecule.org

Keywords:

frontend; GUI, in silico drug discovery; HTML, Phyton

Summary:

This master thesis will focus on the development and release of a new graphic user interface (GUI) for PlayMolecule, a popular platform for biomolecular-related applications with hundreds of unique users every month available at https://www.playmolecule.org. PlayMolecule is a drug discovery platform used by thousands of scientists worldwide and pharmas and biotech companies. The platform is based on two main pillars, physical-based molecular simulations on GPUs and machine learning/AI, thus contributing to the company mission of accelerating the transition towards computerized drug discovery process. The platform was born in 2017 and serves as a repository of web applications for molecular modelling tools such as ProteinPrepare [Martínez-Rosell2017; doi:10.1021/acs.jcim.7b00190] and pioneering deep learning applications such as Kdeep [Jiménez2018; doi:10.1021/acs.jcim.7b00650]. The position is based within the informatic and innovation hub of Barcelona (Spain). Acellera values excellence, merits and scientific innovation above anything else. WHAT YOU WILL BE WORKING ON: 1. You will closely work with Acellera developers to develop and improve the GUI capabilities for PlayMolecule. 2. You will mostly work in front-end tasks such as: - Collaborate with medicinal chemists to improve the capabilities and usability of PlayMolecule web interface - Development of 3D novel graphical interface to better access molecular structural information and dynamic plots ready to interface with the PlayMolecule back-end. 3. You may additionally work in back-end tasks such as: - Development of functionalities necessary to make custom GUI molecular selections WHO WE ARE: Founded in 2006, Acellera was one of the first companies worldwide to leverage the use of novel accelerator processor technology (GPU) for molecular simulations. Among our clients, we count 10 of the top 50 pharmaceutical companies. We were selected as one of the Top30 AI Drug Discovery companies in 2019. Our software includes PlayMolecule, ACEMD, HTMD, KDEEP, etc. and it's used by hundreds of users both in academia and the private sector. In particular, PlayMolecule.com is the first platform to democratize the use of molecular dynamics and machine learning applications for drug discovery.

Expected skills::

THIS PROJECT IS FOR YOU IF: You have good programming skills and a background in either chemistry or computer science You are proficient in: HTML, CSS/CSS3, Javascript, AngularJS, Python And maybe also have some knowledge of: Flask, SQL databases, Plotly.js, NGL.js You have very good communication skills in English

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed

Comments:

WHAT WE OFFER: - Real in-company work experience. - Possibility to participate in the development of a real software product - After graduation, possibility to stay in the company.

Personal Information

Supervisor	João Curado
Email	joao.curado@flomics.com
Institution	Flomics Biotech
Website	www.flomics.com
Group	Liquid biopsies group

Project

Web development & bioinformatic tools

Project Title:

Deconvolution of cell-free RNA transcriptome using RNA-seq

Keywords:

Plasma RNA; deconvolution model; sample heterogeneity; diagnosis;

Summary:

Years of literature demonstrate the existence of cell-free RNA, both messenger RNAs (mRNAs) and long noncoding RNAs (lncRNAs) originating from a wide variety of organs, from heart to the brain, and which change in response to external stimuli, namely diseases. Cell-free RNA molecules, circulating in human fluids such as plasma, saliva or urine, are potential windows into the health, phenotype or development stage of a variety of human organs, in a minimally invasive way. Despite this huge promise, the use of RNA sequencing (RNAseq) methods for global profiling of cell-free RNAs is in its infancy. In this project we propose to take advantage of the public available databases of RNA-seq such as Genotype-tissue expression (GTEx) consortium and tissue-specific gene expression databases, to determine the relative RNA contributions of each tissue in a sample using different methods (quadratic programming, least-squares regression, etc.). From a standard plasma RNA-seq experiment, the resulting tool will be used to calculate the relative contributions of the tissues and to monitor unexpected abnormalities that can be used as warning signs for complex disease detection.

References:

Koh, W. et al. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. Proc. Natl. Acad. Sci. 111, 7361–7366 (2014). Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods 12, 453–457 (2015). Everaert, C. et al. Performance assessment of total RNA sequencing of human biofluids and extracellular vesicles. Sci. Rep. 9, 17574 (2019).

Expected skills::

Transcriptomics; statistics; programming; autonomy

Possibility of funding::

Yes

Possible continuity with PhD: :

To be discussed