

> Nantes
2–5
juillet

JOBIM 2019

JOURNÉES OUVERTES
DE BIOLOGIE
INFORMATIQUE
& MATHÉMATIQUES

Thématiques :

Biologie structurale
Biologie des systèmes
Epidémiologie Génétique
Evolution/Phylogénie
Génomique/Métagénomique
Sciences des données

Abstracts

Keynotes :

Chloé-Agathe Azencott, Paris
Alexander Bockmayr, Berlin
Alessandra Carbone, Paris
Olivier Delaneau, Lausanne
Christophe Dessimoz, Lausanne
Juliette Martin, Lyon

<https://jobim2019.sciencesconf.org>

Dear attendees of the 20th edition of JOBIM, welcome in Nantes !

JOBIM is the French national conference dedicated to promoting an active interface between Biology, Computer Sciences, and Mathematics. After a previous visit to Nantes in 2009, JOBIM comes back this year in this same city from western France. Since the last visit, the bioinformatics community has impressively grown, and new fields are today covered. The impressive number of submissions at JOBIM 2019 reflects such an increase in our community. The Program Committee received a total of 260 submissions deciphered as 29 long presentations, 11 flash presentations, 15 demos and 204 posters. As a main novelty of the 2019th edition, JOBIM 2019 will present five additional thematic sessions that will cover particular topics more specialized.

We sincerely thank all the members of the Program Committee who helped us to set up a great scientific program by reviewing all submissions in time. This task would have been impossible without them! We also are grateful to the six invited speakers that have accepted to contribute to the success of the JOBIM edition in Nantes.

We are indebted to the organizing institutions, the SFBI, the GDR BIM, and the IFB. We are also grateful to all our partners and sponsors for their financial support.

Finally, we could also not forget to warmly thank Sophie Girault, Elodie Guidon, Aurore Morvan, and Jérémie Ségard as well as all the members of the organizing committee who worked collectively without counting sweat and tears to welcome the cream of bioinformaticians today in the best conditions.

Damien Eveillard and Audrey Bihouée

Organizing committee

Jérémie Bourdon and Richard Redon

Program committee

Our Partners and Sponsors



Contents

1	Talks	22
1.1	UniFIRE: the UniProt Functional annotation Inference Rule Engine . .	23
1.2	ProteoRE a Galaxy-based platform for the annotation and the interpretation of proteomics data in biomedical research	25
1.3	Redesign of iPPI-DB a database for modulators of Protein-Protein Interactions	26
1.4	A review of different ways to insert known RNA modules into RNA secondary structures	31
1.5	Adaptation to animal sources of <i>Salmonella enterica</i> subsp <i>enterica</i> deciphered by Genome Wide Association Study and Gene Ontology Enrichment Analysis at the pangenomic scale	39
1.6	Allele-specific analysis of epigenetic and transcriptomic data to study <i>Drosophila</i> developmental cis-regulatory architecture	47
1.7	CISPER: Computational Identification of Switch Points (in a Metabolic Network) within an Environmental Range	51
1.8	CONSENT: Scalable self-correction of long reads with multiple sequence alignment	54
1.9	Genotyping Structural Variations using Long Reads data	62
1.10	mCNA : a new methodology to improve high-resolution copy number variation analysis from next generation sequencing using unique molecular identifiers	70
1.11	Novel insight on molecular dynamics trajectories : local equilibrium viewed by kappa-segmentation	78
1.12	Reference-guided genome assembly in metagenomic samples	86
1.13	SPiP: a Splicing Prediction Pipeline addressing the diversity of splice alterations validated on a curated diagnostic set of 2 784 exonic and intronic variants	94
1.14	Architecture and evolution of blade assembly in beta-propeller lectins . .	103
1.15	Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using <i>gyrB</i> amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing	105
1.16	elPrep 4: A high-performance tool for sequence analysis	107
1.17	Exploring the uncharacterized human proteome using neXtProt	109

1.18	How build up soil bacterial co-occurrence networks from wide spatial scale sampling?	111
1.19	Merging of phenotypic information from cytometric profiles at the single-cell resolution	113
1.20	Sequana coverage: detection and characterization of genomic variations using running median and mixture models	114
1.21	Signature analysis of Structural Variants reveals a new subclass of hepatocellular carcinoma characterized by Cyclin A2/E1 alterations	116
1.22	Une nouvelle méthode pour évaluer l'impact des mesures de similarité sémantique sur l'annotation d'un groupe de gènes	118
1.23	A De Novo Robust Clustering Approach for Amplicon-Based Sequence Data	120
2	Flash presentations	142
2.1	Easy-HLA web application: new tools for HLA genotypes studies	143
2.2	Evaluation d'outils de quantification des transcrits alternatifs à partir de données de séquençage longue lecture Nanopore	144
2.3	From genomics to metagenomics: benchmark of variation graphs	145
2.4	Genomic evolution of contralateral breast cancer revealed from whole exome sequencing	146
2.5	Inter-individual variability in healthy human cytokine responses	147
2.6	Panache: a visualization tool for the exploration of plant pangenomes	148
2.7	scViz: a Rshiny app to easily explore scRNAseq data	149
2.8	Using Metabolomic Data to Predict Maize Yields	151
3	Software demonstrations	152
3.1	A precision medicine application: personalized contextualization of patients after kidney transplantation	153
3.2	Allogenomics – pipeline: prediction of the immune response from genetic variants during transplantation	154
3.3	EasyMatch-R: a web application to facilitate donor query in Hematopoietic Stem Cell Transplantation (HSCT)	155
3.4	INEX-MED: a Knowledge Graph to explore and link heterogeneous biomedical data	156
3.5	Leaves : Application d'aide à l'interprétation de variants	157
3.6	Linking structural and evolutionary information using MIToS jl	158
3.7	Omics Visualizer: a Cytoscape App to visualize omics data	159
3.8	Reconstruction of Transcript phylogenies using PhyloSofS	160
3.9	S3A: A Scalable and Accurate Annotated Assembly Tool for Gene Assembly	161
3.10	T1TADB: the database of Type I Toxin-Antitoxin systems	162

4	Platform session	163
4.1	AskOmics: a user-friendly interface to Semantic Web technologies for integrating local datasets with reference resources	164
4.2	DevOps bioinformatics services with Docker GitLab CI and Kubernetes	165
4.3	IFB-Biosphère Portail pour le Déploiement de Services Bioinformatiques sur une Fédération de Clouds	166
4.4	IFB-Biosphère Services Cloud pour l'Analyse des Données des Sciences de la Vie	167
4.5	Shiny and Galaxy interactive software for multi-source data analysis . .	168
4.6	Vers le déploiement continu d'infrastructures de calculs pour la bioinformatique	169
4.7	WAVES: a Web Application for Versatile Enhanced bioinformatic Services	170
5	Posters	171
5.1	A clinical bioinformatics framework for single-cell profiling of rare diseases	172
5.2	A graph theoretical approach to depicting sex-biased dispersal in ancient populations: mitochondrial DNA vs Y-chromosome variation	174
5.3	A novel DNA methylation signature for cell-type deconvolution in immunoncology	175
5.4	A set of methods to study three classes of non-coding RNAs	176
5.5	A state-of-the-art analysis of innovation software tools for primary analysis for Oxford Nanopore sequence data	177
5.6	A tool for very fast taxonomic comparison of genomic sequences	178
5.7	A web server for identification and analysis of coevolution in overlapping proteins	179
5.8	A workflow based on self-organizing map for clustering stable structures of proteins from molecular dynamics simulations	183
5.9	A workflow to analyse single-cell transcriptomes from heterogeneous tumors	184
5.10	A workflow to build a relevant bacterial genome sub-dataset from public databases	185
5.11	Advanced Visualization of Data Comparisons with BiocompR	186
5.12	ALFA: Annotation Landscape For Aligned reads	187
5.13	AllMine a flexible pipeline for allele mining	188
5.14	An Integrative Deep-Learning Framework for Analyzing Native Spatial Chromatin Dynamics	189
5.15	Analyse de longs reads Nanopore avec des k-mers à erreurs	190
5.16	Analyse du métagénome microbien fonctionnel des sols de parcelles paysannes en zone subsahélienne (Burkina Faso)	191
5.17	Analysis of multi-omics data: a comparison of correlation and functional integrative approaches on a cancer dataset	192
5.18	Analysis workflow for low frequency variant detection	193

5.19	Apollo method: statistical inference to reveal hidden data in chromosome contact maps	194
5.20	Assessment of inflammatory and immune pathways in Rheumatoid Arthritis patients using BIOPRED kit	200
5.21	BamCramConverter: Utility for Easy Alignment/Map Data Storage	201
5.22	Benchmarking Hi-C scaffolders	202
5.23	Bioanalysis activities on the ABiMS (Analysis and Bioinformatic for Marine Science) platform	203
5.24	Bioconvert a common bioinformatics format converter library: status and perspectives	204
5.25	Bioinformatic characterization of the role of TRIP12 in pancreatic adenocarcinoma	205
5.26	Biomarkers for neurodegenerative diseases	206
5.27	Burrowing functional and immunogenetic information through the 1000 Genomes Project with Ferret v 3 0	207
5.28	CADBIOM – Un logiciel de modélisation des réseaux de signalisation	208
5.29	Can we detect DNA methylation with Oxford Nanopore reads ?	209
5.30	Caractérisation de CNV (variants de nombre de copies) à partir de données de séquences exoniques simulées	210
5.31	CD4 T cell reprogramming in brain-injured patients	211
5.32	cDNA length improvement is essential to allow better isoform characterization for long read RNA sequencing	212
5.33	Characterization of Hepatitis B Virus genomes identified by viral capture in Hepatocellular Carcinomas from European and African patients	213
5.34	checkMyIndex: a web-based R/Shiny interface for choosing compatible sequencing indexes	214
5.35	ChIPuana: from raw data to epigenomic dynamics	215
5.36	Chronic mood instability cardiometabolic risk and functional impairment in bipolar patients: relevance of a multidimensional approach	216
5.37	Classification of the evolutionary trajectories of cognitive functions	217
5.38	Co-activity networks reveal the structure of planktonic symbioses in the global ocean	219
5.39	Comment annoter et analyser les protéines à motifs répétés : Cas des protéines contenant des répétitions riches en leucine (LRR) chez le riz	220
5.40	Comment prédire un gRNA efficace dans des contextes expérimentaux variés ? En apprenant des gRNA publiés	221
5.41	Comparaison des réseaux métaboliques de bactéries phytopathogènes	224
5.42	Comparative genomics of Rhizophagus irregularis R cerebriforme R diaphanus and Gigaspora rosea zHighlights specific genetic features in Glomeromycotina	225
5.43	Comparative microbial pangenomics to explore mobilome dynamics	226
5.44	Comparison of efficiency of gene regulatory network inference algorithms on genomic and transcriptomic data	227

5.45	Comparison of large insertion variant callers on whole exome sequencing	228
5.46	Comparison of tolerogenic dendritic cells used in clinic with other in vitro-derived myeloid cells by epigenetic and transcriptomic analyses . . .	230
5.47	Conciliation of process description and molecular interaction networks using logical properties of ontology	231
5.48	Conversion from quantitative model in sbml core to qualitative model in sbml qual	232
5.49	CuteVariant: Un visualisateur de variants génétiques pour le diagnostic médical	233
5.50	Cypascan: an online tool for star allele calling in pharmacogenetics . . .	234
5.51	De-centralized database: new challenges to design innovative contextualization algorithms	235
5.52	Deciphering the activation states of plasmacytoid dendritic cells their dynamical relationships and their molecular regulation	236
5.53	Detection of transcriptional regulatory motifs specific to plant gene responses in stress conditions	237
5.54	Detection of unknown genetically modified organisms (GMO) by statistical analysis of high-throughput sequencing data	238
5.55	Development and validation of an alloscore in kidney transplantation . .	239
5.56	Development of a complete HLA analysis pipeline: HLA-Functional Immunogenomic eXploration (HLA-FIX)	240
5.57	Development of a novel multi-scale integrative computational method dedicated to the analysis of heterogeneous omics data	242
5.58	Divergent Clonal CD8+ T Cell Differentiation Establishes a Repertoire of Distinct Memory T Cell Clones Following Human Viral Infections . . .	243
5.59	Dynamic cell population modeling with UpPMABoSS	245
5.60	Développement et validation de pipelines pour l'analyse de données NGS dans le cadre du diagnostic en oncogénétique somatique	246
5.61	Easy16S : a user-friendly Shiny interface for analysis and visualization of metagenomic data	247
5.62	Eoulsan workflows for tag-based and full-transcript single-cell RNA-seq protocols	248
5.63	Epigenome-wide association study reveals immunogenetic targets of DNA methylation modification by HIV-1	249
5.64	Error Correction Schemes for DNA Storage with Nanopore Sequencing .	250
5.65	Etude de la composante auto-immune de la Polyarthrite Rhumatoïde . .	251
5.66	Etude de la trajectoire de fréquences alléliques pathogènes à travers le temps et l'espace	252
5.67	Evolution of the angiotensin II receptors AT1 and AT2: Insights from molecular dynamics simulations	253
5.68	Exome sequencing in Hereditary Hypophosphatemic Rickets with Hypercalciuria	254
5.69	Exploring relationship between to neuro-inflammatory diseases	255

5.70	Exploring white matter hyperintensities genetic associations through the use of external transcriptomic data	256
5.71	Fast neutron variants detection in TILLING crop populations	257
5.72	Feedback on a comparative metatranscriptomic analysis	260
5.73	Flexible analysis of WGS of bacterial genomes using wgMLST approach	261
5.74	Formatage et annotation des variants structuraux - Présentation du logiciel Svagga	262
5.75	French Guiana Severe Syndromes a metagenomics analysis of unknown dark clinical samples	263
5.76	From primary to tertiary structure analyses of experimentally proven O-GlcNACylated sites for an optimised prediction	264
5.77	GARDEN-NET: a tool for chromatin 3D interaction network visualization	265
5.78	Genetic determinants of intracranial aneurism in autosomal dominant polycystic kidney disease	266
5.79	Genome-scale metabolic networks from two asian brown algae : integrating targeted pathways analyses and metabolomic data	267
5.80	GSAAn : Une alternative aux analyses statistiques des groupes de gènes .	268
5.81	Hermès : a management tool for Next-Generation Sequencing analysis on a genomic platform	269
5.82	High-Throughput Sequencing from preservative ethanol and bulk of specimens to jointly assess species and population genetic diversity of colonial ascidians	270
5.83	How to involve repetitive regions in scaffolding improvement	271
5.84	IBENS Genomics core facility	272
5.85	Identification des proies de gastéropodes venimeux (Conoidea) par approche de métabarcoding	273
5.86	Identification of a common transcriptional signature for regulatory B cells in Humans and Mice	274
5.87	Identification of causal signature from omics data integration and network reasoning-based analysis	275
5.88	Identification of genomic regions for high-resolution taxonomic profiling using long-read sequencing technology	276
5.89	Identifying predictive biomarkers for breast cancer treatment using an integrative transcriptomic analysis	277
5.90	Ignoring the optimal set of tissue-specific metabolic networks can bias the interpretation of data	278
5.91	Impact de la manipulation thermique embryonnaire sur le méthylome de caille japonaise	279
5.92	In-silico benchmark of methods for detecting differentially abundant features between metagenomics samples	281
5.93	Industrial NGS analysis processes from sequencing to variant interpretation on MOABI platform	282

5.94	INEX-MED: INtegration and EXploration of heterogeneous bio-MEDical data	283
5.95	Integration of transcriptomic and proteomic data for biomarker discovery in Lassa fever	284
5.96	Interactions de SNPs d'ordre N par pattern mining	285
5.97	Joint analysis of multiple compositional data	286
5.98	Large-scale RNA-seq datasets enable the detection of genes with a differential expression dispersion in cancer	287
5.99	LC-MS/MS tool and interactive visualizations integration on Galaxy Workflow4Metabolomics infrastructure	288
5.100	LeAFtool: Lesion Area Finding tool	289
5.101	Linking Allele-Specific Expression And Natural Selection In Wild Populations	290
5.102	Long-read pacbio amplicon analysis From raw data to final results . . .	291
5.103	Longitudinal analysis of immune cells in kidney transplantation rejection by single-cell RNA-seq	292
5.104	Mechanism of mechanosensation mediated by the angiotensin II receptor 1: a molecular dynamics approach	293
5.105	MetaChick: assembly and analysis of chicken cecal microbiome reveals wide variations according to the production methods	294
5.106	Metagenomic analysis of an African beer ecosystem using FoodMicrobiomeTransfert application	295
5.107	MetagWGS: an automated Nextflow pipeline for metagenome	296
5.108	Metavisitor-2 a suite of Galaxy tools for simple and rapid detection and discovery of viruses in Deep Sequence Data	297
5.109	METdb: a genomic reference database for marine species	298
5.110	MiBiOmics a shiny application for graph-based multi-omics analysis . .	299
5.111	Microbial communities from deep-lake sediments of Lake Baikal Siberia .	301
5.112	MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic and metabolic comparative analysis	302
5.113	Mise en place d'un LIMS enrichi par une organisation harmonisée des métadonnées	303
5.114	Mise en place d'un pipeline automatisé d'analyses multivariées pour la cytométrie en flux multi-couleurs	304
5.115	MobiDL: next generation family of WDL DNA-NGS pipelines	305
5.116	Modelling the differentiation dynamics of monocytes in contact with CLL B cells	306
5.117	Molecular Modeling of the Asc-1 Transporter: Insights into the first steps of the transport mechanism	307
5.118	Multi-factor Data Normalization enables the detection of LOH in amplicon sequencing data	308
5.119	Multi-omics approach to predict drug response in liver cancer cell lines .	310

5.120	MYC-MACS (MYCétes pour une Meilleure Acquisition des Connaissances Scientifiques)	311
5.121	mzLabelEditor: un outil pour annoter des spectres de masse	312
5.122	Navigating the treacherous waters of HLA imputation with the SHLARC (SNP-HLA Reference Consortium)	313
5.123	OLOGRAM : Modeling the distribution of overlap length between genomic regions sets	314
5.124	Omics Data Analysis Facilities in a Biomedical Research Institute	315
5.125	Palimpsest: an R package for studying mutational and structural variants signatures along clonal evolution in cancer from single or multiple samples sequencing	316
5.126	PanGBank: depicting microbial species diversity via PPanGGOLiN	317
5.127	Pathway analysis from time course gene expression experiments to unveil the dynamic of cellular responses	318
5.128	Performance evaluation of bioinformatics tools for predicting allergenic proteins in food	319
5.129	Pioneer data-driven methods generating synthetic data: the HLA “avatars” are shifting paradigms in data sharing	320
5.130	Pipeline d’analyse et de visualisation avancés de single cell RNAseq (SChnurR)	321
5.131	Polygenic Risk Scores for Autism spectrum disorder and Alzheimer’s disease enable the identification of new white matter tract biomarkers	322
5.132	Population demographic estimation using simulated data	323
5.133	Positive Multistate Protein Design	324
5.134	Predicting isoform transcripts: lessons from human mouse and dog	326
5.135	Prediction of candidate disease genes through deep learning on multiplex biological networks	327
5.136	PREDIdicting bacterial PATHogenicity on plant: PREDIPATH	328
5.137	PrivAS: a tool to perform Privacy-Preserving Association Studies	329
5.138	ProteoCardis: an intestinal metaproteome-wide association study of coronary artery disease	335
5.139	PSH une fonction de hachage issue du domaine du traitement d’images permettant l’indexation et la comparaison de séquences ADN	336
5.140	R: Ecology Met A Data Language	344
5.141	RandomRead : a sequence-read simulator program for metagenomic shotgun	345
5.142	Recherche par clustering de gènes impliqués dans le syndrome PTLD	346
5.143	ReClustOR a Re-Clustering tool using an Open-Reference method that improves OTU definition	347
5.144	Recommendation system embedded in metabolic network visualization: a new way of looking at metabolomics results	350
5.145	Recurrent deletions of 3q13 31 in human osteosarcoma commonly affect TUSC7 and LINC00901	351

5.146	Reducing your NGS dataset using a set of targets : how to optimize storage space compute time and analysis accuracy	352
5.147	Refract-Lyma and CHU hub: from a research cohort to a regional electronic medical record system and back	353
5.148	REGULOUT software identifies regulatory outliers that have unexpected transcription profile inside a group of ortholog genes	354
5.149	repeatsFinder: a web-based R/Shiny interface for visualizing and characterize genomic repeated regions	355
5.150	RGCCA with block-wise missing structure	356
5.151	RPG: fast and efficient in silico protein digestion	357
5.152	RSAT var-tools: an accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding	358
5.153	Régulation par les miARN des gènes régulant la fécondité et le développement embryonnaire précoce chez le poisson medaka (<i>Oryzias latipes</i>)	359
5.154	Réseaux de co-expression pour l'analyse de données de protéomique pour la compréhension des mécanismes d'action de contaminants chez une espèce non-modèle <i>Gammarus fossarum</i>	360
5.155	Sex-specific differences in microglia inflammatory response during brain development	361
5.156	Simulating the impact of Serological-Test-and-Treat measures to target the hidden <i>P. vivax</i> reservoir: public health impact and primaquine overtreatment	363
5.157	Single cell transcriptomic analysis for a better understanding of human CD8 regulatory T cells	364
5.158	Single-cell analysis of human intestinal organoids reveals the ENS progenitor cells contribution on the gut mesoderm development	365
5.159	SIStemA : Gene expression database of human Stem Cell and their differentiated derivative	366
5.160	SpecOMS: découverte des modifications portées par les protéines	367
5.161	srnaMapper: a mapping tool for short RNA reads	368
5.162	Statistical inference of immunogenetic parameters reveals an HLA allele associated with pediatric Focal Segmental Glomerulosclerosis	369
5.163	Stratégie de compression de données de séquençage cliniques	370
5.164	Stratégie de priorisation de variants après séquençage ciblé de l'ADN	371
5.165	Structuration et consolidation de résultats d'analyses de RNAseq et Polymorphisme	374
5.166	Study of sperm epigenetic contribution for the regulation of embryonic gene transcription in early development	375
5.167	Supervised contact prediction in proteins	376
5.168	Symmetries of the hypercube : a tool for regulatory networks analysis	377
5.169	Séquençage d'ADN natif dédié à l'étude du microbiome sur le MinION ® : retour d'expérience de la paillasse à l'assignation taxonomique	378

5.170	The ClermonTyper: an easy-to-use and accurate in silico tool for Escherichia genus strain phylotyping	381
5.171	The extra mile of Gene Set Enrichment Analysis: seeing the data	382
5.172	The limit of cell specification concept: a lesson from scRNA-Seq on early human development	383
5.173	The Migale bioinformatics platform	384
5.174	The relationship between gene co-expression network connectivity and phenotypic prediction sheds light at the core of the omnigenic theory . . .	385
5.175	The role of the LNR domain-containing protein explosion in Oithona nana male differentiation (Crustacea Cyclopoida)	386
5.176	The SeCoNeMo approach and its application to ICE annotation in Firmicutes	387
5.177	The SIRP gene family: widespread conservation in animals haplotypic polymorphisms in humans and its therapeutic consequences for monoclonal antibody reactivity	388
5.178	Transcript-aware Clustering of Orthologous Exons Shed Light on Alternative Splicing Evolution	389
5.179	Transcriptional and functional analyzes of symbiotic coral micro-algae in the framework of Tara Pacific expedition	393
5.180	Transcriptome analysis to identify co-expressed gene networks as a molecular signature for childhood trauma-related mood disorders	394
5.181	Transcriptomic analysis of habenular asymmetries in the catshark S canicula	396
5.182	Transcriptomics Signature of Type I Narcolepsy (T1N)	397
5.183	UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries . .	398
5.184	Understanding Chemical-Genetic Interactions	399
5.185	Unraveling the rules of the Exon Junction Complex deposition with CLIP-seq	400
5.186	Unveiling the neo-antigen landscape of malignant mesothelioma using computational predictions and multi-omics data	401
5.187	Using residues coevolution to search for protein homologs through alignment of Potts models	402
5.188	VCF2Table : a VCF prettifier for the command line	403
5.189	Vidjil une plateforme pour l'analyse des répertoires immunitaires	405
5.190	ViSEAGO: Easier data mining of biological functions organized into clusters using Gene Ontology and semantic similarity	406
5.191	Visualizing metadata change in networks and / or clusters	407
5.192	Which genome browser to use for my data ?	408
5.193	Évaluation de la qualité et comparaison des assemblages des génomes . .	409

A web server for identification and analysis of coevolution in overlapping proteins

Elin Teppa¹, Alessandra Carbone^{1,2}

¹ Laboratory of Computational and Quantitative Biology (LCQB), Sorbonne Université, CNRS, IBPS, UMR7238, 4, place Jussieu 75005 Paris, France.

² Institut Universitaire de France (IUF) 75005 Paris, France

Corresponding author: elinteppa@gmail.com

Abstract

Overlapping genes exist in all domains of life and are especially abundant in viral genomes. The existence of overlapping reading frames increases the rising of deleterious mutations for one of the proteins, since a single nucleotide substitution may affect both proteins. Molecular coevolution may be seen as a mechanism to tolerate or compensate unfavorable mutations, decreasing the evolutionary constraints in the overlapping region. For instance, a favorable mutation in one reading frame may be unfavorable in the other reading frame and additional mutations may be needed to compensate the first mutation. Although molecular coevolution was widely used in viral genomes, the “overlap problem” was disregarded. Here, we present a server that facilitates the analysis of coevolution in overlapping proteins and of the impact of mutations in another ORF.

Keywords: coevolution; compensatory mutations, virus, overlapping proteins

Introduction

Multiple studies of coevolving positions in viral sequences have been useful to understand functionally significant residues [1,2], to predict protein-protein interactions [3], to modulate viral fusion [4] and to identify drug resistance mutations [5–9] among others.

The genomes of most viral species have overlapping genes—two or more proteins coded for by the same nucleotide sequence. ORFs may overlap in various manners considering the type, the direction of transcription and the ORFs’ phase (Fig 1). Sequence analysis in overlapping ORFs represents a challenge due to changes in the nucleotide sequence that may simultaneously affect both proteins within their overlapping region.

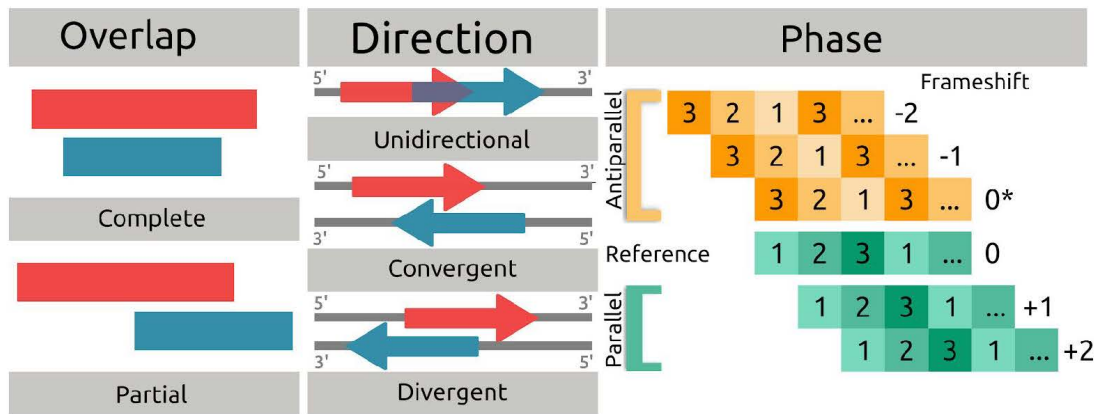


Figure 1: Definitions on ORFs overlap.

An overlap between two ORFs can be complete (if an ORF is nested within the other) or partial (if only the 3' or 5' end are overlapping). ORFs can overlap on the same strand, or in the case of a double-stranded genome, on the reverse complementary strand. Hence, three directions are possible: unidirectional, convergent and divergent. The reference ORF, in a pair of overlapping ORFs, is called phase 0. Overlaps in a parallel strand can be in two phases whereas antiparallel-strand overlaps can be in three phases.

Given that coevolution may be seen as a mechanism to tolerate or compensate unfavorable mutations, molecular coevolution in the overlapping region may help to decrease evolutionary constraints. As far as we know, there is no study of coevolution that considers both overlapped proteins.

In the overlapping region, coevolution in an ORF: may be mirrored by coevolution in the other ORF; may generate a non-synonymous substitution which in turn may be compensated by other mutations (inside or outside the overlapping region); may generate synonymous substitutions (Fig 2).

The motivation for this server is to provide a tool to facilitate the analysis of coevolution in overlapped protein and of the impact of mutations in another ORF. To do that we combine information at protein and nucleotide levels.

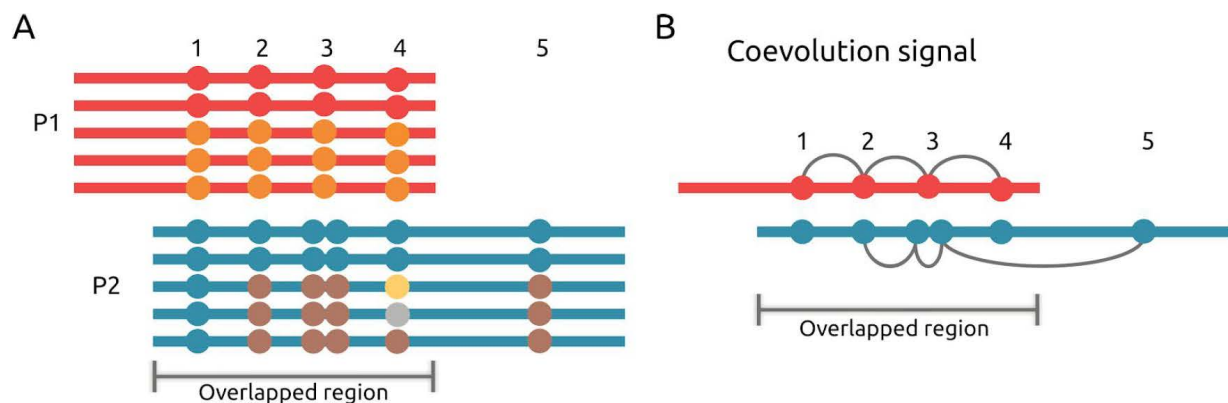


Figure 2: Coevolution pattern in overlapping region.

Different effects of four coevolving positions in the overlapped region of two proteins (P1 and P2). **A:** A cluster of four coevolving positions is represented in P1's alignment where two sequences maintain the wild-type residues (red circles) and three sequences show mutations on all positions (orange circles). A

mutation in P1 may be coupled by synonymous substitutions in P2 (column 1); the same non-synonymous substitution (column 2), two non-synonymous substitutions in adjacent positions (column 3); a variety of non-synonymous substitutions (column 4). The cluster of coevolving positions may also contain positions outside the overlapping region (column 5). **B:** P1 shows a coevolution signal between the first four positions (gray lines) which partially coincides with coevolution detected in P2.

Methods

Input

The input is a nucleotide alignment of the pair of overlapping protein sequences to be analyzed. It will contain the overlapped and non-overlapped regions of both proteins, as well as the start and end positions of the proteins and their corresponding DNA strand (parallel or antiparallel) (Figure 1).

Workflow

Given a DNA alignment and its associated distance tree that can be provided or optionally generated automatically, all subsets of sequences corresponding to the subtrees of the tree are systematically considered for coevolution analysis. For each subset, the ORF1 and ORF2 sequences (Fig 2) are translated into amino acids and the resulting protein alignments are used as input to predict coevolving positions using the BIS2 algorithm [10,11]. Our iterative strategy allows applying BIS2 in a large number of conserved sequences. As part of the result, the clusters of coevolving positions detected for both proteins are provided. If coevolution is detected in the overlapping region for one of the proteins, the effect of variation is analyzed in the other protein. By analyzing the subset of sequences where the cluster is detected for the first protein, we identify if the coevolving positions are accompanied by one or more synonymous/non-synonymous substitution(s) and if these positions also show coevolution in the second protein.

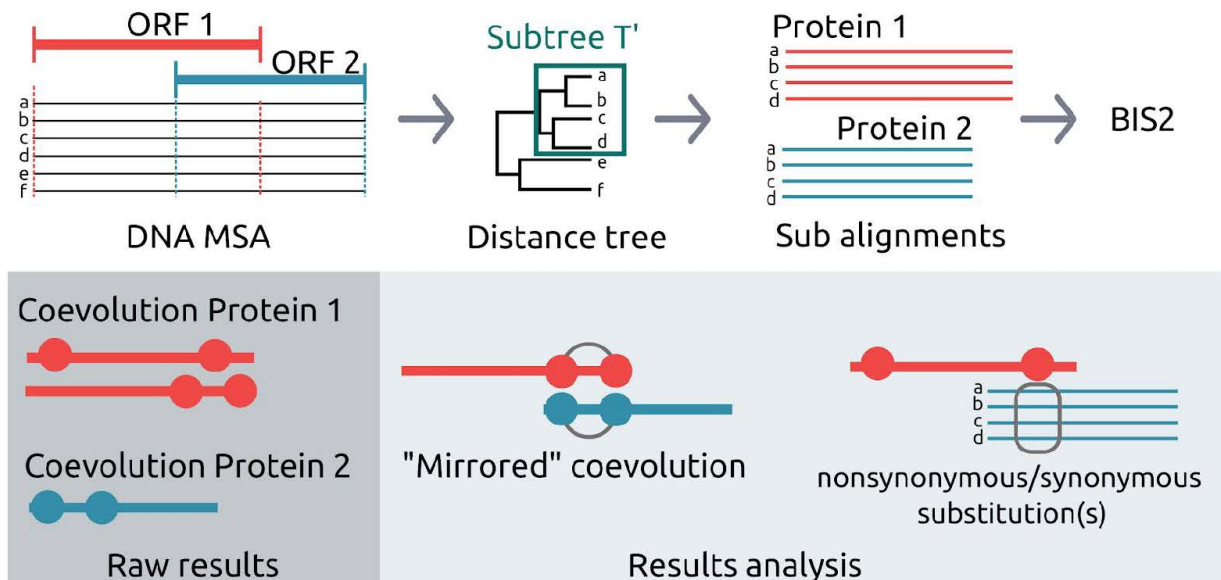


Figure 3: Schematic representation of the workflow from the input sequences to the results.

The DNA alignment covering both ORFs to be analyzed is used to generate a distance tree, optionally the tree may be provided by the user. Then, the tree is partitioned in all possible subtrees. The protein sequences corresponding to the subtrees are used as input to compute coevolution using BIS2 algorithm. The results include the coevolution of each of the proteins, as well as the effect of the mutations of one protein on the other. It is also indicated if both proteins show coevolution in equivalent positions ("mirrored" coevolution) or if the mutation of the co-evolved position in a protein is accompanied by synonymous or nonsynonymous mutations in the other.

Conclusions

We have developed an interactive web server providing an intuitive representation of the coevolved residues predicted in overlapping proteins. To the best of our knowledge, this is the only publicly available method designed to analyze coevolution in overlapping protein sequences. The server is simple to use and it provides a powerful tool the virologist and the biologist to compute coevolution and analyze the effect of mutations in overlapping regions. Its results should help to elucidate the evolutionary constraints found in overlapping ORFs.

Acknowledgements

This work was supported by the French "Agence Nationale de la Recherche sur le SIDA et les hépatites virales" (ANRS CSS4 ECTZ25224 – 2017-19 to AC; www.anrs.fr).

References

1. Le L, Leluk J. Study on phylogenetic relationships, variability, and correlated mutations in M2 proteins of influenza virus A. *PLoS One*. 2011;6: e22970.
2. Jain J, Mathur K, Shrinet J, Bhatnagar RK, Sunil S. Analysis of coevolution in nonstructural proteins of chikungunya virus. *Virol J*. 2016;13: 86.
3. Champeimont R, Laine E, Hu S-W, Penin F, Carbone A. Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. *Sci Rep*. 2016;6: 26401.
4. Douam F, Fusil F, Enguehard M, Dib L, Nadalin F, Schwaller L, et al. A protein coevolution method uncovers critical features of the Hepatitis C Virus fusion mechanism. *PLoS Pathog*. 2018;14: e1006908.
5. Rhee S-Y, Liu TF, Holmes SP, Shafer RW. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput Biol*. 2007;3: e87.
6. Handel A, Regoes RR, Antia R. The role of compensatory mutations in the emergence of drug resistance. *PLoS Comput Biol*. 2006;2: e137.
7. González-Ortega E, Ballana E, Badia R, Clotet B, Esté JA. Compensatory mutations rescue the virus replicative capacity of VIRIP-resistant HIV-1. *Antiviral Res*. 2011;92: 479–483.
8. Bloom JD, Gong LI, Baltimore D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science*. 2010;328: 1272–1275.
9. Tanaka MM, Valckenborgh F. Escaping an evolutionary lobster trap: drug resistance and compensatory mutation in a fluctuating environment. *Evolution*. 2011;65: 1376–1387.
10. Dib L, Carbone A. Protein fragments: functional and structural roles of their coevolution networks. *PLoS One*. 2012;7: e48124.
11. Oteri F, Nadalin F, Champeimont R, Carbone A. BIS2Analyzer: a server for co-evolution analysis of conserved protein families. *Nucleic Acids Res*. 2017;45: W307–W314.