Epilepsy Subgroup Identification Based On Shared Comorbidity Risk Factors And Pleiotropic Findings, Using Data-Mining And Disease-Modeling Approaches

Nora Filep
Fraunhofer SCAI

Epilepsies are characterized by recurrent seizures that affect people all around the world. Increasing evidence shows that seizure is not the only concern in epilepsy treatment. Comorbid health conditions are common among people with epilepsy. One of the possible explanations for this phenomenon is that common pathogenic mechanism or mechanisms mediate the co-occurrence of epilepsy and the comorbid condition.

To identify such common mechanisms between epilepsies and different comorbidities, a computer readable epilepsy disease model has been built based on prior knowledge. Text mining and knowledge extraction methods were used to extract epilepsy-related triples from MEDLINE literature sources and various databases. Those findings were imputed into our Epilepsy BEL-model, which is the base of NeuroMMSig_Epilepsy, the novel prior knowledge based mechanism annotation tool. Epidemiological and pleiotropic findings were compared to see whether shared SNPs, genes and mechanisms indicate higher comorbidity rates in the epilepsy population. Supervised classification was used to cluster four epilepsy comorbidity groups based on ICD-10, namely Neoplasms; Endocrine, nutritional and metabolic diseases; Mental and behavioral disorders; and Diseases of the nervous system. We enriched our disease-model with AED-related pharmacogenetic and pharmacokinetic knowledge, and classified 19 AEDs into seven groups, based on drug mechanism of action.

The highest pleiotropic score was found between epilepsies and breast carcinoma (0.52) however, epidemiologic findings do not show significantly higher prevalence ratio for those diseases. Most of the diseases, which share high rate of genes with epilepsies, do not show significant comorbid association with them. We have to conclude, that shared genes do not necessarily indicate comorbid conditions, and epigenetic factors play a key role in comorbidities. However shared molecular mechanism detection may have a promising future in drug repurposing and in personalized medicine for epilepsy patients with multimorbidities.

Neuroimaging technologies are the most important diagnostic tools in today´s epilepsy treatment. Future research in computational biology should take the challenge to map molecular signatures to epilepsy-related imaging features.

Keywords : Epilepsy, pleiotropy, computational biology, disease-modeling, molecular mechanisms, BEL, comorbidity, knowledge extraction

Ligand-protein interaction analysis for drug-target interaction prediction in drug repositioning

Daniele Parisi and **Yves Moreau**
KU Leuven

INTRODUCTION. The computational analysis brought in by Bioinformatics allows to manage massive amount of data, in order to fill sparse matrices about compounds-target interaction. The drug-target interaction predictions are inferred using different techniques (ligand-based, target-based), with specific pros and cons. They have been recently tackled with new chemogenomic methods, called Proteochemometric techniques, by integrating both the information related to ligands and targets. An example is the ligand-protein interaction profile, a unique and highly informative characteristic that can be easily analysed and compared[8]. In this work I studied and compared drug-target interaction profiles generated with a tool made for bioisosterical replacement, KRIPO, to understand their predictive power in drug repositioning. Furthermore I used those predictions in a real case to support a process of hit discovery for new immunosuppressant drugs.

METHODS. This work aims to use drug-target interaction similarity for interaction prediction and drug repositioning. Starting from 68.000 PDB complexes, the interaction profiles have been calculated on the pharmacophores of the interactions between the ligand and the pocket, and stored as fingerprints[9]. Then, each fingerprint has been compared with all the others using a modified Tanimoto Coefficient[9]. The assumption is that pairs ligand-protein with similar interaction profiles would be able to cross-interact, representing a starting point for a repositioning protocol. To evaluate the predictive power, the new predicted couples ligand-protein have been plotted with both their interaction similarity from KRIPO and their experimental activity from CHEMBL. Furthermore this predictive approach has been used to suggest new hit compounds able to inhibit specific kinases involved in the immuno response by the B-cell activation. About hundred compounds with similar interactions have been provided to the medicinal chemists who selected 20 molecules to test against 5 kinases, measuring the level of inhibition.

RESULTS. The large scale analysis of predictions showed the weak points of KRIPO in drug-target interaction prediction, due to the excessively long and descriptive fingerprints which reduce the range of similarity to 3 units out of 10. However, the in vitro tests showed that half of the predicted compounds had a very high level of inhibition ,>75%. In conclusion, the approach showed here gave interesting results when coupled with experience of chemists but more work is needed to improve the quality of the prediction.

Keywords : drug-target interaction, interaction prediction, pharmacophore fingerprinting, drug discovery, drug repositioning, big data analysis

Analysis of the molecular network underlying left ventricular remodeling after myocardial infarction

Marie Cuvelliez[1], Christophe Bauters[1,2], Thomas Kelder[3], Marijana Radonjic[3], Philippe Amouyel[1] and Florence Pinet[1]

[1]Institut Pasteur de Lille, [2]CHRU de Lille, [3]EdgeLeap

Cardiovascular diseases are one of the most important causes of mortality in occidental countries. Following a myocardial infarction (MI), 30% of patients develop a left ventricular remodeling (LVR) that could lead to heart failure (HF).

REVE-2 is a multicentre study of 246 patients hospitalized for a first anterior wall MI. An echocardiographic follow-up and serial blood samples were realized at 4 time-points after MI: at inclusion (baseline), 1 month, 3 months and 1 year. At each time-point, 25 molecular variables were measured in the patient's plasma (19 proteins and 6 non coding RNAs). These variables expression was compared between patients with high LVR (>20%) and low LVR (<20%) measured between the echocardiography at 1 year and inclusion.

A network based on known molecular interactions was built using all the molecular variables measured in the REVE-2 patient's plasma. The EdgeBox platform (EdgeLeap), composed of 12 public databases (ENCODE, EnsemblGenes, HMDB, Microcosm, miRBase, miRecords, miRTarBase, Reactome, STRING, TargetScan, Tfe and WikiPathways), was used as a resource of public knowledge on molecular interactions. The network includes the REVE-2 variables, the direct neighbors of these variables and the molecules that are part of the shortest paths up to 3 edges between 2 REVE-2 variables. The network, called REVE-2, is composed of 1310 molecules (nodes), including 1263 proteins, 24 miRNAs, 22 metabolites and a long non coding RNA, linked by 8639 interactions (edges). The network analysis identified 40 clusters that have been annotated to biological processes from Gene Ontology (GO). The network is visualized using Cytoscape (version 3.2.1). An analysis of the active modules was realized at each time-points (baseline, 1 month, 3 months and 1 year) using the REVE-2 variables that are significantly modulated between the 2 groups of patients (Pinet et al., 2017). The majority of changes of the expression of the network's molecules were observed at baseline and 3 months after MI, corresponding respectively to the post-MI phase and to the LVR development. A centrality analysis of each node allowed determining their importance in the network, a high centrality suggesting a crucial role of the node in the pathophysiological process. The transcription factors EP300, CTCF and ESR1 are expressed in the heart and are known to regulate SOD2 activity, a protein involves in the oxidative stress regulation in cardiomyocytes. Four non coding ARNs, miR-26b-5p, miR-17-5p, miR-335-5p and miR-375, have a high centrality. The first two are expressed in the heart, unlike the last two. The plasmatic levels of miR-26b-5p are decreased in patients with acute HF and miR-17-5p is involved in cardiomyocyte apoptosis and cardiac fibrosis.

The REVE-2 network analysis allows to study the physiopathological mechanisms underlying LVR after MI in a time-dependent manner, and to identify new potential biomarkers to detect LVR associated to a high risk of heart failure.

Keywords : Left ventricule remodeling, systems biology, molecular network, biomarkers

Black box revelation of linkage between two disparate levels of pathophysiology “Genetics/Imaging” using integrated multi-scale modelling

Sepehr Golrizkhatami, Anandhi Iyappan and Martin Hofmann-Apitius
Fraunhofer Institute

Alzheimer disease (AD) is one of the most complex neurological diseases. The aetiology of the disease is not fully understood, presumably due to the complex dysregulation of different biology levels. Dysregulation events may take place at multiple scales; from the molecular level to the cellular and organ level, finally resulting in structural and functional alterations of the brain. As a consequence, for a profound mechanistic understanding of neurodegeneration, readouts representing different biological entities at different scales need to be investigated and their relationships and dependencies have to be described and modelled at systems level. Considering the above facts, neuroimaging is widely applied to provide important insights into disease pathophysiology at brain region and organ level. Neuroimaging readouts are often referred to as endophenotypes for brain disorders like AD. In the genotype-disease arena, an endophenotype is proposed to serve as a measurable component along the pathway between genes and disease phenotype (Irving I. Gottesman, 2003). In analogy to this concept, neuroimaging endophenotypes are quantitative markers of brain structure and function that manifest a link between molecular dysregulation events and a organ-specific pathophysiology context. The hope is, that linking molecular processes to these endophenotypes will significantly improve our understanding of the functional consequences of genetic variation indicators such as single nucleotide polymorphisms (SNPs) at neuroscience systems level. Following that paradigm, during the last ten years, the number of studies that focus on imaging phenotypes and selecting and associating SNPs to imaging readouts in the AD context, has increased. However, the existing genetics imaging studies are purely association-based, which means, that these studies provide statistical evidence for a possible link between a SNP and an imaging feature. The real mechanistic context that links genes to imaging readouts remains unclear with this sort of association study.

In the course of this work, we have tried to work out new routes to establish cause-and-effect models linking genetic variation information at genome scale to imaging readouts at the organ level. Systematic modeling of causes-and-effects bears great potential to assess the functional impact of SNPs in the context of molecular pathophysiology processes (Naz et al., 2016). Although, there is a huge gap between information at genome (genetics) level and organ (neuroimaging) level, we have stretched the limits of knowledge-based modeling to bridge the gap between genome-scale and imaging/clinical level data in the context of neurodegenerative disease. We combine mechanistic information representing genetic variation with protein–protein interaction and pathway information and reach out to the clinical and organ level including neuro-imaging readouts. The resulting models allow us to decipher, how a gene function modifier (SNP) could impact on molecular processes that ultimately result in cellular and organ-level changes that can be detected in neuro-imaging scans.

Keywords : Neuroimaging, System biology modelling, Alzheimer Disease, Single nucleotide polymorphisms

Genome-wide SNP analysis reveals distinct origins of *Trypanosoma evansi* and *Trypanosoma equiperdum*

Bart Cuypers[1], Frederik Van den Broeck[2], Nick Van Reet[2], Conor J. Meehan[2], Julien Cauchard[3], Jonathan M. Wilkes[4], Filip Claes[5], Bruno Goddeeris[6], Hadush Birhanu[7], Jean-Claude Dujardin[2], Kris Laukens[1], Philippe Büscher[2] and Stijn Deborggraev[2]

[1]University Of Antwerp, [2]Institute of Tropical Medicine Antwerp, [3]Anses Dozulé Laboratory for equine diseases, [4]Wellcome Trust Centre of Molecular Parasitology, University of Glasgow, [5]Food and Agriculture Organization of the United Nations (FAO), Regional Office for Asia and the Pacific, [6]Faculty of Bioscience Engineering, KU Leuven, [7]College of Veterinary Medicine, Mekelle University

Trypanosomes cause a variety of diseases in man and domestic animals in Africa, Latin America and Asia. In the Trypanozoon subgenus, Trypanosoma brucei gambiense and T. b. rhodesiense cause human African trypanosomiasis (sleeping sickness), while T. b. brucei, T. evansi and T. equiperdum are responsible for animal trypanosomiasis. The genetic relationships between T. evansi and T. equiperdum and other Trypanozoon species remain unclear because the majority of phylogenetic analyses have been based on only a few genes. Therefore, we have conducted a phylogenetic analysis based on genome-wide SNP analysis comprising 56 genomes from the Trypanozoon subgenus. Briefly, sequencing reads of all strains were aligned to the T. b. brucei TREU927 reference genome with Bowtie2 and variants were called with Samtools' mpileup. We then used 194,566 high-quality SNPs to: 1) generate a maximum likelihood tree with RaxML, 2) create a split network with SplitsTree and 3) examine the population genetic structure with fineSTRUCTURE. Our data reveal that T. equiperdum has emerged in Eastern Africa and T. evansi at two independent occasions in Western Africa. The genomes within the T. equiperdum and T. evansi monophyletic clusters show extremely little variation, probably due to the clonal spread linked to their tsetse fly independent transmission. Additionally, we could identify hundreds of unique SNPs belonging to T. equiperdum and to T. evansi, opening new possibilities to facilitate their identification and distinction. Interestingly, based on our observations, the classification of T. evansi and T. equiperdum as species does not reflect their evolutionary background. Indeed, when looking at their relation with T. brucei, the genetic distance between Western and Eastern African T. b. brucei strains is larger than the differences between many T. b. brucei and T. evansi or T. equiperdum. This genomic information clearly shows that T. evansi and T. equiperdum can be regarded as subspecies of T. brucei (T. b. evansi and T. b. equiperdum). In summary, we have presented the first whole genome SNP analysis of T. evansi and T. equiperdum and provided new insights in the origin of both species and their relation with the different Trypanosoma brucei subspecies.

Keywords : Trypanosoma brucei, Sleeping Sickness, Genomics, Whole genome sequencing, Phylogeny

Design and modeling of a database dedicated to Medicinal Plants - Herbal Medic

Ouissam El Andaloussi[1], Badr Din Rossi Hassani[2] and Mhamed Ait Kbir[2]
[1]Centre des Etudes en Sciences et Techniques, [2]Faculté des Sciences et Techniques Tanger

Objectives: The aim of this work is to develop a database for medicinal plants ‘’HerbalMedic”.
Methods: Data embedded in HerbalMedic (HM) was manually imported from scientific literature, while MySQL was integrated by programming with PHP. The structure of HM is organised under three major sections; Plant management (Scientific name, Vernacular name, Taxonomy, and Photo of the plant), Recipes (Recipe name, reference, publication date, treated diseases, author, Scientific position, establishment, preparation procedure, and contraindication), and Disease management (List of diseases treated by the plant and a link to suggested treatment recipes).
Results: HerbalMedic contains many data of medicinal plants linked with disease and recipes used to treat it; it is still unpublished on the internet.
Conclusion and perspectives: HerbalMedic is a useful database to study, manage, and conserve medicinal plants data. Now we are thinking of developing a solution for the automatic collection of information from the web and analyze it and integrate it in HerbalMedic.

Keywords : Bioinformatics, Database, Herbalmedic, Medicial Plant

Study of recent coalescence events in contemporaneous landscapes : C++ template library for Approximate Bayesian Computation

Arnaud Becheler[1], **Camille Coron**[2] **and Stéphane Dupas**[1]
[1]EGCE, [2]Laboratoire de Mathématiques d'Orsay

1 Introduction

The study of the recent evolution of contemporaneous populations allows us to understand the environmental impacts of global changes. For example, the genetic patterns of an invasive species like the yellow-legged hornet (Vespa velutina) in Europe can keep track of the species demographic reactions (i.e. growth and dispersion patterns) to its new environment.

Genetic data can be linked to ecological processes by coupling demographic models accounting for spatio-temporal landscape heterogeneity with coalescence models reconstructing the genetic history of the sampled genes copies. The models emerging from this problematics are analytically intractable, therefore the previous studies [1] used Approximate Bayesian Computation (ABC) methods for estimating quantities of interest (dispersal law and/or niche function parameters) thanks to numerous simulations [2].

However, previous studies reconstruct the coalescent until the most recent common ancestor (MRCA) of the sampled genes copies [3]. This can be a problem if we are interested in recent history and if the MRCA is located in a distant spatio-temporal window. First, it forces us to inform ancient and far-off historical processes in which we are not interested. Second, quantity and quality of data drop when going further in past. Last but not least, taking ancient history into account can unnecessarily increase ABC rejection rate, leading to prohibitive computation time or poor-quality estimation of the parameters. We propose a new method toavoid these costs by focusing only on the very recent coalescence events underlying the modern invasive history of the yellow-legged hornet.

2 Method

The necessity to tackle various models with different complexity levels, jointly with the performance issue imposed by ABC, lead us to develop a C++ template library for simulation of coalescence processes. This generic, modular, extensible (and in short-time open-source) library offers large degree of freedom and high efficiency in the definition of the simulation model, which is to our knowledge unique and unmatched.

During each simulation, the coalescence can be stopped at the time when data are to sparse to inform the process, leading to a forest of genes genealogies partitioning the genetic dataset. Simulated forest and observed dataset are then converted to fuzzy partitions without loss of information, after which the ABC procedure allows to accept the parameters values being used by the simulation if the fuzzy transfer distance [4] computed between simulated partition and observed partition is less than a given threshold.

References

[1] Estoup, Arnaud, et al. Combining genetic, historical and geographical data to reconstruct the dynamics of bioinvasions: application to the cane toad Bufo marinus. Molecular ecology resources (10.5):886-901, 2010.

[2] He, Qixin, Danielle L. Edwards, and L. Lacey Knowles. Integrative testing of how environments from the past to the present shape genetic structure across landscapes. Evolution (67.12): 3386-3402, 2013.

[3] Currat, Mathias, Nicolas Ray, and Laurent Excoffier. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Molecular Ecology Notes (4.1): 139-142, 2004.

[4] Campello, Ricardo JGB. Generalized external indexes for comparing data partitions with overlapping categories. Pattern Recognition Letters (31.9): 966-975, 2010.

Keywords : C++ template library, Coalescence, Landscape, ABC

Evaluation of methodologies for the characterization of plant mosaic genomes

Aurélien Cottin[1], Benjamin Penaud[1], Guillaume Martin[1], Joao Santos[1], Franck Curk[2], Angélique D'Hont[1], Manuel Ruiz[1], Jean-Christophe Glaszmann[1], Nabila Yahiaoui[1] and Mathieu Gautier[2]

[1]CIRAD - UMR AGAP, [2]INRA - UMR AGAP

Hybridization events between species and subspecies are considered as major evolutionary steps, possibly contributing to the advent of new phenotypes. These events are widespread in several crop species and are expected to produce genomes with a mosaic structure of sequence blocks from different ancestry. Characterizing the inter(sub)specific mosaic structure of crop plant genomes that result from recent hybridization events can help understanding how they were formed, their domestication history, and possibly the ancestral origin of phenotypic traits.

With the development of NGS genotyping technologies, several population genomics approaches have been proposed to infer the ancestry of genome segments, by comparing polymorphism patterns across individuals along chromosomes. However, these Local Ancestry Inference (LAI) methods have mainly been developed for applications in animal models, and human most particularly. They are based on assumptions which do not always fit plant models due to more complex genome structures (e.g. different ploidy levels, variable heterozygosity levels within species) or different reproductive systems (e.g., vegetative propagation, selfing). In this context, there is a need to evaluate available methods on plant models.

To that end, we developed a small and flexible R program to simulate data under a wide variety of scenarios representative of plant model characteristics. We use this tool to evaluate two main types of LAI methods: exploratory approaches (based on multivariate analysis) and full probabilistic approaches (based on Hidden Markov Model). First results will be presented and discussed.

Keywords : evolution, hybridation, crop plants, genome, NGS, bioinformatics

Bi-Objective Integer Programming For RNA Secondary Structure Prediction With Pseudoknots

Audrey Legendre, **Eric Angel and Fariza Tahi**
IBISC/Université d'Evry/Genopole/Université Paris Saclay

RNA structure prediction is an important field in Bioinformatics, and numerous methods and tools have been proposed to tackle this problem. Pseudoknots are specific motifs of RNA secondary structures that are difficult to predict. Several models exist for RNA secondary structure prediction, based on the minimum free energy (MFE) structure, the maximum expected accuracy (MEA) structure or the comparative approach. However, it is clear that a given model alone can generate prediction only close to the real structures. For example, it has been established that the real structure has indeed a very low energy, but not necessarily the minimum one, since many factors are involved, such as the environment. It is therefore interesting to propose approaches able to combine different models. To our knowledge, combination of different models was used only in the prediction of consensus structure of several homologous sequences. Also, proposing only the optimal solution of a given model, is limiting, for the same reason given above. Our goal is to propose a method to combine different models and to return both optimal and sub-optimal solutions.

We propose an original method for predicting RNA secondary structure with pseudoknots, based on the integer programming appproach. An integer program corresponds to a mathematical formalization of a problem with an objective function to optimize on a set of variables, subject to a set of linear constraints. The flexibility of this method allows to model in a mathematical way diverse problems. This method is used in various domains, from economy to industry. To our knowledge, only one team used integer programming for RNA pseudoknotted secondary structure prediction. They first developed an integer program [3] to find the MFE structure using the stacking energy parameters of Mfold 3.0 [5]. Then they provided the IPknot software [4] based on a MEA model.

We propose an algorithm which allows us to combine two models into a single bi-objective integer program (BOIP) from which we can get optimal solutions having the best tradeoff between the two criteria. As stated before, sub-optimal solutions are equally of great interest from a biological point of view. Then our algorithm can retrieve the k-best (sub-)optimal solutions for any BOIP.

In this work, we consider two thermodynamic models to predict pseudoknotted RNA secondary structure inspired by [3] and [4]. We combined those two models in a BOIP. The resulting tool is called BiokoP (Bi-objective programming pseudoknot Prediction) and is available on our EvryRNA platform at https://evryrna.ibisc.univ-evry.fr.

We evaluated BiokoP on a dataset of 161 sequences gathered from PseudoBase++, and corresponding to RNAs pseudoknotted secondary structures. The obtained results show the relevance of our approach: the real structure is often given by a sub-optimal solution. BiokoP was compared with other existing methods for RNA pseudoknotted secondary structure prediction that propose several solutions: pKiss [2] and McGenus [1]. The comparison shows that BiokoP gives better results (with respect to F1-score and sensitivity) than pKiss and McGenus.

[1] Michaël Bon, Cristian Micheletti, and Henri Orland. McGenus: a Monte Carlo algorithm to predict RNA secondary structures with pseudoknots. Nucleic acids research, pages 1895–1900, 2012.

[2] Stefan Janssen and Robert Giegerich. The RNA shapes studio. Bioinformatics, pages 423–425, 2014.

[3] Unyanee Poolsap, Yuki Kato, and Tatsuya Akutsu. Prediction of RNA secondary structure with pseudoknots using integer programming. BMC bioinformatics, 10(Suppl 1):S38, 2009.

[4] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu, and Kiyoshi Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics, 27(13):i85–i93, 2011.

[5] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research, 31(13):3406–3415, 2003.

Keywords : RNA, Secondary Structure, Pseudoknot, Integer programming, Bi-objective