# Phylogenetic profiling, an untapped resource for the prediction of secreted proteins and its complementation with sequence-based classifiers in bacterial type III, IV and VI secretion systems

## Abstract

In the establishment and maintenance of the interaction between pathogenic or symbiotic bacteria with a eukaryotic organism, protein substrates of specialized bacterial secretion systems called effectors play a critical role once translocated into the host cell. Proteins are also secreted to the extracellular medium by free-living bacteria or directly injected into other competing organisms to hinder or kill. In this work, we explore an approach based on the evolutionary dependence that most of the effectors maintain with their specific secretion system that analyzes the co-occurrence of any orthologous protein group and their corresponding secretion system across multiple genomes. We compared and complemented our methodology with sequence-based machine learning prediction tools for the type III, IV and VI secretion systems. Finally, we provide the predictive results for the three secretion systems in 1606 complete genomes at <http://www.iib.unsam.edu.ar/orgsissec/>.

# How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives

## Abstract

The sequencing of the transcriptome of single cells, or single-cell RNA-sequencing, has now become the dominant technology for the identification of novel cell types in heterogeneous cell populations or for the study of stochastic gene expression. In recent years, various experimental methods and computational tools for analysing single-cell RNA-sequencing data have been proposed. However, most of them are tailored to different experimental designs or biological questions, and in many cases, their performance has not been benchmarked yet, thus increasing the difficulty for a researcher to choose the optimal single-cell transcriptome sequencing (scRNA-seq) experiment and analysis workflow. In this review, we aim to provide an overview of the current available experimental and computational methods developed to handle single-cell RNA-sequencing data and, based on their peculiarities, we suggest possible analysis frameworks depending on specific experimental designs. Together, we propose an evaluation of challenges and open questions and future perspectives in the field. In particular, we go through the different steps of scRNA-seq experimental protocols such as cell isolation, messenger RNA capture, reverse transcription, amplification and use of quantitative standards such as spike-ins and Unique Molecular Identifiers (UMIs). We then analyse the current methodological challenges related to preprocessing, alignment, quantification, normalization, batch effect correction and methods to control for confounding effects.

# Modeling biological problems in computer science: a case study in genome assembly

## Abstract

As computer scientists working in bioinformatics/computational biology, we often face the challenge of coming up with an algorithm to answer a biological question. This occurs in many areas, such as variant calling, alignment and assembly. In this tutorial, we use the example of the genome assembly problem to demonstrate how to go from a question in the biological realm to a solution in the computer science realm. We show the modeling process step-by-step, including all the intermediate failed attempts. Please note this is not an introduction to how genome assembly algorithms work and, if treated as such, would be incomplete and unnecessarily long-winded.

# Upstream analysis of alternative splicing: a review of computational approaches to predict context-dependent splicing factors

## Abstract

Alternative splicing (AS) has shown to play a pivotal role in the development of diseases, including cancer. Specifically, all the hallmarks of cancer (angiogenesis, cell immortality, avoiding immune system response, etc.) are found to have a counterpart in aberrant splicing of key genes. Identifying the context-specific regulators of splicing provides valuable information to find new biomarkers, as well as to define alternative therapeutic strategies. The computational models to identify these regulators are not trivial and require three conceptual steps: the detection of AS events, the identification of splicing factors that potentially regulate these events and the contextualization of these pieces of information for a specific experiment. In this work, we review the different algorithmic methodologies developed for each of these tasks. Main weaknesses and strengths of the different steps of the pipeline are discussed. Finally, a case study is detailed to help the reader be aware of the potential and limitations of this computational approach.

# Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey - Kevin

## Abstract

Computational prediction of drug–target interactions (DTIs) has become an essential task in the drug discovery process. It narrows down the search space for interactions by suggesting potential interaction candidates for validation via wet-lab experiments that are well known to be expensive and time-consuming. In this article, we aim to provide a comprehensive overview and empirical evaluation on the computational DTI prediction techniques, to act as a guide and reference for our fellow researchers. Specifically, we first describe the data used in such computational DTI prediction efforts. We then categorize and elaborate the state-of-the-art methods for predicting DTIs. Next, an empirical comparison is performed to demonstrate the prediction performance of some representative methods under different scenarios. We also present interesting findings from our evaluation study, discussing the advantages and disadvantages of each method. Finally, we highlight potential avenues for further enhancement of DTI prediction performance as well as related research directions.

# Characteristics and evolution of the ecosystem of software tools supporting research in molecular biology – Nancy

## Abstract

Daily work in molecular biology presently depends on a large number of computational tools. An in-depth, large-scale study of that ‘ecosystem’ of Web tools, its characteristics, interconnectivity, patterns of usage/citation, temporal evolution and rate of decay is crucial for understanding the forces that shape it and for informing initiatives aimed at its funding, long-term maintenance and improvement. In particular, the long-term maintenance of these tools is compromised because of their specific development model. Hundreds of published studies become irreproducible de facto, as the software tools used to conduct them become unavailable. In this study, we present a large-scale survey of >5400 publications describing Web servers within the two main bibliographic resources for disseminating new software developments in molecular biology. For all these servers, we studied their citation patterns, the subjects they address, their citation networks and the temporal evolution of these factors. We also analysed how these factors affect the availability of these servers (whether they are alive). Our results show that this ecosystem of tools is highly interconnected and adapts to the ‘trendy’ subjects in every moment. The servers present characteristic temporal patterns of citation/usage, and there is a worrying rate of server ‘death’, which is influenced by factors such as the server popularity and the institutions that hosts it. These results can inform initiatives aimed at the long-term maintenance of these resources.

# Recent development of Ori-Finder system and DoriC database for microbial replication origins

## Abstract

DNA replication begins at replication origins in all three domains of life. Identification and characterization of replication origins are important not only in providing insights into the structure and function of the replication origins but also in understanding the regulatory mechanisms of the initiation step in DNA replication. The Z-curve method has been used in the identification of replication origins in archaeal genomes successfully since 2002. Furthermore, the Web servers of Ori-Finder and Ori-Finder 2 have been developed to predict replication origins in both bacterial and archaeal genomes based on the Z-curve method, and the replication origins with manual curation have been collected into an online database, DoriC. Ori-Finder system and DoriC database are currently used in the research field of DNA replication origins in prokaryotes, including: (i) identification of oriC regions in bacterial and archaeal genomes; (ii) discovery and analysis of the conserved sequences within oriC regions; and (iii) strand-biased analysis of bacterial genomes.

Up to now, more and more predicted results by Ori-Finder system were supported by subsequent experiments, and Ori-Finder system has been used to identify the replication origins in > 100 newly sequenced prokaryotes in their genome reports. In addition, the data in DoriC database have been widely used in the large-scale analyses of replication origins and strand bias in prokaryotic genomes. Here, we review the development of Ori-Finder system and DoriC database as well as their applications. Some future directions and aspects for extending the application of Ori-Finder and DoriC are also presented.

# SEGreg: a database for human specifically expressed genes and their regulations in cancer and normal tissue – Asha

## Abstract

Human specifically expressed genes (SEGs) usually serve as potential biomarkers for disease diagnosis and treatment. However, the regulation underlying their specific expression remains to be revealed. In this study, we constructed SEG regulation database (SEGreg; available at <http://bioinfo.life.hust.edu.cn/SEGreg>) for showing SEGs and their transcription factors (TFs) and microRNA (miRNA) regulations under different physiological conditions, which include normal tissue, cancer tissue and cell line. In total, SEGreg collected 6387, 1451, 4506 and 5320 SEGs from expression profiles of 34 cancer types and 55 tissues of The Cancer Genome Atlas, Cancer Cell Line Encyclopedia, Human Body Map and Genotype-Tissue Expression databases/projects, respectively. The cancer or tissue corresponding expressed miRNAs and TFs were identified from miRNA and gene expression profiles, and their targets were collected from several public resources. Then the regulatory networks of all SEGs were constructed and integrated into SEGreg. Through a user-friendly interface, users can browse and search SEGreg by gene name, data source, tissue, cancer type and regulators. In summary, SEGreg is a specialized resource to explore SEGs and their regulations, which provides clues to reveal the mechanisms of carcinogenesis and biological processes.

# Metabolomics technology and bioinformatics for precision medicine

**Abstract**

Precision medicine is rapidly emerging as a strategy to tailor medical treatment to a small group or even individual patients based on their genetics, environment and lifestyle. Precision medicine relies heavily on developments in systems biology and omics disciplines, including metabolomics. Combination of metabolomics with sophisticated bioinformatics analysis and mathematical modeling has an extreme power to provide a metabolic snapshot of the patient over the course of disease and treatment or classifying patients into subpopulations and subgroups requiring individual medical treatment. Although a powerful approach, metabolomics have certain limitations in technology and bioinformatics. We will review various aspects of metabolomics technology and bioinformatics, from data generation, bioinformatics analysis, data fusion and mathematical modeling to data management, in the context of precision medicine.

# Drug knowledge bases and their applications in biomedical informatics research Rani & Kevin

## Abstract

Recent advances in biomedical research have generated a large volume of drug-related data. To effectively handle this flood of data, many initiatives have been taken to help researchers make good use of them. As the results of these initiatives, many drug knowledge bases have been constructed. They range from simple ones with specific focuses to comprehensive ones that contain information on almost every aspect of a drug. These curated drug knowledge bases have made significant contributions to the development of efficient and effective health information technologies for better health-care service delivery. Understanding and comparing existing drug knowledge bases and how they are applied in various biomedical studies will help us recognize the state of the art and design better knowledge bases in the future. In addition, researchers can get insights on novel applications of the drug knowledge bases through a review of successful use cases. In this study, we provide a review of existing popular drug knowledge bases and their applications in drug-related studies. We discuss challenges in constructing and using drug knowledge bases as well as future research directions toward a better ecosystem of drug knowledge bases.