Title

Automation of Rotifer Genome Assembly

Authors

Luis A. Aguirre,[1] Aaron Robbins,[2] Patrick D. Brown,[2] Mario Davila,[1] Jonathon E. Mohl,[1,3,4] and Elizabeth J. Walsh[1,2,3]

[1]Bioinformatics Program, [2]Department of Biological Sciences, [3]Border Biomedical Research Center, and [4]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Genome assembly refers to the process of taking large numbers of short DNA sequences and putting them back together to create a representation of the chromosomes from where the sequences originated. Recent technological advancements in DNA sequencing have reduced the cost to produce short DNA reads of non-model species, while software to analyze, process, and assemble these short DNA sequences is readily available. In addition, genomes are still lacking in several important groups, including the phylum Rotifera. Rotifers are globally distributed aquatic microscopic animals that play important roles in aquatic food webs. To date, only a handful of rotifer species have their genomes sequenced. To assist in assembling rotifer genomes, we developed a semi-automated genome assembly pipeline, which combines commonly used programs in the genome assembly process, reduces the amount of time the user has to interact with a command line interpreter, and organizes output in an orderly manner. The pipeline is designed to give quality control checks on sequencing reads, trim or delete reads of poor quality, assemble contigs from trimmed reads, and assess the quality and completeness of the assembly. This pipeline was tested on a small bacterial genome and is now being implemented for several rotifer species.

Title

Determining evolutionary hotspots in the human AMPylase FICD

Authors

Kristen Arce and Amanda Bataycan*

Bioinformatics Program, The University of Texas at El Paso, El Paso, TX

Abstract

The objective of this study was to identify which parts of the FICD enzyme are the most conserved among species. The FICD enzyme function includes regulating proteostasis and is implicated in regulating protein aggregation in aging-associated diseases, more specifically observed in this research is the FICD AMPylation process.

The initial investigation was aligning full protein sequence length, however later shifted into focusing on the independent domain of a particular motif "HPXDGNGR". This motif is found within the active site that acts as a catalyst for FICD-mediated AMPylation. After retrieving the conserved region, an analysis of the variability was performed to help identify the frequency changes among groups of species, with the intent to find commonality of species-specific mutations and correlations associated with their life expectancy. When altering an amino acid, there is a potential cascading effect that can occur which can modify the activity, functionality, and the interaction when folding into a protein. A parallel investigation was conducted to determine the evolutionary rate of this motif. To assist with the evolutionary interpretation, a phylogenetic analysis was performed to cluster species into clades.

The methodology for both parts of this study were fairly similar, concentrating on the ratio of non-synonymous to synonymous substitution rates. Nonsynonymous substitution occurs when a change in a protein coding sequence is altered resulting in a change in the original amino acid sequence. Whereas synonymous substitution is similar in the sense that a change has occurred, however no change in the original amino acid sequence occurs. Analyses were implemented by developing and utilizing various software pipelines for given conditions. Evolution is driven by mutations, determining the type of selection that has occurred provides insight into the mutations that alter the protein AMPylation and its impact.

*Kristen Arce is presently pursing PhD in Data Science at UTEP and Amanda Bataycan has also started her PhD in Computational Science. This work is part of their internship projects in summer 2021 with Dr. Matthias Truttman, UM Geriatrics Center, Department of Molecular and Integrative, University of Michigan, Ann Arbor, MI.

Title

Computational insights into the binding of the same ligand to GPCRs of different families

Authors

Kwabena Owusu Dankwah,[1] Jonathon E Mohl,[1,2,3] Khodeza Begum,[1,2] and Ming-Ying Leung[1,2,3]

[1]Computational Science Program, [2]Bioinformatics Program, [2]Border Biomedical Research Center, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

G protein-coupled receptors (GPCRs) are the largest class of cell-surface receptor proteins with important functions in signal transduction and often serve as therapeutic drug targets. With the rapidly growing public data on three dimensional (3D) structures of GPCRs and GPCR-ligand interactions, computational prediction of GPCR ligand binding becomes a convincing option to high throughput screening and other experimental approaches during the beginning phases of drug discovery. Such predictions are cost-efficient and can be important aides for planning wet lab experiments. In this work, we set out to uncover and understand the binding of the same ligand to multiple GPCRs of different families. We observed that: 1) only a few GPCRs share a conserved sequence motif which was expected; 2) the GPCRs shared local 3D structural similarities and local sequence similarities; 3) the GPCRs shared similar binding pockets for the same ligand; 4) binding pockets that are similar share similar binding affinities; 5) and finally, molecular docking revealed that a ligand can bind to GPCRs of different families with varying conformations. These findings can be important aides for drug discovery as it can help improve protein function inference, drug repurposing and drug toxicity prediction, and expedite the development of new drugs.

Title

A Python script to count genetic marker sequences to identify gaps in taxonomic coverage

Authors

Mario E. Davila,[1] Elizabeth J. Walsh,[1,2,3] and Jonathon E. Mohl[1,3,4]

[1]Bioinformatics Program, [2]Department of Biological Sciences, [3]Border Biomedical Research Center, and [4]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Rotifers are microscopic organisms that are found globally and play a role in aquatic food chain. Databases, like NCBI's GenBank, are used to store nucleotide sequences of rotifers and many other organisms that are used in the analysis of their evolutionary history. Many laboratories use the GenBank website to download sequences for specific genetic markers of their organisms of interest. With the vast number of genetic sequences continually growing, it is important to retrieve and organize the sequences in an efficient manner to identify gaps in knowledge. We have constructed a Python script that automates the process of creating a comma-seperated value file that reports the number of times a sequence of a genetic marker was found for each of the desired taxanomic group. The script reads in the sequences files from both the lab and downloaded fasta files from GenBank, filters out unwanted sequences based on size and taxanomic classification, and then counts the sequences at the species level for a variety of genetic markers. This process has allowed the lab to identify rotifer genera that have not been sequenced for a set of genetic markers and highlights them for preference in future analysis.

Title

Review Presentation: “Liftoff: accurate mapping of gene annotation”

Presenter

Luis Antonio Gracia Mazuca, Bioinformatics program, The University of Texas at El Paso, El Paso, TX

Abstract

This is a review of the article “Liftoff: accurate mapping of gene annotations” by Shumate *et al*. (2021) published in the journal *Bioinformatics*. The abstract originally published by the authors is as follows:

“**Motivation:** Improvements in DNA sequencing technology and computational methods have led to a substantial increase in the creation of high-quality genome assemblies of many species. To understand the biology of these genomes, annotation of gene features and other functional elements is essential; however, for most species, only the reference genome is well-annotated.

Results: One strategy to annotate new or improved genome assemblies is to map or ‘lift over’ the genes from a previously annotated reference genome. Here, we describe Liftoff, a new genome annotation lift-over tool capable of mapping genes between two assemblies of the same or closely related species. Liftoff aligns genes from a reference genome to a target genome and finds the mapping that maximizes sequence identity while preserving the structure of each exon, transcript and gene. We show that Liftoff can accurately map 99.9% of genes between two versions of the human reference genome with an average sequence identity >99.9%. We also show that Liftoff can map genes across species by successfully lifting over 98.3% of human protein-coding genes to a chimpanzee genome assembly with 98.2% sequence identity.

Availability and implementation: Liftoff can be installed via bioconda and PyPI. In addition, the source code for Liftoff is available at https://github.com/agshumate/Liftoff.

Contact: alainashumate@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.”

Title

Review Presentation: “Design of the MCAW compute service for food safety bioinformatics”

Presenter

Luisa Veronica Gracia Mazuca, Bioinformatics program, The University of Texas at El Paso, El Paso, TX

Abstract

This poster is a review of the paper “Design of the MCAW compute service for food safety bioinformatics” by Edlund *et al*. (2016) published in the *IBM Journal of Research and Development.* The abstract, as originally published by the authors is as follows:

“The techniques of microbe community genome sequencing as applied to environmental samples—metagenomics—offer powerful insight into microbial community structure and ecology that can affect food safety decisions for public health security. In this paper, the design and characteristics of a new informatics service, the Metagenomics Computation and Analytics Workbench (MCAW), are presented and illustrated with reference to the analysis of metagenomics data. The service is designed to meet the requirements for analyzing metagenomic and metatranscriptomic sequence data to assess microbial hazards and food authentication in the supply chain. Moreover, MCAW provides for reliable storage and management of raw genomic sequences and analysis results, high-volume informatics processing, meticulous tracking of data provenance and processing steps, and function-rich visualization of results.”

Title

Locating potential pathogenic DNA sequence variants within G protein-coupled receptor genes in prostate cancer patients

Authors

Axel Hidalgo,[1] Kwabena Owusu Dankwah,[2] Khodeza Begum,[1,2] Jonathon E. Mohl,[1,2,3] and Ming-Ying Leung[1,2,3]

[1]Bioinformatics Program, [2]Computational Science Program, [2]Border Biomedical Research Center, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

Prostate cancer is the second most common cancer in the male population of the United States with an estimate of 1 in 8 men developing this cancer within their lifetime. Cancer cells' proliferating abilities rely on both the dysregulation of intercellular signaling mechanisms and aberrant intracellular signaling cascades. G protein-coupled receptors (GPCRs) are a large family of cell surface receptors that play critical roles in several signaling pathways. Currently, GPCRs are the targets for ~40% of FDA-approved drugs. In addition, a variety of GPCRs are known to be directly involved in the regulation of prostate cancer, highlighting the need to investigate their roles in cancer signaling. For this purpose, we selected a prostate cancer dataset from The Cancer Genome Atlas (TCGA) describing single nucleotide substitutions of tumor and non-tumor samples from 498 patients. The substitutions resulting in protein-level alterations were identified and submitted to the programs PROVEAN and FATHMM-XF for prediction of deleterious effects caused by the mutations. Those with highest pathogenic and deleterious potential were selected for further analyses of their influence to the GPCR molecular structure and ligand binding capabilities. These structural aberrations could potentially affect the signal transduction functions of the GPCRs leading to advantageous cancer phenotypes.

Title

Identifying differentially methylated regions between Alzheimer's patients and healthy controls

Authors

Raisa Nkweteyim[1] and Liang Ma[2]

[1]Bioinformatics Program, The University of Texas at El Paso, El Paso, TX, and [2]Department of Pharmacology, The University of Texas Health Science Center at San Antonio, San Antonio, TX

Abstract

Alzheimer's disease (AD) is the most common cause of dementia, a general term for memory loss and other loss of cognitive abilities. Prior studies suggest that DNA methylation alterations in AD are dependent upon the cell type studied and are region-specific. This research aims to determine the genome-wide methylation of nucleotide bases from DNA derived from the dorsolateral prefrontal cortex of the brain in AD patients and the differentially methylated regions in the AD samples compared to normal samples.

Chip Analysis Methylation Pipeline (CHAMP) was used to analyze methylation microarray data of 66 AD patients and normal samples downloaded from the NCBI Sequence Read Archive. Raw, encrypted .idat files were converted into readable methylation information by filtering out data of low quality, doing quality control and normalization. Methylated bases are located, annotated and the differentially methylated regions between the AD samples and normal controls identified. This pipeline was also automated to process many samples using a high-performance cluster.

We expect to see a characteristic trend in the DNA methylation of bases in the AD samples that will not be observed in the normal samples. This methylation pattern analysis may lead to better understanding of the mechanism in which AD occurs.

Title

Analyses of genetic sequence variants in whole exome sequencing data from patients with prostate cancer

Authors

Kelvin Ofori-Minta,[1] Bofei Wang,[2] Jonathon E. Mohl,[1,2,3] and Ming Ying Leung[1,2,3]

[1]Department of Mathematical Sciences, [2]Computational Science Program, [3]Bioinformatics Program, and [3]Border Biomedical Research Center, The University of Texas at El Paso, El Paso, TX

Abstract

The repercussions of mutations in the base pairs of genetic sequence variants (GSVs) often lead to the realization of undesired rather than desired characteristics, hence the dire need for biomedical and statistical bioinformatics research to uncover genetic triggers of prostate cancer (PrCa). This research seeks to elicit and discover patterns and insights from a set of whole exome sequencing data from patients diagnosed with PrCa obtained from The Cancer Genome Atlas (TCGA). Using multiple approaches in statistical data mining strategies and machine learning procedures such as hierarchical (agglomerative and divisive) and partitional (k-medoids) clustering, our results have indicated that (i) single nucleotide substitutions frequently change from strong bases to weak bases, (ii) GSV counts and densities among different chromosomes are rather heterogeneous, and (iii) there are three pairs of frequently occurring variants that are constantly grouped together by multiple clustering methods. These results will be reported to biomedical scientists for further bioinformatics analyses and wet lab studies.

Title

Prevalence and predictors of alternative medicine use among pregnant women in North Central Nigeria and relationship with microbial communities in breast milk

Authors

Dayo Shittu[1] and Bolarinwa Oladimeji[2]

[1]Bioinformatics Program, The University of Texas at El Paso, El Paso, TX, and [2]Epidemiology and Community Health Department, University of Ilorin, Kwara State, Nigeria

Abstract

Alternative medicine (AM) is a topic of increasing public health importance in the developing world. Maternity care is an area where AM use has attracted interest due to the potential clinical impacts of increased use of AM by pregnant women.

The purpose of this research was to determine the prevalence and identify predictors of AM use among pregnant women in North Central Nigeria by a hospital based cross-sectional descriptive study. Study population was pregnant women (15-49 yr) attending primary healthcare. Using systematic sampling, 200 pregnant women were selected and necessary information was obtained by a semi-structure questionnaire. The study reported a high prevalence of awareness and good knowledge of AM (97% and 51.5% respectively). Education, socioeconomic background, knowledge and perception are some predicting variables that significantly influenced the use of AM. I further plan to investigate the microbial communities of breast milk and compare with published literature to look for possible difference when AM is a factor.

The study reported high prevalence of AM use among pregnant women and disclosure to healthcare professionals, highlighting the importance of further research into the safety and efficacy of AM and its overall effect on the health of pregnant women and their unborn children.

Title

Visualization and analysis of giant virus structure

Authors

Yifan Wang,[1] Mary Mackay,[2] Martin Chacon,[3] Son-Young Yi,[2,4] Art Duval,[1,2] and Chuan Xiao[1,3,5]

[1]Bioinformatics Program, [2]Department of Mathematical Sciences, [3]Department of Chemistry and Biochemistry, [4]Computational Science Program, and [5]Border Biomedical Research Center, The University of Texas at El Paso, El Paso, TX

Abstract

Recent discovery of giant viruses (girus), some of whose sizes exceed that of small cells, has ignited debate about the tree of life, the definition of a virus, and the evolutionary relationship between viruses and cellular organisms. In this research, we use a marine girus CroV as the first experimental model. The pseudo-atomic structure of CroV has 120 million atoms by using cryo-EM reconstructed map. However, it is hard to visualize and analyze the model on high performance computing clusters with a large number of atoms. Our research is to improve cryo-EM resolution, develop a new mathematical model to reduce computation, and develop new parallel algorithms optimized for visualizing and analyzing this supramolecular system. The new mathematical model and parallel algorithms will facilitate the studies of the self-assembly of giruses critical to understand their life cycles. Understanding the details of viral assembly will improve our ability to fight against viral infections, and facilitate bioengineering of virus-like nanoparticles. Methods developed through the project can also be applied to future studies of other supramolecular cellular structures or even the entire cell.

Title

Workflows to automate determination of physicochemical properties for large sets of affinity-selected molecules using structure- and simulation-based modeling

Authors

Huanhuan Zhao,[1] Payam Kelich,[2] and Lela Vukovic[1,2]

[1]Bioinformatics Program and [2]Department of Chemistry and Biochemistry, The University of Texas at El Paso, El Paso, TX

Abstract

Using machine learning in chemical discovery is of great interest to chemists. Our group is developing machine learning (ML) models for predicting peptidomimetic ligands that bind to target proteins. One challenge in developing predictive ML models is finding valuable features to improve their prediction accuracy. Here we develop a workflow for finding useful features using principal component analysis (PCA), molecular dynamics (MD), and density plots. We used peptidomimetic ligands data sets that target four proteins: Streptavidin, Kelch-like ECH-associated protein 1(Keap1), Sonic Hedgehog, and Bovine carbonic anhydrase (BCA). The first three datasets showed good separation in PCA space and trained ML models attaining prediction accuracy around 0.92, whereas little separation and the prediction accuracy of 0.596 was observed for the BCA dataset. However, further explorations of the BCA dataset indicated that it may be possible to find useful features by MD simulation. We plan to use the OMEGA and Rosetta software to improve the conformational space of binding and non-binding ligands and predict structures of BCA protein-ligands complexes for MD simulations. Then, we will analyze the simulations of ligands and the protein-ligands complexes and seek for features that can be used as input for training more predictive ML models.