WABI 2018 - 18th Workshop on Algorithms in Bioinformatics
Helsinki, 20-22 August 2018
http://algo2018.hiit.fi/wabi/

Poster Abstracts – WABI 2018

The poster session is on Monday, Aug 20, starting at 17:00

Poster list

P01

Bio-Express: Cloud Service for high throughput analysis of biological big data from Korean Bioinformation Center

Pan-Gyu Kim, GunHwan Ko, Gukhee Han, Wangho Song and Byungwook Lee
Korean Bioinformation Center, KRIBB, Daejeon 34141, Korea

Contact: pgkim@kribb.re.kr

Abstract:

Because of the exponential growth of genomic data since the introduction of next generation sequencing technology, the analysis of large bio-data is becoming more and more complex and difficult problem. The enormous amount of genomic data requires suitable computing resources such as network and storage devices, computational power, security system, and computer management expertise, as well as analysis pipelines. One of the easy ways to overcome this challenge is to use cloud service, which provides access to computing resources such as large amounts of storage and computation as a service.

We developed a cloud service for high throughput analysis of biological big data, Bio-Express, to provide large-scale analysis service for massive genomic sequence data via workflow editor, CLOSHA. The Bio-Express provides 22 pipelines and 120 programs to analysis biological big data, and we continue to add new pipelines and programs. CLOSHA, workflow editor of Bio-Express, offers a user-friendly graphical user interface which allows users to create multi-step analysis using drag and drop functionality, and modify parameters of pipeline tools. Bio-Express runs on the hybrid cluster system which can run Hadoop and Linux programs. It is based on HDFS which can run not only Hadoop programs but also Linux programs applying the disk caching technique which transfers a file in HDFS into general Linux file system on each request. As a result, users can use both analysis programs for general purposes (mainly Linux-based) and the Hadoop-based big data analysis programs in a single pipeline simultaneously.

To maximize the throughput, Bio-Express performs job scheduling by assigning jobs via SGE after checking available computing resources through YARN. We developed a high-speed data transmission solution, KoDS, to transmit a large amount of data at a fast rate. KoDS has a file transferring speed up to 10 times than normal FTP and HTTP protocols. Computer hardware for Bio-Express is 660 CPU cores and 800Tb, which enable 500 jobs to run at the same time.

Bio-Express is a scalable, cost-effective, and publicly available web service. For past a year after launching Bio-Express service, it has successfully completed over 1,000 requested analysis service. By keeping datasets and analytic tools up-to-date, Bio-Express will continue to serve as a major cloud computing service for massive genomic data.

P02

Methods for Identifying Tumor …

WITHDRAWN

P03

TrioBinning: Trio-based assembly

Sergey Koren[1], Arang Rhie[1], Brian P. Walenz[1], Alexander T. Dilthey[1,2], Derek M. Bickhart[3], Sarah B. Kingan[4], Stefan Hiendleder[5,6], John L. Williams[5], Timothy P. L. Smith[7], Adam M. Phillippy[1]

[1] Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA
[2] Institute of Medical Microbiology, Heinrich-Heine-University Düsseldorf, Düsseldorf, North Rhine-Westphalia, Germany
[3] Cell Wall Biology and Utilization Laboratory, ARS USDA, Madison, Wisconsin, USA
[4] Pacific Biosciences, Menlo Park, California, USA
[5] Davies Research Centre, School of Animal and Veterinary Sciences, The University of Adelaide, Roseworthy SA, Australia
[6] Robinson Research Institute, The University of Adelaide, Adelaide SA, Australia
[7] US Meat Animal Research Center, ARS USDA, Clay Center, Nebraska, USA

Contact: sergey.koren@nih.gov

Abstract:

Reference genome projects have historically selected inbred individuals to minimize heterozygosity and simplify assembly. We challenge this dogma and present a new approach designed specifically for heterozygous genomes. Prior approaches for assembling heterozygous diploid genomes only phase small variants or partially reconstruct the haplotypes. Our trio binning method uses short reads from two parental genomes to partition long reads from an offspring into haplotype-specific sets. Each haplotype is then assembled independently. The output of this process is a complete genome for each parental haplotype, containing all classes of haplotype variation assembled from the long reads, including single nucleotide, structural, and copy number variants.

To demonstrate the effectiveness of trio binning on a heterozygous genome, we sequenced an F1 cross between cattle subspecies Bos taurus taurus and Bos taurus indicus, and assembled both parental haplotypes with NG50 haplotig sizes >20 Mbp each, surpassing the quality of current cattle reference genomes. In fact, these haplotype-specific contigs (haplotigs) are larger than the scaffolds of many prior inbred or haploid reference genome assemblies. In addition to their high continuity, both haplotypes approach 99.999% accuracy at the base level using PacBio data alone. Further application of this method to a benchmark human trio (NA12878 and parents) also achieved high accuracy and recovered complex structural variants missed by alternative approaches such as 10x Genomics linked-read sequencing.

Trio binning of both the human and cattle haplotypes successfully reconstructed highly heterozygous loci important for immunity and adaptation. For example, in human, both parental haplotypes of the Major Histocompatibility Complex (MHC) were accurately assembled and showed perfect human leukocyte antigen (HLA) gene typing accuracy. For cattle, many heterozygous regions between the newly assembled Angus and Brahman haplotypes intersected with previously identified quantitative trait loci (QTL). For example, in one structurally diverse region relative to the Brahman haplotype, the Angus haplotype is missing a ~140 kbp duplication containing GBP2, while containing its own duplicated GBP6-like sequence. This region intersects with previous QTLs linked to muscularity and visual conformation score, making it a suggestive candidate for adaptation among the cattle breeds.

Given the quality of the assemblies we were able to achieve with this approach, we propose trio binning as a new best practice for diploid genome assembly that will enable platinum-quality reference genomes and new studies of haplotype variation and inheritance.

P04

Performance and Prediction Algorithmic Methodology Applied To Assisted Procreation Technology

N. Nafati[1], O. Ait-Ahmed[1], and S. Hamamah[1,2]

[1:] Unité 1203-INSERM. Hôpital Saint Eloi. 80 Av Fliche Augustin. FRANCE.
[2:] CHU Arnaud de Villeneuve. 371 Av Doyen Gaston Giraud. 34295 Montpellier Cedex 05. FRANCE.

Contact: nicolas.nafati@inserm.fr

Abstract:

In the field of assisted reproduction, several studies suggest that the genes involved in the crosstalk of the oocyte-cumulus cell could represent candidate gene biomarkers to select embryos competent in terms of implantation. To achieve this objective and before any processing of the acquired transcriptomic data from of RT-qPCR (Real-Time Quantitative Chain Polymerase) physical system, it is necessary to ensure the reliability of the acquired data. Indeed, these acquired data are generally provided with noise from various sources, such as experimental, technical noise, etc. The transcriptomic data correspond to 21 biomarker genes and 102 embryonic/cumulus cell samples from patients undergoing in vitro fertilization. So our goal is to test whether this genomic signature could be used as a biomarker or not. If so, we can state that this transcriptome is predictable and could generate a reliable mathematical model. We give a typical algorithm whose task is to verify the validity or not of the RT-qPCR transcriptomic data, and then to decide their validation and exploitation.

The originality of this paper is to combine the performance, namely Binary and Multiple Logistic Regression (BMLR), and the Likelihood criterions. This is in order to maximize the probability that the test will correctly predict or not the event of interest (pregnancy). In this framework, the Odds-Ratio (OR) will be also analysed to be sure that the obtained transcriptomic data is open or not for the use.

P05

Algorithm to assess the evolutionary history of distantly related protein domains

Sridhar Hariharaputran
National Centre for Biological Sciences, TIFR, Bangalore, Karnataka, India

Contact: sridharh@ncbs.res.in

Abstract:

We have been developing several tools, databases, information systems PASS2, PASS2+, SInCRe, MyCompare, MyGOnets to the community hosting multiple and diverse information. In our works we discovered distantly related protein domain “outliers” and for our analysis we mapped this structural information with Gene Ontology and Enzyme number. Using a semi-automatic analysis we discovered outliers following a pattern.

Now we have extended the work to analyze these patterns followed by multiple superfamilies consisting of several thousand protein domains. The algorithm developed compares and assesses the evolutionary relationship of the domains using the variety of information. Using a case study of alpha/beta hydrolase superfamily which is associated with a variety of functions and related to human diseases we again see a pattern or a motif of information that is generated from the distantly related protein domains and outliers. This pattern is either associated at the superfamily/family level and also crosses the border and shares the information with member/s of other superfamily or superfamilies. Analysis of entire SCOPe/PASS2 domains reveal a global and a local motif at the superfamily and family levels. We also incorporate the distance measures from multiple information associated with the protein domains for our analysis.

References

1. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification ofproteins database for the investigation of sequences and structures, J. Mol. Biol. 247(4):536-40.
2. Murzin AG (1998) How far divergent evolution goes in proteins, Curr. Opin. Struct. Biol.8:380-387.
3. Arumugam G, Nair AG, Hariharaputran S, Ramanathan S (2013) Rebelling for a Reason:Protein Structural “Outliers”. PLoS ONE 8(9): e74416.
4. Metri R, Hariharaputran S, Ramakrishnan G, Anand P, Raghavender US, Ochoa-Montaño B, Higueruelo AP, Sowdhamini R, Chandra NR, Blundell TL, Srinivasan N. SInCRe structural interactome computational resource for Mycobacterium tuberculosis. Database (Oxford). 2015 Jun 30;2015:bav060.
5. Das S, Dawson NL, Orengo CA.Diversity in protein domain superfamilies.Curr Opin Genet Dev. 2015 Dec;35:40-9.
6. Gandhimathi A, Ghosh P, Hariharaputran S, Mathew OK, Sowdhamini R.PASS2 database for the structure-based sequence alignment of distantly related SCOP domain superfamilies: update to version 5 and added features. Nucleic Acids Res. 2016 Jan 4;44(D1):D410-4.
7. Lees JG, Dawson NL, Sillitoe I, Orengo CA.Functional innovation from changes in protein domains and their combinations.Curr Opin Struct Biol. 2016 Jun;38:44-52.
8. Chandonia JM, Fox NK, Brenner SE.SCOPe: Manual Curation and Artifact Removal in the Structural Classification of Proteins - extended Database. J Mol Biol. 2017 Feb 3;429(3):348-355.
9. Sridhar Hariharaputran, Ramanathan Sowdhamini, Tom L Blundell, Narayanaswamy Srinivasan, Nagasuma R Chandra. MyGOnets - Mycobacterium species and Gene Ontology Based Networks. (Abstract Submitted to Nucleic Acids Research (NAR), 2015).
10. Sridhar Hariharaputran, Sumanta Mukherjee, Nagasuma R Chandra, Ramanathan Sowdhamini, Tom L Blundell, Narayanaswamy Srinivasan. MyCompare - Mycobacterium tuberculosis Network Comparer. (Extended Abstract Submitted to Nucleic Acids Research (NAR), 2015).
11. Sridhar Hariharaputran, Tom L Blundell, Nagasuma R Chandra, Narayanaswamy Srinivasan, Ramanathan Sowdhamini. PASS2+ A Resource To Understand Distantly Related Structural Domains Using Multiple Information (Extended Abstract Submitted to NAR, 2015).

P06

PowerExplorer: An R package for simulation-based power analysis

Xu Qiao, Tomi Suomi, Mikko Venäläinen, Laura L. Elo
University of Turku

Contact: xu.qiao@utu.fi

Abstract:

Introduction. The objective of this study is to develop a power analysis tool for both RNA sequencing (RNA-seq) and mass spectrometry-based (MS-based) proteomics data. At present, there are few tools available for both RNA-seq and proteomics. Our tool is capable of both prospective and retrospective power analyses. Not only can users evaluate the reliability of the detected differential expressions, but also predict the sufficient sample size for a desired power.

Methods. We model the RNA-seq read counts using negative binomial distribution and deploy existing method DESeq2 [1,2] to estimate the corresponding parameters. For MS-based proteomics and log-transformed RNA-seq data, we model the data using normal distribution and use maximum likelihood method to estimate the parameters. Based on the estimated parameters, we carry out Monte Carlo simulations [3] under both null and alternative hypotheses and perform statistical tests (t-test or Wald-test). Furthermore, we calculate the power based on the difference between null and alternative distributions. To test the performance of our methods for RNA-seq data, we use an RNA-seq dataset from RNA Sequencing Quality Control project [4]. The ERCC spike-ins were grouped based on the expected log2 fold changes (LFC) (Group I: 2, Group II: 0, Group III: -0.58 and Group IV: -1). In addition, we use an MS-based proteomics dataset [5] consisting of two hybrid samples (A and B) which are human proteomes mixed with E. coli and yeast proteomes in different proportions (1:1 for human, 2:1 for yeast and 1:4 for E. coli proteins). Furthermore, we predicted the power of sample sizes from 5 to 50.

Results. As expected, in both RNA-seq and proteomics datasets, we observed the highest power estimates (power e 0.8) among spike-in genes and proteins with large LFCs. The proportions of high power ERCC spike-ins were substantially greater in large LFC groups (Group I: 80.95%, Group IV: 59.09%) as compared to low LFC groups (Group II: 13.63%, Group III: 27.27%). Similarly, in the proteomics dataset, the fraction of high power proteins was substantially higher for spike-in non-human proteins (Yeast: 88.66%, E.coli: 95.28%) as compared to background proteins (Human: 25.16%). It was observed that in both datasets, the power estimates increased with larger sample sizes. For RNA-seq data, the improvements in power were greatest for groups with large LFCs whereas, for groups with small LFCs, only minor improvements were observed. In groups with large LFCs, a high percentage of spike-ins already showed high power estimates with 15 replicates (Group I: 90% and Group IV: 86%), while the fraction of high power genes in Groups II and III remained only at 50% and 55%. For proteomics data, the fraction of high power proteins was observed to be high for yeast (89%) and E. coli (96%) proteins already with five replicates whereas most human proteins would have required more than 50 replicates to reach high power.

Conclusion. In this study, we developed a power analysis tool for RNA-seq and MS-based proteomics data. The benchmarking results showed that our method gives distinct power estimates without depending on prior differential expression (DE) analysis and assumptions (proportion of prognostic genes or proteins, fold changes or expected mean expressions).

References

1. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:1-21.
2. Anders S, Huber W. DESeq: Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.
3. Fang Z, Cui X. Design and validation issues in RNA-seq experiments. Brief Bioinform 2011;12:280-7.
4. Consortium S-I, Su Z, Aabaj PP et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol 2014;32:903-14.
5. Kuharev J, Navarro P, Distler U et al. In-depth evaluation of software tools for data-independent acquisition based label-free quantification. Proteomics 2015;15:3140-51.

P07

A visualization tool to evaluate pairwise protein structure alignment algorithms

Shalini Bhattacharjee and Asish Mukhopadhyay
University of Windsor

Contact: asish.mukerji@gmail.com

Abstract:

The alignment of two protein structures is a fundamental problem in structural bioinformatics. Their structural similarity carries with it the connotation of similar functional behaviour that could be exploited in various applications. A plethora of algorithms, including one by us, is a testament to the importance of the problem. We propose a novel approach to measure the effectiveness of a sample of three such algorithms, DALI, TM-align and EDAlign$_{sse}$, for detecting structural similarities among proteins. The underlying premise is that structural proximity should translate into spatial proximity.

To verify this, we carried out extensive experiments with five different datasets, each consisting of proteins from two to six different families. For each dataset, we computed a distance matrix, where each distance is the cRMSD distance of a pair of protein structures. For each distance matrix, we used Principal Component Analysis to obtain an embedding of a set of points (each representing a protein) that realize these distances in a two-dimensional space. To compare the clustering of the families, we used the k-means clustering algorithm to cluster the points, sans family labels.

The main contribution of this work is the development of a visual tool for comparing the effectiveness of various protein structure alignment algorithms in capturing structural proximity by mapping this proximity to an Euclidean distance space. As a test case, of the three algorithms we compared, TM-align appears to capture structural proximity most effectively.

P08

π-cyc: A Reference-free SNP Discovery Application using Parallel Graph Search

Reda Younsi, Jing Tang, and Liisa Holm
University of Helsinki

Contact: reda.younsi@helsinki.fi

Abstract:

Bubbles discovery in coloured de Bruijn graph for de novo genome assembly is a problem that can be translated to cycles enumeration in graph theory. In coloured de Bruijn graphs, bubble paths coverages are used in downstream SNP calling analysis. Working with a large number of genomes simultaneously is of great interest in genetic population and comparative genomics research. Cycle enumerations algorithms in big and complex de Bruijn graphs are time consuming. Specialised fast algorithms for efficient bubble search are in need for coloured de Bruijn graph variant calling applications.

In this poster, we show the results of a new parallel graph search method using a combined multi-node and multi-core design to speeds up cycles enumeration. The search algorithm uses an index extracted from the raw assembly of a coloured de Bruijn graph stored in a hash table. The index is distributed across different CPU-cores, in a shared memory HPC compute node, to build undirected subgraphs then search independently and simultaneously specific cycle sizes. This same index can also be split between several HPC compute nodes to take advantage of as many cpu-cores available to the user. The local neighbourhood parallel search approach reduces the graph's complexity and facilitate cycles search of a multi-colour de Bruijn graph. The search algorithm is incorporated into Π-cyc application and tested on a number of Pombe genomes.

P09

Lep-MAP3: Robust Linkage Mapping even for Low Coverage Data

Pasi Rastas
University of Helsinki

Contact: pasi.rastas@gmail.com

Abstract:

Linkage map shows positions of genetic markers along chromosomes based on how often they are inherited together. The information to build linkage maps can be obtained from a genotype data of simple cross(es) of parents and their offspring. Physical and linkage positions are highly correlated (and their relation is often visualised using Marey maps), making linkage mapping one of the best tools to detect errors in de novo genome assemblies. Linkage maps also anchor assembled contigs to and within chromosomes. Other uses of linkage maps include, e.g. family-based linkage and association studies, quantitative trait locus (QTL) mapping and analysis of genome synteny. Computationally linkage map construction is very similar to finding solution to a travelling salesperson problem (TSP). From all potential marker orders, the one minimising the number of recombinations (distance between cities in TSP) is preferred. However, when the data is not completely informative and/or contains errors, the problem becomes more challenging and complicated that a TSP. Brute force algorithms evaluating all about n!/2 marker marker orders are not feasible for >15 markers.

The number of markers and individuals affect linkage mapping resolution. A mapping cross of ten individuals can detect many assembly errors. With more individuals and markers, even more local errors can be detected and more contigs can be anchored. As the marker density increases, resulting map will have multiple markers at most positions. This will anchor contigs more reliable by locally pinpointing each recombination and even the shortest contigs. However, the tools that are currently available for linkage mapping are not well suited for very large number of markers nor individuals.

Here we present linkage mapping software Lep-MAP3, capable of analysing large datasets. It is fast ($O(mn^2)$) and has small memory ($O(mn)$) footprint (for m individuals and n markers). It can simultaneously analyse multiple families and requires little manual work and data curation. It can analyse low-coverage whole genome sequencing datasets on millions of markers and thousands of individuals. Such cost-efficient data enables comprehensive validation and anchoring of genome assemblies.

We demonstrate that Lep-MAP3 obtains very good performance already on 5x sequencing coverage and outperforms the fastest available software on accuracy and often on speed. We also construct de novo linkage maps with millions of markers on real low coverage whole-genome sequencing data. We will also discuss how to automatically combine physical maps (genome assemblies) and linkage maps. Lep-MAP3 is freely available with the source code under GNU general public license from http://sourceforge.net/projects/lep-map3.

P10

Comprehensive Extraction of Structural Variations from Long-read DNA Sequences

Tim White
Berlin Institute of Health

Contact: tim.white@bihealth.de

Abstract:

Structural variations (SVs) in DNA -- insertions, deletions, inversions and more complex events spanning tens of bases to multiple kilobases -- have come into focus as important correlates of disease processes in humans. Detecting SVs accurately is complicated by several factors, particularly the presence of repetitive genomic DNA. Long-read DNA sequencing technologies from Pacific Biosciences and Oxford Nanopore have the potential to radically improve the accuracy of SV discovery and genotyping in repetitive genomes, but they come with challenges of their own: an abundance of indel errors, as well as occasional long "garbage segments" that can confound traditional SV callers. Existing SV discovery tools for long-read data use fast ad hoc approaches to estimate SV locations based on read mapping information.

Our new tool, LRSV, aims to improve accuracy by considering how the reads overlapping a candidate SV site relate to one another: we build a multiple sequence alignment (MSA) from pieces of reads overlapping each candidate SV, extract a consensus sequence and align this to the reference. The challenges are to prevent the MSA from being corrupted by mismapped reads or to incorporate garbage segments at this stage, leading to erroneous calls.

LRSV takes a principled approach based on the intuition that if a set of sequences would result in a "bad" MSA, then there most likely exists a pair of sequences in that set with a "bad" pairwise alignment. Specifically, we form a "conflict graph" containing a vertex for every read segment overlapping the site of a potential deletion as suggested by read mapping CIGAR information. We add an edge between every pair of vertices that map close enough (by default, 1kbp) to each other on the genome: either a "conflict edge" (when a pairwise alignment of the relevant parts of the two reads has a low score) or a "good edge" (when this alignment has a high score). We then solve the Vertex Cover problem: We look for the smallest set of vertices to delete that would result in the elimination of all conflict edges. Finally, for each connected component of “good” edges that remains, we use partial order alignment to build a multiple sequence alignment from the surviving sequences in that component, extract a consensus sequence, and split-align this to the genome to infer the deletion breakpoints.

We present early results showing that at a minimum coverage level of 5, LRSV finds 5115 deletions in a human genome long-read dataset, including over 95% of the 3573 deletions found by the state-of-the-art tool Sniffles, while Sniffles finds only 67% of the deletions found by LRSV. Manual inspection of randomly chosen novel deletion calls suggests that most are correct, even when the minimum coverage is lowered to 3. We believe that our tool extends the utility and practicality of long-read data for SV discovery.