Program of the Monday, September 8

Monday, September 8

7:45	Registration opens		
8:45	**Keynote 2: Patrick ALOY.** ***A network biology approach to novel therapeutic strategies.***		Auditorium ERASME
9:35	Distribution in two parallel sessions		
	Auditorium ERASME	Room SCHUMANN	
	Session Mon1: Pathways and Molecular Networks (1)	**Session Mon5: Evolution and Population Genetics (1)**	
9:40	*PP01 - HubAlign: An accurate and efficient method for global alignment of protein-protein interaction networks.* Daniela Boernigen	*PP11 - Polytomy Refinement for the Correction of Dubious Duplications in Gene Trees.* Manuel Lafond	
10:05	*PP02 - Alignment-free protein interaction network comparison.* Waqar Ali	*PP12 - RidgeRace: Ridge regression for continuous ancestral character estimation on phylogenetic trees.* Christina Kratsch	
10:30	*PP03 - Fast randomisation of large genomic datasets while preserving alteration counts.* Francesco Iorio	*PP13 - Point estimates in phylogenetic reconstructions.* Philipp Benner	
10:55	Coffee Break		Main Floor & 1st Floor
	Session Mon3: Sequencing and Sequence Analysis for Genomics (1)	**Session Mon6: Evolution and Population Genetics (2)**	
11:15	*PP06 - Lambda: The local aligner for massive biological data.* Hannes Hauswedell	*PP14 - ASTRAL: Genome-scale coalescent-based species tree estimation*. Siavash Mirarab	
11:40	*PP07 - Fiona: a parallel and automatic strategy for read error correction.* Marcel Schulz	**Highlight Talk:** *HP03 - Patterns of positive selection in seven ant genomes.* Julien Roux	
12:05	**Highlight Talk**: *HP02 - Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data.* Ségolène Caboche		
12:30	LUNCH		Dining Room Contades
13:30	**Industrial and Demo Track**		See corresponding pages
14:35	**Keynote 3: Alice McHardy.** ***Gaining Insight into the Uncultured Microbial World by Computational Metagenome Analysis.***		Auditorium ERASME
15:25	Distribution in two parallel sessions		
	Auditorium ERASME	Room SCHUMANN	
	Session Mon4: Sequencing and Sequence Analysis for Genomics (2)	**Session Mon7: Structural Bioinformatics (1)**	
15:30	*PP08 - FastHap: fast and accurate single individual haplotype reconstruction using fuzzy conflict graphs.* Sepideh Mazrouee	*PP15 - Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane).* Gabriel Studer	
15:55	*PP09 - Probabilistic single-individual haplotyping*. Volodymyr Kuleshov	*PP16 - A new statistical framework to assess structural alignment quality using information compression*. James Collier	
16:20	*PP10 - cnvOffSeq: detecting intergenic copy number variation using off-target exome sequencing data*. Evangelos Bellos	*PP17 - Entropy driven partitioning of the hierarchical protein space*. Nadav Rappoport	
16:45	Coffee Break		Main Floor & 1st Floor

	Session Mon2: Pathways and Molecular Networks (2)	**Session Mon8: Structural Bioinformatics (2)**
17:05	*PP04 - Identifying transcription factor complexes and their roles*. Thorsten Will	*PP18 - PconsFold: Improved contact predictions improve protein models*. Mirco Michel
17:30	*PP05 - Personalized identification of altered pathways in cancer*. Taejin Ahn	*PP19 - Microarray R-based analysis of complex lysate experiments with MIRACLE*. Markus List
17:55	**Highlight Talk:** *HP01 - Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles*. Enrica Calura	**Highlight Talk:** *HP04 - Comprehensive analysis of DNA polymerase III alpha subunits and their homologs in bacterial genomes.* Česlovas Venclovas
18:20	Break – Moving to Poster Session	
18:30	**Poster Session (Odd numbers)**	1st Floor
19:30	**Poster Session (Even numbers)**	1st Floor
20:30	Launching Ice-Breaking Event	Main Floor
20:45	Ice-breaking Event	

Mon1 (Area C): Pathways and Molecular Networks (1)

Chairs: Ralf Zimmer, Anaïs Baudot

PP01 - HubAlign: An accurate and efficient method for global alignment of protein-protein interaction networks

Daniela Boernigen, Somaye Hashemifar and Jinbo Xu

Toyota Technological Institute at Chicago, IL 60637, USA.

ABSTRACT

Motivation: High-throughput experimental techniques have pro-duced a large amount of protein-protein interaction (PPI) data. The study of PPI networks, such as comparative analysis, shall benefit the understanding of life process and diseases at the molecular level. One way of comparative analysis is to align PPI networks to identify conserved or species-specific subnetwork motifs. A few methods have been developed for global PPI network alignment, but it still remains challenging in terms of both accuracy and efficiency.

Results: This paper presents a novel global network alignment algorithm, denoted as HubAlign, that makes use of both network topology and sequence homology information, based upon the observation that topologically important proteins in a PPI network usually are much more conserved and thus, more likely to be aligned. HubAlign uses a minimum-degree heuristic algorithm to estimate the topological and functional importance of a protein from the global network topology information. Then HubAlign aligns topologically important proteins first and gradually extends the alignment to the whole network. Extensive tests indicate that HubAlign greatly out-performs several popular methods in terms of both accuracy and efficiency, especially in detecting functionally similar proteins.

Availability: HubAlign is available freely for non-commercial purposes at http://ttic.uchicago.edu/~hashemifar/software/HubAlign.zip

Contact: jinboxu@gmail.com

PP02 - Alignment-free protein interaction network comparison

Waqar Ali[1], Tiago Rito[1], Gesine Reinert[1], Fengzhu Sun[2] and Charlotte M. Deane[1]

[1]Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK. [2]Molecular and Computational Biology Program, University of Southern California, California, USA.

ABSTRACT

Motivation: Biological network comparison software largely relies on the concept of alignment where close matches between the nodes of two or more networks are sought. These node matches are based on sequence similarity and/or interaction patterns. However due to the incomplete and error prone data sets currently available, such methods have had limited success. Moreover, the results of network alignment are in general not amenable for distance based evolutionary analysis of sets of networks. In this paper we describe Netdis, a topology based distance measure between networks, which offers the possibility of network phylogeny reconstruction.

Results: We first demonstrate that Netdis is able to correctly separate different random graph model types independent of network size and density. The biological applicability of the method is then shown by its ability to build the correct phylogenetic tree of species based solely on the topology of current protein interaction networks. Our results provide new evidence that the topology of protein interaction networks contains information about evolutionary processes, despite the lack of conservation of individual interactions. As Netdis is applicable to all networks due to its speed and simplicity we apply it to a large collection of biological and non-biological networks where it clusters diverse networks by type.

Availability: The source code of the program is freely available at http://www.stats.ox.ac.uk/research/proteins/resources.

Contact: w.ali@stats.ox.ac.uk

From Area J: Methods and Technologies for Computational Biology

PP03 - Fast randomisation of large genomic datasets while preserving alteration counts

Andrea Gobbi[1*], Francesco Iorio[2, 3*], Kevin J. Dawson[3], David C. Wedge[3], David Tamborero[4], Ludmil B. Alexandrov[3], Nuria Lopez-Bigas[4], Mathew J. Garnett[3], Giuseppe Jurman[1] and Julio Saez-Rodriguez[2].

[1]Fondazione Bruno Kessler, Trento, Italy. [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK. [3]Wellcome Trust Sanger Institute, Cambridge, UK. [4]Universitat Pompeu Fabra, Barcelona, Spain. *Equally contributing authors.

ABSTRACT
Motivation: Studying combinatorial patterns in cancer genomic datasets has recently emerged as a tool for identifying novel cancer driver networks. Approaches have been devised to quantify, for example, the tendency of a set of genes to be mutated in a 'mutually exclusive' manner. The significance of the proposed metrics is usually evaluated by computing p-values under appropriate null models. To this end, a Monte Carlo method (the switching-algorithm) is used to sample simulated datasets under a null-model that preserves patient- and gene-wise mutation rates. In this method, a genomic dataset is represented as a bipartite network, to which Markov chain updates (switching-steps) are applied. These steps modify the network topology, and a minimal number of them must be executed in order to draw simulated datasets independently under the null model. This number has previously been deducted empirically to be a linear function of the total number of variants, making this process computationally expensive.
Results: We present a novel approximate lower bound for the number of switching-steps, derived analytically. Additionally we have developed the R package BiRewire, including new efficient implementations of the switching-algorithm. We illustrate the performances of BiRewire by applying it to large real cancer genomics datasets. We report vast reductions in time requirement, with respect to existing implementations/bounds and equivalent pvalue computations. Thus, we propose BiRewire to study statistical properties in genomic datasets, and other data that can be modeled as bipartite networks.
Availability: BiRewire is available on BioConductor at
http://www.bioconductor.org/packages/2.13/bioc/html/BiRewire.html
Supplementary information: Available on Bioinformatics online and at http://www.ebi.ac.uk/_iorio/BiRewire
Contact: iorio@ebi.ac.uk

Mon2 (Area C): Pathways and Molecular Networks (2)

Chairs: Ralf Zimmer, Anaïs Baudot

PP04 - Identifying transcription factor complexes and their roles

Thorsten Will and Volkhard Helms

Center for Bioinformatics, Campus Building E2.1, Saarland University, D-66123 Saarbrücken, Germany.

ABSTRACT
Motivation: Eukaryotic gene expression is controlled through molecular logic circuits that combine regulatory signals of many different factors. In particular, complexation of transcription factors and other regulatory proteins is a prevailing and highly conserved mechanism of signal integration within critical regulatory pathways and enables us to infer controlled genes as well as the exerted regulatory mechanism. Common approaches for protein complex prediction that only use protein interaction networks, however, are designed to detect self-contained functional complexes and have difficulties to reveal dynamic combinatorial assemblies of physically interacting proteins.
Results: We developed the novel algorithm DACO that combines protein-protein interaction networks and domain-domain interaction networks with the cluster-quality metric cohesiveness. The metric is locally maximized on the holistic level of protein interactions and connectivity constraints on the domain level are used to account for the exclusive and thus inherently combinatorial nature of the interactions within such assemblies. When applied to predicting transcription factor complexes in the yeast *S.cerevisiae*, the proposed approach outperformed popular complex prediction methods by far. Furthermore, we were able to assign many of the predictions to target genes, as well as to a potential regulatory effect in agreement with literature evidence.
Availability: A prototype implementation is freely available at https://sourceforge.net/projects/dacoalgorithm/.
Contact: volkhard.helms@bioinformatik.uni-saarland.de

PP05 - Personalized identification of altered pathways in cancer

Taejin Ahn[1, 2, 3], Eunjin Lee[1, 2], Nam Huh[1] and Taesung Park[3]

[1]Samsung Advanced Institute of Technology, [2]Samsung Genome Institute, Republic of Korea. [3]Interdisciplinary Program in Bioinformatics, Seoul National University, Republic of Korea.

ABSTRACT
Motivation: Identifying altered pathways in an individual is important for understanding disease mechanisms and for the future application of custom therapeutic decisions. Existing pathway analysis techniques are mainly focused on discovering altered pathways between normal and cancer groups and are not suitable for identifying the pathway aberrance that may occur in an individual sample. A simple way to identify individual's pathway aberrance is to compare normal and tumor data from the same individual. However, the matched normal data from the same individual is often unavailable in clinical situation. We therefore suggest a new approach for the personalized identification of altered pathways, making special use of accumulated normal data in cases when a patient's matched normal data is unavailable. The philosophy behind our method is to quantify the aberrance of an individual sample's pathway by comparing it to accumulated normal samples. We

propose and examine personalized extensions of pathway statistics, Over-Representation Analysis (ORA) and Functional Class Scoring (FCS), to generate individualized pathway aberrance score (iPAS).
Results: Collected microarray data of normal tissue of lung and colon mucosa is served as reference to investigate a number of cancer individuals of lung adenocarcinoma and colon cancer, respectively. Our method concurrently captures known facts of cancer survival pathways and identifies the pathway aberrances that represent cancer differentiation status and survival. It also provides more improved validation rate of survival related pathways than when a single cancer sample is interpreted in the context of cancer-only cohort. In addition, our method is useful in classifying unknown samples into cancer or normal groups. Particularly, we identified 'amino acid synthesis and interconversion' pathway is a good indicator of lung adenocarcinoma (AUC 0.982 at independent validation). Clinical importance of the method is providing pathway interpretation of single cancer even though its matched normal data is unavailable.
Availability: The method was implemented using the R software, available at our website: http://bibs.snu.ac.kr/ipas.
Contact: tspark@stat.snu.ac.kr
Supplementary information: Available at *Bioinformatics* online.

Highlight Talk: *HP01 - Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles.*

Enrica Calura, Paolo Martini, Gabriele Sales and Chiara Romualdi.

Department of Biology, University of Padova, via U. Bassi 58/B, 35121 Padova, Italy.

ABSTRACT
The production rate of gene expression data is nothing less than astounding. However, with the benefit of hindsight we can assert that, since we completely ignored the non-coding part of the transcriptome, we spent the last decade to study cell mechanisms having few data in our hands. In this scenario, microRNAs, which are key post-trascriptional regulators, deserve special attention. Currently, miRNA and gene circuits are identified through the combination of binding prediction and expression correlation analyses, MAGIA, the web tool we developed, is an example to feel this aim (Sales et al NAR 2010, Bisognin et al NAR 2012). Although effective in many cases the simple correlation does not imply a causal relationship and a lot of false positive miRNA-mRNA interactions are still found. Moreover, miRNA and target genes are characterized by many-to-many relationships and they should be considered as part of a much more complex system of cellular interactions. Recently, to analyze the cellular circuits we developed a new web tool dedicated to topological pathway analyses called Graphite Web (Sales et al NAR 1013). Given the state of knowledge about the biogenesis of miRNAs, their mechanisms of action and the numerous experimentally validated target genes, miRNAs are also gradually appearing in the formal pathway representations such as KEGG and Reactome maps. However, the number of miRNAs annotated in pathway maps is very small and pathway analyses exploiting this new regulatory layer are still lacking. To fill these gaps, we developed micrographite a new pipeline to perform topological pathway analysis integrating gene and miRNA expression profiles. Micrographite analysis of gene and miRNA integrated transcriptome is used to study and dissect the epithelial ovarian cancer gene complexity and miRNA transcriptome defining and validating a new regulatory circuits.
Publications:
Calura E, Fruscio R, Paracchini L, Bignotti E, Ravaggi A, Martini P, Sales G, Beltrame L, Clivio L, Ceppi L, Di Marino M, Fuso Nerini I, Zanotti L, Cavalieri D, Cattoretti G, Perego P, Milani R, Katsaros D, Tognon G, Sartori E, Pecorelli S, Mangioni C, D'Incalci M, Romualdi C, Marchini S. MiRNA landscape in stage I epithelial ovarian cancer defines the histotype specificities. Clin Cancer Res. 2013 Aug 1;19(15):4114-23.

Calura E, Martini P, Sales G, Beltrame L, Chiorino G, D'Incalci M, Marchini S, Romualdi C. Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. Nucleic Acids Res. 2014;42(11):e96.

Contact: enrica.calura@unipd.it

Mon3 (Area A): Sequencing and Sequence Analysis for Genomics (1)

Chairs: To be announced

PP06 - Lambda: The local aligner for massive biological data

Hannes Hauswedell, Jochen Singer and Knut Reinert

Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany.

ABSTRACT
Motivation: Next-generation sequencing technologies produce unprecedented amounts of data, leading to

completely new research fields. One of these is metagenomics, the study of large-size DNA samples containing a multitude of diverse organisms. A key problem in metagenomics is to functionally and taxonomically classify the sequenced DNA, to which end the well known BLAST program is usually used. But BLAST has dramatic resource requirements at metagenomic scales of data, imposing a high financial or technical burden on the researcher. Multiple attempts have been made to overcome these limitations and present a viable alternative to BLAST.
Results: In this work we present Lambda, our own alternative for BLAST in the context of sequence classification. In our tests Lambda often outperforms the best tools at reproducing BLAST's results and is the fastest compared to the current state-of-the art at comparable levels of sensitivity.
Availability: Lambda was implemented in the SeqAn open source C++ library for sequence analysis and is publicly available for download at http://www.seqan.de/projects/lambda .
Contact: hannes.hauswedell@fu-berlin.de or knut.reinert@fu-berlin.de

PP07 - Fiona: a parallel and automatic strategy for read error correction

Marcel Schulz[1, 2, §], David Weese[3, §], Manuel Holtgrewe[3, §], Viktoria Dimitrova[4, 5], Sijia Niu[4, 5], Knut Reinert[3] and Hugues Richard[4, 5, §]

[1]Cluster of Excellence "Multimodal Computing and Interaction", Saarland University & Max Planck Institute for Informatics, Saarbrücken, Germany. [2]Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, USA. [3]Department of Mathematics and Computer Science, Freie Universit ¨ at Berlin, Berlin, Germany. [4]Université Pierre et Marie Curie, UMR7238, CNRS-UPMC, Paris, France. [5]CNRS, UMR7238, Laboratory of Computational and Quantitative Biology, Paris, France. [§]These authors contributed equally to this work.

ABSTRACT
Motivation: Automatic error correction of high throughput sequencing data can have a dramatic impact on the amount of usable base pairs and their quality. It has been shown that the performance of tasks such as de novo genome assembly and SNP calling can be dramatically improved after read error correction. While a large number of methods specialized for correcting substitution errors as found in Illumina data exist, few methods for the correction of indel errors, common to technologies like 454 or Ion Torrent, have been proposed.
Results: We present Fiona, a new stand-alone read error correction method. Fiona provides a new statistical approach for sequencing error detection, optimal error correction and estimates its parameters automatically. Fiona is able to correct substitution, insertion, and deletion errors and can be applied to any sequencing technology. It uses an efficient implementation of the partial suffix array to detect read overlaps with different seed lengths in parallel. We tested Fiona on several real data sets from a variety of organisms with different read lengths and compared its performance to state-of-the-art methods. Fiona shows a constantly higher correction accuracy over a broad range of data sets from 454 and Ion Torrent sequencers, without compromise in speed.
Conclusion: Fiona is an accurate, parameter-free read error correction method that can be run on inexpensive hardware and can make use of multi-core parallelization whenever available. Fiona was implemented using the SeqAn library for sequence analysis and is publicly available for download at http://www.seqan.de/projects/_ona.
Contact: mschulz@mmci.uni-saarland.de , hugues.richard@upmc.fr

Highlight Talk: *HP02 - Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data*

Ségolène Caboche[1], Christophe Audebert[2], Yves Lemoine[3] and David Hot[2,3]

[1]FRE 3642 Molecular and Cellular Medecine, CNRS, Institut Pasteur de Lille and University of Lille Nord de France. [2]Genes Diffusion, Douai, France. [3]Transcriptomics and Applied Genomics, Center for Infection and Immunity of Lille, Inserm U1019, Lille, France.

ABSTRACT
A fundamental step in High-throughput sequencing (HTS) data analysis is the mapping of reads onto reference sequences. Choosing a suitable mapper is a subtle task because of the difficulty of evaluating mapping algorithms. We present a benchmark procedure to compare mappers using both real and simulated datasets and considering computational resource and time requirements, robustness of mapping, ability to report positions for reads in repetitive regions, and ability to retrieve true genetic variation positions. To measure robustness, a new definition for a correctly mapped read was introduced. We developed CuReSim, a read simulator, and CuReSimEval, a tool to evaluate the mapping quality of the simulated reads. The benchmark procedure was applied to evaluate mappers in the context of whole genome sequencing of small genomes with Ion Torrent data. These results were used to develop a pipeline to quickly and automatically characterize pathogens during an episode of infection.
Publication:
Caboche S, Audebert C, Lemoine Y, Hot D. Comparison of mapping algorithms used in high-throughput

sequencing: application to Ion Torrent data. BMC Genomics. 2014 Apr 5;15:264.

Contact: segolene.caboche@pasteur-lille.fr

Mon4 (Area A): Sequencing and sequence analysis for genomics (2)

Chairs: To be announced

PP08 - FastHap: fast and accurate single individual haplotype reconstruction using fuzzy conflict graphs

Sepideh Mazrouee and Wei Wang.

Computer Science Department, University of California Los Angeles (UCLA), 3551 Boelter Hall, Los Angeles, CA 90095-1596, USA.

ABSTRACT

Motivation: Understanding exact structure of an individual's haplotype plays a significant role in various fields of human genetics. Despite tremendous research effort in recent years, fast and accurate haplotype reconstruction remains as an active research topic, mainly due to the computational challenges involved. Existing haplotype assembly algorithms focus primarily on improving accuracy of the assembly, making them computationally challenging for applications on large high-throughput sequence data. Therefore, there is a need to develop haplotype reconstruction algorithms that are not only
accurate but also highly scalable.

Results: In this paper, we introduce FastHap, a fast and accurate haplotype reconstruction approach, which is up to one order of magnitude faster than the state-of-the-art haplotype inference algorithms while also delivering higher accuracy than these algorithms. FastHap leverages a new similarity metric that allows us to precisely measure distances between pairs of fragments. The distance is then utilized in building the fuzzy conflict graphs of fragments. Given that optimal haplotype reconstruction based on minimum error correction (MEC) is known to be NP-hard, we use our fuzzy conflict graphs to develop a fast heuristic for fragment partitioning and haplotype reconstruction.

Availability: An implementation of FastHap is available for sharing upon request.

Contact: sepideh@cs.ucla.edu

PP09 - Probabilistic single-individual haplotyping

Volodymyr Kuleshov

Department of Computer Science, Stanford University, Stanford, CA, 94305, USA.

ABSTRACT

Motivation: Accurate haplotyping – determining from which parent particular portions of the genome were inherited – is still mostly an unresolved problem in genomics. Only recently have modern long read sequencing technologies begun to offer the promise of routine, cost-effective haplotyping. Here, we introduce ProbHap, a new haplotyping algorithm targeted at such technologies. ProbHap is based on a probabilistic graphical model; it is highly accurate and provides useful confidence scores at phased positions.

Results: On a standard benchmark dataset, ProbHap makes 11% fewer errors than current state-of-the-art methods. This accuracy can be further increased by excluding low-confidence positions, at the cost of a small drop in haplotype completeness.

Availability: Our source code is freely available at https://github.com/kuleshov/ProbHap .

Contact: kuleshov@stanford.edu

From Area J: Methods and Technologies for Computational Biology

PP10 - cnvOffSeq: detecting intergenic copy number variation using off-target exome sequencing data

Evangelos Bellos[1] and Lachlan Coin[1,2]

[1]Department of Genomics of Common Disease, Imperial College London, London W12 0NN, UK. [2]Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072, Australia.

ABSTRACT

Motivation: Exome sequencing technologies have transformed the field of Mendelian genetics and allowed for efficient detection of genomic variants in protein-coding regions. The target enrichment process that is intrinsic to exome sequencing is inherently imperfect, generating large amounts of unintended off-target sequence. Off-target data is characterized by very low and highly heterogeneous coverage and is usually discarded by exome analysis pipelines. We posit that off-target read depth is a rich but overlooked source of information that could be mined to detect intergenic copy number variation (CNV). We propose cnvOffseq, a novel normalization

framework for off-target read depth that is based on local adaptive singular value decomposition (SVD). This method is designed to address the heterogeneity of the underlying data and allows for accurate and precise CNV detection and genotyping in off-target regions.
Results: cnvOffSeq was benchmarked on whole-exome sequencing samples from the 1000 Genomes Project. In a set of 104 gold standard intergenic deletions, our method achieved a sensitivity of 57.5% and a specificity of 99.2%, while maintaining a low FDR of 5%. For gold standard deletions longer than 5kb, cnvOffSeq achieves a sensitivity of 90.4% without increasing the FDR. cnvOff-Seq outperforms both whole-genome and whole-exome CNV detection methods considerably and is shown to offer a substantial im-provement over naïve local SVD.
Availability and Implementation: cnvOffSeq is available at http://sourceforge.net/p/cnvoffseq/
Contact: evangelos.bellos09@imperial.ac.uk ; l.coin@imb.uq.edu.au

Mon5 (Area F): Evolution and Population Genomics (1)

Chairs: To be announced

PP11 - Polytomy refinement for the correction of dubious duplications in gene trees

Manuel Lafond[1], Cedric Chauve[2,3], Riccardo Dondi[4] and Nadia El-Mabrouk[1]

[1]Department of Computer Science, Université de Montreal, Montreal (QC), Canada. [2]LaBRI, Université Bordeaux 1, Bordeaux, France. [3]Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada. [4]Universitá degli Studi di Bergamo, Bergamo, Italy.

ABSTRACT
Motivation: Large scale methods for inferring gene trees are errorprone. Correcting gene trees for weakly supported features often results in non-binary trees, i.e., trees with polytomies, thus raising the natural question of refining such polytomies into binary trees. A feature pointing toward potential errors in gene trees are duplications that are not supported by the presence of multiple gene copies.
Results: We introduce the problem of refining polytomies in a gene tree while minimizing the number of created non-apparent duplications in the resulting tree. We show that this problem can be described as a graph-theoretical optimization problem. We provide a bounded heuristic with guaranteed optimality for well characterized instances. We apply our algorithm to a set of ray-finned fish gene trees from the Ensembl database to illustrate its ability to correct dubious duplications.
Availability: The C++ source code for the algorithms and simulations described in the paper are available at http://wwwetud.iro.umontreal.ca/lafonman/software.php .
Contact: lafonman@iro.umontreal.ca , mabrouk@iro.umontreal.ca

PP12 - RidgeRace: Ridge regression for continuous ancestral character estimation on phylogenetic trees.

Christina Kratsch and Alice McHardy.

Department for Algorithmic Bioinformatics, Heinrich Heine University, Universitätsstr. 1, 40225 Düsseldorf, Germany.

ABSTRACT
Motivation: Ancestral character state reconstruction describes a set of techniques for estimating phenotypic or genetic features of species or related individuals that are the predecessors of those present today. Such reconstructions can reach into the distant past and can provide insights into the history of a population or a set of species when fossil data are not available, or they can be used to test evolutionary hypotheses e.g. on the co-evolution of traits. Typical methods for ancestral character state reconstruction of continuous characters consider the phylogeny of the underlying data and estimate the ancestral process along the branches of the tree. They usually assume a Brownian motion model of character evolution or extensions thereof, requiring specific assumptions on the rate of phenotypic evolution.
Results: We suggest using ridge regression to infer rates for each branch of the tree and the ancestral values at each inner node. We performed extensive simulations to evaluate the performance of this method and have shown that the accuracy of its reconstructed ancestral values is competitive to reconstructions using other state-of- the-art software. Using a hierarchical clustering of gene mutation profiles from an ovarian cancer dataset, we demonstrate the use of the method as a feature selection tool.
Availability: The algorithm described here is implemented in C++ as a standalone program, and the source code is freely available at http://algbio.cs.uni-duesseldorf.de/software/RidgeRace.tar.gz .
Contact: mchardy@hhu.de

PP13 - Point estimates in phylogenetic reconstructions

Philipp Benner[1], Miroslav Bacak[1] and Pierre-Yves Bourguignon[1,2]

[1]Max-Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig, Germany. [2]Isthmus SARL, 81 rue Réaumur, 75002 Paris, France.

ABSTRACT

Motivation: The construction of statistics for summarizing posterior samples returned by a Bayesian phylogenetic study has so far been hindered by the poor geometric insights available into the space of phylogenetic trees, and adhoc methods such as the derivation of a consensus tree make up for the ill-definition of the usual concepts of posterior mean, while bootstrap methods mitigate the absence of a sound concept of variance. Yielding satisfactory results with sufficiently concentrated posterior distributions, such methods fall short of providing a faithful summary of posterior distributions if the data does not offer compelling evidence for a single topology.

Results: Building upon previous work of Billera et al. (2001), summary statistics such as sample mean, median, and variance are defined as the geometric median, Fréchet mean and variance respectively. Their computation is enabled by recently published works (Bačák, 2013; Miller et al., 2012), and embeds an algorithm for computing

shortest paths in the space of trees (Owen and Provan, 2011). Studying the phylogeny of a set of plants, where several tree topologies occur in the posterior sample, the posterior mean balances correctly the contributions from the different topologies, where a consensus tree would be biased. Comparisons of the posterior mean, median, and consensus trees with the ground truth using simulated data also reveals the benefits of a sound averaging method when reconstructing phylogenetic trees.

Availability: We provide two independent implementations of the algorithm for computing Fréchet means, geometric medians, and variances in the space of phylogenetic trees.

TFBayes: https://github.com/pbenner/tfbayes , TrAP: https://github.com/bacak/TrAP .

Contact: philipp.benner@mis.mpg.de

Mon6 (Area F): Evolution and Population Genomics (2)

Chairs: To be announced

PP14 - ASTRAL: Genome-scale coalescent-based species tree estimation

Siavash Mirarab[1], Rezwana Reaz Rimpi[1], Md. Shamsuzzoha Bayzid[1], Théo Zimmermann[1], Shel Swenson[2] and Tandy Warnow[1]

[1]Department of Computer Science, The University of Texas at Austin, Austin TX, USA. [2]Department of Electrical Engineering, The University of Southern California, Los Angeles CA, USA.

ABSTRACT

Motivation: Species trees provide insight into basic biology, including the mechanisms of evolution and how it modifies biomolecular function and structure, biodiversity, and co-evolution between genes and species. Yet gene trees often differ from species trees, creating challenges to species tree estimation. One of the most frequent causes for conflicting topologies between gene trees and species trees is incomplete lineage sorting (ILS), which is modelled by the multi-species coalescent. While many methods have been developed to estimate species trees from multiple genes, some which have statistical guarantees under the multi-species coalescent model, existing methods are too computationally intensive for use with genome-scale analyses or have been shown to have poor accuracy under some realistic conditions.

Results: We present ASTRAL, a fast method for estimating species trees from multiple genes. ASTRAL is statistically consistent, can run on datasets with thousands of genes, and has outstanding accuracy – improving upon MP-EST and the population tree from BUCKy, two statistically consistent leading coalescent-based methods. ASTRAL is often more accurate than concatenation using maximum likelihood, except when ILS levels are low or there are too few gene trees.

Availability: ASTRAL is available in open source form at https://github.com/smirarab/ASTRAL /. Datasets studied in this paper are available at http://www.cs.utexas.edu/users/phylo/datasets/astral .

Contact: warnow@illinois.edu

Highlight Talk: *HP03 - Patterns of positive selection in seven ant genomes.*

Julien Roux[1], Eyal Privman[2], Sébastien Moretti[1], Josephine Daub[3], Marc Robinson-Rechavi[1] and Laurent Keller[1]

[1]University of Lausanne, Switzerland. [2]University of Haifa, Israel.[3]University of Bern, Switzerland.

ABSTRACT

The evolution of ants is marked by remarkable adaptations that allowed the development of very complex social systems. To identify how ant-specific adaptations are associated with patterns of molecular evolution, we searched for signs of positive selection on amino-acid changes in proteins. We identified 24 functional categories of genes which were enriched for positively selected genes in the ant lineage. We also

reanalyzed genome-wide datasets in bees and flies with the same methodology, to check whether positive selection was specific to ants or also present in other insects. Notably, genes implicated in immunity were enriched for positively selected genes in the three lineages, ruling out the hypothesis that the evolution of hygienic behaviors in social insects caused a major relaxation of selective pressure on immune genes. Our scan also indicated that genes implicated in neurogenesis and olfaction started to undergo increased positive selection before the evolution of sociality in Hymenoptera. Finally, the comparison between these three lineages allowed us to pinpoint molecular evolution patterns that were specific to the ant lineage. In particular, there was ant-specific recurrent positive selection on genes with mitochondrial functions, suggesting that mitochondrial activity was improved during the evolution of this lineage. This might have been an important step toward the evolution of extreme lifespan that is a hallmark of ants.

Publication:

Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. Patterns of positive selection in seven ant genomes. Mol Biol Evol. 2014 Jul;31(7):1661-85.

Contact: julien.roux@unil.ch

Mon7 (Area E): Structural Bioinformatics (1)

Chairs: Torsten Schwede, Anna Tramontano

PP15 - Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane)

Gabriel Studer[1,2], Marco Biasini[1,2] and Torsten Schwede[1,2]

[1]Biozentrum, University of Basel, Basel, 4056, Switzerland. [2]SIB Swiss Institute of Bioinformatics, Basel, 4056, Switzerland.

ABSTRACT

Motivation: Membrane proteins are an important class of biological macromolecules involved in many cellular key processes including signalling and transport. They account for one third of genes in the human genome and more than 50% of current drug targets. Despite their importance, experimental structural data is sparse, resulting in high expectations for computational modelling tools to help filling this gap. However, as many empirical methods have been trained on experimental structural data, which is biased towards soluble globular proteins, their accuracy for transmembrane proteins is often limited.

Results: We developed a local model quality estimation method for membrane proteins ("QMEANBrane") by combining statistical potentials trained on membrane protein structures with a per-residue weighting scheme. The increasing number of available experimental membrane protein structures allowed us to train membrane-specific statistical potentials that approach statistical saturation. We show that reliable local quality estimation of membrane protein models is possible, thereby extending local quality estimation to these biologically relevant molecules.

Availability: Source code and data sets are available on request.

Contact: torsten.schwede@unibas.ch

PP16 - A new statistical framework to assess structural alignment quality using information compression

James Collier[1], Lloyd Allison[1], Arthur Lesk[2], Maria Garcia de La Banda[1] and Arun Konagurthu[1]

[1] Clayton School of Information Technology, Monash University, Clayton, VIC 3800 Australia. [2] Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802 USA.

ABSTRACT

Motivation: Progress in protein biology depends on the reliability of results from a handful of computational techniques, structural alignments being one. Recent reviews have highlighted substantial inconsistencies and differences between alignment results generated by the ever-growing stock of structural alignment programs. The lack of consensus on how the quality of structural alignments must be assessed has been identified as the main cause for the observed differences. Current methods assess structural alignment quality by constructing a scoring function that attempts to balance conflicting criteria, mainly alignment coverage and fidelity of structures under superposition. This traditional approach to measuring alignment quality, the subject of considerable literature, has failed to solve the problem. Further development along the same lines is unlikely to rectify the current deficiencies in the field.

Results: This paper proposes a new statistical framework to assess structural alignment quality and significance based on lossless information compression. This is a radical departure from the traditional approach of formulating scoring functions. It links the structural alignment problem to the general class of statistical inductive inference problems, solved using the information-theoretic criterion of minimum message length. Based on this, we developed an efficient and reliable measure of structural alignment quality, I-value. The performance of I-value is demonstrated in comparison with a number of popular scoring functions, on a

large collection of competing alignments. Our analysis shows that I-value provides a rigorous and reliable quantification of structural alignment quality, addressing a major gap in the field.
Availability: http://lcb.infotech.monash.edu.au/I-value
Supplementary Information: http://lcb.infotech.monash.edu.au/I-value/suppl.html
Contact: arun.konagurthu@monash.edu

From Area J: Methods and Technologies for Computational Biology

PP17 - Entropy driven partitioning of the hierarchical protein space

Nadav Rappoport[1], Amos Stern[1], Nathan Linial[1] and Michal Linial[2]

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel. [2]Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Israel.

ABSTRACT
Motivation: Modern protein sequencing techniques have led to the determination of over 50 million protein sequences. *ProtoNet* is a clustering system that provides a continuous hierarchical agglomerative clustering tree for all proteins. While ProtoNet performs unsupervised classification of all included proteins, finding an optimal level of granularity for the purpose of focusing on protein functional groups remain elusive. Here, we ask whether knowledge-based annotations on protein families can support the automatic, unsupervised methods for identifying high quality protein families. We present a method that yields within the ProtoNet hierarchy an optimal partition of clusters, relative to manual annotation schemes. The methods principle is to minimize the entropy-derived distance between annotation-based partitions and all available hierarchical partitions. We describe the *best front* (BF) partition of 2,478,328 proteins from UniRef50. Out of 4,929,553 ProtoNet tree clusters, BF based on Pfam annotations contain 26,891 clusters. The high quality of the partition is validated by the close correspondence with the set of clusters that best describe thousands of keywords of Pfam. The BF is shown to be superior to naïve cut in the ProtoNet tree that yields a similar number of clusters. Finally, we used parameters intrinsic to the clustering process to enrich a-priori the BF's clusters. We present the entropy-based method's benefit in overcoming the unavoidable limitations of nested clusters in ProtoNet. We suggest that this automatic information-based cluster selection can be useful for other large-scale annotation schemes, as well as for systematically testing and comparing putative families derived from alternative clustering methods.
Availability: A catalogue of BF clusters for thousands of Pfam keywords is provided at: http://protonet.cs.huji.ac.il/bestFront/
Contact: michal.linial@huji.ac.il

Mon8 (Area E): Structural Bioinformatics (2)

Chairs: Torsten Schwede, Anna Tramontano

PP18 - PconsFold: Improved contact predictions improve protein models

Mirco Michel[1,2], Sikander Hayat[3], Marcin J. Skwark[4], Chris Sander[5], Debora S. Marks[3] and Arne Elofsson[1,2]

[1]Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden, [2]Science for Life Laboratory, Box 1031, 17121 Solna, Sweden, [3]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA, [4]Department of Information and Computer Science, Aalto University, PO Box 15400, FI-00076 Aalto, Finland, and [5]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, USA.

ABSTRACT
Motivation: Recently is has been shown that the quality of protein contact prediction from evolutionary information can be improved significantly if direct and indirect information is separated. Given sufficiently large protein families the contact predictions contain sufficient information to predict the structure of many protein families. However, since the first studies contact prediction methods have improved. Here, we ask how much the final models are improved if improved contact predictions are used.
Results: In a small benchmark of 15 proteins we show that the TM-scores of top ranked models are improved by on average 33% using PconsFold compared to the original version of EVfold. In a larger benchmark we find that the quality is improved with 15-30% when using PconsC in comparison to earlier contact prediction methods.Further, using Rosetta instead of CNS does not significantly improve global model accuracy but the chemistry of models generated with Rosetta is improved.
Availability: PconsFold is a fully automated pipeline for ab-initio protein structure prediction based on evolutionary information. PconsFold is based on PconsC contact prediction and uses the Rosetta folding protocol. Due to its modularity, the contact prediction tool can be easily exchanged. The source code of

PconsFold is available on GitHub at https://www.github.com/ElofssonLab/pcons-fold under the MIT license. PconsC is available from http://c.pcons.net/ .
Contact: arne@bioinfo.se
Supplementary information: Supplementary data are available at Bioinformatics online.

From Area J: Methods and Technologies for Computational Biology

PP19 - Microarray R-based analysis of complex lysate experiments with MIRACLE

Markus List[1,2,3,§], Ines Block[1,2,§], Marlene Lemvig Pedersen[1,2], Helle Christiansen[1,2], Steffen Schmidt[1,2], Mads Thomassen[1,3], Qihua Tan[3,4], Jan Baumbach[5] and Jan Mollenhauer[1,2]

[1]Lundbeckfonden Center of Excellence in Nanomedicine NanoCAN, University of Southern Denmark, Odense, Denmark. [2]Molecular Oncology, Institute of Molecular Medicine, University of Southern Denmark, Odense, Denmark. [3]Institute of Clinical Research, University of Southern Denmark, Odense, Denmark. [4]Epidemiology, Biostatistics and Biodemography, Institute of Public Health, University of Southern Denmark, Odense, Denmark. [5]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. [§] joint first authorship.

ABSTRACT
Motivation: Reverse phase protein arrays (RPPAs) allow sensitive quantification of relative protein abundance in thousands of samples in parallel. Typical challenges involved in this technology are antibody selection, sample preparation and optimization of staining conditions. The issue of combining effective sample management and data analysis, however, has been widely neglected.
Results: This motivated us to develop MIRACLE, a comprehensive and user-friendly web application bridging the gap between spotting and array analysis by conveniently keeping track of sample information. Data processing includes correction of staining bias, estimation of protein concentration from response curves, normalization for total protein amount per sample and statistical evaluation. Established analysis methods have been integrated with MIRACLE, offering experimental scientists an end-to-end solution for sample management and for carrying out data analysis. In addition, experienced users have the possibility to export data to R for more complex analyses. MIRACLE thus has the potential to further spread utilization of RPPAs as an emerging technology for high-throughput protein analysis.
Availability: Project URL: http://www.nanocan.org/miracle/
Contact: mlist@health.sdu.dk

Highlight Talk: *HP04 - Comprehensive analysis of DNA polymerase III alpha subunits and their homologs in bacterial genomes*

Kęstutis Timinskas[1], Monika Balvočiūtė[2], Albertas Timinskas[1] and Česlovas Venclovas[1]

[1]Institute of Biotechnology, Vilnius University, Lithuania. [2]University of Otago, New Zealand.

ABSTRACT
Bacteria, unlike archaea and eukaryotes, use distinct C-family DNA polymerases for genome replication. Unfortunately, except for a few species, bacterial genome replication is poorly characterized. It is not known whether all bacteria use C-family DNA polymerases for DNA replication, how many distinct C-family groups are there, and how many different replication systems they form. In order to address these questions, we performed extensive computational analysis of C-family polymerases in nearly 2000 complete bacterial genomes. We found that all the genomes without exception encode at least one C-family polymerase implying the universal use of this polymerase family for bacterial DNA replication. Our analysis revealed four distinct groups of C-family polymerases. Based on their properties and distribution in genomes we discovered a novel, so far experimentally uncharacterized, replication system in Clostridia. Computational results also indicated that one of the C-family groups might be responsible for shaping genomic G+C content.
Publication:
Timinskas K, Balvočiūtė M, Timinskas A, Venclovas Č. Comprehensive analysis of DNA polymerase III α subunits and their homologs in bacterial genomes. Nucleic Acids Res. 2014 Feb;42(3):1393-413.

Contact: venclovas@ibt.lt