ECCB'14

PROCEEDINGS PAPERS ABSTRACTS

Area A: Sequencing and Sequence Analysis for Genomics

Lambda: The Local Aligner for Massive Biological Data.

Hannes Hauswedell, Jochen Singer and Knut Reinert.
Department of Mathematics and Computer Science, Freie Universit ¨ at 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

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

Towards a piRNA prediction using multiple kernel fusion and support vector machine.

Jocelyn Brayet[1,2], Farida Zehraoui[1], Laurence Jeanson-Leh[2], David Israeli[2] and Fariza Tahi[1].
[1]IBISC, UEVE/Genopole, IBGBI, 23 bv. de France, 91000 Evry, France. [2]Genethon, 1, bis rue de l'Internationale, 91002 Evry Cedex, France.

ABSTRACT
Motivation: Piwi interacting RNA (piRNA) is the most recently discovered and the least investigated class of AGO/Piwi protein interacting small non-coding RNAs. PiRNAs are mostly known to be involved in protecting the genome from invasive transposable elements. But recent discoveries suggest their involvement in the pathophysiology of diseases, such as cancer. Their identification is therefore an important task, and computational methods are needed. However, the lack of conserved piRNA sequences and structural elements makes this identification very challenging and difficult.
Results: In the present study, we propose a new modular and extensible machine learning method based on multiple kernels and a support vector machine (SVM) classifier for piRNA identification. Very few piRNA features are known to date. The use of a multiple kernels approach allows editing, adding or removing piRNA features that can be heterogeneous in a modular manner according to their relevance in a given species. Our algorithm is based on a combination of the previously identified features (sequence features (k-mer motifs and a uridine at the first position) and piRNAs cluster feature) and a new telomere/centromere vicinity feature. These features are heterogeneous and the kernels allow to unify their representation. The proposed algorithm, named piRPred, gives very promising results on Drosophila and Human data and outscores previously published piRNA identification algorithms.
Availability: piRPred is freely available to non-commercial users on our Web server EvryRNA: http://EvryRNA.ibisc.univ-evry.fr
Contact: tahi@ibisc.univ-evry.fr

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

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

AREA B: Gene Expression

Two-dimensional segmentation for analyzing HiC data.

Celine Levy-Leduc[1], Maud Delattre[1], Tristan Mary-Huard[1,2] and Stephane Robin[1].
[1]AgroParisTech/INRA MIA 518, 16 rue Claude Bernard, 75005 Paris, France. [2]UMR de Génétique Végétale, INRA/Université Paris-Sud/CNRS, 91190 Gif-sur-Yvette, France.

ABSTRACT

Motivation: The spatial conformation of the chromosome has a deep influence on gene regulation and expression. HiC technology allows the evaluation of the spatial proximity between any pair of loci along the genome. It results in a data matrix where blocks corresponding to (self-)interacting regions appear. The delimitation of such blocks is critical to better understand the spatial organization of the chromatin. From a computational point of view, it results in a 2D-segmentation problem.

Results: We focus on the detection of cis-interacting regions, which appear to be prominent in observed data. We define a block-wise segmentation model for the detection of such regions. We prove that the maximization of the likelihood with respect to the block boundaries can be rephrased in terms of a 1D-segmentation problem, for which the standard dynamic programming applies. The performance of the proposed methods are assessed by a simulation study on both synthetic and re-sampled data. A comparative study on public data shows good concordance with biologically confirmed regions.

Availability: The HiCseg R package is available from the Comprehensive R Archive Network (CRAN) and from the web page of the corresponding author.

Contact: celine.levy-leduc@agroparistech.fr

Broad-Enrich: Functional interpretation of large sets of broad genomic regions.

Raymond Cavalcante[1], Chee Lee[1], Ryan Welch[1,2], Snehal Patil[3], Terry Weymouth[3], Laura Scott[2] and Maureen Sartor[1,2,3].

[1]Department of Computational Medicine and Bioinformatics, [2]Department of Biostatistics, and [3]Center of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA.

ABSTRACT

Motivation: Functional enrichment testing facilitates the interpretation of ChIP-seq data in terms of pathways and other biological contexts. Previous methods developed and used to test for key gene sets affected in ChIP-seq experiments treat peaks as points, and are based on the number of peaks associated with a gene or a binary score for each gene. These approaches work well for transcription factors, but histone modifications often occur over broad domains, and across multiple genes.

Results: To incorporate the unique properties of broad domains into functional enrichment testing, we developed Broad-Enrich, a method that uses the proportion of each gene's locus covered by a peak. We show that our method has a well-calibrated false positive rate, performing well with ChIP-seq data having broad domains compared to alternative approaches. We illustrate Broad-Enrich with 55 ENCODE ChIP-seq datasets using different methods to define gene loci. Broad-Enrich can also be applied to other datasets consisting of broad genomic domains such as copy number variations.

Availability: http://broad-enrich.med.umich.edu for web version and R package.

Contact: sartorma@umich.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Estimating the activity of transcription factors by the effect on their target genes.

Theresa Schacht[1,2,3], Marcus Oswald[1,2], Roland Eils[3,4], Stefan Eichmüller[5] and Rainer Koenig[1,2,3].

[1]Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, D-07747 Jena, Erlanger Allee 101, Germany. [2]Network Modeling, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute Jena, Beutenbergstrasse 11a, 07745 Jena, Germany. [3]Theoretical Bioinformatics, German Cancer Research Center, INF 580, 69121 Heidelberg, Germany. [4]Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology, and Bioquant, University of Heidelberg, Im Neuenheimer Feld 267, Germany. [5]Translational Immunology, German Cancer Research Center (DKFZ), INF 280, 69120 Heidelberg, Germany.

ABSTRACT

Motivation: Understanding regulation of transcription is central for elucidating cellular regulation. Several statistical and mechanistic models have come up the last couple of years explaining gene transcription levels using information of potential transcriptional regulators as transcription factors (TFs) and information from epigenetic modifications. The activity of TFs is often inferred by their transcription levels, promoter binding and epigenetic effects. However, in principle, these methods do not take hard-to-measure influences such as post-transcriptional modifications into account.

Results: For TFs, we present a novel concept circumventing this problem. We estimate the regulatory activity of TFs using their cumulative effects on their target genes. We established our model using expression data of 59 cell lines from the National Cancer Institute. The trained model was applied to an independent expression dataset of melanoma cells yielding excellent expression predictions and elucidated regulation of melanogenesis.

Implementation: Using mixed integer linear programming (MILP), we implemented a switch like optimization enabling a constrained but optimal selection of TFs and optimal model selection estimating their effects. The method is generic and can also be applied to further regulators of transcription.

Contact: Rainer.Koenig@uni-jena.de

Modeling DNA methylation dynamics with approaches from phylogenetics.

John A. Capra[1] and Dennis Kostka[2].
[1]Center for Human Genetics Research and Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, 37232, USA. [2]Departments of Developmental Biology and Computational & Systems Biology, University of Pittsburgh, Pittsburgh, PA, 15201, USA.

ABSTRACT

Motivation: Methylation of CpG dinucleotides is a prevalent epigenetic modification that is required for proper development in vertebrates. Genome-wide DNA methylation assays have become increasingly common, and this has enabled characterization of DNA methylation in distinct stages across differentiating cellular lineages. Changes in CpG methylation are essential to cellular differentiation; however, current methods for modeling methylation dynamics do not account for the dependency structure between precursor and dependent cell types.

Results: We developed a continuous-time Markov chain approach, based on the observation that changes in methylation state over tissue differentiation can be modeled similarly to DNA nucleotide changes over evolutionary time. This model explicitly takes precursor to descendant relationships into account and enables inference of CpG methylation dynamics. To illustrate our method, we analyzed a high-resolution methylation map of the differentiation of mouse stem cells into several blood cell types. Our model can successfully infer unobserved CpG methylation states from observations at the same sites in related cell types (90% correct), and this approach more accurately reconstructs missing data than imputation based on neighboring CpGs (84% correct). Additionally, the single CpG resolution of our methylation dynamics estimates enabled us to show that DNA sequence context of CpG sites is informative about methylation dynamics across tissue differentiation. Finally, we identified genomic regions with clusters of highly dynamic CpGs and present a likely functional example. Our work establishes a framework for inference and modeling that is well-suited to DNA methylation data, and our success suggests that other methods for analyzing DNA nucleotide substitutions will also translate to the modeling of epigenetic phenomena.

Availability: Source code is available at www.kostkalab.net/software .

Contact: tony.capra@vanderbilt.edu , kostka@pitt.edu

Area C: Pathways and Molecular Networks

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

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.

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

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

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

Area D: Computational Systems Biology

Experimental design schemes for learning Boolean network models.

Nir Atias, Michal Gershenzon, Katia Labazin and Roded Sharan.
Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel.

ABSTRACT
Motivation: A holy grail of biological research is a working model of the cell. Current modeling frameworks, especially in the protein-protein interaction domain, are mostly topological in nature, calling for stronger and more expressive network models. One promising alternative is logic-based, or Boolean network modeling, which was successfully applied to model signaling regulatory circuits in human. Learning such models requires observing the system under a sufficient number of different conditions. To date, the amount of measured data is the main
bottleneck in learning informative Boolean models, underscoring the need for efficient experimental design strategies.
Results: We developed novel design approaches that greedily select an experiment to be performed so as to maximize the difference or the entropy in the results it induces with respect to current best-fit models. Unique to our maximum difference approach is the ability to account for all (possibly exponential number of) Boolean models displaying high fit to the available data. We applied both approaches to simulated and real data from the EFGR and IL1 signaling systems in human. We demonstrate the utility of the developed strategies in substantially improving on a random selection approach. Our design schemes highlight the redundancy in these data sets, leading up to 11-fold savings in the number of experiments to be performed.
Availability: Source code will be made available upon acceptance of the manuscript.
Contact: roded@post.tau.ac.il

TEMPI: Probabilistic modeling Time-Evolving differential PPI networks with MultiPle Information.

Yongsoo Kim[1], Jin-Hyeok Jang[1], Seungjin Choi[2] and Daehee Hwang[1,3].
[1]School of Interdisciplinary Bioscience and Bioengineering and [2]Department of Computer Science and Engineering, Pohang University of Science and Technology, Pohang 790-784, Korea, [3]Center for Systems Biology of Plant Senescence and Life History, Institute for Basic Science, Daegu Gyeongbuk Institute of Science and Technology, Daegu 711-873, Korea.

ABSTRACT
Motivation: Time-evolving differential protein-protein interaction (PPI) networks are essential to understand serial activation of differentially regulated (up- or down-regulated) cellular processes (DRPs) and their interplays over time. Despite developments in the network inference, current methods are still limited in identifying temporal transition of structures of PPI networks, DRPs associated with the structural transition, and the interplays among the DRPs over time.
Results: Here, we present a probabilistic model for estimating Time-Evolving differential PPI networks with MultiPle Information (TEMPI). This model describes probabilistic relationships among network structures, time-course gene expression data, and Gene Ontology biological processes (GOBPs). By maximizing the likelihood of the probabilistic model, TEMPI estimates jointly the time-evolving differential PPI networks (TDNs) describing temporal transition of PPI network structures together with serial activation of DRPs associated with transiting networks. This joint estimation enables us to interpret the TDNs in terms of temporal transition of the DRPs. To demonstrate the utility of TEMPI, we applied it to two time-course datasets. TEMPI identified the TDNs that correctly delineated temporal transition of DRPs and time-dependent associations between the DRPs. These TDNs provide hypotheses for mechanisms underlying serial activation of key DRPs and their temporal associations.
Availability: Source code and sample data files are available at http://sbm.postech.ac.kr/tempi/sources.zip .
Contact: seungjin@postech.ac.kr or dhwnag@dgist.ac.kr

Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations.

Tommi Suvitaival[1], Simon Rogers[2] and Samuel Kaski[1,3].
[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, FI-00076, Espoo, Finland. [2]School of Computing Science, University of Glasgow, Glasgow, G12 8QQ, UK. [3]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland.

ABSTRACT
Motivation: Data analysis for metabolomics suffers from uncertainty due to the noisy measurement technology and the small sample-size of experiments. Noise and the small sample-size lead to a high probability of false findings. Further, individual compounds have natural variation between samples, which in many cases renders

them unreliable as biomarkers. However, the levels of similar compounds are typically highly correlated, which is a phenomenon that we model in this work.
Results: We propose a hierarchical Bayesian model for inferring differences between groups of samples more accurately in metabolomic studies, where the observed compounds are collinear. We discover that the method decreases the error of weak and non-existent covariate effects, and thereby reduces false positive findings. To achieve this, the method makes use of the mass spectral peak data by clustering similar peaks into latent compounds, and by further clustering latent compounds into groups that respond in a coherent way to the experimental covariates. We demonstrate the method with three simulated studies and validate it with a metabolomic benchmark data set.
Availability and Implementation: An implementation in R is available at http://research.ics.aalto.fi/mi/software/peakANOVA/ .
Contact: tommi.suvitaival@aalto.fi , simon.rogers@glasgow.ac.uk , samuel.kaski@aalto.fi .

Causal network inference using biochemical kinetics.

Chris Oates[1], Frank Dondelinger[2], Nora Bayani[3], James Korkola[4], Joe Gray[4] and Sach Mukherjee[2,5].
[1]Department of Statistics, University of Warwick, Coventry, UK. [2]MRC Biostatistics Unit, Cambridge, UK. [3]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, USA. [4]Knight Cancer Institute, Oregon Health and Science University, Portland, USA. [5]School of Clinical Medicine, University of Cambridge, Cambridge, UK.

ABSTRACT
Motivation: Networks are widely used as structural summaries of biochemical systems. Statistical estimation of networks is usually based on linear or discrete models. However, the dynamics of biochemical systems are generally nonlinear, suggesting that suitable nonlinear formulations may offer gains with respect to causal network inference and aid in associated prediction problems.
Results: We present a general framework for network inference and dynamical prediction using time-course data that is rooted in nonlinear biochemical kinetics. This is achieved by considering a dynamical system based on a chemical reaction graph with associated kinetic parameters. Both the graph and kinetic parameters are treated as unknown; inference is carried out within a Bayesian framework. This allows prediction of dynamical behavior even when the underlying reaction graph itself is unknown or uncertain. Results, based on (i) data simulated from a mechanistic model of mitogen activated protein kinase signaling and (ii) phosphoproteomic data
from cancer cell lines, demonstrate that nonlinear formulations can yield gains in causal network inference and permit dynamical prediction and uncertainty quantification in the challenging setting where the reaction graph is unknown.
Availability: MATLAB R2014a software is available to download from http://warwick.ac.uk/chrisoates .
Contact: c.oates@warwick.ac.uk ; sach@mrc-bsu.cam.ac.uk

Effects of small particle numbers on long-term behaviour in discrete biochemical systems.

Peter Kreyssig[1], Christian Wozar[1], Stephan Peter[1], Tomas Veloz[2,3,4], Bashar Ibrahim[1,5,6] and Peter Dittrich[1].
[1]Bio Systems Analysis Group, Department of Mathematics and Computer Science and Jena Centre for Bioinformatics, Friedrich Schiller University Jena, 07743 Jena, Germany. [2]Mathematics Department, University of British Columbia, Kelowna, BC V1V 1V7, Canada. [3]Instituto de Filosofía y Ciencias de la Complejidad - IFICC, Los Alerces 3024 Ñuñoa, Santiago, Chile. [4]Center Leo Apostel, Vrije Universiteit Brussel, Krijgskundestraat 33, B-1160 Brussels, Belgium. [5]Umm Al-Qura University, 1109 Makkah Al-Mukarramah, Kingdom of Saudi Arabia. [6]Al-Qunfudah Center for Scientific Research (QCSR), 21912 Al-Qunfudah, Kingdom of Saudi Arabia.

ABSTRACT
Motivation: The functioning of many biological processes depends on the appearance of only a small number of a single molecular species. Additionally, the observation of molecular crowding leads to the insight that even a high number of copies of species does not guarantee their interaction. How single particles contribute to stabilising biological systems is not well understood yet. Hence we aim at determining the influence of single molecules on the long-term behaviour of biological systems, i.e. whether they can reach a steady state or not.
Results: We provide theoretical considerations and a tool to analyse SBML models for the possibility to stabilise due to the described effects. The theory is an extension of chemical organisation theory which we called discrete chemical organisation theory. Furthermore we scanned the BioModels Database for the occurrence of discrete chemical organisations. To exemplify our method we describe an application to the Template model of the mitotic spindle assembly checkpoint mechanism.
Availability: http://www.biosys.uni-jena.de/Services.html
Contact: bashar.ibrahim@uni-jena.de , dittrich@minet.uni-jena.de
Supplementary Information: Supplementary data are available at Bioinformatics online.

Area E: Structural Bioinformatics

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.

CRISPRstrand: Predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci.

Omer S. Alkhnbashi[1], Fabrizio Costa[1], Shiraz A. Shah[2], Roger A. Garrett[2], Sita J. Saunders[1] and Rolf Backofen[1,3].

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany, [2]Archaea Centre, Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK2200 Copenhagen, Denmark, [3]BIOSS Centre for Biological Signalling Studies, Cluster of Excellence, University of Freiburg, Germany.

ABSTRACT

Motivation: The discovery of CRISPR-Cas systems almost 20 years ago rapidly changed our perception of the bacterial and archaeal immune systems. CRISPR loci consist of several repetitive DNA sequences called repeats, inter-spaced by stretches of variable length sequences called spacers. This CRISPR array is transcribed and processed into multiple mature RNA species (crRNAs). A single crRNA is integrated into an interference complex, together with CRISPR-associated (Cas) proteins, to bind and degrade invading nucleic acids. Although existing bioinformatics tools can recognize CRISPR loci by their characteristic repeat-spacer architecture, they generally output CRISPR arrays of ambiguous orientation and thus do not determine the strand from which crRNAs are processed. Knowledge of the correct orientation is crucial for many tasks, including the classification of CRISPR conservation, the detection of leader regions, the identification of target sites (protospacers) on invading genetic elements, and the characterization of protospacer-adjacent motifs (PAMs).

Results: We present a fast and accurate tool to determine the crRNA-encoding strand at CRISPR loci by predicting the correct orientation of repeats based on an advanced machine learning approach. Both the repeat sequence and mutation information were encoded and processed by an efficient graph kernel to learn higher order correlations. The model was trained and tested on curated data comprising more than 4,500 CRISPRs and yielded a remarkable performance of 0.95 AUC ROC (area under the curve of the receiver operator characteristic). In addition, we show that accurate orientation information greatly improved detection of conserved repeat sequence families and structure motifs. We integrated CRISPRstrand predictions into our CRISPRmap web server of CRISPR conservation and updated the latter to version 2.0.

Availability: CRISPRmap and CRISPRstrand are available at http://rna.informatik.uni-freiburg.de/CRISPRmap

Contact: backofen@informatik.uni-freiburg.de

Identification of structural features in chemicals associated with cancer drug response: A

systematic data-driven analysis.

Suleiman Ali Khan[1], Seppo Virtanen[1], Olli Kallioniemi[2], Krister Wennerberg[2], Antti Poso[2,3] and Samuel Kaski[1,4].

[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, 00076 Espoo, Finland. [2]Institute for Molecular Medicine Finland FIMM, University of Helsinki, 00014 Helsinki, Finland. [3]School of Pharmacy, Faculty of Health Sciences, University of Eastern Finland, 70211 Kuopio, Finland. [4]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland.

ABSTRACT

Motivation: Analysis of relationships of drug structure to biological response is key to understanding off-target and unexpected drug effects, and for developing hypotheses on how to tailor drug therapies. New methods are required for integrated analyses of a large number of chemical features of drugs against the corresponding genome-wide responses of multiple cell models.

Results: In this paper, we present the first comprehensive multi-set analysis on how the chemical structure of drugs impacts on genome-wide gene expression across several cancer cell lines (CMap database). The task is formulated as searching for drug response components across multiple cancers to reveal shared effects of drugs and the chemical features that may be responsible. The components can be computed with an extension of a very recent approach called Group Factor Analysis (GFA). We identify 11 components that link the structural descriptors of drugs with specific gene expression responses observed in the three cell lines, and identify structural groups that may be responsible for the responses. Our method quantitatively outperforms the limited earlier methods on CMap and identifies both the previously reported associations and several interesting novel findings, by taking into account multiple cell lines and advanced 3D structural descriptors. The novel observations

include: previously unknown similarities in the effects induced by 15-delta prostaglandin J2 and HSP90 inhibitors, which are linked to the 3D descriptors of the drugs; and the induction by simvastatin of leukemia-specific response, resembling the effects of corticosteroids.

Availability: Code http://research.ics.aalto.fi/mi/software/GFAsparse

Contact: suleiman.khan@aalto.fi, samuel.kaski@aalto.fi

Supplementary Information: Available at *Bioinformatics* online

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

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

Area F: Evolution and Population Genomics

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

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

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

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

Area G: Bioinformatics of Health and Disease

OncodriveROLE classifies cancer driver genes in Loss of Function and Activating mode of action.

Michael P Schroeder[1], Carlota Rubio-Perez[1], David Tamborero[1], Abel Gonzalez-Perez[1,*] and Nuria Lopez-Bigas [1,2]

[1] Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Dr. Aiguader 88, Barcelona, Spain. [2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys, 23, Barcelona, Spain.

ABSTRACT

Motivation: Several computational methods have been developed to identify cancer drivers genes – genes responsible for cancer development upon specific alterations. These alterations can cause the loss of function of the gene product, for instance in tumor suppressors, or increase or change its activity or function, if it is an oncogene. Distinguishing between these two classes is important to understand tumorigenesis in patients and has implications for therapy decision making. Here, we assess the capacity of multiple gene features related to the pattern of genomic alterations across tumors to distinguish between activating and loss of function cancer

genes and we present an automated approach to aid the classification of novel cancer drivers according to their role.
Result: OncodriveROLE is a machine learning-based approach that classifies driver genes according to their role, using several properties related to the pattern of alterations across tumors. The method shows an accuracy of 0.93 and Matthew's Correlation Coefficient of 0.84 classifying genes in the Cancer Gene Census. The OncodriveROLE classifier, its results when applied to two list of predicted cancer drivers and TCGA-derived mutation and copy number features used by the classifier are available at http://bg.upf.edu/oncodrive-role.
Contact: abel.gonzalez@upf.edu , nuria.lopez@upf.edu

Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning.

Mehmet Gönen and Adam A. Margolin.
Sage Bionetworks, 1100 Fairview Avenue North, Seattle, WA 98109, USA. Present address: Department of Biomedical Engineering, Oregon Health & Science University, 3303 SW Bond Avenue, Portland, OR 97239, USA.

ABSTRACT
Motivation: Human immunodeficiency virus (HIV) and cancer require personalized therapies due to their inherent heterogeneous nature. For both diseases, large-scale pharmacogenomic screens of molecularly characterized samples have been generated with the hope of identifying genetic predictors of drug susceptibility. Thus, computational algorithms capable of inferring robust predictors of drug responses from genomic information are of great practical importance. Most of the existing computational studies that consider drug susceptibility prediction against a panel of drugs formulate a separate learning problem for each drug, which cannot make use of commonalities between subsets of drugs.
Results: In this study, we propose to solve the problem of drug susceptibility prediction against a panel of drugs in a multi-task learning framework by formulating a novel Bayesian algorithm that combines kernel-based nonlinear dimensionality reduction and binary classification (or regression). The main novelty of our method is the joint Bayesian formulation of projecting data points into a shared subspace and learning predictive models for all drugs in this subspace, which helps us to eliminate off-target effects and drug-specific experimental noise. Another novelty of our method is the ability of handling missing phenotype values due to experimental conditions and quality control reasons. We demonstrate the performance of our algorithm via cross-validation experiments on two benchmark drug susceptibility datasets of HIV and cancer. Our method obtains statistically significantly better predictive performance on most of the drugs compared to baseline single-task algorithms that learn drug-specific models. These results show that predicting drug susceptibility against a panel of drugs simultaneously within a multi-task learning framework improves overall predictive performance over single-task learning approaches.
Availability: Our Matlab implementations for binary classification and regression are available at https://github.com/mehmetgonen/kbmtl .
Contact: mehmet.gonen@sagebase.org

Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm.

Jingwen Yan[1,2], Lei Du[2], Sungeun Kim[2], Shannon Risacher[2], Heng Huang[3], Jason Moore[4], Andrew Saykin[2] and Li Shen[2], and the Alzheimer's Disease Neuroimaging Initiative[§].
[1]BioHealth, Indiana University School of Informatics & Computing, Indianapolis, IN, 46202, USA.
[2]Radiology & Imaging Sciences, Indiana University Sch. of Medicine, Indianapolis, IN, 46202, USA.
[3]Computer Science & Engineering, The University of Texas at Arlington, TX, 76019, USA. [4]Genetics, Community & Family Medicine, Dartmouth Medical School, Lebanon, NH, 03756, USA. [§]A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how to apply/ADNI Acknowledgement List.pdf.

ABSTRACT
Motivation: Imaging genetics is an emerging field that studies the influence of genetic variation on brain structure and function. The major task is to examine the association between genetic markers such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from neuroimaging data. The complexity of these data sets have presented critical bioinformatics challenges that require new enabling tools. Sparse canonical correlation analysis (SCCA) is a bi-multivariate technique used in imaging genetics to identify complex multi-SNP-multi-QT associations. However, most of the existing SCCA algorithms are designed using the soft thresholding method, which assumes that the input features are independent from one another. This assumption clearly does not hold for the imaging genetic data. In this paper, we propose a new knowledge-guided SCCA algorithm (KG-SCCA) to overcome this limitation as well as improve learning results by incorporating valuable prior knowledge.
Results: The proposed KG-SCCA method is able to model two types of prior knowledge: one as a group structure (e.g., linkage disequilibrium blocks among SNPs) and the other as a network structure (e.g., gene co-

expression network among brain regions). The new model incorporates these prior structures by introducing new regularization terms to encourage weight similarity between grouped or connected features. A new algorithm is designed to solve the KG-SCCA model without imposing the independence constraint on the input features. We demonstrate the effectiveness of our algorithm with both synthetic and real data. For real data, using an Alzheimer's disease (AD) cohort, we examine the imaging genetic associations between all SNPs in the *APOE* gene (i.e., top AD gene) and amyloid deposition measures among cortical regions (i.e., a major AD hallmark). In comparison with a widely used SCCA implementation, our KG-SCCA algorithm produces not only improved cross-validation performances but also biologically meaningful results.
Availability: Software is freely available upon request.
Contact: shenli@iu.edu

ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples.

Rui Tian[1], Malay Basu[1,2] and Emidio Capriotti[1,2,3].
[1]Division of Informatics, Department of Pathology, University of Alabama at Birmingham, 619 19th St. South, 35249 Birmingham, AL, USA. [2]Department of Clinical and Diagnostic Sciences, University of Alabama at Birmingham, 1705 University Boulevard, 35249 Birmingham, AL, USA. [3]Department of Biomedical Engineering, University of Alabama at Birmingham, 1075 13th Street South, 35249 Birmingham, AL, USA.
ABSTRACT
Motivation: The recent advance in high-throughput sequencing technologies is generating a huge amount of data that are becoming an important resource for deciphering the genotype underlying a given phenotype. Genome sequencing has been extensively applied to the study of the cancer genomes. Although a few methods have been already proposed for the detection of cancer-related genes, their automatic identification is still a challenging task. Using the genomic data made available by The Cancer Genome Atlas Consortium (TCGA), we propose a new prioritization approach based on the analysis of the distribution of putative deleterious variants in a large cohort of cancer samples.
Results: In this paper, we present ContastRank, a new method for the prioritization of putative impaired genes in cancer. The method is based on the comparison of the putative defective rate of each gene in tumor versus normal and 1000 genome samples. We show that the method is able to provide a ranked list of putative impaired
genes for colon, lung and prostate adenocarcinomas. The list significantly overlaps with the list of known cancer driver genes previously published. More importantly, by using our scoring approach, we can successfully discriminate between TCGA normal and tumor samples. A binary classifier based on ContrastRank score reaches an overall accuracy higher than 90% and the Area Under the Curve (AUC) of Receiver Operating Characteristics (ROC) higher than 0.95 for all the three types of adenocarcinoma analysed in this paper. In addition, using ContrastRank score we are able to discriminate the three tumor types with a minimum overall accuracy of 77% and AUC of 0.83.
Conclusions: We describe ContrastRank, a method for prioritizing putative impaired genes in cancer. The method is based on the comparison of exome sequencing data from different cohorts and can detect putative cancer driver genes. ContrastRank can also be used to estimate a global score for an individual genome about the risk of adenocarcinoma based on the genetic variants information from a whole-exome VCF (Variant Calling Format) file. We believe that the application of ContrastRank can be an important step in genomic medicine to enable genome-based diagnosis.
Availability: The lists of ContrastRank scores of all genes in each tumor type are available as supplementary materials. A webserver for evaluating the risk of the three studied adenocarcinomas starting from whole-exome VCF file is under development.
Contact: emidio@uab.edu

Area H: Biological Knowledge Discovery from Data, Text and Bio-images

Unveiling new biological relationships using shared hits of chemical screening assay pairs.

Xueping Liu[1,2] and Monica Campillos[1,2].
[1]Institute of Bioinformatics and Systems Biology and [2]German Center for Diabetes Research, Helmholtz Center Munich, 85764, Neuherberg, Germany.

ABSTRACT
Motivation: Although the integration and analysis of the activity of small molecules across multiple chemical screens is a common approach to determine the specificity and toxicity of hits, the suitability of these approaches to reveal novel biological information is less explored. Here, we test the hypothesis that assays sharing selective hits are biologically related.
Results: We annotated the biological activities (i.e. biological processes or molecular activities) measured in assays and constructed chemical hit profiles with sets of compounds differing on their selectivity level for 1,640

assays of ChemBank repository. We compared the similarity of chemical hit profiles of pairs of assays with their
biological relationships and observed that assay pairs sharing nonpromiscuous chemical hits tend to be biologically related. A detailed analysis of a network containing assay pairs with the highest hit similarity confirmed biological meaningful relationships. Furthermore, the biological roles of predicted molecular targets of the shared hits reinforced the biological associations between assay pairs.
Contact: monica.campillos@helmholtz-muenchen.de

Large-Scale Automated Identification of Mouse Brain Cells in Confocal Light Sheet Microscopy Images.

Paolo Frasconi[1], Ludovico Silvestri[2], Paolo Soda[3], Roberto Cortini[1], Francesco Pavone[2] and Giulio Iannello[3].
[1]Department of Information Engineering (DINFO), Università di Firenze, Italy. [2]European Laboratory for Nonlinear Spectroscopy (LENS), Università di Firenze, Italy. [3]Integrated Research Centre, Università Campus Bio-Medico di Roma, Italy.

ABSTRACT
Motivation: Recently, confocal light sheet microscopy has enabled high-throughput acquisition of whole mouse brain 3D images at the micron scale resolution. This poses the unprecedented challenge of creating accurate digital maps of the whole set of cells in a brain.
Results: We introduce a fast and scalable algorithm for fully automated cell identification. We obtained the whole digital map of Purkinje cells in mouse cerebellum consisting of a set of 3D cell center coordinates. The method is very accurate and we estimated an F_1 measure of 0.96 using 56 representative volumes, totaling 1.09 GVoxel and containing 4,138 manually annotated soma centers.
Availability and implementation: Source code and its documentation are available at http://bcfind.dinfo.unifi.it/. The whole pipeline of methods is implemented in Python and makes use of ylearn2 (Goodfellow et al., 2013) and modified parts of Scikitlearn (Pedregosa et al., 2011). Brain images are available on request.
Contact: paolo.frasconi@unifi.it
Supplementary information: Coordinates of predicted soma centers of a whole mouse cerebellum and additional figures.

Integration of molecular network data reconstructs Gene Ontology.

Vladimir Gligorijevic, Vuk Janjic and Natasa Przulj.
Department of Computing, Imperial College London, SW7 2AZ, UK.

ABSTRACT
Motivation: Recently, a shift was made from using Gene Ontology (GO) to evaluate molecular network data to using these data to construct and evaluate GO: Dutkowski et al. [2013] provide the first evidence that a large part of GO can be reconstructed solely from topologies of molecular networks. Motivated by this work, we develop
a novel data integration framework that integrates multiple types of molecular network data to reconstruct and update GO. We ask how much of GO can be recovered by integrating various molecular interaction data.
Results: We introduce a computational framework for integration of various biological networks using Penalized Non-negative Matrix Tri-Factorization (PNMTF). It takes all network data in a matrix form and performs simultaneous clustering of genes and GO terms, inducing new relations between genes and GO terms (annotations) and between GO terms themselves. To improve the accuracy of our predicted relations, we extend the integration methodology to include additional topological information represented as the similarity in wiring around non-interacting genes. Surprisingly, by integrating topologies of bakers yeasts protein-protein interaction, genetic interaction and co-expression networks, our method reports as related 96% of GO terms that are directly
related in GO. The inclusion of the wiring similarity of non-interacting genes contributes 6% to this large GO-term association capture. Furthermore, we use our method to infer new relationships between GO terms solely from the topologies of these networks and validate 44% of our predictions in the literature. In addition, our integration
method reproduces 48% of cellular component, 41% of molecular function and 41% of biological process GO terms, outperforming the previous method in the former two domains of GO. Finally, we predict new GO annotations of yeast genes and validate our predictions through genetic interactions profiling.
Supplementary information: Supplementary Tables of new GO term associations and predicted gene annotations are available at: http://bio-nets.doc.ic.ac.uk/GO-Reconstruction/ .
Contact: natasha@imperial.ac.uk

Extracting patterns of database and software usage from the bioinformatics literature.

Geraint Duck[1], Goran Nenadic[1,2], Andy Brass[1,3], David Robertson[3] and Robert Stevens[1].

The University of Manchester, United Kingdom.
[1]School of Computer Science, University of Manchester, UK. [2]Manchester Institute of Biotechnology, University of Manchester, UK. [3]Computational and Evolutionary Biology, Faculty of Life Sciences, University of Manchester, UK.

ABSTRACT

Motivation: As a natural consequence of being a computer-based discipline, bioinformatics has a strong focus on database and software development, but the volume and variety of resources are growing at unprecedented rates. An audit of database and software usage patterns could help provide an overview of developments in bioinformatics and community common practice, and comparing the links between resources through time could demonstrate both the persistence of existing software and the emergence of new tools.

Results: We study the connections between bioinformatics resources and construct networks of database and software usage patterns, based on resource co-occurrence, that correspond to snapshots of common practice in the bioinformatics community. We apply our approach to pairings of phylogenetics software reported in the literature, and argue that these could provide a stepping-stone into the identification of scientific best practice.

Availability: The extracted resource data, the scripts used for network generation and the resulting networks are available at: http://bionerds.sourceforge.net/networks/

Contact: robert.stevens@manchester.ac.uk

The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective.

Yuxiang Jiang[1], Wyatt Clark[1], Iddo Friedberg[2,3] and Predrag Radivojac[1].
[1]Department of Computer Science and Informatics, Indiana University, Bloomington, Indiana, USA.
[2]Department of Microbiology, Miami University, Oxford, Ohio, USA. [3]Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, USA.

ABSTRACT

Motivation: The automated functional annotation of biological macro-molecules is a problem of computational assignment of biological concepts or ontological terms to genes and gene products. A number of methods have been developed to computationally annotate genes using standardized nomenclature such as Gene Ontology (GO). However, questions remain about the possibility for development of accurate methods that can integrate disparate molecular data as well as about an unbiased evaluation of these methods. One important concern is that experimental annotations of proteins are incomplete. This raises questions as to whether and to what degree
currently available data can be reliably used to train computational models and estimate their performance accuracy.

Results: We study the effect of incomplete experimental annotations on the reliability of performance evaluation in protein function prediction. Using the structured-output learning framework, we provide theoretical analyses and carry out simulations to characterize the effect of growing experimental annotations on the correctness and stability of performance estimates corresponding to different types of methods. We then analyze real biological data by simulating the prediction, evaluation, and subsequent re-evaluation (after additional experimental annotations become available) of GO term predictions. Our results agree with previous observations that incomplete and accumulating experimental annotations have the potential to significantly impact accuracy assessments. We find that their influence reflects a complex interplay between the prediction algorithm, performance metric, and underlying ontology. However, using the available experimental data and under realistic assumptions, our results also suggest that current large-scale evaluations are meaningful and almost surprisingly reliable.

Contact: predrag@indiana.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Area J: Methods and Technologies for Computational Biology

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 cancerdriver 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

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

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

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