SBW

swedish bioinformatics workshop

2018 ÖREBRO

25–26 October 2018
Aula Nova, Örebro University

ÖREBRO UNIVERSITY

PROGRAM

THURSDAY, 25 OCTOBER

09.00–10.00	Registration
10.00–12.00	GSMM Tutorial
	Adil Mardinoglu *"Generation of human cell/tissue specific GEMs"*
	Partho Sen *"Hands-on-lab: Genome-Scale metabolomic modeling"*
12.00–13.30	Lunch
12.30–13.30	MedBioInfo internal meeting
13.30–15.00	Georg Fuellen *"Bioinformatics of Ageing"*
	Anne-Laure Boulesteix *"A plea against fishing for significance and for open science with focus on omics data analysis"*
15.00–15.30	Coffee (Setting up poster latest)
15.30–16.30	Jon Robinson *"Integrative systems biology through genome-scale metabolic models"*
	Dieter Maier *"Knowledge Management at the Proms – surviving the tooth and claw world of commercial bioinformatics, and having fun"*
16.30–18.00	Poster session I
19.30	Dinner (Örebro Castle)

FRIDAY, 26 OCTOBER

08.00–10.00	Fredrik Boulund *"Analyzing human microbiomes at the Center for Translational Microbiome Research"*
	Stephen Rush *"Capturing context-specific regulation in molecular interaction networks"*
	Mika Gustafson *"Deep auto-encoders for the identification of disease modules"*
10.00–10.30	Coffee
10.30–11.30	Cecilia Engel Thomas *"Multiomics signatures predictive of early diabetes remission following bariatric surgery: A Direct study"*
	Ola Spjuth *"Towards intelligent drug discovery safety testing"*
11.30–12.00	Poster session II and refreshments
12.00–13.00	Arne Elofsson *"Using deep learning and direct coupling analysis for protein structure prediction"*
	Pär Engström *"National Bioinformatics Infrastructure Sweden (NBIS)"*
13.00–14.00	Lunch
14.00–15.00	Career Panel Discussion organised by RSG-Sweden Join us for a panel discussion on essential skills and networking advice for students and early-career researchers.
15:00	Closing

GOOD TO KNOW DURING THE CONFERENCE

Questions during the conference

- Registration desk in the Nova Building Lounge
- E-mail to konferensinfo@oru.se
- Telephone +46 701 88 50 33

Conference buses

From Nova Building to Clarion Hotel and Scandic Grand Hotel
Thursday 25 October at 18:10

From Scandic Grand Hotel through Clarion Hotel
Friday 26 October at 07:30

From Nova Building to Central station
Friday 26 October at 15:10

Taxi

Örebro läns taxi +46 19 124 300

Wifi

Guestnet
Username: SBW2018
Password: SBW2018

Conference web site

https://www.oru.se/SBW2018/

ABSTRACT BOOK

The Swedish Bioinformatics workshop
25-26 October 2018 Örebro

Title
PΨfinder: Identification of novel PΨ in DNA sequencing data

Authors (presenting author underlined)
Sanna Abrahamsson[1], Anna Rohlin[2], Frida Eiengård[2], Marcela Davila Lopez[1]

Affiliations
[1]Bioinformatics Core Facility, Gothenburg, Sweden; [2]Department of Molecular and Clinical Genetics, Gothenburg, Sweden

Email (presenting author): sanna.abrahamsson@gu.se

Abstract text

Pseudogenes, are typically defined as non-functional genomic sequences derived from functional genes. This non-functionality is often caused by the lack of functional promoters or regulatory elements, accumulating frame disruptions such as frameshifts or in-frame stop codons.

Processed pseudogenes, one of the main three groups of pseudogenes, are created by retro transposition of mRNA from functional protein-coding loci back into the genome. The formation of processed pseudogenes has shown to represent a new class of mutation occurring during cancer development. Typically they are characterized by a complete lack of introns, the presence of small flanking direct repeats, and a polyadenine tail near the 3′-end.

There have been several efforts to identify and characterize these pseudogenes for the entire human genome. These approaches relay on locus-specific transcription evidence and high throughput sequencing data.

Here we present an automatic pipeline that uses targeted sequencing data to identify processed pseudogenes as a complement to SNP analysis. The output of this method encompasses a list of candidate pseudogenes and visualization plots together with an html report summarizing the findings. Currently we are screening 120 samples from hereditary colorectal cancer. We have identified several new processed pseudogenes and its validations is ongoing.

The Swedish Bioinformatics workshop
25-26 October 2018 Örebro

Title

In silico Analysis for Comparative Binding of Human Frizzled-6 with WNT Family Members to Explore Hereditary Nail Dysplasia

Authors

Muhammad Muzammal Adeel[1,2], Sajid Rashid.[1]

Affiliations
[1]Qauid-i-Azam university, Islamabad, Pakistan; [2]College of informatics HZAU, Wuhan , China.

Email:
m.muzammal.adeel@outlook.com

Abstract text

Hereditary nail dysplasia is rare and belongs to heterogeneous group of developmental abnormalities including ectodermal dysplasia. It may occur as isolated and/or syndromic ectodermal conditions. Worldwide, prevalence rate of isolated nail dysplasia is 1/10,000 to 1/1,000,000. Nail dysplasia caused by mutations in WNTS and its receptor FZD6, several mutations in WNTs and FZD6 have been reported responsible for ectodermal appendages disorders like nail dysplasia. This is because WNTs are involved in critical processes like regulation of the cell cycle, apoptosis and angiogenesis, in addition to embryonic development, growth and survival, all of which are hallmarks of nail dysplasia. The current research was based upon in silico analysis of comparative binding mutational analysis of WNTs, in order to characterize its nail disorders causing abilities. For this purpose, 4 reported substitutions of WNTs (WNT2, WNT2B, WNT10A and WNT10B) were extracted, the structures of FZD6-CRD and WNTs (wild type and mutated WNTs) were modeled and their binding was examined and compared with binding of normal WNT-FZD6-CRD complex. It was seen that for all the nail dysplasia causing mutations, WNT-FZD6-CRD bind differently as compared to normal binding. The present results emphasize the effect of mutations on WNT-FZD6 binding which lead to disruption of WNT-FZD6 signaling pathway in nail development causing nail dysplasia. Additionally, our research also suggests the evolutionarily conserved superimposed residues, which could probably define the major hallmark of the WNT signaling pathway. Furthermore, these residues could also be used to design pharmacophore model of novel drugs for different diseases that are linked to WNT-FZD signaling pathway, especially nail dysplasia

Title

Identification and reconstruction of novel antibiotic resistance genes from metagenomes

Authors (presenting author underlined)
Fanny Berglund[1,2], Tobias Österlund[1,2], Fredrik Boulund[3], Nachiket P. Marathe[2,4], Carl-Fredrik Flach[2,4], D. G. Joakim Larsson[2,4], Erik Kristiansson[1,2]

Affiliations
[1]Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden.
[2]Centre for Antibiotic Resistance Research (CARe), University of Gothenburg
[3]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden
[4]Department of Infectious Diseases, Institute of Biomedicine, the Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.

Email (presenting author):
fannyb@chalmers.se

Abstract text

Infections caused by antibiotic resistant bacteria are increasing globally, providing a major threat to public health. New antibiotic resistance genes are constantly discovered in pathogenic bacteria isolated in clinical settings and are hypothesized to originate from commensal and environmental bacteria. Identification of new antibiotic resistance genes is therefore of great importance in order to keep them under surveillance and to understand how they evolve, mobilize and spread to pathogens. Metagenomics enables cultivation-independent characterization of bacterial communities, but the resulting data is noisy and highly fragmented, making the assembly and identification of previously unknown antibiotic resistance genes challenging. This has so far limited the scope when studying the diversity of ARGs and their dissemination routes towards pathogens. We have therefore developed fARGene, a method for identification and reconstruction of ARGs from shotgun metagenomic data. fARGene uses optimized gene models and can therefore identify previously uncharacterized ARGs with high sensitivity and specificity. By performing the analysis directly on the metagenomic fragments, fARGene circumvents the need for a high-quality assembly. The applicability of fARGene was demonstrated by reconstruction of β-lactamases from more than five billion metagenomic reads, resulting in 221 full-length ARGs whereof 102 were previously undescribed. Based on 38 ARGs reconstructed by fARGene, experimental verification showed that 81% provided a resistance phenotype in *Escherichia coli*. fARGene furthermore includes the capability to create and optimize customized gene models, enabling the identification of any well-defined class of antibiotic resistance genes. We conclude that fARGene provides an efficient and reliable way to explore the unknown resistome in bacterial communities. The method is freely available via GitHub (https://github.com/fannyhb/fargene) under the MIT license.

Title
Functional Analysis of Circulating microRNAs in Pancreatic Cancer

Authors
Emmy Borgmästars[1], Hendrik de Weerd[2], Zelmina Lubovac-Pilav[2] Malin Sund[1]

Affiliations
[1]Department of Surgical and Perioperative Sciences, Umeå University, Umeå, Sweden; [2]Department of Systems Biology, University of Skövde, Skövde, Sweden

Email: emmy.borgmastars@umu.se

Abstract

MicroRNAs (miRNAs) are small RNA species that regulate gene expression at a post-transcriptional level and are emerging as potentially important biomarkers for various disease states. To understand the underlying role of miRNAs, *in silico*-based functional analysis can be performed to enable discovery of novel molecular mechanisms that contribute to the pathogenesis of the disease and potential novel therapeutic targets. Functional analysis of miRNA consists of miRNA target prediction and functional enrichment analysis of miRNA target genes. MiRNA target predictions often generate many potential target genes and to validate all of these in an experimental setup is not possible. Hence, a validation approach is valuable to narrow down interesting candidate target genes. One method commonly used is to correlate miRNA expression to mRNA expression to assess the regulatory effect of a particular miRNA.

Finding novel non-invasive biomarkers for pancreatic cancer is highly desirable as the only clinically used biomarker today, carbohydrate antigen-19-9 (CA19-9), is not sensitive or specific enough. MiRNAs are emerging as potential biomarkers in pancreatic cancer and a dataset of 15 circulating miRNAs identified as differentially expressed in pancreatic cancer was used in this study. These 15 miRNAs were shown to outperform CA19-9 in terms of area under curve (AUC).

This study aimed to develop a bioinformatics pipeline in R that performs miRNA target prediction, functional enrichment analysis and *in silico* evaluation of predicted miRNA target genes by correlation analyses between miRNA expression levels and its targets on mRNA and protein expression levels. For miRNA target prediction, DIANA-TarBase v7, DIANA-microT-CDS and TargetScan v7.1 were downloaded and queried from the pipeline. MiRNA, mRNA and protein expression data were downloaded from the cancer genome atlas (TCGA) and the cancer proteome atlas (TCPA). The miRNA isoform expression data from the cancer genome atlas was utilized to obtain the specific expression pattern for each mature miRNA isoform. Fifteen significantly altered circulating miRNAs detected in pancreatic cancer patients were queried separately in the pipeline.

Predicted miRNA target genes, enriched gene ontology (GO) terms and Kyoto encyclopedia of genes and genomes (KEGG) pathways were generated for 15 miRNAs. Predicted miRNA targets were evaluated by correlation analyses between each miRNA and its target genes. miR-885-5p showed strong correlations (Pearson's correlation coefficient ≥ 0.7 or ≤ -0.7) to some of its target genes and might have a prognostic value in pancreatic cancer.

Title

A plea against fishing for significance and for open science with focus on omics data analysis

Authors (presenting author underlined)
Anne-Laure Boulesteix[1]

Affiliations
[1]Ludwig-Maximilians University, Munich, France

Email (presenting author): boulesteix@ibe.med.uni-muenchen.de

Abstract text

There are usually plenty of conceivable approaches to statistically analyze data that both make sense from a substantive point of view and are defensible from a theoretical perspective. The data analyst has to make a lot of choices, a problem sometimes referred to as “researcher’s degree of freedom”. This leaves much room for (conscious or subconscious) “fishing for nice results“: the researcher (data analyst) sometimes applies several analysis approaches successively and reports only the results that seem in some sense more satisfactory, for example in terms of statistical significance. This may lead to apparently interesting but false research findings that fail to get replicated in independent studies. This problem is enhanced in the so-called n<<p setting (i.e. when the number of considered features is much larger than the sample size), where results are particularly instable. In this talk I describe and illustrate these problems in the context of omics research and discuss possible strategies related to "open science" in a broad sense to (partially) address them. In particular, systematic validation strategies using independent data, the increased development of guidance documents, and the publication of negative research findings, analysis plans, data and code may be important steps towards more replicable research.

Title

TC-Hunter: Identifying insertion sites of a transgenic construct within its host

Authors

Marcela Dávila[1], Jelena Milosevic[2], John Inge Johnsen[2], Per Kogner[2]

Affiliations

[1]Bioinformatics Core Facility, Sahlgrenska Academy, University of Gothenburg, Sweden
[2]Insitutionen för kvinnors och barns hälsa, Karolinska Institutet, Sweden

Abstract

Genetically manipulated animal models are considered essential for studying gene functions in whole animals. Typically a construct, containing critical elements for gene expression (such as promoters, introns, the protein coding sequence of interest and a poly-A site), is microinjected into a host. It is common to evaluate the presence of the transgene by using Polymerase chain reaction (PCR) and Southern blotting.

Today, thanks to whole genome sequencing data we are now capable to identify the exact insertion

Title
MODifieR: An R package for disease module inference

Authors
Hendrik Arnold de Weerd[1], Tejaswi Badam[1,2], Mika Gustafsson[2], Zelmina Lubovac[1]

Affiliations
[1]School of bioscience, Systems biology research center, University of Skövde
[2] Linköping University, Bioinformatics, Department of Physics, Chemistry and Biology

Email: hendrik.arnold.de.weerd@his.se

Abstract

Cells are organized in a modular fashion, where essential functions are carried out by functional modules. Modules can be described as clusters of genes, gene products or metabolites that interact together, are co-regulated or physically interacting. Complex diseases rarely arise from a single causal factor but rather from multiple factors, with large individual variation. This leads to dysregulation of parts of functional modules and thereby gives rise to a disease phenotype. The underlying perturbation in parts of the functional modules and the connectivity between them makes up a disease module.

To better understand complex diseases it is crucial to identify disease modules. Genes present in the module might not have a significant impact on the disease on its own. However, the cumulative effect of multiple low penetrance genes could play a major role in the pathogenesis of complex diseases. Network-based approaches could be key in detecting these low penetrance genes, which could potentially be novel biomarkers or therapeutic targets.

Various disease module inference methods have been proposed earlier, using different approaches. Validation of disease modules is a challenging aspect of network medicine as complete disease modules are unknown. In this study, genomic concordance with GWAS studies has been used as a metric for validating disease modules. Preliminary results show that integrating the results of different inference methods lead to more robust results.

Here, MODifieR is presented, an R package that bundles 9 different disease module inference methods into a single package. The 9 methods methods can be classified into 3 different algorithmic classes: seed-based, clique based and co-expression based methods. MODifieR is available under the GNU GPL open source license and can be freely downloaded from https://gitlab.com/Gustafsson-lab/MODifieR.

Title
Multiomics signatures predictive of early diabetes remission following bariatric surgery: A DIRECT study

Authors
Cecilia Engel Thomas[1,2,3], Violeta Raverdy[4], Helle Krogh Pedersen[1], Ali Syed[2], Ana Viñuela[5,6,7], Anna Artati[8], Birgitte Nilsson[2], Caroline Brorsson[2], Cedric Howald[5,6,7], Christelle Stouder[5,6,7], Cornelia Prehn[8], Federico De Masi[2], Francesca Frau[9], Han Wu[10], Harald Grallert[11,12,13,14], Helene Verkindt[4], Johann Gassenhuber[9], Karina Banasik[1], Konstantinos Rouskas[4,15], Mark Farmen[10], Mark Haid[8], Mathias Gebauer[9], Mickaël Canouil[16], Mun-Gwan Hong[3], Peter Davidsen[1], Peter Løngreen[2], Ramneek Gupta[2], Robert Caiazzo[4], Sapna Sharma[11,12], Stephane Dupuis[16,5], Stéphane Lobbens[16], Valborg Gudmundsdottir[2], Bernd Jablonka[9], Emmanouil T. Dermitzakis[5,6,7], Ewan R. Pearson[17], Hartmut Ruetten[18], Jerzy Adamski[7,11], Jochen M. Schwenk[3], Melissa K. Thomas[10], Philippe Froguel[16,19], Søren Brunak[1,2 §], Francois Pattou[4 §] for the IMI DIRECT Consortium

Affiliations
[1] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2100 Copenhagen, Denmark.
[2] Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark.
[3] Affinity Proteomics, Science for Life Laboratory, School of Biotechnology, KTH - Royal Institute of Technology, Box 1031, SE-171 21 Solna, Sweden.
[4] Centre Hospitalier Régional Universitaire de Lille 2, F-59037 Lille Cedex, France.
[5] Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health, D-85764 Neuherberg, Germany.
[6] Department of Genetic Medicine and Development, University of Geneva Medical School, CH-1211 Geneva, Switzerland.
[7] Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva Medical School, CH-1211 Geneva, Switzerland.
[8] Swiss Institute of Bioinformatics, CH-1211 Geneva, Switzerland.
[9] R&D Translational Med. and Early Development, Biomarkers and Clinical Bioanalyses, Sanofi-Aventis Deutschland GmbH, Frankfurt am Main, Germany.
[10] Lilly Research Laboratories, Eli Lilly and Company, US-46285 Indianapolis, Indiana, USA.
[11] Research Unit of Molecular Epidemiology, Institute of Epidemiology II, Helmholtz Zentrum München, D-85764 Neuherberg, Germany.
[12] German Center for Diabetes Research (DZD), Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany.
[13] Clinical Cooperation Group Type 2 Diabetes. Helmholtz Zentrum München and Ludwig-Maximilians Universität München, D-85764 Neuherberg, Germany.
[14] Clinical Cooperation Group Nutrigenomics and Type 2 Diabetes, Helmholtz Zentrum München and Technische Universität München, D-85764 Neuherberg, Germany.
[15] Biomedical Sciences Research Center Alexander Fleming, 16672 Vari, Greece.
[16] Université de Lille, CNRS, Institut Pasteur de Lille, UMR 8199 - EGID, F-59000 Lille, France.
[17] Medical Research Institute, University of Dundee, DD1 4HN Dundee, UK.
[18] R&D Translational Med. and Early Clinical, Sanofi-Aventis, Deutschland GmbH, Industriepark Höchst, D-65926 Frankfurt am Main, Germany.
[19] Division of Population Health & Genomics, School of Medicine, University of Dundee, DD1 9SY Dundee, UK.

Email (presenting author): Cecilia.thomas@scilifelab.se

Abstract text

Bariatric surgery can be an effective treatment for type 2 diabetes and has gained increased interest in recent years. Nevertheless, not all patients experience improvements in glycemic control following surgery, and factors predicting diabetic remission remain elusive. Identification of such factors could facilitate patient stratification and optimize treatment decisions as well as lead to mechanistic insights into surgery-induced diabetes remission. In this study, we combine clinical patient information with the so far most diverse omics data types to rank molecular signatures that may play a role in remission and train artificial neural network models to stratify patients into likely remitters and non-remitters. The study includes obese (BMI > 35 kg/m^2) patients with preoperative type 2 diabetes undergoing bariatric surgery and is part of the IMI-DIRECT Consortium and derives from the Biological Atlas of Severe Obesity (ABOS). The study is presently the largest within this field when taking the number of patients (n=249) and omics data types (n=9) into account. The study consists of two independent cohorts. One cohort was used for feature selection and model construction while the other was kept aside for the final validation of the prediction models. We used forward-feature selection to identify seven clinical and 77 omics variables important for discrimination between remitters and non-remitters. We find that signatures from transcriptomics of multiple tissues (liver and visceral fat) and from targeted and untargeted metabolomics have especially high predictive value. Further, functional analysis of the selected omics signatures showed that they are interconnected in biological networks and convene on common pathways. Many of the identified omics signatures have known functions in glucose homeostasis and metabolism, but only few of them have previously been linked to diabetes remission following bariatric surgery.

Title
InVi: Integration and visualization of genomic data

Authors (presenting author underlined)
Luciano Fernandez-Ricaud[1] and
Marcela Dávila López[1]

Affiliations
[1]Bioinformatics Core Facility, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

Email (presenting author): luciano.fernandez@gu.se

Abstract text

Visualization of genomic data is vital to allow researchers to explore and understand their experimental data or large-scale datasets. A common process involves processing genomics files (E.g. NGS files) to be used by visualization software like Circos (http://circos.ca/). This involves the creation of complex configuration files amongst other steps. This task is time consuming, repetitive and requires constant input from the user to adjust the views to fit the demands.
We developed InVi an integration and visualization tool that aims at combining already established technologies to facilitate the process of producing circular visualizations using intuitive point and click. This tool allows for data queries and the filtering of functional genomics data, and the generation of graphical representations and tabular data displays with ease.
Additionally we developed CiGui which provides only a GUI for Circos without the database component.

The Swedish Bioinformatics workshop
25-26 October 2018 Örebro

ABSTRACT

Title
"Utilities for doing Bioinformatics analyses of Health, Ageing, Senescence and Rejuvenation"

Authors (presenting author underlined)
Georg Fuellen

Affiliations
Institute for Biostatistics and Informatics in Medicine and Ageing Research, University of Rostock, Germany

Email (presenting author):
fuellen@alum.mit.edu

Abstract text

Do we know what we are talking about? Given the variety in the usage of terms such as "biological age" or "biomarker of aging", and a lack of clean definitions, there may be doubts.
Thus, in the first part, following an introduction to health and aging research, an overview of work on a generic framework of definitions of important terms will be presented, including an operational definition of health. How can we systematically investigate the molecular underpinnings of health and cellular senescence? Systematic work could deliver a bird's-eye view of the molecular biology that may be most relevant for health, for undoing its deterioration and for precision prevention. Thus, in the second part, I will be presenting prototypical "pathway maps" of health and of cellular senescence, and showcase how these may be utilized.

Title
InterPep – Predicting Protein-Peptide interaction sites

Authors (presenting author underlined)
Isak Johansson-Åkhe[1], Claudio Mirabello[1], Björn Wallner[1]

Affiliations
[1]Linköpings Universitet, Linköping, Sverige

Email (presenting author): isak.johansson-akhe@liu.se

Abstract text

Protein-peptide interactions play an important role in major cellular processes, and are associated with several human diseases. To understand and potentially regulate these cellular function and diseases it is important to know the molecular details of the interactions. However, because of peptide flexibility and the transient nature of protein-peptide interactions, peptides are difficult to study experimentally. Thus, computational methods for predicting structural information about protein-peptide interactions are needed. We present InterPep, a pipeline for predicting protein-peptide interaction sites. It is a novel pipeline that, given a protein structure and a peptide sequence, utilizes structural template matches, sequence information, random forest machine learning, and hierarchical clustering to predict what region of the protein structure the peptide is most likely to bind. When tested on its ability to predict binding sites, InterPep successfully pinpointed 255 of 502 (50.7%) binding sites in experimentally determined structures at rank 1 and 348 of 502 (69.3%) among the top five predictions using only structures with no significant sequence similarity as templates. InterPep is a powerful tool for identifying peptide-binding sites; with a precision of 80% at a recall of 20% it should be an excellent starting point for docking protocols or experiments investigating peptide interactions.

Title

Novel transcriptome assembly method identifies novel MADS-box isoforms during early bud development in *Picea abies*

Authors (presenting author underlined)

Warren W. Kretzschmar[1], Shirin Akhter[2], Veronika Nordal[2], Nicolas Delhomme[2], Nathaniel R. Street[3], Ove Nilsson[2], Olof Emanuelsson[1] and Jens F. Sundström[2]

Affiliations

[1]KTH Royal Institute of Technology, Stockholm, Sweden; [2]Swedish University of Agricultural Sciences, Sweden; [3]Umeå University, Umeå, Sweden

Email (presenting author): warrenk@kth.se

Abstract text

The genomes and transcriptomes of several gymnosperm species are more complex in certain gene families compared to angiosperms. One such example is the SOC1 sub-clade of the MADS-box family of transcription factors. We have previously identified a member of this sub-clade, the conifer gene DEFICIENS AGAMOUS LIKE 19 (DAL19), as being specifically upregulated in cone-setting shoots.

We present abeona: a novel short-read transcriptome assembly method that combines naive De Bruijn graph traversal with kallisto to create a parsimonious set of transcript isoforms. We used this method to recapitulate DAL19 transcripts that we identified through Sanger sequencing of *Picea abies* and mapping to assembled conifer genomic sequences. Current transcriptome assembly methods did not properly assemble these transcripts from short reads. Furthermore, we applied our method to 42 putative MADS-box core regions in nine P. abies bud samples, from which we identified 1084 novel transcripts. We manually curated these transcripts to arrive at 933 potential MADS-box isoforms of 38 putative MADS-box sequences. 152 of these transcripts, which we assigned to 28 putative MADS-box genes, were differentially expressed across eight female, male, and vegetative buds. Finally, 16 out of 38 putative MADS-box sequences have PacBio IsoSeq circular consensus reads derived from pooled sample sequencing that only map to assembled transcripts associated with a single putative MADS-box sequence. Our method is highly sensitive to transcript isoforms of the MADS-box gene family, and it may provide new insight into other genes expressing many splicing variants.

ABSTRACT

Title
Resolving the weak mRNA-protein correlation by cell-type specific splice-variant models in human and mice

Authors (presenting author underlined)
Rasmus Magnusson[1], Olof Rundquist[1],
Mika Gustafsson[1]

Affiliations
[1]Department of physics, chemistry and biology, Linköping University, Sweden

Email (presenting author): rasmus.magnusson@liu.se

Abstract text

Almost all cellular functions are driven by proteins, and understanding their function has long been of high priority. Despite this profound importance of proteins, most studies rely on gene expression to draw conclusions on cellular status. Indeed, measuring RNA expression in order to unravel the status of biological systems is a fundamental part of biology research, and the focal point of several research fields. However, the correlation between gene expression and protein abundance is infamously poor, with a potentially profound impact on all conclusions drawn from mRNA expression.

A longstanding goal in bioinformatics has been to understand how gene expression should be analyzed, with Wilhelm et al., 201X, Nature and Fortelny et al., Nature brief communications arising, making clear marks in the scientific discussion. Nevertheless, the answer to the question “can we predict proteins from mRNA” remains inconclusive at the most.

Herein, we show that the prediction of protein levels from mRNA expression can indeed be performed by making use of gene splice variants. Instead of, as presented in Wilhelm et al, analyze multiple human cell types at one time we measured human T-cells undergoing polarization into Th1. We performed triplicates at 6 time points of mass spectrometry derived protein abundance and RNA-seq of gene expression, and found a median correlation of 0.21. We also observed a broader correlation profile for genes than what would be generated under the null hypothesis. In other words, there were enrichments both of genes correlating negative and positive with their protein expression.

These findings prompted us to construct a mathematical model for predicting protein abundance from splice variant expression. We observed that combining splice variants and allowing for a time delay from gene to proteinwhich increased the median correlation to 0.79 from 0.22. Next, we validated these findings by observing similar results in human Treg (Schmidt et al. BMC Biology 2018) and mouse B-cells. Lastly, we aimed to assert the biological importance of estimating cellular status via mRNA expression, and applied our model to RNAseq data of a set of case-control studies. Using this approach, we found an increase in differential expression among predicted proteins, as compared to the ingoing RNA-seq data.

Title
Bridging the gap between the proteome characteristics of Prokaryotes and Eukaryotes

Authors

Mahajan Mayank[1], Benjamin Yee[1], Emil Hägglund[1], Lionel Guy[1], John Fuerst[2] and Siv G. Andersson[1]

Affiliations
[1]Science for Life Laboratory, Uppsala University, Sweden; [2]School of chemistry and Molecular Biosciences, Queensland, Australia

Email (presenting author): mayank.mahajan@icm.uu.se

Abstract text

Studies have shown that protein lengths vary among the three domains of life and proteins from Prokaryotes tend to be only two thirds as long as those from Eukaryotes. Similar systematic variation has been observed in the proportion of disordered region per protein length among Eukaryotes and Prokaryotes. Eukaryotes contain a significantly larger proportion disordered region per protein length as compared to the Prokaryotes. Prokaryotes have smaller genome size, less number of activities in the cell and are generally less complex than the Eukaryotes. Thus, Prokaryotes do not need the complex and elaborate signalling and regulation systems as found in the Eukaryotes.

The basic structure of a protein can be described with the secondary structural elements such as helices, beta-sheets and the disordered segments that consist of turns, bends, coils and binding motifs. The Structural Classification of Proteins (SCOP) database released in Feb 2009 shows about 1200 unique structural elements (referred to as 'folds') in total, which make up a more or less comprehensive set of conserved structural units that are found in the proteins within the three domains of life. The regions in-between the conserved folds and at the termini of the protein are usually disordered and lack a fixed structure. The disordered regions in proteins have been observed to be dynamic in structure and have been referred to as fluid or fuzzy. Disordered regions help the protein fold and also makes it available for more interactions with other biomolecules.

We discovered that a few bacterial taxa have large genomes and large functional repertoire; moreover, they are comparable to some of the Eukaryotes in terms of proteome characteristics, and genomic complexity, thus, bridging the gap between Prokaryotes and Eukaryotes. Our results show an increase in protein interaction and signalling networks in these bacteria, therefore, we expect an increase in regulation and organization of cellular activities as compared the other Prokaryotes.

Proportion of the disordered region varies between different proteins and this variation seems to be somewhat correlated with the protein size and total genome size.

Our knowledge of life forms in hindered by our methods of search for new organism and it is believed that the diversity of life forms on earth is many folds larger than what we know of at the moment. Thus, we propose that many more of such exceptional cases exist among the prokaryotes.

The Swedish Bioinformatics workshop
25-26 October 2018 Örebro

ABSTRACT

Authors (presenting author underlined)
Dieter Maier

Affiliations

Email (presenting author):
dieter.maier@biomax.com

Abstract text

Biomax is a Munich based SME that, since 1997, provides services and computational solutions for better decision making and knowledge management in
the life sciences with a special focus on semantic knowledge representation, data integration, knowledge aggregation and machine learning.
Our staff therefore have a front seat on the galley's rowing-benches of the technological and knowledge advances in biology, from the transformation
of industrial chemical production by synthetic biology to the personalisation of clinical practice by Systems Medicine.
Within Healthcare the network-based bioinformatics frameworks advanced by Biomax allow solutions, which integrate electronic health records, clinical
and environmental data, experimental (omics) results and general knowledge on the affected morbidities to provide individually tailored care
suggestions, improve the quality of life for diseased patients and apply outcome based monitoring. Over the last 12 years in more than 10 public
research projects we developed an approach to Systems Medicine which today allows us to run a clinic in full and support its transformation towards
personalised, integrated care.

What is knowledge management about, how does that lead to applicable advances in a clinical setting and what is a "typical" bioinformaticians
contribution to that at Biomax? Just a few aspects to be covered by the presentation.

ABSTRACT

Authors (presenting author underlined)
Adil Mardinoglu

Affiliations
KTH-Royal Institute of Technology

Email (presenting author):
adilm@scilifelab.se

Abstract text

Altered metabolism is linked to the appearance of various human diseases and a better understanding of disease-associated metabolic changes may lead to the identification of novel prognostic biomarkers and the development of new therapies. Genome-scale metabolic models (GEMs) have been employed for studying human metabolism in a systematic manner, as well as for understanding complex human diseases. In the past decade, such metabolic models – one of the fundamental aspects of systems biology – have started contributing to the understanding of the mechanistic relationship between genotype and phenotype. In my presentation, I focus on the construction of the Human Metabolic Reaction database, the generation of healthy cell type- and cancer-specific GEMs using different procedures, and the potential applications of these developments in the study of human metabolism and in the identification of metabolic changes associated with various disorders. I further discuss how in silico genome-scale reconstructions can be employed to simulate metabolic flux distributions and how high-throughput omics data can be analyzed in a context-dependent fashion. Insights yielded from this mechanistic modeling approach can be used for identifying new therapeutic agents and drug targets as well as for the discovery of novel biomarkers. Finally, recent advancements in genome-scale modeling and the future challenge of developing a model of whole-body metabolism will be presented. The emergent contribution of GEMs to personalized and translational medicine will also be discussed.

The Swedish Bioinformatics workshop
25-26 October 2018 Örebro

ABSTRACT

Title
Combined Lipidomics and Informatics for Human Disease Research

Authors
Aidan McGlinchey[1], Dawei Geng[2], Olivier Govaere[3], Ann Daly[3], Tuulia Hyötyläinen[2], Quentin Anstee[3], Matej Oresic[1]

Affiliations
[1]School of Medical Sciences, Örebro University, 702 81 Örebro, Sweden; [2]Department of Chemistry, Örebro University, 702 81 Örebro, Sweden; [3]Newcastle University, Newcastle upon Tyne, NE2 4HH, UK

Email: Aidan.McGlinchey@oru.se

Abstract text

The rise of big data in human disease research, along with developments in machine learning, provides fertile ground for novel, powerful approaches for human disease research. Such methodologies open the door for the development of prognostic and diagnostic tools and techniques to improve intervention speed and efficacy. Here, we present an example in the form of preliminary data concerning a project dealing with nonalcoholic fatty liver disease (NAFLD). NAFLD is a major risk factor leading to chronic liver disease and type 2 diabetes. Non-invasive diagnostic techniques for the different stages of NAFLD, such as steatosis, nonalcoholic steatohepatitis (NASH) and fibrosis, are currently unavailable and thus are an unmet medical need. In our previous studies, we successfully identified specific serum molecular lipid signatures which associate with the amount of liver fat (1) as well as with NASH (2).

Here, we investigated serum lipidomic profiles in a clinical cohort (n = 688) in the European project Elucidating Pathways of Steatohepatitis (EPoS). The EPoS cohort comprised individuals at various stages of NAFLD (n = 666), including NASH (n = 661) and fibrosis (n = 511). In line with previous studies (1), we found that steatosis grade was strongly associated with the increase of certain triglycerides with low carbon number and double bond content as well as a decrease of specific phospholipids. As NAFLD progresses from an earlier steatosis state to a later, more severe fibrotic stage, fibrosis grades are also used as a clinical measure for assessing progression to and severity of NASH. Preliminary analysis of 511 of the cohort with graded presence of fibrosis versus those without, revealed that distinct relationships also exist between circulating lipids and fibrosis stage, the profile changing appreciably between steatosis and fibrosis.

In summary, our findings suggest that dysregulation of lipid metabolism in progressive stages of NAFLD is reflected in circulation and may thus hold diagnostic value as well as offer new insights about the NAFLD pathogenesis. Further analysis of these markers alone and in combination is warranted and is currently being undertaken to take these preliminary findings further with a view to assessing diagnostic utility of such markers.

(1) Orešič M, et al. Prediction of non-alcoholic fatty-liver disease and liver fat content by serum molecular lipids. Diabetologia. 2013 Oct;56(10):2266-74.
(2) Zhou Y, et al. Noninvasive detection of nonalcoholic steatohepatitis using clinical markers and circulating levels of lipids and metabolites. Clin Gastroenterol Hepatol. 2016 Oct;14(10):1463-1472.e6

ABSTRACT

Title
Deep learning using the Human Protein Atlas: discovery of biomarkers for prostate basal cells

Authors (presenting author underlined)
[1]Benedek Bozoky, [1]Ingemar Ernberg,
[1,2]Andrey Alexeyenko, [2]Laszlo
Szekely, [1,2]Iurii Petrov

Affiliations
[1]Karolinska Institutet, Stockholm, Sweden; [2]SciLifeLab, Stockholm, Sweden;

Email (presenting author):iurii.petrov@ki.se

Abstract text

Cancer research and treatment in the era of personalized cancer medicine increasingly demand prognostic and predictive markers. Prostate cancer is the most common cancer among men and is particularly challenging to differentiate between those patients who will die of the disease vs. those who will live with it. Prostate basal-cell specific biomarkers are already used as negative markers in diagnosis.
The Human Protein Atlas (HPA) online database is a rich source of potential biomarkers as it contains millions of images of immunohistochemically stained tissue microarrays. To identify new biomarkers we analyzed normal prostate images from HPA with a deep learning based classifier. We processed each image, performing colour deconvolution and binarization operations. The classifier performance (AUC) was estimated as 87%, verification on independent data. We identified over 40 specific biomarkers for prostate basal cells, including two of the routinely used biomarkers.
Our method can be used for biomarker discovery for any cell type or tissue structure that is predicted to give a specific staining pattern with immunohistochemistry. The new biomarkers for prostate basal cells could be used to clarify their functional role, and to help stratify prostate cancer patients. As predicted, the majority of our newly identified markers were absent in prostate cancers, but we identified a few that were positive in some of the cancers. These are promising markers for a subtype of prostate cancer.

ABSTRACT

Title:

High-resolution regulatory maps connect risk variants to cardiovascular disease related pathways

Authors Örjan Åkerborg[a], Rapolas Spalinskas[a], Sailendra Pradhananga[a]†, Anandashankar Anil[a], Pontus Höjer[a], Flore-Anne Poujade[b], Lasse Folkersen[c], Per Eriksson[b], Pelin Sahlén[a]

Affiliations

[a] Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, Division of Gene Technology, KTH Royal Institute of Technology, Solna, Sweden

[b] Cardiovascular Medicine Unit, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, Stockholm, Sweden

[c] Department of Bioinformatics, Technical University of Denmark, Copenhagen, Denmark

Email : sailendra.pradhananga@scilifelab.se

Genetic variant landscape of cardiovascular disease (CVD) is dominated by non-coding variants among which many occur within enhancers regulating the expression levels of relevant genes. It is crucial to assign them to their correct gene both to gain insights into perturbed functions and better assess the risk of disease. In this study, we generated high-resolution genomic interaction maps (~600 bases) in aortic endothelial, smooth muscle and THP-1 macrophages using Hi-C coupled with sequence capture targeting 25,429 features including variants associated with CVD. We detected interactions for 466 CVD risk variants obtained by genome-wide association studies (GWAS) and identified functions such as endothelial cilia assembly potentially perturbed in disease. Moreover, we were able to fine-map 589 GWAS variants using interaction networks, thereby identifying additional genes and functions.

The Swedish Bioinformatics workshop
25-26 October 2018 Örebro

ABSTRACT

Title
Epigenetic variations in precision cancer immunotherapy

Authors (presenting author underlined)
Qingyang Xiao[1], Pilar Piñeiro [2,] Isabel Barragan.[1,2]

Affiliations
[1]Karolinska Institute, Stockholm, Sweden; [2] Institute of Biomedical Research in Malaga, Malaga, Spain

Email (presenting author):xiao.qingyang@ki.se

Abstract text

BACKGROUND: anti-PD1 blockade has been increasingly applied in non-small-cell lung cancer (NSCLC) therapy. However, only a fraction of NSCLC patients respond to the therapy and no definite biomarkers are available to predict the clinical response. We explored DNA methylation profiles as biomarker candidates in PD1 blockade responders and non-responders with stage IV NSCLC.
METHODS: The strategy comprised filtering DNA methylation features correlated with clinical parameters in a discovery cohort, followed by validation in another two independent cohorts. A methylation signature (EPIMMUNE) and the methylation status of *FOXP1* were stablished as potential biomarkers in NSCLC patients treated with a PD1 inhibitor. Kaplan-Meier estimator was employed to analyse progression free survival (PFS) and overall survival (OS) according to the methylation signature and status. Statistical significance was calculated using the log-rank test, and Hazard ratios from univariate Cox regressions were used to determine the association between clinicopathological features with survival.
FINDINGS: The EPIMMUNE panel is significantly correlated with PFS in anti-PD1 treated NSCLC patients ((hazard ratio [HR] 0·010, 95% CI $3{\cdot}29 \times 10^{-4}$-0·0282; p=0·0067), as well as correlated with OS ((0·080, 0·017-0·373; p=0·0012). The panel also distinguished PFS in the validation cohort (0·330, 0·149-0·727; p=0·0064). In addition, *FOXP1* methylation status was correlated with longer PFS ((0·415, 0·209-0·802; p=0·0063) and OS (0·409, 0·220-0·780; p=0·0094).
Conclusion: The results demonstrated that epigenetic status is a predictive biomarker candidate for NSCLC patients receiving anti-PD1 blockade.

Title
Integrative systems biology through genome-scale metabolic models

Authors
Jonathan L. Robinson,[1] Jens Nielsen.[1]

Affiliations
[1]Dept. of Biology and Bioengineering, Chalmers University of Technology, Gothenburg, Sweden.

Email (presenting author): jonrob@chalmers.se

Abstract text

The advancement of high-throughput biological profiling technologies has transformed the way in which we can explore biological systems. Systems-level "-omics" data are being generated and disseminated at an increasing rate and with decreasing cost. However, the analysis and interpretation of such data in the context of the underlying biology is nontrivial, and often acts as a bottleneck. This bottleneck can be alleviated through the use of genome-scale metabolic models (GEMs), which serve as a scaffold for the integration and interpretation of large biological datasets. Here, we demonstrate the utility of GEMs by using the Human Metabolic Reaction (HMR) GEM to aid in the analysis of tumor mutation profiles from ~9,000 patients spanning 30 cancer types in the context of metabolism. Specifically, colorectal and gastric tumors were found to exhibit substantial metabolic reprogramming associated with mutation burden, which was largely focused in one-carbon metabolism and nucleotide biosynthesis.

In addition, we present a new approach for incorporating proteomic and enzyme kinetic data into a human GEM, thus constraining the flux solution space to a more physiologically relevant region. This integration improves the accuracy of cellular metabolic simulations without requiring nutrient uptake rates, which are often difficult or infeasible to quantify. Finally, we describe new methods for integrating single-cell RNA-sequencing profiles with GEMs, which seek to address the challenges of large variability and high drop-out rates often associated with these datasets. Overall, these approaches demonstrate the ongoing evolution of GEMs and their ability to serve as a central "hub" through which increasingly diverse biological datatypes can be integrated and analyzed.

Title
RSG Sweden - Developing community of computational biologists

Authors (presenting author underlined)
Martin Rydén [1,2], Nazeefa Fatima [1,2]

Affiliations
[1] RSG-Sweden; [2] Lund University, Lund, Sweden

Email (presenting author): rsg-sweden@iscbsc.org

Abstract text

The Regional Student Group for Sweden (RSG-Sweden) is a non-profit organisation for students and researchers in the field of computational biology. By promoting interactions among researchers from both academia and industry through networking and knowledge exchange activities, the RSG want to broaden participation in the Swedish STEM community through bioinformatics and computational biology.

RSG-Sweden will participate at the Swedish Bioinformatics Workshop by organising a panel discussion on the topic "Essential skills and networking advice for students", where the participants will discuss the experiences that have shaped their careers: what do they currently do, how did they get their current position, what skills they use from their degree. There will be time for the audience to ask questions and contribute to the discussion with comments.

Metabolic modelling on a genome-scale: A hands-on workshop to predict and understand the metabolic capabilities of organisms

Partho Sen

Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, FI-20520 Turku, Finland.

email: partho.sen@utu.fi

Abstract

Rapid advancement of cutting-edge technologies urges development of integrative methods and computational models. These approaches when applied at the systems level could mechanistically relate entities like gene, proteins and metabolites that might unveil the biological markers and related pathways at the systems level. Genome-scale metabolic modeling (GSMM) is a constraint-based mathematical modeling approach evolving over the past 30 years. GSMM integrates biochemical, genetic and genomic informations within a computational framework. Thus, can help to understand metabolic genotype-phenotype relationship of an organism. Today, GSMM have been used to study cell, tissue and organ specific metabolism in the context of various diseases such as cancer, non-alcoholic fatty liver disease (NAFLD), and diabetes. Recently, it has been used to study the metabolic role of gut microbiota and its association with host. Furthermore, GSMM as an integrative tool has been used to model diet-tissue and multi-tissue interactions in humans.

In this workshop, we will highlight several aspects of GSMM applied to microbes, lower eukaryotes and human. We will get accustom with the structure of genome-scale models, flux balance analysis (FBA), web-based[1] tools and standalone resources[2-4] for automatic/manual model reconstruction. We will apply GSMM to a simple model (microbe), and thereby ask several relevant biological questions. We will discuss, how these models could guide us to address these questions, and how to interpret such results. In this context, the genome-scale models (GEMs) will be contextualize and constrained for different conditions such as healthy and disease states, given omic(s) data. Furthermore, we will discuss, how GSMM aided to analyze human gut microbiome. The workshop aims to give a mechanistic overview of GSMM and its plethora of applications. It also aims to get accustom with the formulation, structure, quality of the GEMs, and integration of multiomics datasets suited for functional and downstream analysis. Moreover, it will inform about the limitations and proper usage of these complex models.

1. Arkin AP, Stevens RL, Cottingham RW, et al. The DOE Systems Biology Knowledgebase (KBase). bioRxiv 2016:096354.
2. Becker SA, Feist AM, Mo ML, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc 2007;2:727-38.
3. Agren R, Liu L, Shoaie S, et al. The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. PLoS Comput Biol 2013;9:e1002980.
4. Magnúsdóttir S, Heinken A, Kutt L, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nat Biotechnol 2016.

ABSTRACT

Title
Effects of Sequencing Depth in Human Gut Microbiome Analysis

Authors (presenting author underlined)
[1]Sukithar Rajan, [1]Mårten Lindqvist,[1]Ida Schoultz, [1]Robert Jan Brummer, [1]Drik Repsilber

Affiliations
[1]School of Medical Sciences, Örebro University, Sweden

Email (presenting author): sukithar.rajan@oru.se

Abstract text

Random Shotgun sequencing of DNA obtained directly from the environment has revealed profound microbial novelty and diversity. Culture-independent Next Generation Sequencing method presents an exciting means to elucidate microbial dynamics, which invariably determines the human health and global biogeochemical processes (1); making it an important tool for microbiome analysis regarding composition and function prediction. However, sequencing costs often set limits to the amount of sequences that can be generated and, consequently, the biological interpretation that can be achieved from any such attempt (2).

The aim of this study is to address the effects of varying sequencing depth on taxonomy classification and protein prediction using different methods. This study uses deep sequenced shotgun metagenome sequences collected from faecal sample. Metagenome sequences were subsampled to stimulate analyses at varying sequencing depth and taxonomically classified using three different classifiers; Metaphlan-II, Kraken, and Clark (3-5). Taxonomy information from this step was used in predicting function on the level of protein using PICRUSt (6).

Our study will help to (a) elucidate the impact of sequencing depth on taxonomy and protein prediction and (b) understand classification resolution of methods by comparing them at varying sequencing depths. The outcome of this study would serve as a reference for estimating an adequate sequencing depth and appropriate classifiers in understanding taxonomy composition of gut microbiota.

Reference:
1. Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, et al. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nat Rev Microbiol. 2015;13(6):360-72.
2. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121-32.

3. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16:236.
4. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9(8):811-4.
5. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46.
6. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31(9):814-21.

The Swedish Bioinformatics workshop
25-26 October 2018 Örebro

Title
A Deep Learning Model Based on Sparse Auto-encoder for Prioritizing Cancer-related Genes and Drug Target Combination

Authors
Muhammad Tahir ul Qamar, Ji-Wei Chang, Yuduan Ding, Yin Shen, Junxiang Gao, Ling-Ling Chen

Affiliations
[1]Huazhong Agricultural University, Wuhan, P.R. China

Email: m.tahirulqamar@webmail.hzau.edu.cn

Abstract text

Prioritization of cancer-related genes from gene expression profiles and proteomic data sets is vital to improve the targeted therapies research. In this current study, we introduced a deep learning model based on a sparse auto-encoder to learn the specific characteristics of protein interactions in cancer cell lines integrated with gene expression data. Our model outperformed differential expression and network-based methods in predicting cancer-related genes and their combination. Comparing to other reported methods, our model got high AUC value (>0.8). The model showed learning ability to identify cancer-related genes from the input of different gene expression profiles by extracting the characteristics of protein interaction information, which also could predict cancer-related gene combination. Our study prioritizes ~500 high confidence cancer related proteins, among these proteins 211 already known cancer drug targets were found, which supports the accuracy of our method. The above results indicated that the current auto-encoder model could computationally prioritize candidate genes involve in cancer and improve the targeted therapies research.

Title:
Omic Network Modules as tools for Personalized cancer chemotherapy in NSCLC

Authors (presenting author underlined)
Tejaswi Badam [1,2] , Mika Gustafsson[2], Zelmina Lubovac[1]

Affiliations
1.Department of Bioinformatics, University of Skövde , Skövde , Sweden
2.Institute for Physics, Chemistry & Biology / Bioinformatics , Linköping University , Linköping , Sweden

Email (presenting author):tejaswi.venkata.satya.badam@his.se

Abstract :

Non-small cell lung cancer (NSCLC) constitutes the most common type of lung cancer. Even though there are targeted chemotherapies existing for the treatment, the hospital admissions due to adverse drug reactions are on the rise. The range of individual therapeutic window varies due to high variability in drug response for the given fixed dose of the drug. Hence a need for the personalized cancer therapy achieved through synergistic integration of germ line genetics decreases the risk of severe ADR's .Powerful hypothesis generation from whole-genome sequencing analysis require synergistic and collective analysis of SNV using their combined effect, which could be performed using network and systems biology. we hypothesized that genetic variants in genes encoding proteins that regulate the elimination and distribution of drugs are likely to correlate with both drug exposure and the occurrence of ADR's. Omic data of SNVs was derived from whole genome-sequencing of 96 lung cancer patients treated with gemcitabine/carboplatin, where about 50 suffered from induced leukopenia (leu), thrombocytopenia (thr), and neutropenia (npk) respectively and they were mapped to their closest gene for the downstream analysis and thereby identified 896, 995 and 936 genes for npk, tpk and lpk. From these we constructed disease modules using clique-based clustering method for each of the traits and thereby identified gene modules of size 357, 320, and 347. Interestingly, 245 of those genes were shared across at least two modules, which we refer to as the *shared toxicity module*. Gene expression data of human cells from 300 microarrays treated with Carboplatin and Gemcitabine respectively filtered for bone marrow expression , which corresponded to 120 Carboplatin and 109 Gemcitabine genes. We then performed enrichment analysis of the expression gene set validation lists of the modules, which showed no enrichment for any of the trait lists, but significant enrichments ($P<0.05$) for each of the modules on both lists (odds ratio (OR) = 2.5- 3.2). However, the shared module showed higher significant overlaps than each of the traits.(OR=4.1- 4.5, P= 2.2-3.7 x 10^{-3}). The major significant pathways which could be looked upon effectively are Non-small cell lung cancer, RAS Calcium ,ErbB and estrogen signaling pathways. Most of the terms stated above are known to be affected in case of the K-RAS mutated or EGFR mutated NSCLC's. We found NSCLC genes to be highly enriched in the modules which suggested cancer genes highly interact with toxicity genes.

Title

Integration of Data Sources and Omics to Identify Susceptibility Variants for Bipolar Disorder

Authors (presenting author underlined)

K Truvé[1], D Vizlin-Hodzic[2], TZ Parris[3], S Illes[3,4,5], S Salmela[3], H Ågren[6], K Funa[3,7]

Affiliations

1 Bioinformatics Core Facility, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden. 2 Sahlgrenska University Hospital, Clinical Pathology and Genetics, Gothenburg. 3 Department of Oncology, Institute of Clinical Sciences, Sahlgrenska Cancer Center, Sahlgrenska Academy at University of Gothenburg. 4 Institute of Neuroscience and Physiology, Department of Physiology, Sahlgrenska Academy at University of Gothenburg. 5 Cellectricon AB, Mölndal, Sweden. 6 Institute of Neuroscience and Physiology, Section of Psychiatry and Neurochemistry, Sahlgrenska Academy at University of Gothenburg. 7 Sahlgrenska Cancer Center at the Sahlgrenska Academy, Institute of Biomedicine, University of Gothenburg, Gothenburg, Sweden.

Email (presenting author):
Katarina.truve@gu.se

Abstract text

Bipolar disorder (BD) is a severe chronic psychiatric disorder affecting >1% of the population worldwide. The disease is characterized by recurrent episodes of mania and depression. About 15% of patients with bipolar disorder are expected to die from suicide. Thus, early detection, diagnosis and initiation of correct treatment are critical.
In an earlier study aiming to identify shared mis-regulated genes or pathways, we combined induced pluripotent (iPSC) technology and neural stem cells with RNA-seq to investigate differences in the global transcriptome of neural stem cells (NSC) between BD patients and healthy controls.

We found the *NLRP2* gene to be the most significant differentially expressed gene with clear differences in expression for all cases and controls. To find the causative mutation explaining the upregulation of *NLRP2* and to explore and identify susceptibility variants for bipolar disorder, we performed Whole Genome Sequencing (WGS). We used several criteria to extract variants most likely to be involved in disease development, focussing on rare amino acid changing variants predicted to be deleterious with diverse bioinformatics tools. Identified variants were genotyped in a cohort of 47 cases and 44 controls. A dbGAP dataset (phs0008666.v1.p1) containing WGS data for six large families affected with BD was further used to evaluate variants in identified genes, taking into account family structure. The focus of our current research is to identify variants and genes that are involved in development of bipolar disorder (BD). Ultimately knowledge gained could be used to develop diagnostic tests, and improve current treatments.

Title
Miodin: software infrastructure for multi-omics data integration

Authors
Benjamin Ulfenborg[1]

Affiliations
[1]University of Skövde, Skövde, Sweden

Email benjamin.ulfenborg@his.se

Abstract text

Rapid developments of high-throughput biotechnology instruments have given us an unprecedented ability to measure molecular entities in biological systems. This opens up possibilities to investigate complex mechanisms that span multiple molecular layers, with the promise to unlock novel biological insights. Studies on multiple modalities of omics data are therefore growing in popularity, and it is recognized that progress in personalized medicine and many other fields depends on our ability to go from multi-omics data to understanding of processes underlying health and disease. Several bioinformatics tools have appeared that can analyze different types of omics data, but data integration remains a key challenge since extensive technical expertise is required to combine different tools or packages into a coherent analysis pipeline. To address this challenge, and promote streamlined, transparent and reproducible biomedical data science, the *miodin* R package was developed. The package allows users to integrate data from multiple omics modalities on the same samples, or across studies on the same variables. Data analysis with *miodin* is performed by declaring a study design and executing a workflow of sequentially connected modules. The study design allows the user to define all information required for data analysis (such as sample characteristics and assay data files) in one place, reducing the risk for clerical errors in the analysis script. The package provides a streamlined syntax for building workflows and modules can easily be combined using the pipe operator, improving readability of the script. Workflows promote transparency by automatically documenting processing and analysis steps carried out, and provided that the input data and parameters are the same, the same results will be generated. This makes the workflow-based analysis reproducible on different systems. Workflows currently support analysis of data from microarrays, sequencing and mass spectrometry. Supported omics modalities include transcriptomics, genomics, epigenomics and proteomics. The *miodin* package with extensive documentation is freely available on GitLab (https://gitlab.com/algoromics/miodin) under the GPL-3 license.

Title
Multiple testing corrections in a data driven age

Authors (presenting author underlined)
Anders Wirén[1]

Affiliations
[1]Clinical Research Centre, Region Örebro county, Örebro, Sweden

Email (anders.wiren@regionorebrolan.se):

Abstract text

The hypothesis driven approach to science that now is traditional in many fields of research relies on assessments, statistical tests, of the likelihood of seeing a certain research result by pure chance. A principle in probability theory is that, the more independent statistical tests we do at the same time, the larger is the likelihood of seeing false positive results/false discoveries. This increased likelihood is commonly managed by the application of multiple testing corrections, an upwards adjustment of p-values based on the number of independent tests made at the same time.

In data driven science, such as genomics and transcriptomics, we frequently work with thousands or tens of thousands of variables (e.g. read counts for mRNAs derived from tissue samples of some organism) and try to detect differences in these variables between groups. The vast number of independent tests involved in this kind of analysis – for good or bad – leads to sometimes-drastic reductions in the number of tests that are considered interesting enough to try to validate and replicate.

This type of situation raises several questions:

- Do multiple testing corrections mean that more information leads to less knowledge?
- Which strategies, in addition to dimensionality reduction, may we employ to find the "needles" in a scientific haystack?
- If data driven science is seen as a way of coming up with new hypotheses to be tested and findings replicated, in which situations should we be more worried about false positives/discoveries than about false negatives?
    - When validation/replication is expensive?
    - When important decisions may be taken based on our results?
- What do we mean by "independent tests" and what do we mean by "the same time"? Should there be a "career-wise p-value"?
- Which criteria, including previous knowledge of our study system, may we use to assess which hypotheses to follow up, regardless of statistical significance in preliminary tests?
- How can we think about multiple testing corrections in machine learning methods such as neural networks?
- Which practices are in use in your own field of study, and why?

This abstract is meant to kindle an open-ended discussion about the way in which we interpret data in a data driven age. Please come to my poster to share your thoughts and experiences! From those who leave written comments, one person will be drawn (at random) to receive a box of chocolate.

ABSTRACT

Title
National Bioinformatics Infrastructure Sweden (NBIS)

Presenter
Pär Engström, Bioinformatics Long-term Support Manager, NBIS, SciLifeLab

Affiliations
NBIS (National Bioinformatics Infrastructure Sweden), forming the SciLifeLab Bioinformatics Platform, provides a single point of contact for users needing bioinformatics support, facilitating contacts and enabling efficient service provision. NBIS is a distributed infrastructure formed as a consortium of major Swedish universities with staff placed at all sites. NBIS also constitutes the Swedish node in Elixir, the European infrastructure for bioinformatics.

In this presentation, I will describe the services offered by NBIS, as outlined below, and highlight a few recent research projects where NBIS has assisted with data analyses.

NBIS constitutes a sustainable infrastructure enabling bioinformatics analyses in life science, offering advanced expertise for analyses in genomics, proteomics, metabolomics and systems biology, including provision of efficient tools for large-scale analyses. NBIS provides support in organising and analysing complex omics data, with a focus on enabling integrative and systems biology approaches. NBIS further aims to catalyse transitions of large-scale omics approaches into clinical use. NBIS provides access to genome annotation and assembly expertise, supports data publication, and engages in advanced bioinformatics training. NBIS coordinates the Swedish Elixir node and engages in Nordic and European collaborations. The majority of the user groups are at academic institutions, but NBIS also interfaces with hospitals and industry for mutual benefits.

NBIS has staff at all large universities providing an easy access point to the infrastructure. We arrange weekly bioinformatics drop-in sessions at all our sites across Sweden, allowing researchers to get feedback and guidance on experimental design, choice of analysis methods, software etc. Alternatively, one can book a consultation meeting with one of our experts, also free of charge.

For in-depth project support, NBIS currently offers two tracks. In the Short- and Medium-term Support track, support is provided for an hourly fee and can usually start within a few weeks from first contact with a user, if data is available. This track is mainly targeting projects with a reasonably well-defined bioinformatics problem, but also longer and more open-ended projects are welcome to apply. In the Long-Term Support track, enabled by a grant from the Knut and Alice Wallenberg Foundation, extensive support is provided free of charge to a limited set of scientifically outstanding projects. Applications are submitted in open calls three times per year, and projects are selected based on scientific peer-review by a national committee.

Application forms and more information are available at http://www.nbis.se.

ABSTRACT

Title
Deep auto-encoders for the identification of disease modules

Authors
Mika Gustafsson

Affiliations
Institutionen för fysik, kemi och biologi (IFM), Bioinformatik, Linköpings Universitet

Email: mika.gustafsson@liu.se

Abstract

A fundamental problem of systems biology is how to aggregate thousands of measurements to a minimal set of variables describing the cell. Network analysis of protein-interaction networks has been suggested as a key tool for coarse-grain interpretation of omics data at a systems biological level. However, a fundamental problem is then how to generate new knowledge from data? For example, these networks are biased with many false positives at well-studied disease associated proteins and many false negatives around unknown proteins.

One promising approach is to use deep neural networks (NNs) for unbiased compression into latent variables describing the cellular state. The LINCS project pre-identified 1000 marker genes from which they found to predict ~90% of the variance in the expression of all 20,000 genes using deep neural networks, which supports that gene expression is redundant. In this work we have aimed to identify a best unbiased hidden representation of gene expression from global gene expression measurements in terms of deep auto-encoders (DAE), and analyse the characteristics of the DAE in the context of complex diseases. Our work show a further compression onto 500 hidden nodes could explain ~95% of the variance. Analysis of this representation showed that genes associated to the same hidden node also co-localised in the protein-protein interaction network, thus supporting highly interconnected gene sets explaining the global gene expression pattern. Furthermore, unsupervised cluster analysis of the hidden node representation found that they clustered in both cell types and diseases, and supervised classification based on this learned representation also showed separation of cell types and diseases. These findings led us to test if our learned representation was able to detect unbiased upstream genetic associations to diseases, for which disease modules and upstream transcription factor analysis has been successfully been used. For this purpose, we learned another neural network mapping expression to diseases based. Strikingly, this new NN based on the DAE identified mRNA signatures which were highly significant disease associated for ten diseases at the GWAS levels, which indicates an upstream involvement.

The ISCB's Regional Student Group in Sweden (**RSG-Sweden**) features an informative panel discussion on

Essential skills and networking advice for students

Join us to discuss the important skills needed for a successful career in industry and academia.

Anna Reznichenko
Associate Principal Scientist,
AstraZeneca R&D,
Gothenburg, Sweden

Warren Winfried Kretzschmar
Postdoctoral Researcher,
KTH Royal Institute of Technology,
Stockholm, Sweden

Dieter Maier
Head of Project Management,
Biomax Informatics AG,
Planegg, Germany

Friday, October 26 | 14:00-15:00 | Örebro University

Organisers: Nazeefa Fatima and Martin Rydén

ISCB Student COUNCIL
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

ISCB REGIONAL Student GROUP
Sweden

SBW swedish bioinformatics workshop 2018 ÖREBRO

iscbsc.org **rsg-sweden.iscbsc.org** **oru.se/sbw2018**