1st European Conference on Translational Bioinformatics

Copenhagen

April 26-27, 2016

Organized by:

Novo Nordisk Foundation
Center for Protein Research

Co-organized by:

DET SUNDHEDSVIDENSKABELIGE FAKULTET
KØBENHAVNS UNIVERSITET

SYNAPSE
RESEARCH MANAGEMENT PARTNERS

MedBioinformatics

elixir
DENMARK

The work presented in the framework of the MedBioinformatics project has received funding from the European Union's Horizon 2020 Research and Innovation Programme 2014-2020 under Grant Agreement no 634143

Abstracts, posters

Network-Attacking Mutations Detected in Structural Diseasomes Using Hot-spot Prediction

Didier Barradas Bautista[2], Juan Fernández Recio[1,2]
didier.barradas@bsc.es ,juanf@bsc.es

[1]*Joint BSC-CRG-IRB Research Program in Computational Biology*, [2]*Barcelona Supercomputing Center.*

Next generation sequencing projects have demonstrated that mutations like the non-synonymous single nucleotide polymorphisms(nsSNPs) are responsible for population diversity and how individual are affected. Protein-protein interactions(PPIs) are involved in almost all essential cellular processes,and may be affected by nsSNPs causing a pathology[1], [2]. However, in spite of their importance, there is no available 3D structure for the vast majority of known PPIs[3]. Computational methods, such as protein-protein docking, can complement existing experimental efforts and help building the human structural interactome[4]. The correct prediction of protein complexes by docking is still very challenging for many cases. However the identification of hot spot interface residues, based on sequence conservation or on physico-chemical properties, is more accurate and can be applied at more large scale. When characterizing PPI interfaces, it would be important to identify hot-spot residues, which are those that contribute significantly to the binding energy[5]. A method developed in our group ,called pyDockNIP[6], is able to identify interface hotspots with high precision, and has the clear advantage of not needing prior information of the complex structure. We have developed and validated a variation of this method called "pyDockNIP extended" that can be applied to identify pathological mutations that are involved in altering the PPIs. Our method have 40% recall with 75% precision to identify network affecting mutations. We constructed PPI individual networks with all the 3D protein structures available for key elements of RASopathies network and six monogenic inheritable diseases. We included complex phenotypes with big networks like Colorectal cancer and Myocardial infarction. We found 292 nsSNPS that could be altering the network and causing a pathology thus providing leads for new therapeutic targets. 20 of these nsSNPs change the PPIs in the Ras signaling cascade without eliciting major changes in the binding energy and can only be detected by our method.

[1] X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin, and H. Yu, "Three-dimensional reconstruction of protein networks provides insight into human genetic disease," *Nat Biotechnol*, vol. 30, no. 2, pp. 159 – 164, 2012.
[2] R. Mosca, J. Tenorio-Laranga, R. Olivella, V. Alcalde, A. Céol, M. Soler-López, and P. Aloy, "dSysMap: exploring the edgetic role of disease mutations," *Nat. Methods*, vol. 12, no. 3, pp. 167–168, Feb. 2015.
[3] R. Mosca, A. Céol, and P. Aloy, "Interactome3D: adding structural details to protein networks," *Nat. Methods*, vol. 10, no. 1, pp. 47–53, Jan. 2013.

[4] R. Mosca, C. Pons, J. Fernández-Recio, and P. Aloy, “Pushing Structural Information into the Yeast Interactome by High-Throughput Protein Docking Experiments,” *PLoS Comput. Biol.*, vol. 5, no. 8, p. e1000490, Aug. 2009.
[5] O. Keskin, B. Ma, and R. Nussinov, “Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot-spot residues,” *J Mol Biol*, vol. 345, pp. 1281 – 1294, 2005.
[6] S. Grosdidier and J. Fernández-Recio, “Identification of hot-spot residues in protein-protein interactions by computational docking,” *BMC Bioinformatics*, vol. 9, no. 1, p. 447, 2008.

Large-scale prediction of disease risk using 18 years of registry data from 6.8 million Danes

Anders Boeck Jensen[1], Pope L. Moseley[1,2], Søren Brunak[1]

[1]NNF Center for Protein Research, University of Copenhagen Denmark, [2]University of Arkensas Medical Sciences, United States of America

A crucial goal in P4 medicine is to be able to predict disease progression from the patient’s current state given the patient’s disease history. An important source for patient history data is the patient registries that systematically gather diagnoses and other medical data on patients. In recent large-scale disease correlation studies they have been shown to be a promising source for exploring disease patterns. However, the studies have focused mainly on pair-wise co-morbidities or predicting single diseases. So far there have been no attempted at building predictive models encompassing all the patient’s hospitalization history on a large scale.

Here, we present a large-scale discovery-driven study of temporal disease progression (disease trajectories) and the work towards obtaining predictive models from this data. In a recent study, we published an analysis of disease trajectories for the entire spectrum of diseases that used health registry data for a 15 years period. With the addition of 3 more years to the data, we have systematically identified all branching points in our disease trajectories where there are multiple common outcomes. Using machine learning techniques, we will predict each patient’s probability of future diagnoses from these points.

From Genome-wide RNA Sequencing to Diagnostic Tool for Inflammatory Bowel Disease

Jette Bornholdt[1,2], Mette Boyd[1,2], Malte Thodberg[1,2], Morana Vitezic[1,2], Kristoffer Vitting Serup[1,2], Mehmet Coskun[1,2], Robin Andersson[1], Kerstin Skovgaard[3], Thilde Bagger Terkelsen[1,2], Anders Gorm Pedersen[4], Jesper T. Troelsen[5], Jakob B. Seidelin[6], Jacob T. Bjerrum[6], Ole H. Nielsen[6], Albin Sandelin[1,2]

[1]Section for Computational and RNA Biology, Institute of Biology, UCPH, [2]Biotech Research and Innovation Centre, UCPH, [3]Section for Immunology and Vaccinology, National Veterinary Institute, DTU, [4]Centre for Biological Sequence Analysis, DTU, [5]Department of Science, Systems and Models (NSM), Roskilde University, [6]Department of Gastroentrology, Medical Section, Herlev Hospital.

Inflammatory bowel disease (IBD) is a common inflammatory condition of the intestine, causing abdominal pains, diarrhea and rectal bleeding. The two principal types of IBD are Crohn's Disease (CD) and Ulcerative Colitis (UC). CD can affect the whole gastrointestinal tract and is often transmural, while UC is characterized by mucosal inflammation in the colon and rectum. In practice, the subtypes are hard to distinguish with current methods (typically based on colonoscopy and histology) and patients are often classified as "indeterminate colitis". As treatment regimes differ between CD and UC, correct diagnosis is crucial for an efficient medication.

Here we use state-of-the-art high-throughput, genome-wide RNA expression methods to screen gut biopsies from 94 subjects, with the goal to identify differentially expressed promoters and enhancers which in turn will allow us to distinguish the IBD subtypes. By combining the expression data with a machine learning approach, we were able to classify the patients with an overall accuracy of 80% (analyzed via 5-fold cross validation). We then proceeded to validate the top 144 features by real-time PCR using a nano-fluidics system (Fluidigm). Using the machine learning/cross validation approach on these data enabled us to reduce the feature set to consist of only 39 transcripts/primer assays, retaining an accuracy of 86%. The qPCR primer sets resulting from our study provides a key starting point for the development of kits to classify IBD and its subtypes with high confidence.

The Promoter Landscape of Inflammatory Bowel Disease: Finding Predictors of Inflammation and Disease State

Mette Boyd[1,2], Jette Bornholdt[1,2], Malte Thodberg[1,2], Morana Vitezic[1,2], Kristoffer Vitting-Serup[1,2], Mehmet Coskun[1,2], Robin Andersson[1], Anders Gorm Pedersen[3], Katja Dahlgaard[4], Jesper T. Troelsen[4], Jakob B. Seidelin[5], Jacob T. Bjerrum[5], Ole H. Nielsen[5], Albin Sandelin[1,2]

[1]Section for Computational and RNA Biology, Institute of Biology, UCPH, [2]Biotech Research and Innovation Centre, UCPH, [3]Centre for Biological Sequence Analysis, DTU, [4]Department of Science and Environment (INM), Roskilde University, [5]Department of Gastroentrology, Medical Section, Herlev Hospital.

Inflammatory bowel disease (IBD) is a common chronic inflammatory bowel disorder with an increasing incidence and prevalence worldwide. It is classified into two major entities: ulcerative colitis (UC), characterized by mucosal inflammation restricted to the colon, and Crohn's disease (CD) where transmural inflammation can occur in any part of the gastrointestinal tract. The distinction between CD and UC is critical for correct management, especially options of surgery and personalized treatment, yet the diagnosis is challenging, approximately 10% of patients are classified as 'indeterminate colitis'. Therefore, novel biomarkers for stratifying patients and to improve diagnostics are highly needed - a first step towards personalized medicine in IBD.

Here, we have applied a unique RNA sequencing technique, Cap Analysis of Gene Expression (CAGE), on intestinal biopsies from 94 patients with IBD as well as healthy controls. This provided a genome wide, nucleotide-resolution atlas of active transcription start sites (TSSs), for both known and novel genes. We show that highly expressed immune cell related transcripts were powerful predictors of the degree of inflammation (controls vs CD & UC), whereas, low expression of epithelial cell related transcripts, including previously unidentified long non-coding RNAs and alternative TSSs of known genes, were far more powerful predictors of UC *vs.* CD. Our findings suggest that the general inflammatory response of IBD is related to the regulation of immune-related genes, but that the diagnoses of CD and UC are distinguished by disruption of resident epithelial intestinal cell function, such as tight junction integrity and metabolite transport.

In addition to insights into the disease transcriptome of IBD, our study provides ample ground for screening and development of new biomarkers for IBD in general, and CD and UC in particular.

The Hematological Relapse Project
Implementation of individualized care for therapy resistant cancer – a health technology assessment

R.F. Brøndum[1], J.S. Bødker[1], A. Schmitz[1], M.D. Bendtsen[1,2], C. Baggesen[1], M. Sommer[1], J.G. Frausing[1], M. Bøgsted[1,2,3], K. Dybkær[1,2,3], T.C. El-Galaly[1], and H.E. Johnsen[1,2,3]

[1]Department of Hematology, Aalborg University Hospital, Denmark; [2]Department of Clinical Medicine, Aalborg University; [3]Clinical Cancer Research Center, Aalborg University Hospital, Denmark

Patients suffering from hematological malignancies are usually offered one or more chemotherapeutic drugs as first line treatment. However, a large proportion of these patients

experience relapse, which in many cases leave physicians with no evidence-based treatment options. We perceive the future of cancer treatment as personalized medicine strategies, and relapse is an excellent example where this approach can be tested and validated in a translational research program.

The Hematological Relapse Project (PROGENE) at Aalborg University Hospital is a personalized medicine initiative that aims to use a health assessment approach to address this problem from several angles: technology, patient welfare, economy, and organization. The primary goal of the project is to use -omics technologies and clinical data to understand the mechanisms of treatment and to use this information to design specific therapeutic strategies aimed at overcoming them.

Ideally, this approach should lead to less under- and over treatment, resulting in fewer patients experiencing a second relapse and an increase in health-related quality of life, but it might also give a reduction in medicine expenses. We aim to explore the patients' need for information and counseling as well as ethical concerns for incidental findings and heritable risks, to investigate if individual treatment plans actually result in higher health-related quality of life. We also plan to do a full economical analysis of this approach versus conventional diagnostics since implementation has associated costs for new staff and machinery.

Finally, we need to establish a new organization, the tumor board, where bioinformaticians and molecular biologists aid physicians in planning the individualized treatment for relapsed patients. Implementation and validation of the precision medicine concept will run for two years with collection of data and no intervention in the treatment of patients, followed by a three- year clinical study where patients are offered either the doctor's best choice for second line treatment or individualized treatment guided by molecular data.

EU Data Law Adopted into Interoperability Standards Can Make Compliance Easier

Sangzi Chen

EU data privacy law applies to anyone using the personal data of EU residents. The new General Data Protection Regulation, written in 2012, passed into law in 2015. Scientists and others should take this opportunity to get in compliance with the new rules. The paper proposes that technical interoperability standards, particularly HL7, can be used to promulgate compliance with EU data law.

Under the old EU Directive 95-46, the law failed to consider common issues in data use, most notably the internet. Further, it was enforced by individual member states. The new General Data Protection Regulation, in addition to uniform regulation across the EU, now explicitly considers scientific research, and defines new categories for genetic data and data concerning health.

The EU also actively encourages technical standards for compliance with the law. HL7 is a technical interoperability standard for electronic healthcare records (EHR), widely adopted by the EHR industry in Europe and the US. By incorporating elements of EU data law into technical standards, compliance can be made easier for researchers, healthcare systems, and other data users.

Characterizing Age-dependent Regulatory Variation in the Human Frontal Lobe Region

Maria Dalby[1], Cornelis Blauwendraat[2], Sarah Rennie[1], Margherita Francescatto[2], Patrizia Rizzu[2], Peter Heutink[2], Robin Andersson[1]

[1]The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark, [2]German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany

Age is a major risk factor for most common neurodegenerative diseases, such as Alzheimer's disease, Parkinson's disease and cognitive impairment. While a considerable effort has been made to associate gene expression changes with neurodegenerative disease state and etiology, the general impact of aging on human transcription and transcriptional regulation remains to a large degree unexplored. To this end, we have characterized gene expression levels and regulatory activities over the healthy human aging process, with a particular focus on the frontal lobe brain region.

We have systematically characterized the frontal lobe transcriptome using Cap Analysis of Gene Expression (CAGE) on total RNA isolated from post mortem frontal lobe samples of 144 healthy individuals with an age span from 2 to 95 years. From this data, we have determined genome-wide activities of frontal lobe gene promoters and transcriptional enhancers and the effect of aging on transcriptional and regulatory variability.

A detailed characterization and clustering of the major age related trends for promoters demonstrate expression changes across lifespan for a significant proportion of genes and recapitulate known genes associated with senescence in non-human species. By co-expression analysis, we fill infer regulatory architectures and determine the impact of regulatory genetic variants on age-related transcriptional programs. Our detailed map of regulatory variation over the human aging process will further allow a focused analysis of genetic variants associated with major neurodegenerative diseases.

Familial Co-occurrence of Congenital Heart Defects Follows Distinct Patterns

Sabrina G. Ellesøe, MD, PhD[1,*]; Christopher T. Workman, PhD[2,*]; Patrice Bouvagnet, MD, PhD[3]; Christopher A. Loffredo, PhD[4]; Kim L. McBride, MD[5]; Robert B. Hinton, MD[6]; Klaartje van Engelen, MD, PhD[7,8]; Emma C. Gertsen, MD[7]; Barbara J.M. Mulder, MD, PhD[9]; Alex V. Postma, PhD[7, 10]; Robert H. Anderson, BSc, MD, FRCPath[11]; Vibeke E. Hjortdal, MD, PhD, DMSc[12]; Søren Brunak, MSc, PhD[1]; Lars A. Larsen, MSc, PhD[13]

[1]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark; [2]Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark; [3]Laboratoire Cardiogénétique, Hospices Civils de Lyon, Groupe Hospitalier Est, Bron, France; [4]Department of Oncology, Lombardi Cancer Center, Georgetown University, Washington, DC; [5]Center for Cardiovascular and Pulmonary Research, Nationwide Children's Hospital, Department of Pediatrics, Ohio State University, Columbus OH; [6]Division of Cardiology, The Heart Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH; [7]Department of Clinical Genetics, Academic Medical Centre, Amsterdam; [8]Department of Clinical Genetics, VU University, Amsterdam; [9]Department of Cardiology, Academic Medical Centre, Amsterdam, The Netherlands; [10]Department of Anatomy, Embryology & Physiology, Academic Medical Centre, Amsterdam, The Netherlands; [11]Institute of Genetic Medicine, Newcastle University, London, United Kingdom; [12]Department of Cardiothoracic Surgery, Aarhus University Hospital, Skejby, Denmark; [13]Wilhelm Johannsen Centre for Functional Genome Research, Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark.

* These authors contributed equally to the study

Background

Congenital heart defects (CHD) affect almost 1% of all live born children and the number of adults with CHD is increasing. In families where CHD has occurred previously, estimates of recurrence risk and the type of recurring heart defect are important for counseling and clinical decision making, but the recurrence patterns of CHD in families are poorly understood.

Methods and Results

We investigated the co-occurrences of congenital heart defects in 1,163 CHD families, comprising 10,278 individuals, of which 3,080 had a clinically confirmed CHD diagnosis. We calculated rates of concordance and discordance for 42 types of CHD, observing a high variability in the rates of concordance and discordance. By calculating Odds Ratios for each of 1,771 pairs of discordant lesions observed between affected family members, we were able to identify 192 pairs of malformations that co-occurred significantly more or less often than expected in families. The data show that distinct groups of cardiac malformations co-occur in families, suggesting influence from underlying developmental mechanisms. Analysis of data from mouse models with malformations corresponding to familial CHDs showed that susceptibility genes were shared in 21.4% of pairs of

co-occurring discordant malformations but none of malformations that rarely co-occur, suggesting that a significant proportion of co-occurring CHD in families is caused by overlapping susceptibility genes.

Conclusion
Congenital heart defects in families follow specific patterns of recurrence, suggesting a strong influence from genetically regulated developmental mechanisms. Co-occurrence of malformations in familial CHD is caused by shared susceptibility genes.

Drug Adverse Effect Discovery Goes Unsupervised

Emre Guney

Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), c/ Baldiri Reixac 10-12, 08028 Barcelona, Spain

Drug safety issues remain as one of the major bottlenecks in drug development, contributing to more than 20% of the clinical trial failures.
Though effective, experimental screening of drugs for large scale adverse effect detection is currently unattainable. Computational methods, relying on drug and side effect similarity to train classifiers, offer a cost-effective alternative but typically fail to provide a universal solution over different data sets. In this study, we present, ProXide, a purely interactome topology based drug side effect prediction method. ProXide uses the network-based proximity of drug targets to side effect modules (proteins likely to induce the side effects) to quantify the likelihood of the drug-side effect association. Our analysis of 819 FDA approved drugs and 537 side effect modules in the interactome shows that proximity can discover known drug side effects with prediction accuracy comparable to similarity-based approaches. Furthermore, combined with drug chemical and target similarity, proximity based adverse effect detection is robust against data incompleteness and outperforms any single method individually. We demonstrate how ProXide can pinpoint novel drug-side effect associations on several case studies.

Leveraging Text Mining, Expert Curation and Data Integration to Develop a Database on Psychiatric Diseases and Their Genes

Alba Gutiérrez-Sacristán[1], Àlex Bravo[1], Olga Valverde[2], Marta Torrens[3], Ferran Sanz[1] and Laura I. Furlong[1]

[1]*Research Group on Integrative Biomedical Informatics, Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Experimental and Health Sciences (DCEXS), Universitat Pompeu Fabra (UPF),* [2]*Neurobiology of*

Behaviour Research Group (GReNeC), IMIM, DCEXS, UPF, [3]Institute of Neuropsychiatry and Addiction, Parc de Salut Mar, Universitat Autonoma de Barcelona, Barcelona, 08003, Spain

Keywords: text mining, data mining, psychiatric disorder, genetics, curation, database, data integration

During the last years there has been a growing research in the genetics of psychiatric disorders, supporting the notion that most of them display a strong genetic component. However, there is still a limited understanding of the cellular and molecular mechanisms leading to psychiatric diseases, which has hampered the application of this wealth of knowledge into the clinical practice to improve diagnosis and treatment of psychiatric patients. This situation also applies to psychiatric comorbidities, which are a frequent problem in these patients. Some of the factors that explain the lack of understanding of psychiatric diseases etiology are the heterogeneity of the information about psychiatric disorders and its fragmentation into knowledge silos, and the lack of resources that collect this wealth of data, integrate them, and supply the information in an intuitive, open access manner to the community along with analysis tools. PsyGeNET (http://www.psygenet.org/) has been developed to fill this gap, by facilitating the access to the vast amount of information on the genetics of psychiatric diseases in a structured manner, providing a set of analysis and visualization tools. PsyGeNET is focused on mood disorders (e.g. depression and bipolar disorder), addiction to substances of abuse and schizophrenia. In this communication we describe the process to update the PsyGeNET database, which involves i) extraction of gene-disease associations (GDAs) from the literature with state-of-the-art text mining approaches, ii) curation of the text-mined information by a team of experts in psychiatry and neuroscience, iii) integration with data gathered from other publicly available resources. BeFree, a text mining tool to extract gene-disease relationships, is used to identify genes associated to the psychiatric diseases of interest from a corpus of more than 1M publications. BeFree has a performance of 85% F-score for the identification of genes associated to diseases by exploiting morpho-syntactic features of the text. In addition, it normalizes the entities to standard biomedical ontologies and vocabularies. The textmined data is then reviewed by a team of experts to validate the GDAs, following specific curation guidelines. Each expert is assigned a disease area according to her/his area of expertise. A web-based annotation tool was developed to assist the curation process. The tool supports a multi-user environment by user and password assignment. It displays the evidence that supports the association for each GDA to the curator. More specifically, it shows the sentences that support the association between the gene and the disease, highlighted (both the sentence and the entities involved in the association) in the context of the MEDLINE abstract. The curator has to validate the particular association based on the evidence of each publication, and select an exemplary sentence that states the association. We also describe the protocol designed to assign the curation tasks to the different experts and the method to assess the inter-annotator agreement. Finally, we describe the approach to integrate the expert-curated data with GDAs identified from other publicly available resources.

Funding: We received support from ISCIII-FEDER (PI13/00082, CP10/00524), IMI-JU under grants agreements nº 115002 (eTOX), nº 115191 (Open PHACTS)], nº 115372 (EMIF) and nº

115735 (iPiE), resources of which are composed of financial contribution from the EU-FP7 (FP7/2007-2013) and EFPIA companies' in kind contribution, and the EU H2020 Programme 2014-2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

hemaClass.org: An Online Based Diffuse Large B-cell Lymphoma Classification Tool

Lasse Hjort Jakobsen[1,3], Steffen Falgreen[1], Anders Ellern Bilgrau[1,2], Jonas Have[1,3], Kasper Lindblad Nielsen[1,3], Tarec Christoffer El-Galaly[1], Julie Støve Bødker[1], Alexander Schmitz[1], Hans Erik Johnsen[1,3], Karen Dybkær[1,3] and Martin Bøgsted[1,3]

[1]Department of Haematology, Aalborg University Hospital, [2]Department of Mathematical Sciences, Aalborg University, [3]Department of Clinical Medicine, Aalborg University

Dozens of genomics based cancer classification systems have been introduced with prognostic, diagnostic, and predictive capabilities. However, they are often based on complex algorithms only applicable on whole cohorts of patients, making them difficult to apply in a personalized clinical setting. This prompted us to create hemaClass.org which is an online tool providing an easy interface to one-by-one microarray based risk classification of diffuse large B-cell lymphoma (DLBCL) into e.g. ABC/GCB [1], BAGS [2], and REGS [3] classes. However, laboratory specific effects cannot be accounted for in one-by-one normalisation. Hence, hemaClass.org optionally allows the user to supply a reference dataset to increase the accuracy of the classifications. Classification results for one-by-one array processing with and without a user supplied reference dataset were compared to cohort based classifiers in 4 publicly available datasets. Overall, hemaClass.org yields satisfactory inter-method agreements across all datasets. hemaClass.org is relevant for biological and clinical researchers in the lymphoma field as it provides a reliable and swift method for calculation of DLBCL subclasses. Finally, hemaClass.org is based on the R-package hemaClass accompanied with a Shiny server interface, which makes it straightforward to deploy it as a web server. The server works at the moment on Affymetrix HG U-133 plus 2.0 microarrays, but extensions to other high throughput platforms and cancers are planned.

[1] Alizadeh A et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature.
[2] Dybkær K et al. (2015) Diffuse large B-cell lymphoma classification system that associates normal B-cell subset phenotypes with prognosis. Journal of Clinical Oncology.
[3] Falgreen et al. (2015) Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. BMC Cancer.

Identifying Potential Adverse Drug Reactions by Text Mining Electronic Patient Records from Steno Diabetes Center

Niels Erik Olesen, Robert Eriksson & Søren Brunak

Department of Disease Systems Biology, Faculty of Health and Medical Sciences, NNF Center for Protein Research, University of Copenhagen

Correspondence to: nielserik.olesen@cpr.ku.dk

Background
Adverse drug reactions (ADRs) continue to impact a large part of the patient population. Traditionally, the safety of drugs has been assessed through clinical trials during the drug development process. After market approval, spontaneous reporting of ADRs to the authorities and manufacturers is used for surveillance of ADR patterns during daily clinical use and identification of previously undetected ADRs. However, this system has been described to result in underreporting. However such information about ADRs might still be present in the clinical narrative in the patient record and thus text mining the narratives might improve the situation. This framework has the potential to identify ADRs in a clinical reality where the clinical situation might be very different from a well-controlled drug trial with regard to polypharmacy and variable drug adherence.

Methodology
In this project we applied our previously developed text mining pipeline to electronic patient records from Steno Diabetes Center, a highly specialized diabetic care center. The text mining process consists of automatic tagging of the patient record with a predefined vocabulary consisting of drugs and concepts describing potential ADRs, which together allows identification of relevant information in the clinical narrative. We are performing the current analysis to prepare for a major up scaling, namely analyzing all electronic patient records in the Danish society.

Results
The preliminary text mining of the electronic patient records from Steno Diabetes Center resulted in 9826 possible ADRs. As an initial result, we identified drugs causing common well-known ADRs and noted that the spectrum of identified ADRs was in accordance with the expectations for a diabetic population.

Conclusion
The text mining pipeline was successfully applied on the corpus from Steno Diabetes Center. Using the method we are able to describe ADR patterns in a population suffering from diabetes. Future work will be dedicated to improve the vocabulary and efficiency of the pipeline to comprise text mining on even larger corpora of electronic patient records.

DisGeNET: A Discovery Platform to Support Translational Research and Drug Discovery

Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Ferran Sanz and Laura I. Furlong

Integrative Biomedical Informatics Group, Research Programme on Biomedical Informatics; Hospital del Mar
Medical Research Institute; Pompeu Fabra University, Barcelona, Spain.

In the last two decades, technological breakthroughs in the field of biomedical research have resulted in an unprecedented increase of the volume and variety of data concerning disease genetics. The fragmented nature of this process has produced bottlenecks in the analysis and extraction of knowledge from this vast amount of information. Currently, our knowledge on the genetic determinants of diseases is scattered across different catalogs, each focusing on one aspect of the gene-disease relationship.
Additionally, the data are represented and annotated using different schemes, vocabulary and standards, which poses extra hurdles to the analysis and interpretation. The creation of tools that collect and homogeneously annotate our current knowledge on the genetic basis of diseases is of paramount importance to the development of translational bioinformatics.
To target this need we have developed DisGeNET (Piñero *et al*, 2015), a discovery platform that collects and integrates the available information on gene-disease associations (GDAs), covering the whole landscape of human diseases. DisGeNET is organized around the gene-disease relationship, annotated with its supporting evidence that includes the original paper and database reporting the association, a representative sentence from each supporting publication, among other attributes. DisGeNET possesses several unique features that make it a platform of choice for biomedical translational researchers and drug discovery applications. First, it contains one of the largest collections of GDAs arising from both expert-curated knowledge, and information extracted from the scientific literature using NLP-based text-mining techniques, with special attention paid to the explicit provenance of the association. Second, it provides mappings to different biomedical vocabularies annotating diseases, thus facilitating the work of clinical and biomedical researchers. Third, it features a score developed to rate the confidence of each association. Finally, several ways to access the data are available, to better serve the purposes of different types of users: a web interface, aimed at biologists and health-care practitioners; a Cytoscape plugin, intended for systems biology expert; and Semantic Web technologies are available for bioinformatician and software developers.
DisGeNET is also available as a Linked Open Data (LOD) data set through DisGeNET-RDF (Queralt Rosinach *et al*, 2015). DisGeNET-RDF is interlinked to other biomedical databases to support translational research through evidence-based exploitation of a rich and fully interconnected LOD cloud that include data on gene expression, drug targets, biological pathways, kinetic models, to just mention a few examples.

References
Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F & Furlong LI (2015)
DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.
Database **2015:** bav028–bav028
Queralt Rosinach N, Pinero J, Bravo Serrano A, Sanz F & Furlong LI.DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. Bioinformatics, under review.

Rational Design of Cancer Gene Panels with OncoPaD

C.Rubio-Perez1, J.Deu-Pons1, D.Tamborero1, A.Gonzalez-Perez[1]*, N.Lopez-Bigas[1,2]*

[1]*Biomedical Genomics Lab, Research Program on Biomedical Informatics, IMIM Hospital del Mar Medical Research Institute and Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.*
[2]*Institucio Catalana de Recerca i Estudis Avançats, Barcleona, Catalonia, Spain.*

Profiling somatic mutations in the coding sequence of genes or specific mutational hotspot fragments that may inform several aspects of tumor evolution, prognostic and treatment is becoming a standard tool in clinical oncology. Gene panels to address these questions need to be tailored by researchers through laborious search of the literature and cancer genomics resources, with no possibility to estimate their performance on a cohort of patients.

Here, we present OncoPaD, to our knowledge the first tool aimed at the rational design of cancer gene panels by dynamically estimating their efficacy to profile large cohorts of tumors of 28 malignancies. OncoPaD computes the accumulated fraction of samples in cancer type-specific or pan-cancer cohorts of tumors bearing mutations in any of the genes/regions provided by the user or prioritized by the tool. In addition to the estimate of the designed panel on a real cohort of patients, OncoPaD provides reports on the importance of individual mutations for tumorigenesis or therapy to support the interpretation of results obtained with the panel. We demonstrate *in silico* that OncoPaD designed panels are more cost-effective than commercially available tumor panels and four tools that aid researchers in the design of panels to probe malignancies in terms of the fraction of samples they successfully probe and the kbps of DNA sequenced.

OncoPaD will help clinicians and researchers design panels tailored to achieve the best performance in the early detection of tumors in liquid biopsies or make genomic-informed therapeutic decisions among other use cases.

miRandola 2016: the latest version of the circulating RNA database

Francesco Russo[1,2,3], Sebastiano Di Bella[4], Giovanni Nigita[5], Federica Vannini[6], Gabriele Berti[6], Flavia Scoyni[7], Alessandro Laganà[8], Alfredo Pulvirenti[4], Rosalba Giugno[4], Marco Pellegrini[1], Kirstine Belling[3], Søren Brunak[3], Alfredo Ferro[4]

[1]Institute of Informatics and Telematics (IIT) , National Research Council (CNR), Italy; [2]Department of Computer Science, University of Pisa, Italy; [3]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark; [4]Department of Clinical and Experimental Medicine, University of Catania, Italy; [5]Department of Molecular Virology, Immunology, The Ohio State University, USA; [6]Institute of Clinical Physiology (IFC), National Research Council (CNR), Italy; [7]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Denmark; [8]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, NYC, USA

Non-coding RNAs (ncRNAs) such as for example microRNAs (miRNAs) are frequently dysregulated in cancer and have shown great potential as tissue-based markers for cancer classification and prognostication. ncRNAs are present in membrane-bound vesicles, such as exosomes, in extracellular human body fluids. Circulating miRNAs are also present in human plasma and serum cofractionate with the Argonaute2 (Ago2) protein and the High-density lipoprotein (HDL). Since miRNAs and the other ncRNAs circulate in the bloodstream in highly stable, extracellular forms, they may be used as blood-based biomarkers for cancer and other diseases. A knowledge base of non-invasive biomarkers is a fundamental tool for biomedical research.

Data is manually collected from ExoCarta, a database of exosomal proteins, RNA and lipids and PubMed. Articles containing information on circulating RNAs are collected by querying PubMed database using keywords such as “microRNA”, “miRNA”, “extracellular” and “circulating”. Data is then manually extracted from the retrieved papers. General information about miRNAs is obtained from miRBase. The aim of miRandola is to collect data concerning RNAs contained not only in exosomes but in all extracellular types functionally enriched with information such as diseases, processes, functions, associated tissues, and their potential roles as biomarkers.

Here, we present an updated version of the miRandola database, a comprehensive manually curated collection and classification of extracellular circulating RNAs. The first version of the database has been published in 2012 and it contained 89 papers, 2132 entries and 581 unique mature miRNAs. Now, we have updated the database with 271 papers, 2695 entries, 673 miRNAs and 12 long noncoding

RNAs. RNAs are classified into several categories, based on their extracellular form: RNAAgo2, RNA-exosome, RNA-microvesicles, RNA-HDL and RNA-circulating. Moreover, the database contains several tools, allowing users to infer the potential biological functions of circulating miRNAs, their connections with phenotypes and the drug effects on cellular and extracellular miRNAs.

miRandola is the first online resource which gathers all the available data on circulating RNAs in a unique environment. It represents a usufeul reference tool for anyone investigating the role of extracellular RNAs as biomarkers as well as their physiological function and their involvement in pathologies. miRandola is constantly updated (usually once a year) and the online submission system is a crucial feature which helps ensuring that the system is always up-to-date. The future direction of the database is to be a resource for all the potential non-invasive biomarkers such as cell-free DNA, circular RNA and circulating tumor cells (CTCs). miRandola is available online at: http://mirandola.iit.cnr.it/

References

1) Francesco Russo, Sebastiano Di Bella, Giovanni Nigita, Valentina Macca, Alessandro Laganà, Rosalba Giugno, Alfredo Pulvirenti, Alfredo Ferro. miRandola: Extracellular Circulating microRNAs
Database. PLoS ONE 7(10): e47786. doi:10.1371/journal.pone.0047786

2) Francesco Russo*, Sebastiano Di Bella*, Vincenzo Bonnici, Alessandro Laganà, Giuseppe Rainaldi,
Marco Pellegrini, Alfredo Pulvirenti, Rosalba Giugno, Alfredo Ferro. A knowledge base for the discovery of
function, diagnostic potential and drug effects on cellular and extracellular miRNAs. BMC Genomics 2014,
15(Suppl 3):S4. doi:10.1186/1471-2164-15-S3-S4

NEXT Bioinformatics

Anna Amanda Schönherz and Martin Bøgsted

Aalborg University and Aalborg University Hospital, on behalf of NEXT Bioinformatics

NEXT, the National Experimental Therapy Partnership, is a Danish public-private partnership between six public institutions and five pharmaceutical companies working to strengthen Denmark as a preferred country for early clinical research trials. For the time being NEXT has established two Clinical Centre's of Intelligence one within dermatology and one within oncology & haematology with the potential to expand within other diseases with international competitive medical expertise.

The goal of NEXT is to improve the development of novel medical therapies and implement precision medicine to benefit the individual patient by making it easier and more attractive for the pharmaceutical industry to collaborate with hospital and university units and invest in early clinical trials in Denmark.

In addition to the Clinical Centre´s of Intelligence, NEXT has established a bioinformatics unit that teams experts in the clinical application of bioinformatics at Aalborg University Hospital, Rigshospitalet in Copenhagen, Odense University Hospital and Aarhus University Hospital.

NEXT Bioinformatics focus on precision medicine within oncology and haematology by supporting the generation, validation and implementation of common standards for patient classification based on genomic analyses. Furthermore, NEXT Bioinformatics is implementing a national bioinformatics database bringing together clinical, molecular and genomic patient profiles, which provides a unique possibility to recruit suitable patient candidates for early clinical trials.

Benchmarking of Biomedical Software Applications

Luís A. Bastião Silva[1], José Luis Oliveira[1,2]

[1]*BMD Software, Aveiro, Portugal*
[2]*DETI/IEETA, University of Aveiro, Portugal*

Due to the ever increasing number of publicly available bioinformatics tools, the expectations put on translational bioinformatics have been growing, pushing the development of these tools into a new level of requirements. By allowing perform many distinct processing tasks, these applications are seen as an important key to disclosure of the knowledge behind the biomedical data that are being accumulated in the last years. There is an increasing interest from physicians and medical institutions in applying these research tools in their daily practice, which also create new challenges regarding their quality for professional use. This is a major aim of the MedBioinformatics project, i.e. the development of several bioinformatics tools that are not only targeted for research, but, most of all, for clinical practice.
We aim defining a benchmarking methodology for biomedical applications, based on a set of performance indicators (such as quality, cost, time, productivity, etc.), that might allow the creation of the MedBioinformatics stamp – a seal of quality to guarantee that a series of methods have been used to develop these tools and that the final result respects a groups of quality requirements. This framework will be used to evaluate how methods and tools are positioned to maximise its effective use in translational research and clinical practice.

A Microkernel Architecture for Biomedical Software Integration

Luís A. Bastião Silva[1], José Luis Oliveira[1,2]

[1]*BMD Software, Aveiro, Portugal*
[2]*ETI/IEETA, University of Aveiro, Portugal*

A major issue in translational bioinformatics is still, many times, the lack of integration between data and different tools. Several solutions are already available to integrate and extract relevant information from different sources such as literature, omics databases, electronic health records, medical images, and clinical reports. Moreover, most of them already provide web based portals, with programmatic web services and easy-to-use graphical interfaces. However, they are often isolated applications which are difficult to use by third-party applications.
We propose an application to describe generic biomedical tools that can easily integrate and support third-party components allowing users without major programming skills to extend its functionalities, with a minimum effort in the integration process. To accomplish this goal, we developed a microkernel software architecture that allows the applications to be easily extended in different ways, mainly from the client-side perspective. Based on this model, we developed a solution for a neuroradiology scenario, where we integrate distinct tools to manipulate different patient data, such as electronic health records and medical images.

Mining of Electronic Patient Records to find Adverse Drug Reactions

Freja Karuna Hemmingsen Sørup[1,2], Robert Eriksson[2], Jesper Hallas[3], Søren Brunak[2] & Stig Ejdrup Andersen[1].

[1]Unit of Clinical Pharmacology, Zealand University Hospital, Roskilde, [2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medicine, University of Copenhagen, [3]Unit of Clinical Pharmacology, University of Southern Denmark.

Background
The Pharmacovigilance system relies on spontaneous adverse drug events (ADEs) reports from clinicians and patients to identify possible new adverse drug reactions (ADRs). Underreporting and selection bias is however a mayor problem.
Adverse drug reactions are very common in hospitalized patients. Mining the text parts of electronic patient records is a feasible approach to identify possible adverse drug reactions, but have so far in Denmark only been tried in a smaller homogenous patient population.
An ADE dictionary, based on the undesirable effects section of 7446 Danish summary of product characteristics was developed in a former ph.d-study.

Objectives
The aim of this ph.d-project is to investigate and validate the use of the ADE dictionary in a large real-life population to find frequencies and patterns of adverse drug reactions related to specific drugs. Furthermore, we want to evaluate the feasibility of using text mining in Pharmacovigilance.

Methods
The electronic patient records are mined for phenotypic data, biochemical data, drug use and possible adverse reactions using the dictionary. Filtering is applied to enhance the specificity of the data. The result can be visualised with Bioinformatic tools such as heatmaps or cluster-mapping or used as input to more conventional epidemiological design.

Planned study
The first study planned is mapping and comparing the adverse drug patterns of warfarin and the new oral anticoagulants (NOACs) in a real-life population. We will focus on adverse reactions associated with shifts between the different anticoagulants and the use of co-medication in people experiencing bleeding while treated with warfarin or a NOAC.

Perspective
The use of text mining of electronic patient records might be a future tool to supplement or replace the Pharmacovigilance based on spontaneous reports from clinicians and patients.

The DrugTargetCommons Initiative – a Community Effort for Building up the Consensus Knowledgebase of Drug-target Interactions

Jing Tang[1], Zia ur Rehman[1], Gretchen Repasky[1], Janica Wakkinen[1], Gopalacharyulu Peddinti[1], Alok Jaiswal[1], Markus Vähä-Koskela[1], Ella Karjalainen[1], Balaguru Ravikumar[1], Arjan Adrichem[1], Liye He[1], Elina Parri[1], Prson Gautam[1], Matti Kankainen[1], Suleiman Khan[1], Gupta Abhishekh[1], John Overington[2], Anne Hersey[2], Anne-Lena Gustavsson[3], Brinton Seashore-Ludlow[3], Ola Engkvist[4], Hassan Al-Ali[5], Lemmon Vance[5], Krister Wennerberg[1], Tero Aittokallio[1]

[1]*Institute for Molecular Medicine Finland (FIMM), University of Helsinki, FI*
[2]*European Bioinformatics Institute (EBI), UK*
[3]*Karolinska Institute, SE*
[4]*AstraZeneca, SE*
[5]*University of Miami, FL, USA*

Comprehensive knowledge of the target space of pharmacologically active substances, drugs and drug candidates, provides important insights into therapeutic potential and possible adverse effects. The existing data for compound-target interactions are relatively sparse and scattered across a number of studies and databases, including ChEMBL and PubChem, two of the most comprehensive compound bioactivity databases. However, an insufficient annotation of the use of different experimental procedures to determine bioactivity values may pose a major challenge for understanding the data heterogeneity. Many databases, such as DrugBank and STITCH, list the primary drug targets only, and lack quantitative bioactivity data covering the spectrum of both on and off-targets, which is needed for in-depth understanding modes of action. Further, the databases

are maintained independently and there is minimal coordinative effort to reach a consensus on the annotations of drug-target interactions.

To address these challenges, we have initiated a community-driven effort, called DrugTargetCommons (DTC), to collectively extract, annotate, manage and curate high-quality drug-target bioactivity data from literature and other resources. To facilitate the exchange of assay information, we developed a specific bioassay ontology that standardizes the description of drug-target interaction assays. Further, the DrugTargetCommons web-interface features a wiki function, enabling an interested investigator to not only upload new data from their own experiments or literature survey, but also to participate in the workflow of annotation, integration and quality control, together with the committed domain experts. Such an open environment ensures that most research discoveries will be manually curated, evaluated sufficiently, and cross-checked before being deposited into the DTC database.

As the first case study, we applied the DTC workflow to the FIMM Oncology Compound Collection, currently containing 461 anticancer agents including mainly kinase inhibitors. We manually-curated quantitative drug-target interaction data from ChEMBL, IUPHAR, canSAR, DrugKiNet and relevant literature, and utilized the recently-developed KiBA data integration method to derive a consensus score for each drug-target interaction. Compared to the existing drug-target databases, such as DrugBank and STITCH, the DTC platform showed an improved accuracy when capturing the polypharmacological profiles of these multi-targeted compounds. The manually curated and annotated bioactivity data are included in DTC, and will be made freely available soon.

The success of DTC relies on effective collaboration and open-data sharing with a common interest in expanding our knowledge of drug-target interactions. With an increasing number of research groups and pharmaceutical companies joining this effort, we anticipate that DTC will provide valuable resources for many exciting applications, including extended target identifications for drug discovery and drug repurposing opportunities.

[1] Abeyruwan,S. et al. (2014) Evolving BioAssay Ontology (BAO): modularization, integration and applications. J. Biomed. Semant., 5, S5.
[2] Bento,A.P. et al. (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res., 42, D1083–1090.
[3] Tang,J. et al. (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J. Chem. Inf. Model., 54, 735–743.

An Enhancer Atlas Explains the Transcriptional Regulation and Genetic Architecture Underlying Inflammatory Bowel Disease (IBD).

Malte Thodberg[1,2], Mette Boyd[1,2], Jette Bornholdt[1,2], Morana Vitezic[1,2], Kristoffer Vitting Serup[1,2], Mehmet Coskun[1,2,5], Robin Andersson[1], Thilde Bagger Terkelsen[1,2], Anders Gorm Pedersen[3], Jesper T. Troelsen[4], Jakob B. Seidelin[5], Jacob T. Bjerrum[5], Ole H. Nielsen[5], Albin Sandelin[1,2]

[1]Section for Computational and RNA Biology, Institute of Biology, UCPH, [2]Biotech Research and Innovation Centre, UCPH, [3] Centre for Biological Sequence Analysis, DTU, [4]Systems and Models, Department of Science and Environment, Roskilde University, [5]Department of Gastroenterology, Medical Section, Herlev Hospital.

Traditionally, much of disease genomics and transcriptomics have focused on the impact of coding or intragenic variation. However, a slew of recent scientific work has highlighted the importance of changes in intergenic regions (DNAse hypersensitive sites, chromatin marks, methylation, enhancers, etc.), which may affect the transcriptional regulation of genes and thereby transcript levels.

Cap-Analysis of Gene Expression (CAGE) is a promising novel sequencing technique which captures the expression of 5'-capped mRNA. While originally developed to measure promoter activity, it has recently been shown that CAGE can be used to reliably detect active enhancers by patterns of weak, but consistent bidirectional transcription of so-called enhancer RNAs (eRNAs).

IBD is a chronic intestinal disease of multifactorial origin. Despite having a strong genetic component, the pathogenesis of IBD is poorly understood, but believed to be a complex interplay of both genetic, luminal, and environmental factors that trigger an abnormal mucosal immune response to the gut microbiome.

In this interdisciplinary study we used CAGE data obtained from intestinal tissue biopsies from 94 IBD patients to build an IBD-specific enhancer atlas. Using this enhancer atlas we show that enhancer transcription can distinguish the inflammatory state of patients and that similarly regulated enhancers and promoters share transcription factor binding activity. Utilizing GWAS data we show that enhancers are more enriched for the heritability of IBD than promoters. Finally we show that putative regulatory interactions of individual enhancers and promoters can be computationally inferred.

In addition to providing insights into IBD pathogenesis and potential IBD biomarkers, this study also illustrates the general usefulness of CAGE for building disease-specific promoter and enhancer atlases.

Towards precision medicine for the treatment of cystinuria

Mark N Wass[1], Kathie Wong[2], Kay Thomas[2]

[1]*School of Biosciences, University of Kent, Canterbury Kent, UK.*
[2]*Urology Centre, Guy's and St. Thomas' NHS Foundation Trust, London, UK.*

Cystinuria is an inherited disease that results in the formation of cystine stones in the kidney. Two genes (SLC7A9 and SLC3A1) that form an amino acid transporter are known to be responsible for the disease. Variants that cause the disease disrupt amino acid transport across the cell membrane, which leads to the build up on relatively insoluble cystinine, leading to the formation of stones. In this project we have sequenced SLC7A9 and SLC3A1 in a cohort of patients from Guy's Hopsital, London, UK [1]. Structural and bioinformatics analysis of the variants identified was performed with the aim of identifying 1) how they alter transporter function and 2) how severe any effect they have on transport be may be to causing cystinuria symptoms [2]. This analysis was linked with the known symptoms of patients in the cohort with the intention that linking specific mutations with disease severity will enable us to infer the likely severity of new patients presenting with cystinuria and subsequently tailor the treatment they receive.

[1] Wong KA, et. al. (2015) The Genetic Diversity of Cystinuria in a UK Population of Patients. *BJU Int* 116(1):109-116.
[2] Wong KA, Wass M, Thomas K (2016) The Role of Protein Modelling in Predicting the Disease Severity of Cystinuria. *Eur Urol* 69(3):543-544.