GENOMIC MEDICINE

Chairs: Niko Beerenwinkel and Yves Moreau

H-1. DISCOVERY: a resource for the rational selection of drug target proteins and leads for the malaria parasite, Plasmodium falciparum

*Odendaal CJ (1), Harrison CM (1), Szolkiewicz MS (1), Joubert F(1, *)*

In the past, the selection of drug targets and lead compounds in malaria has been mostly based on serendipitous discoveries and legacy compounds. Few rational approaches have been successfully followed in especially the selection of promising drug target proteins in the parasite. The emergence of widespread drug resistance, even against current drugs is making the effective selection of new drug targets together with lead compounds essential and urgent, requiring optimal approaches to be put in place for this process.

Materials and Methods

The project is aimed at providing an informatics resource where comprehensive information on the parasite and host proteins are stored, together with the results from relevant 3rd-party investigations as well as from our own high-throughput analysis. Data included in the resource is aimed as wide as possible, including protein, gene-ontology, orthology, metabolic, structural, expression, interactome and chemoinformatics information. This is combined with an interface for researchers to perform the selection of putative drug target protein and lead compounds.

Results

Protein information includes data from the human, mosquito and the various malaria genome projects. Chemical information is from PDB, KEGG and DrugBank. Information includes basic annotations, motifs, domains, binding sites, structural features, orthology information, ontology terms, protein-protein interactions, protein-ligand interactions, pathogen-host interactions and comparative genomics information. Chemical information includes protein interactions and ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties.

Discussion

The researcher accessing the resource is able to perform advanced searching and filtering of proteins and chemical compounds according to possible interactions and the different types of properties described in the database. This may be initiated either from the protein or chemical compound as starting point. Additional work currently being performed includes the development of more accurate statistical scoring methods for the predictions, a literature mining component as well as the inclusion of additional chemical data sources.

URL

http://malport.bi.up.ac.za

Presenting Author

Fourie Joubert (fourie.joubert@up.ac.za)
University of Pretoria

Author Affiliations

(1) Bioinformatics and Computational Biology Unit, University of Pretoria, Pretoria, 0001, South Africa.

Acknowledgements

National Bioinformatics Network, National Research Foundation, Medical Research Council.

H-2. Variance estimators for t-test ranking influence the stability and predictive performance of microarray gene signatures

Touleimat N, Hernández-Lobato D (*), Dupont P

A Student t-Test is a standard statistical approach for ranking differentially expressed genes from microarray data. Shrinkage t-Test and Window t-Test have been proposed to improve the variance estimates when only very small data sets are available. The choice of variance estimator is known to influence gene ranking but we study here the effect on the classification performance of predictive models built from those gene signatures. We further assess the stability of those gene selection methods with respect to sampling variation.

Materials and Methods

We consider seven datasets from different cancer prognosis or diagnosis studies. Repeated 90% (training)-10% (test) resamplings are performed from each microarray dataset. Gene signatures of increasing sizes are estimated using the various t-Test ranking methods. Predictive models are built from the selected signatures with a nearest-centroid classifier using Pearson correlation. Comparative performances are reported in terms of the average between specificity and sensitivity over the test samples while selection stability is measured through the Kuncheva index.

Results

Our results show that signatures built with the three ranking methods provide models that have comparable predictive performances for signature sizes of more than 50 genes. For smaller signature sizes, models using the shrinkage t-Test offer better performance on 3 data sets (with no significant difference in the other cases). Shrinkage and Student t-Test clearly outperform the Window t-Test in terms of selection stability in most cases. Overall, the shrinkage t-Test offers the best ranking method when looking both at the classification performance and the stability of the selection.

Discussion

We study the classification performance of predictive models for cancer diagnosis or prognosis. A Student t-Test is commonly performed to select differentially expressed genes from microarray data. Due to the limited sample size, correcting the variance estimate may be beneficial in terms of predictive performance. We show that, for signature size with fewer than 50 genes, the shrinkage t-Test globally outperforms its competitors. We also observe that the optimal window size for the Window t-Test should be as small as possible, thus questioning the interest of this variant.

Presenting Author

Daniel Hernández-Lobato (daniel.hernandez-lobato@uclouvain.be)
Machine Learning Group, ICTEAM Institute, Université catholique de Louvain

Author Affiliations

Machine Learning Group, ICTEAM Institute, Université catholique de Louvain, Belgium.

H-3. Selecting small subsets of genes for predicting the outcome of chemotherapy treatments: a dynamic programming approach

Natowicz R (1,), Moraes Pataro C-D (2), Incitti R (3), Costa M-A (2), Cela A (1), Souza T (2), Braga A-P (2), Rouzier R (4)*

Accurate and robust predictors of the outcomes of the chemotherapy treatments are important for clinical purposes, for a better understanding of the biological mechanisms of the responses to the treatments, and for the developpement of new treatments dedicated to the patients who are non responders to the treatments that are presently available. The clinical trials provide tens of thousands of expression levels measured on very limited numbers of patient cases, while the expected property of robustness asks the predictors to be made out of very small sets of relevant genes.

Materials and Methods

133 patients with stage I-III breast cancer were included in a clinical trial. The gene pretreatment expression profiling was performed by Affymetrix U133A microarrays. A responder was a case with no histopathologic evidence of any residual invasive cancer cells in the breast. The genes were ranked by a univariate statistics then paired. Let Sn be the first n pairs, S a subset of Sn, and d(S) the minimum distance between any two responder and non responder cases relative to S. We search for subsets S* that maximize this minimum distance over the subsets of Sn.

Results

Upto n=4950 pairs of genes (made out of the 100 top ranked genes in the univariate statistics) no more than 24 genes were part of Sn (3-fold cross validation). In external validation, the predictors made out of these genes outperformed the best predictors reported so far on the same dataset (sensitivity = 0.92, specificity = 0.86). An example of optimal subset S* is : THRAP2 MAPT FBP1 AMFR JMJD2B ATPIF1 MELK HDC GREM1 THUMPD2 X61079 RLN2 SYNCRIP ZNF131 SLC43A3 IL21 SPDEF

Discussion

Up to now, no predictor of chemotherapy outcomes can be said to be used in clinical routine. To this end, the reliability of the predictors is a central concern. We think that high levels of robustness can be achieved by predictors made out of very small sets of highly informative and weakly correlated genes. In this work, we have modeled the selection of such small sets of genes by a combinatorial optimization approach, and we have proposed an efficient dynamic programming process to compute them. The results that we have obtained so far seem to be promising.

URL

http://www.esiee.fr/~natowicr/

Presenting Author

René Natowicz (rene.natowicz@univ-paris-est.fr)
Université Paris-Est - Esiee-Paris

Author Affiliations

(1) ESIEE-Paris, Université Paris-Est, Paris, France (2) Universidade Federal de Minas Gerais, Escola de Engenharia, Belo Horizonte, Brazil (3) INSERM IFR 10, Université de Paris Val-de-Marne, Hôpital Henri Mondor, Créteil, France (4) INSERM UPRES 4053, Université Pierre-et-Marie-Curie, Hôpital Tenon, Paris, France

Acknowledgements

This research is supported by CAPES-COFECUB, a French-Brazilian cooperation program.

H-4. Networks of generalized top scoring pairs for robust phenotype classification

Popovici V (1,), Budinska E (1), Delorenzi M (1)*

Recently, the top scoring pairs classification method has been proposed for building simple, yet robust, decision rules, based on the comparison of pairs of genes. While this simple approach works well on a number of cases, it suffers from the constraint interaction imposed on the variables. However, by exploring the graph structure of the relationships induced by the top pairs, we noticed that improved decision rules can be constructed by considering cliques of genes, all sharing a similar topology. Finally, we are interested in generalizing the bivariate model used.

Materials and Methods

We use the breast cancer dataset from the MAQC-II project, on which two endpoints are to be predicted: the estrogen receptor status and the pathologic complete response. The dataset consists of independent training and validation sets. We use two different version of the TSP algorithm: as originally proposed, and our generalized approach (bivariate logistic regression models). The top N scoring pairs induce a partial ordering of the genes, that can be represented as an oriented graph with special clique topologies. These cliques are used for building stronger decision rules.

Results

A generalized version of the top scoring pairs algorithm is proposed and implemented. Also, a novel method for combining individual pairs of variables is introduced. Its performance, assessed by stratified 5-fold cross-validation, compares favorably with other, more complex classifiers. Finally, the new algorithm was implemented in a C++ library and designed to take advantage of the multi-core/multi-processor architectures, reducing drastically the execution time.

Discussion

The top N scoring pairs induce a partial ordering among all the variable. This ordering can be represented as a directed more or less large directed graph, depending on the difficulty of the classification problem. This graph has a particular structure that enables us building more powerful decision rules. However, these rules retain the orignal interpretability, which makes the proposed approach extremely interesting.

Presenting Author

Vlad Popovici (vlad.popovici@isb-sib.ch)
Swiss Institute of Bioinformatics

Author Affiliations

(1) Bioinformatics Core Facility, Swiss Institute of Bioinformatics

Acknowledgements

VP and MD acknowledge the support of Swiss National Science Foundation NCCR Molecular Oncology.

H-5. Comparing network and pathway based classification for breast cancer: network and pathway based classifiers do not outperform single gene classifiers

*Staiger C (1, 2, *), Kooter R (3), Dittrich M (4), Müller T (4), Klau GW (1, 5), Wessels L (2, 3, 5)*

Recently many methods for the classification of cancer patients based on primary (mRNA) and secondary data (network data derived from protein-protein interaction (PPI) data or pathway data) have appeared. All these methods claim to improve the classification of breast cancer patients compared to single gene classifiers. So far there no study compared all of the methods against each other and against classification based on single genes only in an unbiased fashion. We selected three of the methods and compared them against single gene classifier.

Materials and Methods

We analysed the prediction of patient outcome of six breast cancer studies with distant metastasis free survival (DMFS) and overall survival (OS) as endpoints. We compare the approaches proposed by 1) Lee et al. which finds predictive subsets of genes from predefined gene sets contained in e.g. MSigDB; 2) Chuang et al. which is employs a greedy search to find predictive subnetwork from PPI data and 3) Heinz (Dittrich et al.), an approach based on an optimal search procedure to find the most predictive PPI subnetwork and 4) a single gene classifier. As secondary data we used HPRD and KEGG.

Results

We find that none of the three methods perform better than the single gene classifier on the breast cancer data. In contrast we show that the single gene classifier outperforms all the other network or pathway based approaches. Furthermore we do not find that the network or pathway based feature selection methods are more stable across datasets. The features found by Heinz have slightly higher average overlap as measured by the Jaccard index but also the variance of the overlap across datasets is higher.

Discussion

In contrast to Chuang et al. and Lee et al. we find that their approaches do not improve on single gene classifiers in prediction of breast cancer outcome. In both studies it was claimed that single gene classifiers achieved a lower AUC value than in our study. Another network based approach, Heinz, does not show an improvement over single genes either. In contrast to Chuang et al., we do not find that the network based methods stabilize the signatures. Taken together, our results suggest that the employed network-based approaches do not provide significant benefit in outcome prediction.

Presenting Author

Christine Staiger (c.staiger@cwi.nl)
(1, 2)

Author Affiliations

(1) Life Sciences group, CWI, Science Park 123, 1098 XG Amsterdam, Netherlands (2) Netherlands Cancer Institute, Division of Molecular Biology, Plesmanlaan 121 1066 CX Amsterdam (3) Delft Bioinformatics Lab, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, The Netherlands (4) Department of Bioinformatics, Biocenter, Am Hubland, D-97074 University of Würzburg, Germany (5) Shared last authorship

H-6. Enterotypes of the human gut microbiome

Raes J (1,2,6,), Arumugam M (1,6), The Metahit Consortium, Dore J (3), Weissenbach J (4), Ehrlich SD (3), Bork P (1,5)*

Our knowledge on species and function composition of the human gut microbiome is rapidly increasing, but it is still based on very few cohorts and little is known about their variation across the world. Computational analysis of gut metagenomics data allows first insights in the role of our flora in health and disease.

Materials and Methods

Here, we analysed the fecal metagenomes of 39 individuals from 6 countries at both functional and phylogenetic level.

Results

We identified several robust clusters (enterotypes hereafter) that are not nation or continent-specific. This further indicates the existence of a limited number of well-balanced host-microbial symbiotic states that might respond differently to diet and drug intake. The enterotypes are mostly driven by species composition, but the abundance of molecular functions detected therein does not necessarily correlate with the known abundant species, highlighting the importance of a functional analysis for a community understanding.

Discussion

While individual host properties such as disease state, age, or gender cannot explain the observed enterotypes, data-driven marker species, genes or pathways can be identified for each of these host properties. For example, Eubacterium abundance is linked to nationality, three functional modules significantly correlate with the body mass index and 11 genes change in abundance with age, hinting at a diagnostic potential of microbial markers. This stud shows the power of combined metagenomic and computational approaches in molecular medicine.

Presenting Author

Jeroen Raes (jeroen.raes@gmail.com)
VIB - Vrije Universiteit Brussel

Author Affiliations

1 European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. 2 VIB—Vrije Universiteit Brussel, 1050 Brussels, Belgium. 3 Institut National de la Recherche Agronomique, 78350 Jouy en Josas, France. 4 Commissariat à l'Energie Atomique, Genoscope, 91000 Evry, France. 5 Max Delbrück Centre for Molecular Medicine, D-13092 Berlin, Germany. 6 Contributed equally

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013): MetaHIT, grant agreement HEALTH-F4-2007-201052. JR is supported by the Odysseus programme of the Fund for Scientific Research Flanders (FWO).

H-7. Identifying chemotherapy resistance genes using outlier detection

de Ronde J (1,2,), Mulder L (2), Lips E (2), Rodenhuis S (3), Wessels L (1)*

Breast cancer is a heterogeneous disease and although gene expression analysis or immuno histo chemistry can be used to subclassify tumors into distinct subgroups, even within these subgroups there exist significant differences at the molecular level. When it comes to chemotherapy resistance, it appears highly unlikely that all resistant tumors share the same mechanism of action. Classical approaches have been unsuccessful in finding reliable biomarkers for therapy resistance and this may be due to the fact that these approaches are not sensitive to changes in small subgroups of samples.

Materials and Methods

Traditional analyses approaches, like a t-test or using the SAM approach, compare the two groups of samples for each gene, and identify genes with a significant difference in expression between the two groups. If only a small subset of the resistant tumors would show aberrant expression indicative of a resistance mechanism, then this would not be picked up by such approaches. Using a novel algorithm we try to circumvent this problem and specifically aim to find relatively small subgroups of tumors within the resistant group that show differential expression compared to the sensitive group.

Results

Using a positive control set, where we mixed in a small set of HER2-positive tumors in a larger HER2-negative set and by generating an artificial dataset we show that our algorithm is able to pick up all positive controls. These controls are not picked up by either a t-test or the SAM algorithm. Next, we applied the algorithm on a gene expression set of 195 patients that were neoadjuvantly treated and identified a number of genes that show an expression pattern that suggests a role in chemotherapy resistance in ER+ tumors. We validate our findings on a separate set of 90 patients.

Discussion

We have developed a novel algorithm that allows the identification of genes that show aberrant expression, relative to the reference group, in a small subset of the samples. Using this approach we identified a number of genes that are linked to chemotherapy resistance. The algorithm can be used in any type of analysis that involves the detection of small subsets of samples within one of the labeled classes, where the small subset shows behavior different from the average behavior of the remaining samples. Our approach can therefore be applied to a wide range of problems.

Presenting Author

Jorma J. de Ronde (j.d.ronde@nki.nl)
The Netherlands Cancer Institute, Amsterdam, The Netherlands

Author Affiliations

1. Department of Bioinformatics and Statistics, The Netherlands Cancer Institute, Amsterdam, The Netherlands 2. Department Experimental Therapy, The Netherlands Cancer Institute, Amsterdam, The Netherlands 3. Department of Medical Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Acknowledgements

This study was performed within the framework of CTMM, the Center for Translational Molecular Medicine (www.ctmm.nl), project Breast CARE (grant 03O-104).

H-8. Genome wide association study of non-synonymous single nucleotide polymorphisms for seven common diseases

Surendran P (1,2,), Shields D (1), Stanton A (2)*

Associations of several Single Nucleotide Polymorphisms (SNPs) with seven common diseases were identified in a study conducted by the Wellcome Trust Case Control Consortium. We hypothesize that there are more chances of finding association of rare SNPs with diseases by refined analysis of non synonymous SNPs (nsSNPs) in genome wide association studies and expect better results in WTCCC data by using larger SNPs datasets which were published after the original WTCCC study. In the present study we analyzed the association of 12,660 nsSNPs using a case control study in the WTCCC population.

Materials and Methods

We simulated the genotypes at 10,798 nsSNP loci studied by the Stage 2 HapMap project using the genotype information from WTCCC for all 14,000 individuals studied for seven diseases and in 3000 controls. We performed these simulations or imputations of missing SNPs using two widely used programs called IMPUTE and MACH after standard quality control checks using programs in PLINK genome wide analysis package. Subsequent case control association of 10,798 imputed nsSNPs and 1,862 genotyped nsSNPs was performed using an additive model and genotype model in a frequentist and bayesian framework.

Results

We found 4 nsSNPs associated with Crohn's Disease, 8 with Rheumatoid Arthritis, 5 with Type 1 Diabetes and 1 with Type 2 Diabetes. In total, 18 new associations with the seven diseases ($p < 5 \times 10\text{-}6$) studied by WTCCC. We also developed a pipeline which summarizes quality control measures which should be considered to minimize false associations in genome wide association studies.

Discussion

Among the 18 SNPs which are found associated with diseases, 5 SNPs are found in genetic loci which were not identified in the previous WTCCC study. All the other SNPs are found adjacent to the SNPs previously identified and are in high linkage disequilibrium with the significant SNPs. These non synonymous SNPs can play a significant role in modulating protein-protein interactions with substitutions in interacting regions of proteins like in domains, motifs or disordered regions of proteins. Such relations are of critical importance in direct or indirect disease causation.

Presenting Author

Praveen Surendran (praveen.surendran@ucd.ie)
University College Dublin

Author Affiliations

1 School of Medicine and Medical Science, University College Dublin, Belfield, Dublin 4, Ireland. 2Department of Clinical Pharmacology, Royal College of Surgeons in Ireland, Dublin-1, Ireland.

Acknowledgements

This work was funded by Embark Initiative, Graduate Research Education Program of Irish Research Council for Science Engineering and Technology. Computational resources were provided by ICHEC, Ireland's High-Performance Computing Centre.

H-9. Comparative bioinformatics approach to multiple tumors reveals novel prognostic markers in breast cancer

Krupp M (1,), Marquardt J (2), Maas T (1), Galle P (1), Tresch A (3), Teufel A (1)*

Over the last decade microarray studies have extensively been applied in tumor research and the obtained results have lead to a more systemic view on tumor development. However until now only partial efforts have been performed to analyze this data in context of comparative genomics between multiple tumor entities to unravel similarities in the gene expression profiles and subsequently their biological behaviour. Ultimately, such investigation may lead to a deeper understanding and extended application of target therapeutic approaches as well as to the discovery of potential prognosis markers.

Materials and Methods

1402 tumor associated microarrays were downloaded from the Standford Microarray Database and analysed for enriched pathway modules obtained from the Kyoto Encyclopaedia of Genes and Genomes. Furthermore we adopted survival data on 99 breast cancer patients with a follow up of 6.1 years. This data comprises 7650 probes which can be assigned to 33 patients with bad and 66 with good outcome. Determination of possible pathways associated with patients survival was calculated by applying the gene set enrichment analysis (GSEA) computational method to the data.

Results

We analysed a dataset of 649 tumor associated microarrays, corresponding to 16 tumor entities. Analyzing these data, we found several signaling pathways referenced by KEGG to be significantly enriched. Two pathway categories were remarkably conserved in the tumors leading to the assumption that they are essential key player in tumor progression. On this basis we adopted breast cancer prognostic data to our dataset and the performed GSEA to the differentially expressed genes located in those conserved pathway categories has shown new prognostic markers to this tumor entity.

Discussion

Together, we provide a comparative analysis of pathways enriched in multiple tumor entities and disclosed new prognostic markers by applying gen set enrichment analysis to pathway groups highly conserved among the tumors. Furthermore we summarize the multiple tumor pathway relations in a clearly structured tumor pathway map enabling scientist to get a systematic overview about tumor affinity and central tumor mechanisms.

Presenting Author

Markus Krupp (kruppm@uni-mainz.de)
Department of Medicine I, Johannes Gutenberg University, Mainz, Germany

Author Affiliations

1 Department of Medicine I, Johannes Gutenberg University, Mainz, Germany 2 National Cancer Institute, National Institutes of Health, United States 3 Gene Center Munich, Department of Chemistry and Biochemistry, Ludwig-Maximilians-University, Munich, Germany

H-10. Genomic and epigenomic molecular signatures reveals network mechanisms associated with ovarian cancer prognosis

Ben-Hamo R, Efroni S*

Epithelial ovarian cancer causes more deaths than any other female gynecologic cancer. A better understanding of the molecular mechanisms in advanced ovarian cancer may improve patient treatment. Identification of molecular interactions that stratify prognosis may be the key for such novel treatments. Gene methylation, copy number variation and gene expression are characterizing factors in malignancies. The Cancer Genome Atlas (http://cancergenome.nih.gov/), a large multi center coordinated effort, has recently made available the molecular characteristics of more than 200 patients.

Materials and Methods

All data obtained from The TCGA and validation data obtained from the Duke Cancer Center. Data consists of CNV, Methylation, Gene expression and clinical annotation. Pathway Consistency and pathway Activity metrics, which overlay expression data over network knowledge, were calculated according to (Efroni,et al,2007). Pathway targeting by genomic and epigenomic alterations were calculated according to Fisher's omnibus test. Metrics for the set of these highly significant pathways were then clustered to stratify patients into two groups. Survival analyses was done using Kaplan-Meier.

Results

Using this multi analyte study and a set of computational algorithms we have recently developed, we were able to identify subnetworks that significantly stratify survival rates. Expression levels of single genes do not explain the prognostic stratification. Only when we apply network effects phenotypically distinct groups emerge. Our results demonstrate specific pathway interactions whose synthetic genetic and epigenetic interactions drive prognosis in ovarian cancer through combined alterations in expression. These interactions are associated with disease outcome.

Discussion

Here, by combining CNV, Methylation and gene expression features we were able to efficiently stratify prognosis. The selection of a set of subnetworks whose transcriptional behavior successfully stratifies the prognosis can identify the most influential molecular agents in the network and the key interactions. By isolating specific subnetworks, we were able to handle the NP-hard numeracy of network interactions. Further analysis revealed specific interactions at the core of the phenotypic clustering.

Presenting Author

Rotem Ben-Hamo (rotem830@walla.com)
The Mina and Everard Goodman Faculty of Life Science, Bar Ilan University, Ramat-Gan, 52900, Israel.

Author Affiliations

The Mina and Everard Goodman Faculty of Life Science, Bar Ilan University, Ramat-Gan, 52900, Israel.

H-11. Inferring distributions of trait-associated SNPs with application to genetic association studies

Neuvirth H (1), Aharoni E (1), Azencott C-A (1,2), Farkash A (3), Landau D (1), Rosen-Zvi M (1,), Geiger D (1,4)*

Hundreds of recent genetic association studies provide an increasingly more viable opportunity to derive generalizing principles regarding SNPs' association potential and utilize these inferences for improving new GWAS studies.

Materials and Methods

Several features of SNPs have been identified as being significantly over-represented in trait-associated SNPs (TASs) based on published data for 19 diseases. A regression model based on these features and the published data has been constructed. For each SNP in a given SNP panel, the model provides the probability of being relevant to a trait under study, before the SNP data of this study is examined. The resulting probabilities are used as weights that increase or decrease the needed threshold for the p-value of each SNP using standard association methods.

Results

For 16 out of the 19 diseases, SNPs judged to be associated with the disease in previous studies have been promoted by the computed weights, and were practically unchanged for the remaining disease. Repeating a real study analysis of type-2 diabetes showed an increase in the number of relevant SNPs found for every sample size in this study.

Discussion

This work is the first effort to perform the full supervised analysis. It applies large scale learning models that generalize a sizeable collection of studies in order to generate an a-priori prioritization of all SNPs on a given chip and consequently improve the outcomes of new GWAS studies. Our method allows a smaller sample to be collected for GWAS studies and enhances the statistical power for a fixed sample size.

URL

http://srv-rimon.haifa.il.ibm.com:8080/rimon_web/snpWeights.jsp

Presenting Author

Michal Rosen-Zvi (rosen@il.ibm.com)
Machine Learning and Data Mining group, IBM Haifa research labs

Author Affiliations

(1) Machine Learning and Data Mining Group, IBM Research Laboratory in Haifa, Israel. (2) School of Information and Computer Sciences, Institute for Genomics and Bioinformatics, University of California, Irvine, California 92697-3435 (3) IT for Healthcare and Life Sciences Group, IBM Research Laboratory in Haifa, Israel. (4) Technion - Israel Institute of Technology, Computer Science Department, Haifa, 36000, Israel

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007-2013 under the project HYPERGENES, grant agreement n° 201550. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

H-12. Optimizing exon CGH array designs for robust rearrangements detection

Gambin T(1,), Sykulski M(2), Gambin A(2)*

Array-comparative genomic hybridization (aCGH) technology enables rapid, high-resolution analysis of genomic rearrangements. With the use of it, genome copy number changes and rearrangement breakpoints can be detected and analyzed at resolutions down to a few kilobases. An exon array CGH approach proposed recently accurately measures copy-number changes of individual exons in the human genome. The crucial and highly non-trivial starting task is the design of an array, i.e. the choice of appropriate set of oligos. The success of the whole high-level analysis depends on the design quality.

Materials and Methods

The dataset, we used in our study, come from 60 arrays hybridized with DNA from subjects with epilepsy, autism, heart defects and mental disorders. Each experiment was performed on the 180 K exon targeted oligonucleotide array. Based on those data we generated several smaller (optimized) designs by selecting various subsets of oligos from original one. Aiming in testing the robustness of segmentation we enhance the DNAcopy method by incorporating parametrized noise model. Finally, we used this algorithm to perform the comparison among reduced designs.

Results

For each probe we computed, using a set of Kolmogorov-Smirnov tests, cumulative properties, which reflects the oligo suitability in the context of its surrounding. To investigate the influence of design optimization strategy on segmentation robustness several approaches for probes selection were tested, including uniform sampling and most/least suitable oligo removal. Some of those methods reduced the nr of probes with a little loss of segmentation robustness. One can benefit from this strategy especially for targeted arrays used for the diagnosis of specific chromosomal aberrations.

Discussion

The investigation shows that while optimizing the design it is crucial to find a tradeoff between keeping uniform distribution and selecting the best performing probes. We discovered that the results of designs comparisons greatly depends on the definition of distance between two segmentations. Finally, we found new robustness measure very useful in evaluation of optimized design performance of rearrangement detection, and its resistance to the noise, in comparison to the original array.

URL

http://bioputer.mimuw.edu.pl/papers/robustness/index.php

Presenting Author

Tomasz Gambin (tgambin@gmail.com)
Institute of Computer Science, Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland

Author Affiliations

(1) Institute of Computer Science, Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warsaw, Poland (2) Institute of Informatics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland

Acknowledgements

This research is supported in part by Polish Ministry of Science and Educations grants N301 065236 and N206 356036. It is also supported by the Foundation for Polish Science and the European Social Fund and the State Budget from the Integrated Regional Operational Program, Action 2.6 "Regional Innovation Strategies and Knowledge Transfer", the project of Mazovia Voivodeship "Mazovia Doctoral Scholarship".

H-13. Interactive human brain atlas for a better understanding of diseases

Kremer A (1), Gaenzler C (2) and van der Spek P (1)*

The bioinformatics department within the Erasmus MC is supporting many large scale clinical research projects. One major challenge is the integration of genomics, proteomics and cytogenetic data with medical imaging data to identify genes associated with neuro-development, neuro-degeneration and neuro-oncology. The aim of our research is to gain insight in the molecular and cellular mechanisms driving the development of the face and the brain.

Materials and Methods

The ever increasing complexity and amount of research data require new approaches to data analysis and visualization. Here we describe an application which was developed together with TIBCO Spotfire that facilitates the integration and visualization of gene expression data in the context of the brain anatomy.

Results

These new brain maps are useful research tools and are an example for the successful integration of medical and biological data.

Discussion

Linking molecular with imaging data opens new avenues for image guided diagnosis and intervention.

Presenting Author

Andreas Kremer (a.kremer@erasmusmc.nl)
Erasmus MC

Author Affiliations

(1) Erasmus Medical Center, Department Bioinformatics (2) TIBCO Software GmbH

H-14. Clinical prognosis with transcriptional association networks and regression trees

Nepomuceno I (1,), Azuaje F (2), Devaux Y (2), Nazarov PV (3), Muller A (3), Aguilar-Ruiz JS (4), Wagner DR (2,5)*

The application of information encoded in molecular networks for prognostic purposes is a crucial objective of systems biomedicine. This approach has not been widely investigated in the cardiovascular research area where the prediction of clinical outcomes after suffering a heart attack would represent a significant step forward. We developed a new supervised prediction method for this prognostic problem based on the discovery of clinically-relevant gene association networks. The method integrates model trees and clinical class-specific networks. It can be applied to other clinical domains.

Materials and Methods

Our method consists of two main steps. The first involves the inference of class-specific networks. Each network is inferred from gene expression of patients with the same clinical class. In order to do it, each gene is analyzed by taking into account the remaining genes as inputs to M5P model tree algorithm. M5P focus on localized similarities. The second step involves predicting the clinical class of a new patient through the inferred networks. The prediction is based on the relative error between the true and predicted expression values of those genes involved in the inferred networks.

Results

We assessed the predictive performance of our approach on a benchmark dataset and we compared the predictive capability against other classifiers applied on the network information inferred from several techniques based on co-expressed gene modules. The predictive strength is reflected in a high AUC value (0,85). Furthermore, we analyzed the gene expression profiles of patients with acute MI that is divided into good and bad prognosis groups. We also detected biological processes which may be used to characterise and possibly treat the development of ventricular dysfunction after MI.

Discussion

Our method discovered small biologically-meaningful networks, which facilitate human expert interpretation. Furthermore, we predicted processes that may represent novel therapeutic targets for heart disease, such as the synthesis of leucine and isoleucine. These results encouraged us to investigate our method as a new strategy to discover potential biomarkers of clinical outcome after MI and one of the strengths of this method is that it can be applied to other prognostic problems. As drawback of this work, we can mention the limited number of gene expression profiles under analysis.

Presenting Author

Isabel Nepomuceno (inepomuceno@us.es)
Universidad de Sevilla

Author Affiliations

1 Dpt. Lenguajes y Sistemas Informáticos, Universidad de Sevilla,Spain 2 Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg 3 Microarray Center, CRP-Santé, Luxembourg 4 School of Engineering, Pablo de Olavide University, Spain 5 Division of Cardiology, Centre Hospitalier, Luxembourg

Acknowledgements

In Luxembourg this research was in part supported by: Fonds National de la Recherche (Luxembourg), Société pour la Recherche sur les Maladies Cardiovasculaires, and Ministére de la Culture, de l'Enseignement Supérieur et de la Recherche. In Spain, it was in part supported by the Spanish Ministry of Science and Innovation under Grant TIN2007--68084--C02--00, and by the Plan Propio of the University of Seville.

H-15. Prediction of clinical outcome after cardiac arrest and induced therapeutic hypothermia

Nepomuceno I (1,), Azuaje F (2), Devaux Y (2), Stammet P (3), Aguilar-Ruiz JS (4), Wagner DR (2,5)*

Standard approaches to biomarker discovery are based on the identification of differentially expressed genes. However, it is known that known biomarkers may be encoded by genes that are not highly differentially expressed across control and disease patients. Network-based prognostic approaches have not been widely investigated in the cardiovascular research area where the prediction of clinical outcome would represent a significant contribution to translational research. We developed a new supervised prediction method based on the discovery of clinically-relevant gene association networks.

Materials and Methods

Our method consists of two main steps. The first involves the inference of class-specific networks. Each network is inferred from gene expression profiles of patients with the same clinical class. In order to do it, each gene is analyzed by taking into account the remaining genes as inputs to M5P model tree algorithm. The second step involves predicting the clinical class of a new patient through the inferred transcriptional networks. The prediction is based on the relative error between the true and predicted expression values of those genes involved in the inferred networks.

Results

We analyzed gene expression profiles of blood cells obtained from 35 CA patients treated with therapeutic hypothermia: 21 patients had a normal brain function and 14 patients showed signs of brain dysfunction after hypothermia. Our method discovered small biologically-meaningful networks, which facilitate human expert interpretation. An average classification accuracy of 75% was obtained after leave-one-out cross-validation. We also detected biological processes which may be used to characterize and possibly prevent the adverse neurological outcome in comatose survivors.

Discussion

A key outcome of this investigation was the discovery of new potential biomarkers for predicting neurological outcome after CA. Although the method has been used to predict favourable and unfavourable prognosis, it can be extended to infer more than two clinical outcomes.. The predictions reported here will require to be validated in larger and independent patient cohorts. Further in vitro investigations will be necessary to test the functional significance of the biomarkers discovered in this study and to determine their involvement in the appearance of neurological sequels.

Presenting Author

Isabel Nepomuceno (inepomuceno@us.es)
Universidad de Sevilla

Author Affiliations

1 Dpt. Lenguajes y Sistemas Informáticos, Universidad de Sevilla,Spain 2 Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg 3 Dpt. Anaesthesia and Intensive Care, Centre Hospitalier, Luxembourg 4 School of Engineering, Pablo de Olavide University, Spain 5 Division of Cardiology, Centre Hospitalier, Luxembourg

Acknowledgements

In Luxembourg this research was in part supported by: Fonds National de la Recherche (Luxembourg), Société pour la Recherche sur les Maladies Cardiovasculaires, and Ministére de la Culture, de l'Enseignement Supérieur et de la Recherche. In Spain, it was in part supported by the Spanish Ministry of Science and Innovation under Grant TIN2007--68084--C02--00, and by the Plan Propio of the University of Seville.

H-16. Allele-specific copy number analysis of breast carcinomas

Van Loo P (1,2,3,12,), Nordgard SH (1,4,12), Lingjærde OC (5), Russnes HG (1,6,7), Rye IH (6), Sun W (4,8), Weigman VJ (4), Marynen P (3), Zetterberg A (9), Naume B (10), Perou CM (4), Børresen-Dale A-L (1,7,13), Kristensen VN (1,7,11,13)*

Whole genome SNP genotyping is an expanding technology to measure genomic aberrations in an allele-specific manner. However, to accurately index all genomic aberrations in a cancer sample, both the ploidy of the cancer cells and the infiltration of non-aberrant cells need to be accounted for in the analysis. We infer tumor ploidy, non-aberrant cell admixture and genome-wide allele-specific copy-number profiles from genome-wide SNP data of breast cancers and identify specific signatures of aberrations in breast carcinoma and breast carcinoma subtypes.

Materials and Methods

We performed genotyping of 112 breast carcinoma samples using Illumina 109K SNP arrays and constructed an algorithm (ASCAT, Allele-Specific Copy number Analysis of Tumors) to estimate the fraction of aberrant cells and the tumor ploidy, and to index all genomic aberrations taking both properties into account. This allows calculation of "ASCAT Profiles" (genome-wide allele-specific copy-number profiles) from which gains, losses, copy-number-neutral events and LOH can accurately be determined.

Results

We present the first allele-specific copy number analysis of the in vivo breast cancer genome. We obtained ASCAT Profiles for 91 of the breast carcinomas (81 %). We observe aneuploidy (>2.7n) in 45% of the cases and an average non-aberrant cell admixture of 49%. We obtain first-time genome-wide views of LOH and copy-number-neutral events in breast cancer. In addition, the ASCAT Profiles reveal differences in aberrant tumor cell fraction, ploidy, gains, losses, LOH and copy-number-neutral events between the five previously identified molecular breast cancer subtypes.

Discussion

Basal-like breast carcinomas have a significantly higher frequency of LOH compared to other subtypes, and their ASCAT Profiles show large-scale loss of genomic material during tumor development, followed by a whole-genome duplication, resulting in near-triploid genomes. Finally, from the ASCAT Profiles, we can construct a genome-wide map of allelic skewness in breast cancer, indicating loci where one allele is preferentially lost while the other allele is preferentially gained. We hypothesize that these alternative alleles have a different influence on breast carcinoma development.

Presenting Author

Peter Van Loo (Peter.VanLoo@med.kuleuven.be)
VIB and K.U.Leuven

Author Affiliations

1. Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway; 2. Department of Molecular and Developmental Genetics, VIB, Leuven, Belgium; 3. Department of Human Genetics, University of Leuven, Leuven, Belgium; 4. Department of Genetics, University of North Carolina, Chapel Hill, NC, USA; 5. Department of Informatics, University of Oslo, Oslo, Norway; 6. Department of Pathology, Oslo University Hospital Radiumhospitalet, Oslo, Norway; 7. Institute for Clinical Medicine, University of Oslo, Oslo, Norway; 8. Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA; 9. Department of Oncology–Pathology, Karolinska Institutet, Stockholm, Sweden; 10. The Cancer Clinic, Oslo University Hospital Radiumhospitalet, Oslo, Norway; 11. Institute for Clinical Epidemiology and Molecular Biology (EpiGen), University of Oslo, Oslo, Norway; 12. These authors contributed equally to this work; 13. These authors share senior authorship.

AUTHOR INDEX