OTHER BIOINFORMATICS APPLICATIONS

Chairs: Martin Vingron and Hans-Peter Lenhof

I-1. Universal virtual screening

Onodera K (*), Kamijo S

Structure-based virtual screening is one of the most promising technologies in drug discovery. Many molecular docking programs and scoring functions have been developed, but more improvements are required for practical usages. One way to improve the accuracy of the screening is the consensus scoring method. It combines docking scores from various scoring functions without considering characteristics of the docking scores.

Materials and Methods

In this study, we use the concepts of the consensus scoring to improve the docking scores from docking programs (DOCK, FRED, and GOLD) in virtual screening. While those docking scores are simple sum of each score component, we introduced weight factors of the score components for scoring calculations, and adjusted for better predictions of true actives during virtual screening. After several optimization processes, we found the best optimized scores using a wide variety of 113 target proteins with over 2000 diverse decoys.

Results

The success rates finding true actives were improved by up to 52.4% (e.g. from 36.8% to 56.1% in GOLD) for the test set. Also, the combination of the optimized scores from GOLD and FRED improved success rate by 77.2%. In total, the combination of optimized scores now predict about 70% of true actives for target proteins in the test set with 20 times enrichment.

Discussion

Universal optimization of docking scores certainly reduces chances to enrich non-specific compounds in the hit list erroneously. At the same time, it increases chances to acquire novel drugs for drug discovery. Our optimizing methods enhance performance of virtual screening using molecular docking programs currently available in our community.

Presenting Author

Kenji Onodera (onoderak@iis.u-tokyo.ac.jp)
IIS, University of Tokyo

Author Affiliations

Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo 153-8505, Japan.

Acknowledgements

This work is supported by grant-in-aids for Scientific Research (19710173 and 22651079).

I-2. Accuracy and computational efficiency of a graphical modeling approach to linkage disequilibrium estimation

Abel H J (1), Thomas A (1,)*

The increasing availability of genome-wide dense single-nucleotide polymorphism (SNP) data provides an abundance of information for gene mapping studies. However, along with the benefits this new information provides come the difficulties posed by allelic associations or linkage disequilibrium (LD). Existing methods for modeling LD take time and storage of O(nnm) and O(nm) where n is the number of individuals in the reference sample and m is the number of SNPs being considered. These methods are not usable for current data where n and m are around 1000 and 100,000.

Materials and Methods

We use a Markov chain Monte Carlo search method to fit a graphical model. Parameter values were chosen by cross validation to maximize the predictive accuracy when imputing masked genotypes. We used data sets with 60 individuals genotyped at between 20 and 200,000 SNP loci to check that the programs scale linearly with the number of loci. We used data simulated on up to 1000 individuals to check for linear scaling with sample size.

Results

While our approach to model estimation and imputation is marginally less accurate than existing methods, it is considerably more computationally efficient as shown by several systematic benchmarking experiments. For a reference set of n individuals genotyped at m markers the time and storage required for fitting a graphical model are O(nm) and O(n+m) respectively. To impute the phases and missing data on n individuals using an already fitted graphical model requires O(nm) time and O(m) storage.

Discussion

Two important features contribute to the computational efficiency: separating the fitting and imputation processes into different programs, and holding in memory only the data within a window of loci during fitting. Note that while the time for fitting and imputation are both O(nm) imputation is considerably faster, thus, once a model is estimated from a reference data set, the marginal cost of phasing and imputing further samples is very low. Previous methods combine the estimation and imputation stages in implementations requiring time of O(nnm) and storage of O(nm).

Presenting Author

Alun Thomas (Alun.Thomas@utah.edu)
Division of Genetic Epidemiology, University of Utah

Author Affiliations

(1) Division of Genetic Epidemiology, University of Utah

Acknowledgements

This work was supported by grants NIH R01 GM81417, and DOD W81XWH-07-01-0483 to Alun Thomas.

I-3. Stepwise classifier for heterogeneous genomic data

Wubulikasimu A (*), van de Wiel MA

Combining heterogeneous data types and produce accuracy which cannot be attained by either one of these data types alone is a hot topic in classification problem. However, there is a shortage of effective and efficient statistical and bioinformatic tools for truly integrative data analysis. Existing integrative classifiers have two main disadvantages: Firstly, coarsely combination in which the potential subtle contributions of one data type likely to be overcome by more obvious contributions of other data types. Secondly, huge computational costs because of sophisticated algorithms.

Materials and Methods

We aim to capture the distinct prediction power of each data type in such a way that they are complementary rather than redundant to each other. We first use the easy to get or economically cheaper data in first stage and only turn to more expensive and noisy data type when local error rate of samples below certain threshold. Local error rate and optimal threshold value will be calculated in the training phase and apply for test samples. By this way we can save considerable amount of the samples reclassify by expensive data in second stage, at the same time keep accuracy as high as possible.

Results

Potential of our method is illustrated on publicly available data sets. For the setting with aCGH and GE data we show that including GE to aCGH data can enhance the prediction accuracy significantly, although a fair proportion of the samples need not to be reclassified, and hence no GE measurements would be needed for those. We come to a similar conclusion for a setting with data sets containing both clinical variables and genomic ones, which implies a serious potential economical and practical benefit.

Discussion

The Stepwise classification method we introduced is practically very attractive. It dos not require specific type of classification algorithm, this give user more freedom in terms of algorithm. Experimental results show that even if we choose wrong algorithm, because we don’t know beforehand which algorithm is suitable for the problem at hand, which produce very worse result in first or second stage, we will not get dramatic decrease in our stepwise approach. This is because it will pass more samples to the data type which produces better result by adjusting the threshold value.

Presenting Author

Askar Wubulikasimu (askar.wubulikasimu@vumc.nl)
Department of Epidemiology and Biostatistics ,VU University medical center

Author Affiliations

Department of Epidemiology and Biostatistics VU University medical center Amsterdam The Netherlands

Acknowledgements

This research was supported by the Center for Translational Molecular Medicine (DeCoDe project)

I-4. Fitness differences as an explanation for difference between chronic myeloid leukemia therapies

Lenaerts T (1,2,), Castagnetti F (3), Traulsen A (4), Pacheco JM (5,6), Rosti G (3), Dingli D (7)*

Since the successful introduction of Imatinib as the treatment of chronic myeloid leukemia (CML), several more potent kinase inhibitors (TKI) like Nilotinib have been developed. Nilotinib binds with a higher affinity to the BCR-ABL oncoprotein and it also is able to block many Imatinib-resistant mutations. Despite these molecular effects, in vitro studies suggest that there is no difference in the inhibition of signaling downstream of BCR-ABL. How can one explain then the faster and deeper response observed in the clinical data [1] ?

Materials and Methods

We have developed a computational model of the hematopoietic system in which we can study the population dynamics of different types of cell: healthy, cancer and TKI treated cells. Normal hematopoiesis is represented by a hierarchical model in dynamic equilibrium where cells move along the hematopoietic tree as they become increasingly differentiated. In this work we use clinical data [1] to predict the overall amount of cells affected by the treatment and the effect of the drug on the differentiation rate of the cancer cells.

Results

We show that the fraction of CML cells responding to nilotinib is slightly higher than for imatinib (8.5% versus 5%), capturing the higher affinity of Nilotinib for BCR-ABL. Yet this in itself does not explain the deeper response for this therapeutic agent. By expressing the difference between the two drugs in terms of their fitness in relation to the fitness of normal hematopoietic cells [2] (1.0) we show the fitness of Nilotinib treated cells (0.49) is lower than that for Imatinib treated cells (0.63). As such, Nilotinib treated cells will disappear more quickly from the overall population

Discussion

Since Nilotinib treated cells will disappear more quickly from the overall population of cells than Imatinib treated cells, the response of the patient will reach lower disease levels. This explanation can only be observed by studying the interplay between the different cell populations in the hematopoietic system, providing an illustration of the power of our model to study the dynamics of tumor growth and response to therapy. [1] G. Rosti et al (2009) Blood 114(24):4933-4938 [2] D. Dingli, A. Traulsen, T. Lenaerts and J.M. Pacheco (2010) Genes & Cancer 1(4):309-315

URL

http://www.ulb.ac.be/di/map/tlenaert/

Presenting Author

Tom Lenaerts (tlenaert@ulb.ac.be)
MLG, DI, Université Libre de Bruxelles

Author Affiliations

1 MLG, Département d'Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium. 2 Computer Science Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. 3 Department of Hematology and Oncology, "L. and A. Seràgnoli", St Orsola University Hospital, Via Massarenti, 9 - 40138 Bologna 4 Evolutionary Dynamics Group, Max-Planck-Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany. 5 Departamento de Matematica e Aplicações, Universidade do Minho, 4710-057 Braga, Portugal. 6 ATP-Group, Centro de Matemática e Aplicações Fundamentais, Complexo Interdisciplinar, 1649-003 Lisboa codex, Portugal. 7 Division of Hematology, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA.

Acknowledgements

TL acknowledges the support of the FNRS (Belgium), JMP acknowledges the support of FCT (Portugal), AT is supported by the Emmy-Noether program of the German Research Foundation and DD is supported by an Early Career Development Award. We all acknowledge the GIMEMA CML working party for providing the data.

I-5. ALADYN: a web server for aligning proteins by matching their large-scale motion

Potestio R (1,), Aleksiev T (1), Pontiggia F (2), Cozzini S (3,1), Micheletti C (1,3)*

The characterization of proteins and enzymes has largely benefited from the large-scale application of sequence- and structure-based alignment methods. These tools proved extremely valuable to identify salient features shared by proteins differing at the level of primary sequence or three-dimensional organisation. Our group has recently shown that the comparison of proteins' internal dynamics can be used to pin-point functionally-oriented relationships that would otherwise be elusive to sequence- or structure-based alignments.

Materials and Methods

To detect these common dynamical features we align proteins by matching their large-scale internal motion. The latter is reliably and efficiently computed using elastic network models. The dynamics-based alignment method is offered as a web-server, named ALADYN and is available at: http://aladyn.escience-lab.org . The server takes as input a pair of protein structures and presents an interactive visualization of the amino acids that are in good structural and dynamical correspondence.

Results

We discuss the application of the dynamics-based alignment to various members of enzymatic superfamilies, such as hydrolases. In particular we illustrate the case of HIV1-PR/Beta-secretase and exonuclease III/human adenovirus proteinase.

Discussion

The dynamics-based alignment sever lends naturally to be used as a quantitative tool for investigating the role of dynamics in protein function, and establish nontrivial relations among proteins performing similar functions in spite of different structural organisation.

URL

http://aladyn.escience-lab.org

Presenting Author

Raffaello Potestio (potestio@sissa.it)
SISSA

Author Affiliations

1) Scuola Internazionale Superiore di Studi Avanzati and eLab, via Bonomea 265, 34136 Trieste, Italy 2) Department of Biochemistry, Howard Hughes Medical Institute, Brandeis University, Waltham, Massachusetts - 02454, USA 3) Democritos CNR-IOM, via Bonomea 265, 34136 Trieste, Italy

Acknowledgements

CNR-IOM Democritos and the Italian Ministry of Education.

I-6. Stochastic extinction of leukemic stem cells provides a road to cure chronic myeloid leukemia

Lenaerts T (1,2,), Traulsen A (3), Pacheco JM (4,5), Dingli D (6)*

Tyrosine kinase inhibitors (TKI), of which Imatinib is a first, play a pivotal role in the treatment of chronic myeloid leukemia (CML). Notwithstanding, their therapeutic success, none of the TKI are considered to be curative. This view follows from the experimental observation that they do not affect leukemic stem cells (LSC), which are at the origin of the disease. As a consequence it is assumed that only by removing the LSC, one can cure a patient (e.g. bone-marrow transplant).

Materials and Methods

We have developed a computational model of the hematopoietic system in which we can study the population dynamics of different types of cell: healthy, cancer and TKI treated cells. Normal hematopoiesis is represented by a hierarchical model in dynamic equilibrium where cells move along the hematopoietic tree as they become increasingly differentiated. The parameters of the model are derived from molecular and clinical literature. This model is now split into a stochastic and deterministic part since the populations of stem cells and early progenitors are limited in size.

Results

Through this model we showed in [1] that this view has to be revisited when one considers that stochastic effects at the level of the stem and early progenitor cells will cause the LSC to disappear from the pool of hematopoietic stem cells (HSC). Our simulations involving a large cohort of virtual patients show that in the majority of these patients (~84%) the LSC have disappeared, making the progenitors and not the stem cells the drivers of the disease. Since these progenitors are susceptible to TKI treatment, the different inhibitors may actually cure the disease.

Discussion

Our predictions show that early diagnosis in combination with long-term TKI therapy opens the road to CML eradication when the response of the patients is sufficiently lowered. The results of are simulations are supported by recent clinical studies on long-term Imatinib treatments of CML-CP patients. Overall, our results are fundamental for the understanding of acquired HSC disorders and other derived diseases. [1] T. Lenaerts, J.M. Pacheco, A. Traulsen and D. Dingli (2010) Haematologica 95(6):900-907.

URL

http://www.ulb.ac.be/di/map/tlenaert/

Presenting Author

Tom Lenaerts (tlenaert@ulb.ac.be)
MLG, DI, Université Libre de Bruxelles

Author Affiliations

1 MLG, Département d’Informatique, Université Libre de Bruxelles, Boulevard du Triomphe CP212, 1050 Brussels, Belgium. 2 Computer Science Department, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. 3 Evolutionary Dynamics Group, Max-Planck-Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany. 4 Departamento de Matematica e Aplicações, Universidade do Minho, 4710-057 Braga, Portugal. 5 ATP-Group, Centro de Matemática e Aplicações Fundamentais, Complexo Interdisciplinar, 1649-003 Lisboa codex, Portugal. 6 Division of Hematology, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA

Acknowledgements

TL acknowledges the support of the FNRS (Belgium), JMP acknowledges the support of FCT (Portugal), AT is supported by the Emmy-Noether program of the German Research Foundation and DD is supported by an Early Career Development Award.

I-7. sscMap Perturbation: unbiased candidate therapeutic ranking in connectivity mapping

McArt D(1,), Zhang S-D(1)*

Different biological states have their own characteristic gene-expression profiles. Connectivity mapping is a technique for discovering the underlying biological connections between states, based on gene-expression similarities. The sscMap method has been shown to provide enhanced sensitivity in mapping meaningful connections leading to testable biological hypotheses, and in identifying drug candidates with particular pharmacological and/or toxicological properties. Challenges remain, however, as how to prioritize the large number of significant connections in an unbiased manner.

Materials and Methods

We examined a perturbation approach on datasets from the GEO website. We attained microarray datasets for cervical cancer, acute myeloid leukemia and breast cancer(letrozole treated). For each casestudy the raw data were downloaded and expression analysed using the Bioconductor suite and analyzed with gene-signature perturbation (removal of a single gene systematically from the signature and measuring its impact). We aimed to test whether an identified connection between a disease gene-signature and a particular set of drug-induced reference profiles is stable.

Results

The perturbation approach helps to identify meaningful biological connections that have been ranked according to the ability to be robust to systematic perturbation and thus retrieves the most influential candidate drugs. In the case of AML, we found that the prevalent drugs were retinoic acids and PPARγ activators. For cervical cancer, our results suggested that potential candidates are likely to involve the EGFR pathway, and with the breast cancer dataset, we identified candidates that are involved in prostaglandin inhibition.

Discussion

Thus the gene-signature perturbation approach offers a convenient and effective approach to the connectivity mapping process. It allows for increased specificity in the identification of possible therapeutic candidates. The 'strength' of each candidate derived from checking for robustness against perturbation offers a viable unbiased method of candidate selection for further experimentation. Diseases that have poor prognosis and treatments that require non-invasive criteria will benefit from timely and unbiased therapeutic selections.

Presenting Author

Darragh G. McArt (d.mcart@qub.ac.uk)
Queens University Belfast

Author Affiliations

(1)Cancer Bioinformatics Group, Centre for Cancer Research and Cell Biology (CCRCB), Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7BL, UK

Acknowledgements

This work was supported by the Department for Employment and Learning through its "Strengthening the all-Island Research Base" initiative

I-8. In silico study of expression profile correlation between microRNAs and cancerous genes

Ng K-L (1,), Weng C-W (1), Huang C-H (2)*

We investigate the possibility that microRNA can act as an oncogene or tumor suppressor gene. Experimentally verified microRNA target genes information (TarBase) are integrated with microRNA and mRNA expression data (NCI-60) to study this hypothesis, in which the Pearson correlation and Spearman rank coefficients are used to quantify these relations for nine cancer tissues.

Materials and Methods

In order to investigate the regulatory role of mirna in cancer, we study the expression profiles correlation between mirna and cancer-related genes, in particular the OCG and TSG targets. Mirna target pairs with PCC/SRC or both below a given threshold are filtered for further investigation. These pairs suggest a regulatory relationship between the mirna and its targets. The TAG dataset is used in order to sort out the mirna-OCG and mirna-TSG pairs. These pairs are further annotated by using the OMIM, GO and KEGG terms. Filtered results are further validated by OMIM, KEGG and miR2Diseas.

Results

Correlation coefficients with negative values are used to filter out miRNA targets. Biological annotations of the targets are supplied by using the TAG, GO and KEGG records. Highly correlated cancer-related miRNAs are validated by the OMIM, miR2Disease and KEGG databases. Our analysis indicated the usefulness of the correlation approach in testing the hypothesis. The above information is utilized to provide a platform in identifying potential cancer related miRNAs. A web based interface is set up for information query, http://ppi.bioinfo.asia.edu.tw/mirna_target/index.html

Discussion

Recent studies indicate that mirna could possibly play an important role in human cancer, where mirna targets TSG or OCG. Experimentally verified mirna targeted genes information are ob-tained from TarBase, which are integrated with the mirna and mRNA expression data from NCI-60 to study this hypothesis. Two correlation coefficients, PCC and SRC, are used to quantify the correlation between mirna and its targets expression profiles. The predicted results are evaluated with reference to the OMIM, KEGG and miR2Disease data sets.

URL

http://ppi.bioinfo.asia.edu.tw/mirna_target/index.html

Presenting Author

Ka-Lok Ng (ppiddi@gmail.com)
Dept. of Bioinformatics, Asia University

Author Affiliations

(1) Department of Bioinformatics, Asia University, 500 Lioufeng Road, Wufeng Shiang, Taichung, Taiwan 41354

Acknowledgements

Ka-Lok Ng and Chien-Hung Huang works are supported by the National Science Council of Taiwan. under the grants of NSC 98-2221-E-468-013 and NSC. 98-2221-E-150-062 respectively.

I-9. Harry Plotter: a user friendly program to visualize genome and genetic map features

Moretto M (1,), Cestaro A (1), Troggio M (1), Costa F (1), Velasco R (1)*

Analyzing a genome requires a lot of different software and tools, which also includes, especially for biologists, those that inspect, visualize and print the data. Visualization of the correspondence among genome sequence and a genetic map is very useful to show the coverage and the correctness of the assembled genome. Moreover there are a lot of genomic feature distribution that could benefit of a graphical representation. For instance, gene density, repeated elements distribution, CpG abundance, and QTL significance could be all better understood with a multi-colored gradient representation

Materials and Methods

Harry Plotter is a stand-alone program written in Java which can run easily on all the platforms for which a Java Virtual Machine is provided, like Microsoft Windows, Linux, MacOSX. It has a Graphical User Interface (GUI) and its use only requires input text files in the correct tab-delimited format. No other software or databases are needed.

Results

Harry Plotter has already been used for the Vitis, Malus and Fragaria sequencing genome projects to visually show genomic scaffolds anchored to chromosomes. It has also been used in Malus to show different QTL significance along chromosomes with a multi-colored gradient. Harry Plotter allows the user to adjust the image size and quality and to export images in PNG or JPEG file format.

Discussion

Most of the software used to visualize genomic information like genome sequence anchored scaffolds and DNA features are usually not public scripts or require a complex setup in order to work properly; while most of the time just a simple image that synthesize all the information is needed. Harry Plotter is a small, easy to use program, that does not require installation and has a graphical user interface. The input files are simple tab-delimited text files with the minimum data required to create the images. Following new users requests, we are currently adding new functionalities.

Presenting Author

Marco Moretto (marco.moretto@iasma.it)
Fondazione Edmund Mach - Istituto Agrario di San Michele all'Adige

Author Affiliations

(1) Fondazione Edmund Mach - Istituto Agrario di San Michele all'Adige

Acknowledgements

Provincia Autonoma di Trento

I-10. Inhibitory activity of thiadiazoles on protein kinase PKnB from Mycobacterium tuberculosis: a virtual screening and molecular docking study

Raj U (1,), Singh VB (1), Swati (1), Srivastava A (1), Naqvi SAH (2)*

Lots of deaths are caused by Tuberculosis per year hence it has become a global threat. In order to develop efficient therapeutic strategies it is vital to understand the physiology of the causative organism, Mycobacterium tuberculosis. Protein Kinase B (PKnB) from Mycobacterium tuberculosis is a crucial receptor-like protein kinase involved in signal transduction. M. tuberculosis PKnB is a trans-membrane Ser/Thr protein kinase (STPK) highly conserved in Gram-positive bacteria and apparently essential for mycobacterial viability.

Materials and Methods

We have attempted with the help of virtual screening and docking approach to expound the extent of specificity of protein kinase B towards different classes of Thiadiazoles(an anti-tubercular agent). Total number of Thiadiazoles was 1800 in number with the minimum binding energy of -11.29 kcal/mol. Three different databases were searched for molecules with a Thiadiazole pharmacophore. The selected Thiadiazoles were chosen on the basis of the structural specificity to the enzyme towards its substrate and inhibitors. Later on ADME & Tox studies were conducted to produce much promising results.

Results

The docking result of the study of 1800 Thiadiazoles demonstrated that the binding energies were in the range of -11.29 kcal/mol to -3.61 kcal/mol, with 15 molecules showing hydrogen bonds with the active site residue. Later, ADME & Toxixicty studies of these 15 molecules were conducted and 6 molecules showed excellent results.

Discussion

The protein kinase B peptide contains two types of structural elements (Valine 95, Arginine 97) and basic residue ring constituted of glycine rich residue. The structure of the protein-ligand complex reveals that Thiadiazoles partially occupies the adenine-binding pocket in PKnB, providing a framework for the design of compounds with potential therapeutic applications. The study provides hints for the future design of new derivatives with higher potency and specificity.

Presenting Author

Utkarsh Raj (rajamity1@gmail.com)
Amity Institute of Biotechnology, Amity Unversity

Author Affiliations

(1)Amity University (Lucknow,IN) (2)Bio Discovery- Solutions for future (Agra,IN)

I-11. A microarray format standardization study: meaningful structures

Kocabas F (1,2,), Can T (3), Baykal N (1)*

There is an increasing demand for microarray experiment. Putting into biological context for interpretation makes microarray data exceptionally larger. Microarray data will be difficult to manage if standards are not developed. Although there are standardization initiatives on content, format, and model, repositories have isolated information space and have proprietary formats. At present, they can not exchange and their integration is costly. There is a need for format and data management standards to make disparate microarray data visible, usable, and understandable for scientific community.

Materials and Methods

We have studied on records on GEO. MINiML file which is a descriptive file for each experiment has been extended in structure (metadata card) and semantics within a metadata framework. Two semantic nets are developed in RDF format (Experimenter: encoded in FOAF/RDF; Summary: in RuleML Datalog). SPARQL has been used for information access. They are meaningful structures which are extendable, queriable, portable, integrable, and machine processable. The metadata card and semantic nets can be exchanged with major repositories. This framework can also be used for all biomedical repositories.

Results

With the proposed framework the benefits are, • Producer can deposit meaningful data and consumer can get the intended meaning • backlog is reduced due to automation • Ambiguity and redundancy is reduced with standard format and additional semantics • Visibility, understandability, and usability is enforced • Users can use W3C tools on public domain • The framework promotes standardization leading to optimum structured reporting through use of more controlled vocabularies (Countries, Date Time Group, Names etc) not only to annotate but to encode the metadata and data.

Discussion

Metadata card and semantic nets are format standard contributions. They do not replace any existing work. However, if adopted, they can be a focus for discovery and integration. There is up to %3 monthly increase in records at GEO which will increase the current backlogs (currently 20% in Series records and 80% in GEO Dataset creation). Such standardization studies to promote machine understandability and semantic interoperability are needed. We believe that once such standardization efforts become adopted, the required tools and detailed guidance will follow.

Presenting Author

Fahri S Kocabas (fahri@ii.metu.edu.tr)
PhD student, 1Department of Health Informatics, Middle East Technical University (METU) Informatics Institute

Author Affiliations

1 Middle East Technical University (METU) Informatics Institute, Health Informatics Department, 06531 Ankara, Turkey 2 NATO HQ C3S, Information Services Branch, Blvd Leopold III B1110 Brussels, Belgium 3 Middle East Technical University (METU) Computer E ngineering Department, 06531 Ankara, Turkey

I-12. Normalization of proteomic ratio data

Sherman J (1,2), Molloy M (2,3), Descallar J (4), Lance B (4), Uitto P (3), Wood G (4,)*

Quantitative mass spectrometry techniques are commonly used for comparative proteomic analysis in order to provide relative quantitation between samples. To account for variable loading between samples a normalization procedure is required. A standard approach to normalization is to use internal standards, proteins that are assumed to display only minimal changes in abundance between the samples under comparison. A normalization procedure allows adjustment of the data, so enabling relative quantitation to be reported, for example, as ratios of the two samples. This research examines current normalization practice and offers improved extensions.

Materials and Methods

Normalization is determined by centring symmetrised data. A generic symmetrising transformation is defined and a range of centring properties considered; these together provide a set of possible normalizations, with each normalization being reciprocally invariant, a needed property. This presents the opportunity for alternative normalization procedures. When testing for departure of a ratio from unity we use a t-test, so normality of the symmetrised observations is desirable. The customary logarithm symmetrisation takes lognormally distributed ratios to normality. A symmetrising function which takes a ratio of normal distributions to normality is developed and its properties examined.

Results

This research makes two contributions to an understanding of normalization. First, the customary centring of logarithmically transformed ratios is shown to attend not only to centring but also to minimisation of the spread of the symmetrised data. Second, the normalization problem is set in a broad context, allowing normalization to be achieved based on a symmetrisation which carries the ratios to approximate normality, so at times increasing the power with which differentially expressed samples can be detected. The improved power of the new method is illustrated using simulated data and then applied to real proteomic data.

Discussion

Current practice is seen to be robust. The alternative normalization method proposed requires the user to input (or estimate, from inputted internal standards data) the mean and variance of the normal distributions of the numerator and denominator of the internal standard ratios. Use of such a tailored symmetrisation can increase the power with which we can detect differentially regulated ratios. The method can be extended to handle alternative assumptions concerning the distribution of the numerator and denominator.

Presenting Author

Graham R. Wood (graham.wood@mq.edu.au)
Department of Statistics, Macquarie University

Author Affiliations

(1) Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143, USA. (2) Australian Proteome Analysis Facility, Macquarie University, NSW 2109, Australia. (3) Department of Chemistry and Biomolecular Sciences, Macquarie University, NSW 2109, Australia. (4) Department of Statistics, Macquarie University, NSW 2109, Australia.

Acknowledgements

J. Sherman was supported by NIH NCRR P41RR001614 and J. Descallar by a Macquarie University Safety Net Grant.

I-13. BioCatalogue: the curated catalogue of life science web services

Tanoh F (1), Bhagat J (1), Nzuobontane E (2), Laurent T (2), Wolstencroft K (1), Stevens R (1), Pettifer S (1,), Lopez R (2), Goble C (1)*

Web services have gained a momentum as a means for packaging existing data and computational resources in a form that is amenable for use and composition by third party applications. Many institutions and organisations such as the EBI, NCBI and DDBJ provide Web Services to access and analyse their resources. However, one of the main issues that hinders the wide adoption and use of web services is the difficulty in locating the "appropriate" web service. The descriptions of available web services are often poor providing little information to the scientist about their usefulness.

Materials and Methods

Web Service and annotations in the BioCatalogue are added by its members or automatically harvested from existing registry such as the Embrace registry http://www.embraceregistry.net/), seekda (http://webservices.seekda.com/) and BioMoby (http://www.biomoby.org/).

Results

The BioCatalogue can be accessed via its Web site or via a Web Service API for programmatic access. It currently holds over 1650 Life Science Web Services and has more than 300 registered users. It has had more than 15000 visits since its launch in 2009.

Discussion

The BioCatalogue has established itself as a one-stop-shop for the Life Science community to locate and use Web service that implement the analysis relevant for their scientific experiment. The next phase of the development concentrates on extending the content, improving the quality and coverage of service annotation and integration with other tools.

URL

http://www.biocatalogue.org/

Presenting Author

Steve Pettifer (steve.pettifer@manchester.ac.uk)
University of Manchester

Author Affiliations

(1) School of Computer Science, University of Manchester, UK (2) EMBL European Bioinformatics Institute, Cambridge, UK

Acknowledgements

The BioCatalogue project is funded by the BBSRC [BB/F01046X/1, BB/F010540/1], with additional funding from the European Commission via the EMBRACE project [LHSG-CT-2004-512092]. Development on Search By Data was funded by EMBO [ASTF 338.00-2009].

I-14. Regression system for prediction of errors in the data on gene expression in situ obtained from confocal images

Myasnikova E (1,), Surkova S (1), Samsonova M (1)*

In the recently published paper (Myasnikova et al., Bioinformatics, 2009) we presented an algorithm for estimation of errors in the data on gene expression in situ extracted from confocal images. The application of the method is limited by the requirement of additional information that is not available from the standard procedure of data acquisition, and the special design of experiments is needed. The aim of the work is to create a learning system for the prediction of error size in the data obtained from a confocal image based on the information about the parameters of the microscope.

Materials and Methods

The common way to reduce the noise in confocal images is averaging of multiple frames that however leads to the biased data in case of clipped single frames. For the estimation and correction of this kind of errors a method based on censoring technique is used, which requires the availability of all the confocal scans along with the averaged image. To predict error size in the data extracted from the averaged image we developed a regression system. The learning sample is composed of images obtained at different combinations of microscope parameters, and for each image all the scans are saved.

Results

The regression system was applied to the data on segmentation gene expression in Drosophila blastoderm stored in the FlyEx database (http://urchin.spbcas.ru/flyex/). The predicted errors proved to be of small size not exceeding 5-7% of the mean intensity level in the embryo nucleus.

Discussion

The high quality of confocal images allows to extract high-precision information at cellular resolution. However the data accuracy is limited due to experimental errors that arise in course of confocal scanning. High values of errors are usually caused by improper choice of the microscope settings used in order to increase the image brightness and contrast. An important application of the current work is the possibility to accurately correct this kind of errors thereby allowing to obtain images of the higher dynamical range and thus to extract more detailed quantitative information from them.

URL

http://urchin.spbcas.ru/downloads/step/step.htm

Presenting Author

Ekaterina Myasnikova (myasnikova@spbcas.ru)
St.Petersburg State Polytechnical University

Author Affiliations

St.Petersburg State Polytechnical University

Acknowledgements

This work is supported by NIH grant RR07801, GAP award RUB1-1578 and RFBR grants 08-04-00712-a, 08-01-00315-a and 09-04-01590-a.

I-15. Impact of genetic variations on phosphorylation sites

Via A (1,), Le Pera L (1), Ferré F (1), Tramontano A (1,2)*

Most genes produce a wide range of different products that might change in different tissues/cells, conditions, individuals, and in time. Each product might have different features and roles: genomic variations can affect the protein or transcript functional properties, regulation, efficiency, and specificity of action. In order to investigate how the environment created by genome variations affects functional sites, we are carrying out a systematic analysis of the mutual interactions between phosphorylation sites (Psites), polymorphisms (SNPs), deleterious mutations, and splice variants.

Materials and Methods

We mapped Mendelian Inheritance in Man (MIM) mutations, cancer-related mutations and SNPs, and Psites (from Phospho.ELM) on Ensembl56 alternative splicing (AS) isoforms. Control sites were defined for each case to evaluate statistically significant inter-dependence of these features. Single-nucleotide resolution level expression data were retrieved from publicly available Illumina Genome Analyzer RNA-seq datasets from adult human tissues to measure when and how much individual sites are spliced.

Results

Our study highlights that the co-occurrence of genome variations and functional sites is not randomly distributed, providing clues about the context in which a mutation can be deleterious, and suggesting possible scenarios of protein evolution. In particular, we found that cancer-related mutations are overrepresented, compared with a control set of non-Psites, in the [-5,+5] surrounding of pS and pT and not pY. No statistically significant differences were observed for SNPs. Additionally, Psites are less spliceable than expected in Ensembl isoforms, and less spliced in human tissues.

Discussion

We addressed important questions such as how and why an important class of post-translational modification sites is expressed and spliced. Interestingly, AS does not seem to be extensively used to modulate the presence of Psites, suggesting that other regulation forms are more appropriate for transient interactions such as phosphorylation. Additionally, our findings provide clues on the inter-relation between cancer-associated mutations and Psites.

Presenting Author

Allegra Via (allegra.via@uniroma1.it)
Department of Biochemical Sciences "A. Rossi-Fanelli" Sapienza University of Rome

Author Affiliations

1 Department of Biochemical Sciences "A. Rossi-Fanelli" Sapienza University of Rome 2 Istituto Pasteur, Fondazione Cenci Bolognetti, Sapienza University of Rome

I-16. FuSiGroups: grouping gene products by functional similarity

Welter DN (1,), Gray WA (1), Kille P (2)*

Gene product annotations have become important in the discovery of functionally related gene products in large datasets. Several approaches to quantify annotation similarity have been developed. However, few efforts have addressed the potential applications of these similarity approaches. We propose a novel algorithm to group gene products based on their functional similarity.

Materials and Methods

The FuSiGroups algorithm calculates the functional similarity between all gene products in a list based on the semantic similarity between the GO terms they are annotated with. A number of different semantic and functional similarity measures are available. Then gene products are sorted into groups that represent their different functional aspects. Each group has a “definition” of one or more GO terms that link the gene products. This makes it easy to determine the functional aspect(s) that a set of gene products share.

Results

We determined the minimum recommended semantic and functional thresholds for different measures using ROC curves of three different types of datasets, namely gene expression, protein interaction and phenotypes.

Discussion

An advantage of our algorithm is that it reflects biological reality better than traditional hierarchical clustering as each gene product can be in multiple groups at the same time. The approach also excludes gene products with insufficient functional similarity to all other gene products so that only gene products with common functionality are grouped together. In summary, we exploit existing semantic and functional similarity measures in a new approach for grouping gene products based on their Gene Ontology annotation, which enables easy discovery of functional links between gene products.

Presenting Author

Danielle N.C. Welter (d.n.welter@cs.cf.ac.uk)
Cardiff University

Author Affiliations

1 Cardiff School of Computer Science and Informatics, Cardiff University 2 Cardiff School of Biosciences, Cardiff University

Acknowledgements

This work is funded by AFR grant "TR-PHD BFR06-102" from the "Fonds National de la Recherche (Luxembourg"

I-17. A file system strategy for high-performance sequence interval queries of very large datasets

Karcz SR, Links MG, Parkin IAP*

The huge numbers of variant sequence features arising from genome re-sequencing projects jeopardizes our ability to efficiently query, visualize and make decisions with the information. The critical scalability issue in current storage and query systems is sequence interval query performance. We developed a scalable file system strategy optimized for high performance range queries and a simple method for visualizing the data using the open source genome browser JBrowse. Our goal was to enable efficient utilization of large datasets without extensive computational infrastructure.

Materials and Methods

Using the Arabidopsis thaliana genome as a basis, we populated a mysql database and a GF2S file system with identical simulated SNP feature data ranging in numbers from 32 to 1024 million and measured population time, disk usage and sequence interval query times for both systems. These performance metrics provided a comparison between a RAID-5 spinning disk solution and a solid state device. A 256 million feature GF2S was used as a dynamic data source for the JBrowse genome browser. All tests were run on a Dell R900 server (Centos 5.4, 64GB RAM) using only free and open source software.

Results

Our write tests showed that GF2S can be populated at a rate 5 times faster than mysql and can write 512 million features overnight (16 hours). Sequence interval query tests demonstrated that GF2S could return feature data at a rate 10-20 times faster than mysql using sequence intervals from 1 kb to 100 kb and for dataset sizes ranging from 32 to 256 million features. The file system can store 1024 million features in under 600 GB. JBrowse was also able to be used without modifying its source code to dynamically receive JSON data from the GF2S without noticeable latency.

Discussion

We have demonstrated that the standard XFS filesystem can be efficiently utilized for high performance sequence interval queries on datasets ranging up to 1 billion SNP features. This was accomplished without resorting to any strategies for caching, multi-node clustering, virtualization or cloud computing. The relative ease of adapting the file system to serve data to a standard genome browser application also suggests that the scalability and speed of the GF2S could be made available to other "interval query-centric" visualization applications in computational biology.

Presenting Author

Steven R. Karcz (karczs@agr.gc.ca)
Agriculture and AgriFood Canada

Author Affiliations

Agriculture and AgriFood Canada 107 Science Place Saskatoon, SK, Canada S7N 0X2

Acknowledgements

We gratefully acknowledge funding provided by the Canadian Crop Genomics Initiative

I-18. Evaluation of multivariate data analysis strategies for high-content screening

Kümmel A (1,), Parker C N (1), Gabriel D (1)*

Biological measurements are increasingly becoming multiparametric. In particular, high content screening (HCS) can monitor multiple phenotypic readouts of e.g. cell morphology or protein levels and localization. To analyze such results a multitude of computational methods is available. However, there has been no published comparison of the performance and applicability of such methods. The aim of this study has been to establish a generic and modular data processing pipeline that enables a systematic review of alternative data processing and analysis methods to assess HCS data.

Materials and Methods

Multivariate HCS data from single cells monitoring cell cycle was analyzed using alternative methods for dimension reduction and well summary. (a) The dimensionality was reduced using either factor analysis or a successive elimination of the parameter with highest correlation to others. (b) The cell populations of different wells were compared using either the median, 5 percentile values or the KS statistic. The alternatives were evaluated on the ability of the resulting data sets to distinguish between controls using classification accuracy or the Z' factor using linear discriminant analysis.

Results

If parameters are eliminated that are highly correlated to others the ability to discrimination between positive and negative controls is not significantly reduced but only leads to a limited reduction in dimensionality. Factor analysis reduces the discrimination ability in general but often performs better if a low number of dimensions is chosen to represent the data. Using the KS statistic to compare control samples is not more sensitive than a comparison of the medians. Taking 5 percentile values to summarize a well improves the discrimination in comparison to only the well median.

Discussion

Using the data analysis pipeline presented here, alternative strategies were evaluated for analyzing a HCS data set. The alternatives for reducing the dimensionality or summarizing the cell population were quantitatively evaluated based on positive control samples. Obviously, other methods to those applied here are available. E.g., the cells within each well can be summarized by fractions of subpopulations. The logical and modular assembly of the data analysis pipeline allows for additional methods to be tested when establishing a suitable strategy for a particular data set.

Presenting Author

Anne Kummel (anne.kuemmel@novartis.com)
Novartis Instiues of Biomedical Research

Author Affiliations

(1) Novartis Institutes for Biomedical Research

Acknowledgements

Postdoc fellowship is funded by the NIBR Education Office

I-19. Studies of the binding capacity of cyclooxygenase II and biodistribution of phenols contained in natural products

Paulino Z-M (2,), Aguilera M-S (1), García Jenifer García (2), Iribarne F (2)*

Phenolic compounds, widely distributed and abundant in natural products such as red wine and propolis, are known for their antioxidant properties. Their biochemical traits are also of great interest in relation to the interaction with mediators of inflammatory processes such as cyclooxygenase II (COX-II). In a previous work, we identified phenols in Uruguayan propolis (M. Paulino Z et al. (2010), Portugal (S. Falcão. Et al. (2010)) and wine grapes (K Ali et al (2009)). The detected structures included phenolic acids, flavonols, flavones, flavanones, flavanols, anthocyanins and stilbenes.

Materials and Methods

In this paper the molecular interactions of these phenols with mouse COX-II (Kiefer R et al (2000)) as well as their biodistribution properties were studied using in silico techniques. All calculations were performed with MOE 2007.09 simulation package. Phenol structures were modeled and optimized. The three-dimensional structure of COX-II was obtained from PDB (Protein Data Bank, code 1CVU). All possibilites for binding sites were examined. Flexible docking was performed at all binding sites identified, using for each compound a conformational basis built in advance.

Results

Docking energies (scores) varied from -18 to -8 kcal/mol and were divided into three ranges: high (-18 to -14), medium (-13 to -10) and low (-10 to -8). For the best scoring poses, phenols were sorted as follows: anthocyanins, flavonols and flavanols (high and medium scores); flavones and flavanones (medium scores); stilbenes (medium and low scores) and phenolic acids (low scores). Additionally, for each ligand, the best score conformations we analyzed in terms of the relationship among the binding energy, the number of hydroxylations and the number of contacts.

Discussion

Finally, given that compound biodistribution may define the quality of a candidate to be developed as an anti-inflammatory drug, a QSAR study was conducted to identify which properties associated with structure (hydrophobicity, molecular weight, accessible polar area, number of rotatable bonds and drug-ability) could affect compound bioavailability.

Presenting Author

Margot Paulino Zunini (margot@fq.edu.uy)
Facultad de Química

Author Affiliations

1. Departamento de Química y Farmacia. Departamento de Física. Facultad de Ciencias. Universidad Católica del Norte. Avda. Angamos 0610. Antofagasta. Chile. 2. LaBioFarMol, DETEMA, Facultad de Química, UdelaR, General Flores 2124, 11600, Montevideo, Uruguay

Acknowledgements

Funding: Minera Escondida Limitada (Chile). Acknowledgments: Pablo Carmona Venegas, undergraduate student, UCN.

I-20. Finding logic networks using the Dutch Life Science Grid: handling 15 million jobs

*Bot J (1, 2, *), de Ridder J (1, 2), Reinders M (1, 2)*

Using insertional mutagenesis one can get insight on how a tumor develops, here we try to explain an observed gene expression with a logical combination of observed insertions. Finding such logic networks is computationally intensive. When using grid computing, finding one network can seen as one job whose runtime is unpredictable and ranges from 2 seconds to 14 hours. The number of possible combinations, networks and the need for permutations results in more than 15 Million jobs which the grid middleware is unable to schedule and whose output cannot be stored on standard grid facilities.

Materials and Methods

We developed a method built upon the ToPoS pilot job framework to schedule and keep track of the jobs. For dealing with the outputs we wrote our own XML-RPC based web-service with database back-end. We also implemented clients to easily communicate between the grid nodes and the ToPoS schedular as well as with the database. The web-service receives both the outputs of the individual runs as well as the errors generated while executing. This makes sure that the user can quickly find out whether a job ran successfully or inspect problems on the grid.

Results

Our method was able to deal with the 15 million jobs (all generating output) that ran on 5000+ cores. The algorithm ran for about a month on the Dutch Life Science Grid doing about a century's worth of CPU work. The debugging facilities were very useful as they provided insight in the different errors and irregularities of the grid. The proposed method also allowed us to easily add a desktop cluster to our compute facilities.

Discussion

The method itself is generic so it is able to deal with other bioinformatics applications, in different domains, and with other data types. However, the current implementation needs be revised to make it easier for layman end-users to work with.

Presenting Author

Jan J. Bot (j.j.bot@tudelft.nl)
Delft University of Technology

Author Affiliations

(1) Delft Bioinformatics Lab, Delft University of Technology (2) Netherlands Bioinformatics Centre

I-21. A detailed view on several Model-Based Multifactor Dimensionality Reduction methods for detecting gene-gene interactions in case-control data in the absence and presence of noise

Cattaert T (1,2,), Van Lishout F (1,2), Mahachie John JM (1,2), Van Steen K (1,2)*

When searching for epistasis, parametric approaches have severe limitations. Alternatively, the non-parametric Multifactor Dimensionality Reduction (MDR) method can be applied. It handles the dimensionality problem by pooling multi-locus genotypes into two groups of risk: High Risk and Low Risk. However, MDR involves computationally intensive cross-validations because it is based on prediction. Moreover, MDR is restricted to univariate, dichotomous traits, and it is not possible to flexibly adjust for lower-order effects and confounders.

Materials and Methods

Recently, Calle et al. proposed the Model-Based MDR (MB-MDR) method. Association tests are applied to identify risk cells and to test the final one-dimensional construct, and a third risk category, that of ‘No Evidence for Risk’, is introduced. We evaluate the empirical power of MB-MDR to detect gene-gene interactions in the absence of any noise and in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. We also explore the potential of alternative definitions for risk cell identification, based on ranking of the case to control ratios.

Results

Our simulation results illustrate that MB-MDR has increased power over MDR to identify gene-gene interactions for most considered genetic models, in particular in the presence of genetic heterogeneity, phenocopy, or low minor allele frequencies. Preliminary results indicate that the alternative risk cell definitions based on ranking further improve MB-MDR power. Type-I error rates have been found to be close to the nominal level. Finally, MB-MDR is much less computationally intensive than MDR.

Discussion

Using association tests allows multiple models to be proposed, no longer requires cross-validations, and flexibly deals with different outcome types and covariates. Hence, MB-MDR offers a flexible framework that naturally deals with different outcome types (e.g. categorical, continuous or survival type) and allows for lower order effects corrections or adjustments for important confounding factors. Furthermore, it can easily handle (a combination of) unrelated individuals and families of any size and structure.

Presenting Author

Tom Cattaert (tom.cattaert@ulg.ac.be)
University of Liege

Author Affiliations

(1) Montefiore Institute, University of Liege, Belgium (2) GIGA-Research, University of Liege, Belgium

Acknowledgements

T. Cattaert is a Postdoctoral Researcher of the Fonds de la Recherche Scientifique - FNRS. T. Cattaert, F. Van Lishout, J. M. Mahachie John and K. Van Steen acknowledge research opportunities offered by the Belgian Network BioMAGNet (Bioinformatics and Modelling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme (Phase VI/4), initiated by the Belgian State, Science Policy Office. Their work was also supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence (Pattern Analysis, Statistical Modelling and Computational Learning), IST-2007-216886. In addition, F. Van Lishout acknowledges support by Alma in Silico, funded by the European Commission and Walloon Region through the Interreg IV Program. The scientific responsibility for this work rests with its authors.

I-22. Quality ranking of 16S sequences: an approach based on poset theory

De Smet W (1,), Verslyppe B (1), De Loof K (1), De Vos P (1), De Baets B (1), Dawyndt P (1)*

In the field of microbial taxonomy, the gene sequences coding for the 16S ribosomal RNA (rRNA) are now an integral part of many taxonomic studies. Because of their usefulness in divining the evolutionary past, 16S rRNA sequences of many taxa have proliferated in the International Sequence Database Collaboration (INSDC) databases. Increasingly, several sequences are available for any one species and tools become necessary to support researchers who want to quickly and easily gather available sequences and assess their quality.

Materials and Methods

The StrainInfo project (http://www.straininfo.net/) already provides a useful resource to microbiological researchers, by integrating information about microbiological cultures, available in Biological Resource Centers (BRCs) around the world, on one single strain passport. Included is taxonomic and sequence data that can then be used to automate the sequence selection process. We used a ranking algorithm based on the theory of partially ordered sets to select an appropriate sequence and compare the results with those of the All-Species Living Tree Project.

Results

Exploration of the results of this comparison show some limitations in the sequence retrieval process, caused by a lack of usable annotations. Comparison with the results of a manually curated data set reveal several controversial or surprising picks, depending on the quality criteria used. The approach used shows promise as a way to quickly explore available sequences for particular genera and visualize quality differences.

Discussion

Despite some limitations, automated retrieval often finds enough sequences to rank, visualize and build a first approximate phylogenetic tree of any genus with. Leveraging data available within StrainInfo we can make this a single step process, helping researchers to waste less time searching for sequences and more time doing research. Visualization of the poset by the various used criteria is an especially promising way to quickly get an overview of available sequences and their relative merits.

Presenting Author

Wim P.P. De Smet (Wim.DeSmet@UGent.be)
Ghent University

Author Affiliations

(1) Ghent University

Acknowledgements

Research carried out by the Interuniversity Attraction Poles programme funded by the Belgian Science Policy

I-23. Clustering and kernel methods using R

Adefioye A, De Moor B*

The kernel methods are advantageous, since they are able to take the non-linear aspect of a given dataset into consideration. The ability of the methods are tested on compounds represented either as bit string representations of molecular structures using fingerprints or as physiochemical descriptors. The performances of the kernel based methods are compared to each other as well as to more traditional methods such as those based on partitioning around mediods. The kernel based methods give a better cluster representation of the compound set.

Materials and Methods

Represented the chemical compounds using six different descriptor types (Fingerprints: Estate, Extended, MACCS, Graph, Standard and physiochemical descriptors). Clustered the compounds using these representation, with the hiearachical Ward clustering algorithm. Determined k. With k=4, clustered all the different types of compound representation using k K-Mean, spectral clustering and pam clustering. The clusters are checked using cluster validity methods.

Results

The kernel based methods give a better cluster representation of the compound set.

Discussion

Clustering can answer important questions when using machine learning approaches for chemoinformatics purposes. Initially, the hierarchical clustering algorithm is able to indicate several clusters these clusters, in this case 4, can be used to decide what training set one will use, and which set of compounds will be in the test set. Here we focus on the usefulness of the hierarchical cluster in selecting the number of k's, for the k k-Means run. The clustering done by the Ward hierarchical method has some notable mistakes, in effect showing the necessity for more superior methods.

Presenting Author

Adeshola A Adefioye (tunde.adefioye@esat.kuleuven.be)
Katholieke Universiteit Leuven

Author Affiliations

Katholieke Universiteit Leuven

Acknowledgements

Katholieke Universiteit Leuven

I-24. Ensuring data integrity in large modern integrative genomic studies

Lynch A (1,2,), Dunning M (2), Chin S-F (1,2), Curtis C (1,2)*

This is an age of large, multi-dimensional genomic studies with clinical applications (e.g. TCGA, METABRIC). As a consequence, samples may be drawn in from many clinics, prepared and plated several times for different technologies, and distributed to several laboratories. As a result of these complications, it is possible for errors to occur in the mapping of sample IDs to resultant data, with detrimental consequences for downstream analyses.

Materials and Methods

Where there exist strong relationships between recorded genotypes and recorded expression (be that due to genuine eQTLs or technical biases) then we can use a record of such relationships to establish a score for the agreement between a set of genotypes and a set of expression intensities.

Results

We present a simple iterative algorithm to validate and correct sample ID mappings for experiments running expression microarrays and also generating genotyping data. Its performance in identifying ID mapping errors is exceptional in large studies, and can pick out systematic changes in relatively modest experiments.

Discussion

The approach is naturally dependent on the quality of the data arising from technologies, so may demonstrate a high-false positive rate for poor quality samples - although it is no bad thing to have one's attention drawn to such samples. If the data are then to be used to search for new eQTLs then there are some subtleties and potential biases for which one should be prepared, but these pale in comparison to the difficulties that mis-identified samples would bring.

Presenting Author

Andy G. Lynch (andy.lynch@cancer.org.uk)
University of Cambridge

Author Affiliations

1 - Department of Oncology, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE 2 - Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre Robinson Way, Cambridge, CB2 0RE

Acknowledgements

We acknowledge the support of The University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited.

I-25. Testing branches in dendrograms representing species abundance data.

Calkiewicz J (1), Wlodarska-Kowalczuk M (2), Wrobel B (1,3)*

Hierarchical clustering is one of the most popular techniques to represent ecological data, eg. species abundances in samples. Usually, however, the resulting dendrograms are presented without support for branches. In this work we aim to investigate if some methods devised to find support for branches in phylogenetic trees are useful in the analysis of ecological data. We use BP (bootstrap proportion, Felsenstein test), AU (Shimodaira approximate unbiased test), and two methods that ask a question if the branch length is significantly longer than zero: the Dopazo and WLS-LRT.

Materials and Methods

We reanalysed data from 3 published studies (Wlodarska-Kowalczuk et al. 2004. Deep-Sea Res I51:1903; Wlodarska-Kowalczuk and Pearson 2004. Polar Biol 27:155; Wlodarska-Kowalczuk and Weslawski 2008. Mar Ecol Prog Ser 356:215). Species that occurred in only one sample were removed, and data were transformed using square root, double root or log. We used 2 distance measures (Bray-Curtis or Euclidean), 4 linkage criteria (single, complete, average linkage or neighbor-joining). For all 4 approaches to test branches, we bootstrapped the data by random sampling the species with replacement.

Results

For the datasets we used here, biologically relevant clustering was observed most easily for average linkage and data transformed with double root. Regardless of the method used, the obtained dendrograms contained branches that separated clusters that had biological relevance. However, these branches were judged significant only using the Dopazo test and WLS-LRT. The support for branches obtained using the BP and AU was mostly below 0.70. This indicates that some approaches originally proposed for phylogenetic trees can be applied in the analysis of species abundance data.

Discussion

The fact that the BP and AU approaches result in low support for branches may explain why these methods are rarely used: if branches are not judged significant even for high quality data, it is understandable that a method does not reach acceptance. However, it seems that the methods that are based on asking a question if a particular branch separating clusters is longer than zero can be useful for determining if the clusters observed using a particular distance measure or linkage method have statistical support.

URL

http://www.evosys.org/

Presenting Author

Joanna Calkiewicz (calkiewicz@iopan.gda.pl)
Computational Biology Group, Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland

Author Affiliations

(1) Computational Biology Group, Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland (2) Marine Ecosystems Laboratory, Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland (3) Laboratory of Bioinformatics, Adam Mickiewicz University, Poznan, Poland

Acknowledgements

Polish Ministry of Science and Education (projects N303 291234 and N N519 402037)

I-26. Reliable identification of hundreds of proteins without peptide fragmentation

Bochet P (1,2,), Rügheimer F (1,2), Guina T (3), Brooks P (4,5), Goodlett D (6), Clote P (7,8,9), Schwikowski B (1,2)*

One of the most common approaches for large scale protein identification is Hight Performance Liquid Chromatography, followed by Mass Spectrometry (HPLC-MS). If more that a few proteins have to be identified, the additional fragmentation of individual peptides has been considered essential. Here we present evidence that, by combining high-precision mass measurement and modern retention time prediction algorithms (Krokhin et al 2006) with a robust scoring scheme, hundreds of proteins can be identified without having to rely on peptide fragmentation.

Materials and Methods

For each candidate protein, taken from a relevant protein sequence database, we predict the peptides resulting from the digestion of the protein and sort them according to their computed HPLC retention times. Corresponding peaks are then searched in the same order in the series of spectra from HPLC-MS. For this purpose, a set of peaks with masses corresponding to those of the predicted peptides is searched by a dynamic programming algorithm maximizing an alignment score. Using quantile regression, this alignment score is compared to those obtained for proteins from a suitable decoy database.

Results

The method was tested on HPLC-MS data obtained from the pathogenic bacteria Francisella tularensis, for which HPLC-MS/MS fragmentation spectra were also available. Out of 1719 possible proteins in F. tularensis, we were able to detect 257 proteins with a FDR of 5.7% (The FDR was estimated from the score distribution of the proteins in the decoy database). This is 59% of the number of proteins detected by applying the Mascot tool to the fragmentation spectra from the same sample. 31 additional proteins were seen that had not been detected in the fragmentation data by Mascot.

Discussion

Previous works describe the identification of proteins from MS and measured retention times (Strittmatter et al, 2003) or use the predicted retention times of peptides to filter false identifications from MS/MS (Pfeifer et al. 2009). Relying only on MS and predicted retention times, we detect many more proteins than Palmblad et al 20004 using the same inputs. Using the high accuracy now available in MS, and eliminating the need for the measurement of the retention times of peptides, this method could be an alternative to peptide fragmentation for the protein identification in complex samples.

Presenting Author

Pascal F. Bochet (pascal.bochet@pasteur.fr)
CNRS, Institut Pasteur

Author Affiliations

1 Institut Pasteur, System Biology Laboratory, Dept Génomes et Génétique, Paris, France 2 CNRS, URA2171, Paris, France 3 University of Washington, Dept of Pediatrics, Seattle, WA, USA 4 CNRS, UMR7592, Paris, France 5 Université Paris 7, Paris, France 6 University of Washington, Dept of Medicinal Chemistry, Seattle, WA, USA 7 Biology Dept, Boston College, Boston, MA,USA 8 Ecole Polytechnique, LIX, Palaiseau, France 9 LRI, Université Paris XI,Orsay, France

Acknowledgements

Work of P. Clote was partially funded by the Digiteo Foundation, as well as by the National Science Foundation grants DBI-0543506 and DMS-0817971.

I-27. Exploring unassigned peaks in protein fragment mass spectra with frequent itemset mining techniques

Vu T-N(1,), Valkenborg D(3), Eeckhout D(2), De Jaeger G(2), Goethals B(1), Witters E(3), Lemière F(4), Laukens K(1)*

Mass spectra generated in proteomics experiments are rich sources of information. Often, only a minor fraction of this information is used for protein identification. Of the remaining peaks, a fraction may be assigned to common contaminants, but usually a significant fraction of the peaks remains unexplained. Some of these unassigned masses may represent interesting fragments with hidden valuable knowledge. We employed frequent itemset mining techniques to exploit potentially useful patterns. The patterns may serve as a basis to generate hypotheses that can be validated by a spectrometrist.

Materials and Methods

Original data from over 20.000 “Peptide Mass Fingerprint”(PMF) MALDI TOF measurements of the higher plant "Arabidopsis thaliana" were compared against different sequence databases including TAIR9 and SwissProt on a Mascot server, and unassigned peaks where retained. The corresponding masses, discretized to their lower integer, were considered as the items, and the spectra in which they occur as transactions. Frequent itemset mining and error tolerant itemset mining algorithms were applied. The results are evaluated towards their potential to reveal novel insights.

Results

Transactions were subjected to the frequent itemset mining with different support thresholds, and ranked according to support. Similarly, error tolerant itemset mining algorithm was also applied on the transactions. This approach is less strict than the previous one because transactions that contain an "almost" complete item set are now taken into account. In follow-up experiments, redundant items, that appear in most frequent item sets, were removed from all transactions prior to error tolerant itemset mining. Retrieved patterns are evaluated.

Discussion

This study demonstrates both the feasibility of frequent itemset mining in mass spectrometry (MS) data analysis and its pitfalls. The analysis of unassigned peaks in large numbers of PMF spectra reveals several highly frequent masses. Frequent itemset mining techniques yields a large number of peaks that frequently co-occur in unassigned peak lists of individual MS spectra. Prioritization, redundancy removal and assessment of statistical significance are remaining challenges in MS-based frequent itemset mining. To evaluate retrieved patterns, we will use existing and new tandem-MS analyses.

Presenting Author

Vu Trung Nghia (TrungNghia.Vu@ua.ac.be)
University of Antwerp, dept. Mathematics & Computer Science, Antwerp, Belgium

Author Affiliations

(1)University of Antwerp, dept. Mathematics & Computer Science, Antwerp, Belgium; (2)VIB - Plant Systems Biology / University of Ghent, Gent, Belgium; (3)Flemish Institute for Technological Research, Antwerp, Belgium;(4) University of Antwerp, dept. Chemistry, Antwerp, Belgium

Acknowledgements

IWT SBO Project "Bioframe" and University of Antwerp Interdisciplinary BOF

I-28. Benchmarking a new semantic similarity measure using reference sets and clustering : evaluation and interpretation for the discovery of missing information

Benabderrahmane S (1,), Smaïl-Tabbone M (1), Napoli A (1), Poch O (2), Devignes MD (1)*

Calculating similarity between genes is an important task in post-genomics. Recently, semantic similarity measures were introduced as a complement to sequence similarity measures. Those measures use GO annotations which are organized into three hierarchical controlled vocabularies (molecular function, biological process, cellular component). Moreover, the GO annotation process is traced by evidence codes. Existing semantic similarity measures do not take into account all these GO features, and few studies have been conducted for their evaluation and comparison.

Materials and Methods

We have defined a semantic similarity measure (IntelliGO) based on representing genes in a new vector space model and taking into account the GO terms frequency, their organization in the GO graph, and the annnotation evidence codes. The comparison with other measures involves two collections of reference sets of genes (KEGG pathways, PFAM clans) and inter- and intra-set global similarity measures. The measure was used for functional clustering. Resulting clusters were compared with reference sets using the F-Score method. Overlap analysis leads to possible discovery of missing information.

Results

The IntelliGO semantic similarity measure is the first described measure that allows differential weighting of annotation evidence codes. When compared with three other measures, the IntelliGO intra-set similarity values reflect at best the biological cohesion of refrerence sets of genes. The inter-set values express stronger discrimination between distinct sets of genes. This is confirmed by clustering evaluation which is better with IntelliGO than with other measures. Mismatches between clusters and reference sets are analyzed and exploited for improving set composition or gene annotation.

Discussion

The methodology for comparing semantic similarity measures relies on defining collections of reference sets of genes that share common annotations. KEGG pathways are relevant for testing biological process GO annotations, Pfam clans are relevant for testing molecular function GO annotations. Other sets can be proposed for completing the benchmark. The method itself relates to well-known concepts in data and cluster analysis. Interestingly this method has lead to analyze the overlaps between clusters and reference sets as a mean for discovering missing information.

URL

http://intelligo.loria.fr

Presenting Author

Sidahmed Benabderrahmane (benabdsi@loria.fr)
Loria, Nancy Université

Author Affiliations

(1): LORIA UMR7503, CNRS-Nancy niversité, Orpailleur Team, Nancy. (2):IGBMC UMR 7104, CNRS-INSERM-Université Louis Pasteur, Strasbourg.

Acknowledgements

This work was supported by the French National Institute of Cancer (INCa) and by the Region Lorraine program for research and technology (CPER MISN). Benabderrahmane S. was recipient of an INCa doctoral fellowship.

I-29. Utopia:GPCRDB: a domain-specific PDF reader

Vroling B (1,), Thorne D (2), McDermott P (2), Pettifer S (2), Vriend G (1)*

The amount of data and information present in literature and databases is growing at an ever-increasing pace. There is so much information available that we no longer know what information exists and finding what we would like to know has become an increasingly difficult task. The knowledge we seek is often fragmentary and disconnected, spread thinly across a vast sea of literature and databases.

Materials and Methods

As a step towards a solution for this enormous challenge, we present the domain-specific PDF reader Utopia:GPCRDB. This reader dynamically links concepts in scientific articles to relevant entities in the GPCRDB such as proteins and mutations. The GPCRDB is a molecular class-specific information system that contains a large amount of heterogeneous data on G protein-coupled receptors (GPCRs) and functions as a one-stop resource from GPCR researchers.

Results

With Utopia:GPCRDB it is now possible to retrieve the knowledge stored in the GPCRDB and integrate it with the article you are currently reading, all with the single click of a button.

Discussion

This integration of scientific articles with a domain-specific database greatly enhances the process of turning data into knowledge and offers new ways of thinking about how to deal with information stored in literature and databases.

Presenting Author

Bas Vroling (bvroling@cmbi.ru.nl)
CMBI, NCMLS, Radboud University Medical Centre

Author Affiliations

1 CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands 2 School of Computer Science, University of Manchester, Manchester, M13 9PL, UK

I-30. Locally optimal structures of an RNA

Saffarian A (1,2,), Giraud M (1,2), Touzet H (1,2)*

The folding space of an RNA provides a rich knowledge about the structures and functions of the RNA molecule. Partition functions or sample of optimal and suboptimal structures give some insight on this folding space. An alternative description of this space can also be achieved by studying so-called locally optimal secondary structures. They are structures maximal for inclusion: no base pair can be added without creating a pseudoknot or a base triplet (Clote, 2005). As far as we know, there were no tool to construct all such structures.

Materials and Methods

We defined an algorithm that computes all locally optimal secondary structures of an RNA sequence. Our algorithm works with any folding model, including those that take into account adjacent nucleotide interactions. The input is a sequence together with a set of putative helices. As the number of locally optimal secondary structures could be exponential, we show that the problem divides in two substeps. First, we apply the juxtaposition operations, by dynamic programming, then we build on-the-fly the full structures using the nesting operation.

Results

The algorithm is implemented in a freely available software called REGLISS. We compared REGLISS to MFOLD and showed that REGLISS better describes the folding landscape, finding some structures missed by MFOLD. Our results also illustrate that for some families, the structural RNA has a different folding landscape than random sequences with the same dinucleotide frequency. Finally, we successfully compared our approach to RNALOSS (Clote, 2005). In RNALOSS, the notion of locally optimal structure is reduced to the Nussinov-Jacobson model.

Discussion

Our work shows that all the locally optimal secondary structures of a given RNA can effectively be computed. These structures can also be filtered out using some post-processing criterium such as the free energy or the shape of the structure. This is a fruitful alternative to MFOLD suboptimal structures. Another advantage of the method is that the user can provide its own set of helices, based on the thermodynamic Nearest Neighbor model or on any other model.

URL

http://bioinfo.lifl.fr/RNA/regliss/index.php

Presenting Author

Azadeh Saffarian (azadeh.saffarian@lifl.fr)
Laboratoire d'Informatique Fondamentale de Lille (LIFL), Lille 1 University, France

Author Affiliations

1. Laboratoire d'Informatique Fondamentale de Lille (LIFL), UMR CNRS 8022, Lille 1 University, France 2. INRIA Lille Nord-Europe, France

I-31. Anesthetics and the role of the biological membrane in the molecular regulation of K2P potassium channels

Bernardi RC (1,), Treptow W (1), Klein ML (2)*

Detailing the action mechanism of anesthetics have been the purpose of several studies, since this is not well established. TREK-1, a member of the K2P family, was identified as one key molecular component required for anesthetic effects. The recent atomistic description of the membrane-bound state of TREK-1 (Treptow & Klein, JACS 2010, in press) provide us with an unique opportunity to decipher the anesthetic action on membrane bound species. We have investigated the structure of K2P channels in a membrane environment using MD simulation.

Materials and Methods

Two distinct fully atomistic models for the most studied K2P channel, namely, the TWIK-related (TREK)-1 channel have been built. These constructs were then inserted into a fully hydrated zwitterionic lipid bilayer, and each relaxed by means of MD simulations spanning ~0.3 μs. Further multi-ns simulations of the channel in presence of isoflurane were also carried out. The simulations were performed in the NPT ensemble using the program NAMD2. The simulations used the CHARMM22-CMAP force field with torsional cross-terms for the protein lipids.

Results

The MD study reveals a direct coupling of the C-terminus to the intracellular membrane surface. As a domain structure physically linked to the channel's pore and energetically coupled to the bilayer, the C-terminus appears to form a robust device able to gate the channel in response to membrane stimulation. In presence of isoflurane, the C-terminal coupling onto the membrane increases with anesthetic partition into the bilayer and we observe the partial opening of the channel’s pore after membrane swelling.

Discussion

An atomistic structural description for TREK-1 is provided, which rationalizes the plethora of experimental findings on this channel. We point to an anesthetic action mechanism in which the agonist affects indirectly the channel’s pore gating through modulation of the C-terminal membrane coupling. Given this, further structural and experimental studies aimed at the investigation of the dynamical processes underlying the anesthetic induced protein activation are within reach and should contribute to the deciphering of the key molecular mechanisms underlying the function of K2P channels.

Presenting Author

Rafael C. Bernardi (bernardi@biof.ufrj.br)
Universidade Federal do Rio de Janeiro

Author Affiliations

(1) Laboratório de Biologia Teórica e Computacional, Universidade de Brasilia – Brasília, DF, Brazil. (2) Institute for Computational Molecular Science, College of Science and Technology, Temple University - Philadelphia, PA, USA.

Acknowledgements

We thanks the Brazilian support agencies: CNPq, CAPES and FAPERJ. Our research support came in part from the National Institutes of Health and the National Science Foundation through TeraGrid resources.

I-32. Comparison of validation methods for merging cancer microarray data sets

Taminau J (1,), Meganck S (1,2), Weiss-Solis DY (3,4), Van Staveren WCG (4), Dom G (4), Venet D (3), Bersini H (3), Detours V (4), Nowé A (1)*

There are ~75,000 publicly available gene expression profiles assayed on the human Affymetrix platforms and their reproducibility and validity are increasingly accepted. Whereas the activity of all 20K genes can be measured, there are much fewer samples per-study, limiting statistical power. A solution is to combine multiple studies into larger meta-studies, but no existing methods remove the technical biases in general, eg, different experimental and bioinformatics protocols. Here, we seek a general method by benchmarking existing methods applied to cases of increasing biological complexity.

Materials and Methods

We compare five merging techniques: RAW (do nothing) NORM (gene standardization), BMC (batch mean centering), DWD (Distance Weighted Discrimination) and XPN (Cross Platform Normalization). We examine their results on three test cases of increasing biological complexity. NCI60: a collection of cell lines of nine different types of cancer. THYROID: two publicly available and two in-house thyroid data sets. BREAST: a collection of eight data sets, which is the most complicated case since breast cancer is known as being a very heterogeneous disease.

Results

MDS was used to inspect the clustering with respect to samples' study of origin, technical replicates and biological type. In the NCI60 and THYROID cases the study-bias could not be removed with the simplest methods in contrast to the BREAST case. DWD and XPN showed that samples belonging to the same biological class were clustered together. Comparison of average expressions of housekeeping genes was consistent with the MDS plots. We analyzed cross-study classification accuracy. For this validation NORM and BMC performed best, which is in contradiction with the previous validation outcomes.

Discussion

Most merging techniques have previously only been validated in one specific setting and on specific data. The added value of our work is the validation on three cases of increasing biological complexity. Our results show that no single algorithm exists that outperforms the others in all tasks. Depending on the desired result it is often possible to select some validation criteria that works best for that specific technique. This opens the perspective of the development of an objective method that can be applied generically for merging microarray data.

URL

http://como.vub.ac.be/~jtaminau/ECCB2010/

Presenting Author

Jonatan Taminau (jtaminau@vub.ac.be)
VUB

Author Affiliations

(1) CoMo, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. (2) ETRO, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. (3) IRIDIA, Université Libre de Bruxelles, Avenue F. D. Roosevelt 50, 1050 Brussels, Belgium. (4) IRIBHM, Université Libre de Bruxelles, Route de Lennik 808, 1070 Brussels, Belgium.

Acknowledgements

This research is partially funded by the Institute for the encouragement of Scientific Research and Innovation of Brussels (IRSIB)

I-33. Mapping insertions to putative target genes

De Jong J (1,), De Ridder J (1,2), Sun N (1), Van Uitert M (1), Adams DJ (3), Wessels LFA (1,2)*

Large-scale retrovirus-based, and more recently also transposon-based, insertional mutagenesis screens are of great use in the identification of new cancer genes and in gene therapy. However useful, there are some limitations. One such limitation lies in the difficulties in predicting which genes in the surrounding DNA are affected by insertions, since the range across which insertions influence target genes is unknown. Here, a knowledge-based approach is presented for mapping insertions to such putative target genes.

Materials and Methods

Two sets of same-sample insertion-expression datasets were used: 1) ~2000 MuLV insertions and expression data for 97 tumors; 2) ~10000 Sleeping Beauty insertions and expression data from 138 tumors. For the MuLV dataset, we compared an existing rule-based approach to mapping each insertion to the nearest gene. Then we investigated the effect of insertions on putative target genes. The range of the influence of insertions was estimated by the association of the insertion and the expression data for increasing window sizes around target genes, for both sense and antisense insertions.

Results

For MuLV, the existing rule-based approach produced superior association of insertions with gene expression than the nearest gene approach. For varying window sizes, the largest effect was seen for upstream-antisense and downstream-sense insertions. For SB, this was for sense insertions, although generally the effect was less pronounced. Furthermore, the effect was more local, and for single insertions per gene, there seemed to be a depletion of upregulated genes for ~20kb to ~200mb windows. For both systems, requiring multiple insertions per gene led to a substantial increase in effect.

Discussion

We believe that automated insertion mapping can be a useful complement to CIS analysis, which generally disregards individual properties of insertions. Some limitations of the analysis include the assumption that insertions have a positive effect on gene expression. Furthermore, the association measure and distance-based approach detect mostly direct insertion-expression associations. Ideally, in the future, more insertional mutagenesis systems will be included, and also other data types, such as chromatin lamina association. This will result in more accurate insertion to gene mappings.

Presenting Author

Johann de Jong (j.d.jong@nki.nl)
The Netherlands Cancer Institute

Author Affiliations

(1) Bioinformatics and Statistics, The Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, The Netherlands (2) Delft Bioinformatics Lab, Faculty of EEMCS, TU Delft, Delft, The Netherlands (3) Experimental Cancer Genetics, Wellcome Trust Sanger Institute, Hinxton, United Kingdom.

I-34. The zebrafish embryo in toxicology: a combination of toxicogenomics and other screening techniques

Legradi J (1,), Yang L (1), Alshut R (2), Ho N(1), Klüver N (3), Scholz S (3), Mikut R (2), Reischl M (2), Liebel U (1), Strähle U (1)*

In our work, we aim at obtaining further insights into the mode of action of toxicants. Our new system should allow us to study genotypic and phenotypic toxicological effects chemicals have. With our approach we hope to get a better understanding of how chemicals act in whole organisms, especially during the embryonic development.

Materials and Methods

We used the commercially available Agilent zebrafish arrays and the Agilent eArray system for designing custom expression arrays. Gene expression studies where performed using 10 toxicants, covering several different mode of actions. The arrays where analysed using our Matlab platform Gait-CAD and MeV. Publicly available databases like GO, KEGG, CTD were used for further analysis of obtained gene lists. As high throughput imaging platform, we utilized the standard Olympus Scan R system and several in-house developed microscope systems.

Results

30 Agilent microarray experiments were performed for finding genes acting like biosensors. To better understand the gene regulation, we further developed a microarray system based on transcription factors. In the second part, we developed an automated high throughput microscope screening platform for zebrafish. This enabled us to study concentration and time depended toxicological effects of our compounds.

Discussion

We studied the different mode of actions of 10 Chemicals via a microarray-based toxicogenomics in combination with a phenotype-based microscope approach. As zebrafish presents a good model organism for studying toxicological effects, we think that our approach is suitable for obtaining a better understanding of the different mode of actions of toxicants.

Presenting Author

Jessica Legradi (jessica.legradi@kit.edu)
Strähle Group, Institute of Toxicology and Genetics, Karlsruhe Institute of Technology

Author Affiliations

1 ITG - Institute of Toxicology and Genetics, Karlsruhe Institute of Technology 2 IAI - Institute of Applied Informatics, Karlsruhe Institute of Technology 3 Helmholtz-Zentrum für Umweltforschung UFZ, Leipzig

I-35. Disease gene prediction based on tissue-specific conserved coexpression

Piro RM(1,), Ala U (1), Molineris I (1), Grassi E (1), Damasco C (1), Bracco C (1), Provero P (1), Di Cunto F (1)*

Even with the success of next generation sequencing techniques, the identification of genes involved in human hereditary disease remains a demanding task that can be significantly aided by computational predictions. We discuss a method for disease gene prediction based on high-throughput microarray expression data that uses the conservation of coexpression as a powerful filter for biological significance and allows to specifically focus on tissue-specific relationships between disease and candidate genes, providing novel high-confidence candidates for several genetic disorders.

Materials and Methods

We build both multi-tissue and tissue-specific conserved coexpression networks (CCNs) and exploit the gene coexpression clusters they provide to uncover global and tissue-specific relationships between candidate genes and "reference genes" known to be involved in the given phenotype or similar disorders. Through an explicit integration of phenomics, in particular the concept of phenotype similarity, we can apply our method also to disease phenotypes with so far unknown molecular basis.

Results

We show that the novelty of our approach, the tissue-specificity, provides disease-related information that is highly complementary to the information obtained from multi-tissue coexpression. This leads to biologically meaningful predictions that could not be derived from multi-tissue expression profiles. We discuss the obtained results and present a user-friendly web tool for custom analysis.

Discussion

The conservation of coexpression between human and mouse allows to focus on biologically relevant relationships that have been preserved during evolution. The analysis of tissue-specific coexpression profiles takes into account the dynamic nature of such relationships that may vary over different tissues, cell types and/or conditions. The integration of both concepts, the tissue-specific conserved coexpression, can provide important clues about molecular mechanisms underlying human hereditary disorders and thus provide a powerful tool for disease gene prediction.

URL

http://www.cbu.mbcunito.it/ts-coexp/

Presenting Author

Rosario M. Piro (rosario.piro@unito.it)
Molecular Biotechnology Center, University of Torino, Italy

Author Affiliations

(1) Molecular Biotechnology Center, University of Torino, Italy.

I-36. Improved classification by integrating multiple patient data sets with literature information using co-inertia analysis

Thomas M (1,), Daemen A (1), De Moor B (1)*

Previous studies on tumor classification have shown that classification can be improved by fusing microarray data sets. In our study, we will proceed one step further by incorporating prior literature knowledge specific to each data set gathered from patients.

Materials and Methods

Method Co-inertia Analysis: a multivariate method used in combination with principal component analysis to couple two or more data sets. Datasets Patient data sets: Micro array data sets of DNA and RNA of prostate cancer patients. Document data sets: Each entries on the matrix, corresponds to number of PUBMED abstract in which a gene and cancer related term co-occurs.

Results

a) Principal component analysis of the Patients data sets gives the classification error as 0.4033 b) Co-inertia analysis of patient data sets and document data sets gives the classification error as 0.3578

Discussion

This study shows that classification accuracy can be improved with addition of literature information as prior to microarray data sets. Further research is being undertaken to investigate the effect of integration of additional text information from PUBMED as prior to microarray and clinical data sets.

Presenting Author

Thomas Minta (minta82@gmail.com)
ESAT, K U Leuven

Author Affiliations

1. ESAT, KU Leuven,Belgium

I-37. Time-resolved signatures from weighted visibility graph representation of time-series data

Nikoloski Z (1,2,), Grimbs S (2), Schäfer R (3), Sers C (3), Selbig J (1,2)*

High-throughput technologies could be employed to generate a scope of data necessary for understanding the complexities of living organisms, particularly if the data capture the kinetic resolution of investigated processes. Despite the decreasing costs of the omics technologies, experimental studies are bound to yield relatively short time series due to the inherent problems of gathering sufficient sample material and designing complex, large-scale perturbation experiments. Consequently, there is a pressing need for development of novel methods for analysis of short time-series data.

Materials and Methods

We provide a representation of time series via directed weighted visibility graphs. The proposed representation proves to be an effective lossless approach that both captures the important properties of short time series and is invariant to vertical rescaling and translation. Our contributions include: (1) definition of a polynomially computable distance measure for time series represented by weighted visibility graphs, (2) proof that the proposed distance measure is a generalized metric, and (3) algorithms for finding time-resolved representatives in large-scale experiments based on the proposed measure.

Results

We demonstrate how the proposed representation and measure can be employed in statistical tests that account for the dependencies between not only the time points but also the entities (i.e., genes, proteins, metabolites) from which time-series data were obtained. In addition, we devise algorithms by which we demonstrate that our methods can be used to efficiently determine time-resolved signatures on a gene expression data set from colorectal cancer cell lines.

Discussion

Our weighted visibility graph representation and the proposed generalized metric are particularly suitable for statistical testing which takes into account the dependencies between not only the time points but also the biological entities considered in the study. We demonstrated that, on a transcriptomics data set from colorectal cancer cell lines, our approach finds ontologically related gene profiles whose temporal behaviour is captured by the corresponding cluster representatives. The latter can serve as a time-resolved label which could be employed for classification of new data. In our view, this is a first necessary step towards statistically sound discovery of temporal biomarkers.

Presenting Author

Sergio Grimbs (grimbs@mpimp-golm.mpg.de)
Institute of Biochemistry and Biology, University Of Potsdam, Potsdam, Germany

Author Affiliations

(1) Max-Planck Institute of Molecular Plant Physiology, Potsdam, Germany (2) Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany (3) Laboratory of Molecular Tumorpathology, Institute of Pathology, University Medicine Charité, Berlin, Germany

Acknowledgements

ZN and JS are supported by the GoFORSYS project funded by the Federal Ministry of Education and Research, Grant no. 0313924. SG, RS, CS and JS are supported by the ColoNET project funded by the Federal Ministry of Education and Research, Grant no. 0215417F. ZN is supported by the Max-Planck Society. CS and RS are supported by IBB (PROFIT).

I-38. Genome sequence of the edible cyanobacterium Arthrospira sp. PCC 8005

Janssen PJ (1), Morin N (1), Monsieurs P (1,), Mergeay M (1), Leroy B (2), Wattiez (2), Vallaeys T (3), Waleron K (4), Waleron M (4), Wilmotte A (5), Quillardet P (6), Tandeau de Marsac N (6), Talla E (7), Zhang C-C (7), Médigue C (8), Barbe V (9), Leys N (1)*

Arthrospira are nonheterocystous filamentous cyanobacteria that typically reside in alkaline lakes. They have a high protein content, are rich in carbohydrates and essential fatty acids, and produce a variety of minerals, vitamins, and nutritional pigments such as beta-carotene. Unsurprisingly, they have a long history of human consumption and are used worldwide as industrial feed. PCC 8005 strain was selected by the European Space Agency (ESA) as an oxygen producer and as a nutritional endproduct of the life support system MELiSSA (Micro-Ecological Life Support System Alternative).

Materials and Methods

Whole-genome shotgun sequencing of strain PCC 8005 was performed using 454 pyrosequencing technology (amounting to 400,000 reads) supplemented with Sanger sequencing (up to 96,000 longer reads), leading to a final assembly in 16 scaffolds. The PCC 8005 genome is most likely organised in a single replicon without current evidence for plasmids. These scaffolds were processed by the MaGe annotation platform. Final assembly was facilitated by optical genome mapping and long-PCR. Genome comparison with A. platensis NIES-39 and A. maxima CS-328 were performed using the GeneRage clustering algorithm.

Results

The genome (6,279,260 bases) is highly repetitive, with > 300 kb present as tandem sequences, and contains four CRISPR elements, at least 140 complete IS elements, and eight copies of a putative genomic island. Genes for the production of beta-carotene and two essential fatty acids linoleic acid and gamma-linolenic acid were identified. Genome data confirm the inability of PCC 8005 to fix nitrogen as it lacks essential nif genes and indicate that nitrogen metabolism in PCC 8005 follows classic routes utilizing nitrate, nitrite, urea, and ammonium, with NtcA as the global nitrogen regulator.

Discussion

This draft genome sequence presents 119 contigs rendering 16 scaffolds by assembly. The MaGe annotation system could assign 63% of the CDSs to one or more functional COG groups and currently reports 1,704 conserved hypothetical and 884 hypothetical proteins. Although gap closure and final assembly are very cumbersome due to extended regions of highly repetitive DNA, we were able to place nine scaffolds on a single optical genome map and we soon hope to fully close the genome by multiplex long-PCR and adaptor-assisted genome walking. GenBank submission ADDH00000000. MaGe data available July 2010.

URL

https://www.genoscope.cns.fr/agc/microscope/about/collabprojects.php?P_id=49

Presenting Author

Peter Monsieurs (pmonsieurs@sckcen.be)
Belgian Nuclear Research Center

Author Affiliations

(1) Molecular and Cellular Biology - Unit of Microbiology, Institute for Environment, Health and Safety , Belgian Nuclear Research Centre SCK•CEN, B-2400 Mol, Belgium; (2) Department of Proteomics and Protein Biochemistry, Université de Mons-Hainaut, B-7000 Mons, Belgium; (3) Laboratoire ECOLAG, UMR CNRS-UMII 5119, Université Montpellier II, cc 093, Place E. Bataillon, F-34095 Montpellier Cedex 5, France; (4) Intercollegiate Faculty of Biotechnology, University of Gdańsk & Medical University of Gdańsk, Kładki 24, Gdańsk, Poland; (5) Centre for Protein Engineering, Institute of Chemistry B6, Sart Tilman, University of Liège, B-4000 Liège, Belgium; (6) Unité des Cyanobactéries, CNRS-URA2172, Institut Pasteur, F-75015, Paris, France; (7) Laboratoire de Chimie Bactérienne, CNRS-UPR9043, Université d'Aix-Marseille, 31 chemin Joseph Aiguier, F-13402 Marseille cedex 20, France; (8) Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme (LABGeM), Centre National de la Recherche Scientifique - UMR8030 et

CEA/DSV/IG/Genoscope, F-91057 Evry, France; (9) Génoscope, Centre national de séquençage, CEA/DSV/IG/Genoscope, 2 rue Gaston Cremieux, F-91057 Evry Cedex, France.

Acknowledgements

This work was supported by the European Space Agency ESA/ESTEC under MELGEN-2 contract No. 20397/06/NL/SFe in frame of the MELiSSA project.

I-39. Moa: managing command line bioinformatics

Fiers M (*)

Bioinformatics projects often consist of many separate, mutually dependent steps. A range of advanced software tools exists to organise, maintain and automate such projects. However, in many cases bioinformaticians use the most basic approach possible: the command line and adhoc scripts. This approach is often the most flexible and rapid route to results. A downside to this method is the substantial effort necessary to prevent a project from becoming disorganised and impractical to maintain.

Materials and Methods

Many reliable tools are available that can assist in the organization of a command line project. Rather than re-implementing functionality, Moa incorporates established software into a single integrated tool. To this end Moa has an extensible architecture, with most functionality implemented as a plugin. At the core Moa uses Gnu Make to describe, schedule and execute jobs.

Results

Moa provides a framework to: create, use and share building blocks; build, maintain and execute analysis pipelines; track pipeline history; generate documentation; and provide easy access for non-bioinformaticians. A command line interface provides uniform access to all Moa functionality.

Discussion

Moa aims to be a compromise between large automated frameworks (such as Galaxy) and adhoc scripting pipelines. In comparison to the larger frameworks Moa is less user friendly. This is offset by Moa's greater flexibility. Adhoc scripting does allow flexibility, but requires boilerplate code and does not enforce any structure. Moa offers the best of both worlds, to both bioinformaticians and capable non-bioinformaticians.

URL

http://mfiers.github.com/Moa/

Presenting Author

Mark W.E.J. Fiers (mark.fiers@plantandfood.co.nz)
The New Zealand Institute for Plant & Food Research Limited

Author Affiliations

The New Zealand Institute for Plant & Food Research Limited, Private Bag 4704, Christchurch 8140

I-40. SAAPdb: structural effects of single amino acid polymorphisms

*Baresic A (1, *), Alnumair N (1, *), Martin ACR (1)*

The SAAPdb database was created with two main goals in mind: to collect and unify data on neutral and pathogenic missense mutations; and automatically to process them in terms of their structural effects. This is supposed to fill the gap between mutation annotations at the sequence level, and their pathogenicity/phenotype annotation. This leads to enhanced understanding of protein evolution and will contribute to pharmacogenomics.

Materials and Methods

SNPs, assumed largely to be phenotypically neutral, are obtained from dbSNP and disease-associated mutations (DAMs) from OMIM and several locus-specific mutation databases and mapped to protein structures, where available. An automated analysis tests mutations against a range of structural effects disrupting protein folding, binding, function and stability, yielding an 'structural effects vector' for every mutation. Several subsets of mutations, e.g. cancer-associated kinase mutations are analysed regarding their preference for specific structural effects.

Results

SNPs and pathogenic mutations show clear differences in frequencies of some structural effects, and in their amino acid propensities. Similarly, compensated DAMs on average have a milder effect on the structure, when compared with uncompensated DAMs. Compensated mutations are also preferentially located on the protein surface and in less conserved structural neighbourhoods. A second analysis shows family-specific features of mutations in kinase domains.

Discussion

The work presented here covers several aspects of analysis of the structural effects of SAAPs. First, we address data collection, and the issue of numerous databases in different formats. Furthermore, this update contains an addition to the spectrum of structural analyses: by adding predicted protein-protein interface residues we compensate for the discrepancy in the number of protein complex structures available. Finally, we propose an alternative method for inferring likely structural effects when structure is not available.

URL

http://www.bioinf.org.uk/saap/db/

Presenting Author

Anja Baresic (a.baresic@ucl.ac.uk)
Institute of Structural and Molecular Biology, Division of Biosciences, University College London

Author Affiliations

1. Institute of Structural and Molecular Biology Division of Biosciences University College London Gower Street London WC1E 6BT

Acknowledgements

AB - Overseas Research Awards Scheme; UCL Graduate School Research Scheme NA - The King Faisal Specialist Hospital and Research Center; Royal Embassy of Saudi Arabia Cultural Bureau

AUTHOR INDEX