Oral presentations

13th Annual Meeting of the Bioinformatics Italian Society
June, 2016, University of Salerno, Italy

Session 1: Transcriptomics and Comparative Genomics

INVITED LECTURE – PREPARATA LECTURE

Computational modeling and design of RNA 3D structure and protein-RNA complexes

Janusz M. Bujnicki

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, and Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Umultowska 89, PL-61-614 Poznan, Poland; email: iamb@genesilico.pl

Motivation

Protein-RNA interactions play fundamental roles in many biological processes, such as regulation of gene expression, RNA splicing, and protein synthesis. The understanding of these processes improves as new structures of protein-RNA complexes are solved and the molecular details of interactions analyzed. However, experimental determination of protein-RNA complex structures by high-resolution methods is tedious and difficult. Therefore, studies on protein-RNA recognition and complex formation present major technical challenges for macromolecular structural biology. Alternatively, protein-RNA interactions can be predicted by computational methods. Although less accurate than experimental measurements, theoretical models of macromolecular structures can be sufficiently accurate to guide experimental analyses and aid in the interpretation of their results.

Methods

I will present an overview of strategies and methods for computational modeling of RNA structure and RNA-protein complexes developed in our laboratory (available at http://genesilico.pl), and I will illustrate it with practical examples of structural predictions.

Results - References:

- Rother, K., Rother, M., Boniecki, M., Puton, T. and Bujnicki, J.M. (2011) RNA and protein 3D

Organizing Partners

structure modeling: similarities and differences. J Mol Model, 17, 2325-2336.
- Tuszynska, I. and Bujnicki, J.M. (2011) DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking. BMC Bioinformatics, 12, 348.
- Rother, M., Rother, K., Puton, T. and Bujnicki, J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic Acids Res, 39, 4007-4022.
- Puton, T., Kozlowski, L., Tuszynska, I., Rother, K. and Bujnicki, J.M. (2012) Computational methods for prediction of protein-RNA interactions. J Struct Biol, 179, 261-268.
- Tuszynska I, Matelska D, Magnus M, Chojnowski G, Kasprzak JM, Kozlowski L, Dunin-Horkawicz S, Bujnicki JM (2014) Computational modeling of protein-RNA complex structures Methods 65(3):310-9.
- Smietanski M, Werner M, Purta E, Kaminska KH, Stepinski J, Darzynkiewicz E, Nowotny M, Bujnicki JM (2014) Structural analysis of human 2'-O-ribose methyltransferases involved in mRNA cap structure formation Nature Commun 5:3004, doi:10.1038/ncomms4004
- Głów D, Pianka D, Sulej AA, Kozłowski ŁP, Czarnecka J, Chojnowski G, Skowronek KJ, Bujnicki JM. Sequence-specific cleavage of dsRNA by Mini-III RNase. Nucleic Acids Res. 2015 43(5):2864-73.
- Waleń T, Chojnowski G, Gierski P, Bujnicki JM. ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. Nucleic Acids Res. 2014 42(19)
- Magnus M, Matelska D, Lach G, Chojnowski G, Boniecki MJ, Purta E, Dawson W, Dunin-Horkawicz S, Bujnicki JM. Computational modeling of RNA 3D structures, with the aid of experimental restraints. RNA Biol. 2014;11(5):522-36.
- Tuszynska I, Magnus M, Jonak K, Dawson W, Bujnicki JM. NPDock: a web server for protein-nucleic acid docking. Nucleic Acids Res. 2015 Jul 1;43(W1):W425-30
- Stefaniak F, Chudyk E, Bodkin M, Dawson WK, Bujnicki JM Modeling of RNA-ligand interactions Wiley Interdiscip Rev Comput Mol Sci 2015 Sep 14, doi: 10.1002/wcms.1226
- Dawson WK, Bujnicki JM Computational modeling of RNA 3D structures and interactions Curr Opin Struct Biol. 2015 Dec 12;37:22-28. doi: 10.1016/j.sbi.2015.11.007.
- Boniecki MJ, Lach G, Dawson WK, Tomala K, Lukasz P, Soltysinski T, Rother KM, Bujnicki JM. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic Acids Res. 2015 Dec 19. [Epub ahead of print doi: 10.1093/nar/gkv1479

Quick Links

Useful Infomation

Tutorials

Satellite Events

Impact of modified nucleobases on base pairing in RNA experimental structures

Chawla M(1), Oliva R(2), Bujnicki JM(3), Cavallo L(1)

(1) Kaust Catalysis Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
(2) Department of Sciences and Technologies, University "Parthenope" of Naples, Napoli
(3) International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

Motivation

Posttranscriptional modifications greatly enhance the chemical information of RNA molecules, contributing to explain the diversity of their structures and functions. A significant fraction of RNA experimental structures available to date present modified nucleobases, with half of them being involved in H-bonding interactions with other bases, i.e. 'modified base pairs'. In addition, several non-natural (synthetic) nucleobases are being introduced in RNA and DNA molecules for targeted applications. This prompted us to investigate the impact of natural and non natural modifications on the H-bonding propensity of nucleobases in nucleic acid structures.

Methods

All PDB structures solved by X-ray crystallography at a resolution of 3.5 Å or better and containing RNA molecules with posttranscriptional modifications were analysed using the BPView tool, in order to identify the modified base pairs and classify their geometry. As a result of this analysis, we obtained 573 base pairs containing at least one modified base. Base pairs containing non natural modifications were also modelled starting from available experimental structures in the PDB. The geometries of the base pairs were optimized with a Density Functional Theory (DFT) approach using the hybrid B3LYP functional in connection with the TZVP basis set. Interaction energies, defined as the difference between the energy of the base pair and the energy of the isolated free bases, were evaluated on the DFT optimized geometries at the 2nd-order Moeller-Plesset level of theory.

Results

We present a systematic investigation of modified base pairs, in the context of experimental nucleic acid structures. To this end, we first compiled an atlas of experimentally observed naturally modified base pairs in RNAs, for which we recorded occurrences and structural context. Then, for each base pair, we selected a representative for subsequent quantum mechanics calculations, to find out its optimal geometry and interaction energy. Our structural analyses show that most of the modified base pairs are non Watson-Crick like and are involved in RNA tertiary structure motifs. Similar analyses were also performed on base pairs involving some non natural modifications of particular biotechnological interest, in the context of RNA or DNA structure.

The combined bioinformatics and quantum mechanics studies we performed help provide a rationale for the impact of the different modifications on the geometry and stability of the base pairs they participate in, and possibly predict the effect of newly designed modifications [1-3].

References

- [1] Chawla M., Oliva R., Bujnicki J.M., Cavallo L. (2015) An atlas of RNA base pairs involving modified nucleobases with optimal geometries and accurate energies, Nucleic Acids Res., 43, 6714-29.
- [2] Chawla M., Credendino R., Oliva R., Cavallo L. (2015) Structural and Energetic Impact of Non-Natural 7-Deaza-8-Azaadenine and Its 7-Substituted Derivatives on H-Bonding Potential with Uracil in RNA Molecules. J Phys Chem B 119, 12982-9.
- [3] Chawla M., Credendino R., Chermak E., Oliva R., Cavallo L. (2016) Theoretical Characterization of the H-Bonding and Stacking Potential of Two Non-Standard Nucleobases Expanding the Genetic Alphabet. J Phys Chem B 120, 2216-24.

Searching for sequence motifs that affect splicing of exons regulated by RBM20

Dal Molin A, Lorenzi P, Romanelli MG, Malerba G

Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona

Motivation

RNA binding motif protein 20 (RBM20) has been recently involved in the alternative splicing of several human and rat genes. RBM20 mutations have been associated to human dilated cardiomyopathy. Although a RBM20 RNA-binding site consensus sequence has been proposed and an over-representation of the site was noticed in the intronic sequence of the target genes in close proximity to differentially spliced exons, a reliable rule to recognize a RBM20-RNA-binding site is still missing. We aim to characterize the sequence features of the RBM20-RNA-binding site studying the exon flanking sequences affected by RBM20, through bioinformatic approaches.

Methods

RNA raw sequencing data (RNA-Seq) of rat and human cardiomyocytes were downloaded from public database and used. Rat samples were grouped according their genotype for the RBM20 gene: 3 wild type (RBM20+/+), 3 heterozygotes (RBM20+/-) and 3 gene-deleted (RBM20-/-) samples. Human samples showed 2 different conditions: 2 control individuals (RBM20+/+) and 1 RBM20-mutated individual (RBM20-/-). The rat RGSC3.4 and human GRCh37 reference genomes were used to map the rat and the human RNA sequence data, respectively, and the expression level of each individual exon was measured by digital counting. An analysis was carried out to identify differentially spliced exons between RBM20-positive and RBM20-negative samples, for both species. Only the differentially spliced exons common between the two species were selected. The reported RBM20 RNA-binding site was searched on the flanking sequences of the differentially spliced exons selected and in an equal number of control exons, in combination with a motif enrichment analysis. A Support Vector Machine (SVM) model was used to discriminate between differentially spliced and not differentially spliced exons, incorporating several features such as the number, size and position of single patterns and clusters of patterns, and the nucleotides and dinucleotides frequency for the selected sequences (404 features).

Results

Differential gene expression analysis between RBM20-positive (RBM20+/+) and RBM20-negative (RBM20-/-) samples detected 96 differentially spliced exons belonging to 33 genes, and 1816 differentially spliced exons belonging to 657 genes, in rat and humans, respectively (p. adj < 0.05). A slightly significant increased number of patterns was observed in the post-exon region of selected exons, respect to control exons (Fisher exact test). The enrichment analysis showed 3 novel motifs more common in selected than in control sequences. SVM results are given evaluating the AUC value of the ROC curve resulted from the analyses for single features or groups of features. Neither the RBM20 RNA-binding site or the 3 novel motifs alone can help SVM to clearly distinguish between RBM20 affected and RBM20 not affected exons. Purines and pyrimidines percentage in the close pre-exon region, TC percentage in the close post-exon region and the size of the selected exons, together, resulted in a value of AUC > 0.8 (AUC range: 0-1) of ROC curve.

To improve the classification, additional features based on the 2D and 3D structure of differentially spliced exons flanking sequences and alternative statistical methods will be investigated in the near future.

An integrated multi-level comparison highlights common aspects and specific features between distantly-related species Tomato and Grapevine

Ambrosino L, Bostan H, Ruggieri V, Chiusano ML

Department of Agriculture, University of Naples Federico II, Portici, Italy

Motivation

Even after years from the first completion of genomes by sequencing, comparative genomics still remains a challenge, also enhanced by the availability of numerous draft genomes with still poor annotation quality. The detection of ortholog genes between different species is a key approach for comparative genomics. For example, ortholog gene detection may support investigations on mechanisms that shaped the organization of the genomes, highlighting on gain or loss of function and on gene annotation. On the other hand, the detection of paralog genes is fundamental for understanding the evolutionary mechanisms that drove gene function innovation and support gene families analyses.

Here we report on the gene comparison between two distantly related plants, Solanum lycopersicum (Tomato) (The Tomato Genome Consortium 2012) and Vitis vinifera (Grapevine) (Jaillon et al. 2007), considered as economically important species from asterids and rosids clades, respectively. The strategy was accompanied by integration of multilevel analyses, from domain investigations to expression profiling, to get to the most reliable results and to offer powerful resources, in order to understand different useful aspects of plant evolution and physiology and to dissect traits and molecular aspects that could provide novel tools for agriculture applications and biotechnologies.

Methods

In order to predict best putative orthologs and paralogs between Tomato and Grapevine, and to overcome possible annotation issues, all-against-all sequence similarity searches between genes, mRNAs and proteins collections of both species were performed. A Bidirectional Best Hit approach was implemented to detect the best orthologs between the two species. Moreover we developed a dedicated algorithm in Python programming language able to define more extended alignments between mRNA sequences. NetworkX package (Hagberg et al. 2008) was used to define networks of paralogs and orthologs. Proteins domain prediction was carried out on the entire Tomato and Grapevine protein collection by using InterProScan program (Jones et al. 2014). The enzyme classification was obtained by sequence similarity searches between Tomato and Grapevine mRNA collections and the entire UniProt reviewed protein collection (UniProt consortium 2015). The metabolic pathways associated to the detected enzymes were identified exploiting the KEGG Database (Kanehisa and Goto 2000). Expression level of three developmental stages of Tomato (2 cm fruit, breaker and mature red) and the corresponding stages of Grapevine (post-setting, veraison,

mature berry) was defined on the basis of the iTAG loci (Shearer et al. 2014) and v1 vitis loci, respectively. The expression was normalized by Reads Per Kilobases per Million (RPKM) for each tissue/stage. The identification of similar expression profiles was performed by the K-means clustering method (Soukas et al. 2000), using the Pearson correlation coefficient as distance metric. For each cluster a subsequent clusterization by the Hierarchical Clustering (HCL) (Eisen et al. 1998) using the Euclidean distance grouped genes also on the basis of expression levels. Both the clustering methods used are those from implemented in the MultiExperiment Viewer (MeV) software.

Results

Although Tomato and Grapevine are phylogenetically distant species, they are both model species for understanding fleshy fruit formation. Comparative analyses, though the available annotations are still preliminary, are essential to understand fruit development. We predicted the presence of a strong core of orthologs genes, exploiting an appropriate approach and overcoming the annotation limits.

Networks of ortholog/paralog genes were built between the compared species, offering resources to support studies about the organization and the evolution of gene families in different organisms. By this approach, we detected gene families of one species that underwent an expansion/reduction in the number of their elements when compared to the other species.

Species-specific genes of Tomato and Grapevine were also detected.

The protein domains common to both species, as the ones exclusively detected in Tomato and Grapevine, and the common and the distinctive enzymatic classes associated to related metabolic pathways, were also predicted for the two compared species supporting structure and functional annotations.

Furthermore, the association of RNA-seq data offered an additional information level for comparing gene functionalities from the two species. Thanks to this core collection, we report on similarities and peculiarities between the two genomes.

Session 2: Special Session: Big Data Management, Modeling and Computing

HIGHLIGHT LECTURE

Data-Driven genomic computing: Making sense of the signals from the genome

Stefano Ceri

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano

E-mail: stefano.ceri@ polimi.it

Next Generation Sequencing (NGS) allows the production of the entire human genome sequence at a cost of about 1000 US $; many algorithms exist for the extraction of genome features, or "signals", including peaks (enriched regions), mutations, or gene expression (intensity of transcription activity). The missing gap is a system supporting data integration and exploration, giving a "biological meaning" to all the available information. The GeCo Project (Data-Driven Genomic Computing, ERC Advanced Grant currently undergoing the contract preparation) has the objective of revisiting genomic computing through the lens of basic data management. Starting from an abstract data model, we already developed a system that can be used to query processed ENCODE, TCGA, and Roadmap Epigenomics data; the system employs internally the Spark, Flink, and SciDB data engines, and prototypes can already be accessed from CINECA servers or be downloaded from PoliMi servers.

During the five-years of the ERC project, the system will be enriched with data analysis tools and environments and will be made increasingly efficient. Among the objectives of the project, the creation of an "open source" system available to biological and clinical research; while the GeCo project will provide public services which only use public data (anonymized and made available for secondary use, i.e., knowledge discovery), the use of the GeCo system within protected clinical contexts will enable personalized medicine, i.e. the adaptation of therapies to specific genetic

features of patients. The most ambitious objective is the development, during the 5-year ERC project, of an "Internet for Genomics", i.e. a protocol for collecting data from Consortia and individual researchers, and a "Google for Genomics", supporting indexing and search over huge collections of genomic datasets.

Algorithms and data structures for the compression and indexing of genomic big data

Prezza N(1), Policriti A(1,2)

(1) Department of Computer Science, Mathematics, and Physics, University of Udine, Italy
(2) Istituto di Genomica Applicata, Udine

Motivation

Building the Burrows-Wheeler transform (BWT) and computing the Lempel-Ziv parsing (LZ77) of huge collections of genomes is becoming an important task in bioinformatic analyses as these datasets often need to be compressed and indexed prior to analysis. Given that the sizes of such datasets often exceed RAM capacity of common machines however, standard algorithms cannot be used to solve this problem as they require a working space at least linear in the input size. One way to solve this problem is to exploit the intrinsic compressibility of such datasets: two genomes from the same species share most of their information (often more than 99%), so families of genomes can be considerably compressed. A solution to the above problem could therefore be that of designing algorithms working in compressed working space, i.e. algorithms that stream the input from disk and require in RAM a space that is proportional to the size of the compressed text.

Methods

In this talk I will present algorithms and data structures to compress and index text in compressed working space. These results build upon compressed dynamic data structure, a sub-field of compressed data structures research that is lately receiving a lot of attention. I will focus on two measures of compressibility: the empirical entropy H of the text and the number r of equal-letter runs in the BWT of the text. I will show how to build the BWT and LZ77 using only O(Hn) and (rlog n) working space, n being the size of the collection. For the case of repetitive text collections (such as sets of genomes form the same species), this considerably improves the working space required by state-of-the art algorithms in the literature. The algorthms and data structures here discussed have all been implemented in a public C++ library, available at github.com/nicolaprezza/DYNAMIC. The library includes dynamic gap-encoded bitvectors, run-length encoded (RLE) strings, and RLE FM-indexes.

Results

I will conclude the talk with an overview of the experimental results that we obtained running our algorithms on highly repetitive genomic datasets. As expected, our solutions require only a small fraction of the working space used by solutions working in non-compressed space, making it feasible to compute BWT and LZ77 of huge collections of genomes even on desktop computers with small amounts of RAM available. As a downside of using complex dynamic data structures however, running times are still not practical so improvements such as parallelization may be needed in order to make these solutions fully practical.

TCGA2BED: converting and querying The Cancer Genome Atlas

Cumbo F(1), Fiscon G(1), Ceri S(2), Masseroli M(2), Weitschek E(1,3)

1) Institute of Systems Analysis and Computer Science "Antonio Ruberti" – CNR, Italy
2) Department of Electronics, Information, and Bioengineering - Politecnico di Milano, Italy
3) Department of Engineering – Uninettuno International University, Italy

Motivation

Thanks to the great advances in biomedical technologies, we are faced with huge amounts of genomic and clinical data. A striking example is The Cancer Genome Atlas (TCGA), one of the largest public repositories of genomic and clinical data about cancer. TCGA contains more than 15 TB of genomic and clinical data, whose analysis and interpretation are posing great challenges to the bioinformatics community.

In this work, we focus on data retrieval, conversion, integration and querying of Next Generation Sequencing (NGS) data and their clinical information extracted from TCGA. In particular, we take into account all publicly available Copy Number Variation (CNV), DNA-methylation, DNA-sequencing (DNA-seq), Gene Expression (RNA-seq V1 and V2), microRNA sequencing (miRNA-seq), and meta (clinical and biospecimen) data.

Methods

We propose TCGA2BED, a software tool able to retrieve genomic and clinical data from TCGA and convert them into the tab-delimited BED format. Additionally, it integrates them with external data (e.g., gene coordinates) from other state-of-the-art biological databases and services such as UCSC Genome Browser, HUGO Gene Nomenclature Committee (HGNC), NCBI Gene, and miRBase. TCGA2BED is available with a graphic user interface and includes three different main components:
- the controller, that reads and executes the user's requests (i.e., data download and conversion) through the graphic user interface or a XML configuration file

- the retrieval system, which handles the search and retrieval of the public genomic and clinical data available from TCGA by building ad-hoc queries and send them to the REST service of TCGA

- the BioParser, which converts all TCGA genomic data types (i.e., CNV, DNA-methylation, DNA-seq, miRNA-seq, and RNA-seq V1 and V2) into the tab-delimited BED format, and all their related clinical meta data into a tab-delimited attribute-value text format.

Figure 1 shows the TCGA software architecture.

Results

By using TCGA2BED, we downloaded and converted all publicly available CNV, DNA-methylation, DNA-seq, miRNA-seq, and RNA-seq V1 and V2 experimental and meta data from TCGA. For each patient sample, cancer type and experiment type in TCGA, we create (i) a .bed file, containing the genomic data of the sample converted in BED format, and (ii) a .meta file, including the clinical data of the sample; additionally, (iii) a header.schema file in XML format that describes the structure of the .bed data files, and (iv) a .txt metadata dictionary file that contains all metadata attributes with all the values that each attribute assumes in the metadata.

The TCGA converted data can be easily processed and analysed with wide-spread bioinformatics tools, including the GenoMetric Query Language (GMQL) available at http://www.bioinformatics.deib.polimi.it/GMQL/, a key instrument for the integrative querying of genomic and clinical big data from heterogeneous sources. Here we report an example GMQL query that integrates DNA-seq and RNA-seq data; for each tumor sample of each patient, it searches and returns the DNA mutations that are the closest to expressed genes:

DNA = SELECT(*) DNAseq;

RNA = SELECT(*) RNAseq;

JoinDnaToRna = JOIN(left->bcr_sample_barcode == right->bcr_sample_barcode,

MINDISTANCE(1), left) DNA RNA;

MATERIALIZE JoinDnaToRna;

The use of the BED format reduces the time spent in managing and analyzing the valuable TCGA data: it is possible to efficiently deal with huge amounts of cancer data, and to easily integrate and query them using GMQL. The BED format facilitates the investigators in easily performing knowledge discovery analyses aiming at aiding cancer treatments. For example, the TCGA data in BED format can be straightforwardly analyzed with CAMUR, a tool using a supervised approach able to elicit a high amount of knowledge by computing many rule-based classification models, and therefore able to identify most of the clinical and genomic features related to the predicted cancer type.

Discovering similar (epi)genomics feature patterns in multiple genome browser tracks

Montanari P(1), Ceol A(2), Bartolini I(1), Ciaccia P(1), Patella M(1),

Ceri S(3), Masseroli M(3)

(1) School of Engineering, DISI - Università di Bologna, Mura Anteo Zamboni 7, 40126 Bologna, Italy
(2) Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia (IIT), Via Adamello 16, 20139 Milan, Italy
(3) Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Motivation

Next Generation Sequencing (NGS), with the high amount of heterogeneous data that it is generating, is opening many interesting practical and theoretical computational problems. Genome browsers, e.g. UCSC Genome Browser (Kuhn et al., 2013) or Integrated Genome Browser (IGB) (Nicol et al., 2009), allow visual inspection and identification of interesting patterns on multiple genome browser tracks, i.e. of sets of (epi)genomic regions/peaks at given distances from each other in different tracks. For example, such patterns can describe gene expression regulatory DNA areas including heterogeneous (epi)genomic features (e.g. histone modification and/or different transcription factor binding regions). Yet, once such patterns are visually identified in a genome section, the search of their occurrences along the whole genome is a complex computational task that is currently not supported, despite their discovery along the whole genome is very important for the biological interpretation of NGS experimental results and comprehension of biomolecular phenomena.

We defined an optimized pattern-search algorithm able to find efficiently, within a large set of (epi)genomic data, genomic region sets which are similar to a given pattern. We implemented it within an IGB plugin, which allows intuitive user interaction in both the visual selection of an interesting pattern on the loaded IGB tracks, and the visualization of occurrences of similar patterns identified along the entire genome.

Methods

Pattern matching is a recurrent problem in data science, typically solved by a cost based approach, where lower cost implies high similarity. Although the Best-Matching Problem (BMP) is suspected to be NP-hard, we propose an alternative Root-element approach (R-BMP) and a Dynamic Programming algorithm (DP-BMP) that lowers the complexity to order of O(M*N^2), with M and N the number of elements in a pattern and target track to be compared. Given the properties of the genomic data to which the algorithm will be applied (strictly increasing sequences, M << N), it is possible to obtain the best match for each element of the pattern in the target track through a binary search. With the resulting Windowed DP-BMP algorithm, the complexity can drop down to O(N*log(N)), making it applicable also to (very) large problem instances.

We extended this model to introduce the aspects missing in the base version, but critical for its application to NGS data: interval regions, multiple, partial, and negative tracks, region attribute matching, and top-K distinct matching. Interval regions can either be reduced to their centroid, or analyzed with an asymmetric approach, which takes into account the region length. Negative matching tracks are considered when no region should be present for this track in the area of a result. Such regions are removed from the search space before search start. Partial matching are tracks that can be reasonably missing in the results. The cost for not matching an element in that track is consequently reduced. Region attributes can also be used to alter the cost of matching the elements in a track. Finally, in order to facilitate the discovery of pattern matching and increase the diversity of results, we implemented a top-K version of the algorithm, which compares the results produced and keep the best K disjoint results.

The algorithm has been implemented in Java 8, and integrated as a plugin for IGB.

Results

We extended IGB with a plugin, which provides biologists with a tool to visually inspect the genome browser's tracks to identify and select a pattern of possible interest, and search other instances of this pattern in the same or different tracks. It is also possible to load the pattern from a file or from a selection of tracks, related for instance to histone marks identified by chromatin immunoprecipitation sequencing (ChIP-seq), for which a peak should or should not be present in targeted regions.

A new track is created for each search query: all regions that match the pattern are highlighted and can be easily browsed. Because a similarity search can be repeated on different groups of samples (for instance treated/non treated, or control/disease), it is possible to compare the resulting tracks and to identify differences in (epi)genomic features, suggesting mechanisms for the response to treatments or for the investigated pathology.

Several "chromatin states" identified by different combination of histone marks were inferred during the Roadmap Epigenomics project (Ernst et al., 2012). Such patterns can be submitted to our plugin to infer the regulation state of genomic regions under different conditions. Because the tool is not limited to the analysis of ChIP-seq peaks, but can be applied to any (epi)genomic regions, analyses can be extended by integrating other features, such as differentially expressed genes (DEG), DNase I hypersensitive sites (DHS), transcription start sites (TSS), or single nucleotide polymorphisms (SNP).

References:

- Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 2012; 9(3), 215–216.
- Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief. Bioinform. 2013; 14(2): 144–161.
- Nicol JW, Helt GA, Blanchard SG Jr, Raja A, Loraine AE. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 2009; 25(20), 2730–2731.

PATRI: an integrated platform for genomics data analysis

Bosotti R(1), Melloni GM(2), Ukmar G(1), D'aiuto F(3), Vescovi M(4), Rossi P(5), Pirchio MR(5), Callari M(3), Leone A(1), Somaschini A(1), Cesarini M(2), Raddrizzani L(1), Della Vedova G(2), Dugo M(3), Canevari S(3), Zambon A(2), Corrao G(2), Nepa G(5), Daidone MG(3), Pettenella M(4), Isacchi A(1)

(1) NMS Oncology, Nerviano Medical Sciences Srl, Nerviano (MI)
(2) University of Milano Bicocca, Milano
(3) Fondazione IRCCS Istituto Nazionale dei Tumori, Milano
(4) Parametric Design Biotech, Gessate (MI)
(5) Icona Srl, Cinisello Balsamo (MI)

Motivation

The growing availability of genomic data, accessible through several repositories both from public and private institutions, raised the need for the integration of these different types of data to allow their analysis in correlation with sensitivity to different drug treatments and/or with other preclinical and clinical parameters. The new emerging high-throughput "omic" technologies have generated an enormous amount of molecular information on both cancer cell lines and clinical samples. Different molecular data types can be obtained from either cell line models or clinical tumor samples, including gene expression values, copy number variations and somatic mutations. The integrated information can then be exploited for a more systematic and comprehensive analysis. An efficient data mining strategy might indeed enable the discovery of new connections within the data, ultimately leading to the identification of potentially new sensitivity biomarkers and new pharmacological targets. However organization, integration and interpretation of such heterogeneous genomic data require complex bioinformatics and computational skills, limiting data accessibility.

Here we present PATRI (Platform for the Analysis of TRanslational Integrated data), a platform integrated with statistical tools for the storage and analysis of genomic data.

Methods

PATRI resides on a mySQL database and uses Open Source R libraries for the statistical analysis of molecular data. The database is accessible through a user-friendly graphical interface, developed in Javascript. The Query interface allows the selection of lists of samples (tumors or cell lines), labeled on the basis on their sensitivity to a specific drug treatment, for statistical comparison either directly from the database or alternatively through the upload of a user provided list. The system automatically retrieves and associates the available genomics data (mutations, gene expression, copy number) to the list of provided samples. The user can then select the desired statistical analysis from a panel of pre-defined statistics. Gene expression, copy number and gene variant analyses can be performed separately or combined in a unique run. Results are visualized by means of heat maps, volcano plot and hierarchical clusters.

Results

PATRI is a flexible, user-friendly data integration resource, which allows data mining of different types of genomic data, associated to sample annotation information.

The approach aims at identifying molecular markers discriminating any two conditions (e.g sensitivity or resistance to a drug) starting from molecular data on untreated samples and leading to the detection of potentially new predictive biomarkers. Typical inputs can be lists of samples, classified according to different criteria, such as groups of cell lines with differing/opposite sensitivity to a compound treatment, or categories of clinical samples resulting from the treatment of patients with specific drugs. Sensitivity data are not limited to drug treatments, but can arise from RNA interference screenings or other perturbations as well. Uploaded samples, either cancer cell lines or tumoral tissues, are firstly labeled in categories based on relative sensitivity to a specific treatment or condition; then statistically significant correlations are calculated between molecular data and drug activity profiles.

The capability of identifying markers of sensitivity to a compound, for which the molecular discriminants are known, has been evaluated in PATRI upon treatment of a panel of cancer cell lines with increasing doses of different drugs to generate IC50 values that can be used to classify samples accordingly to their sensitivity to the treatment. PATRI analysis resulted in the correct prediction of the sensitivity biomarkers for each tested drug.

Hash Clone: a new tool to quantify the minimal residual disease during patient follow-up

Cordero F(1), Beccuti M(1), Genuardi E(2), Romano G(1), Calogero RA(3),

Ladetto M(4), Ferrero S(2)

(1) Department of Computer Science, University of Torino, Torino, Italy;
(2) Department of Molecular Biotechnologies and health sciences, Hematology Division, University of Torino, Torino, Italy;
(3) Department of Molecular Biotechnology and Health Sciences, Molecular Biotechnology Center, University of Turin, Turin, Italy.
(4) Division of Hematology, Az Ospedaliera SS Antonio e Biagio e Cesare Arrigo, Alessandria, Italy

Motivation

In cancer, genome studies revealed extensive intra-tumor heterogeneity as essential determinant of disease progression, thus impacting tumor diagnosis and treatment. A high level of heterogeneity affects several cancer types and may contribute to the treatment failure, by initiating phenotypic diversity and enabling more aggressive and drug-resistant clones. In B-cell lymphoproliferative diseases, the detection of cancer cell clonality is used to monitor the therapeutic response, in terms of minimal residual disease (MRD), responsible of treatment failure and disease recurrence.

Next Generation Sequencing (NGS) technology can overcome the limitations of the standardized RQ-PCR-based MRD method thanks to its high sensitivity, specificity, accuracy and reproducibility. More importantly, NGS MRD approach allows a full repertoire analysis through multi-clones

detection at diagnosis and it gives the opportunity to monitor all the neoplastic clones at several follow-up time ponits. This issue is coped by an appropriate computational analysis of the huge volume of complex data obtained by NGS.

Methods

We present an innovative bioinformatics approach, called HashClone, for NGS analysis that can be applied to MRD detection of B-cell neoplasms, through the study of rearrangements of immunoglobulin genes. HashClone is an easy-to-use and reliable bioinformatics algorithm that provides a clonality assessment and the MRD detection over time. HashClone is composed of four C++ applications for the data processing based on the analysis of all set of sample reads simultaneously and returns the corresponding set of clone aligned with respect to V-GENE, J-GENE, and D- GENE sequences, associated with their frequency in all input samples.

HashClone approach can be divided in three different subtasks: (i) the selection of significant k-mers according to their frequencies in each sample; (ii) Generation of read signatures as a combination of information of significant k-mers present in the read, and (iii) identification of putative tumor clones according to the frequencies of the signatures and validation of these putative clones through IMGT reference database.

Results

We present two studies. In the first, we analyzed four technique replicates of a pool of B-lymphoblastic leukemia cell lines, while in the second study we investigated the full repertoire clones in five patients affected by mantle cell lymphoma (MCL). In this second study we also collected data from four follow-up samples for each patient, that we analyzed in order to extrapolate the clones trend overtime.

We demonstrated that HashClone has better performance to investigate clonotype and clone abundance in leukaemia and lymphoma samples respect to Vidjil, the state-of-the-art available algorithm. We applied our algorithm on NGS-data, in order to investigate the different clonal landscape among the five MCL patients in terms of characteristic immunoglobulin-defined clones. Moreover, we investigated tumor clones kinetics in several follow-up samples of all patients ("MRD analysis").

Thanks to the analysis of different studies, we will test the ability of HashClone to identify neoplastic clones with the best performance in sensibility, sensitivity and scalability, with the intention to manage different kind of tumor samples. Moreover, HashClone is able to describe that the NGS-based MRD method overpasses the limitations of the classical RQ-PCR in MRD analysis, revealing its further potential application to study the clonal evolution on a multitude of cancer types.

Session 3: Next-Generation Sequencing

INVITED LECTURE

The FAIR Guiding Principles for scientific data management and stewardship

Susanna-Assunta Sansone

Associate Director,Oxford e-Research Centre

(Life, Natural and BioMedical Sciences)

University of Oxford, Oxford, UK

susanna-assunta.sansone@oerc.ox.ac.uk

http://uk.linkedin.com/in/sasansone

twitter: @SusannaASansone

Single Cell RNAseq reveals RNA editing heterogeneity in human brain

Picardi E(1,2), D'Erchia AM(1,2), Pesole G(1,2)

(1) Dipartimento di Bioscienze, Biotecnologie e Biofarmaceutica, Università di Bari, Bari, Italy
(2) Istituto Biomembrane e Bioenergetica del Consiglio Nazionale delle Ricerche, Bari, Italy

Motivation

A-to-I RNA editing in human is carried out by members of ADAR family of enzymes that act on double strand RNAs and can alter codon identity, splicing sites or base-pairing interactions within higher-order RNA structures. Recoding RNA editing is essential for normal brain development and regulates important functional properties of neurotransmitter receptors [1, 2]. Indeed, its deregulation has been linked to several nervous diseases such as epilepsy, schizophrenia, Alzheimer, major depression and amyotrophic lateral sclerosis [3, 4]. Recently we have profiled RNA editing in six different human tissues using whole transcriptome sequencing and detected more than three million events [5]. Interestingly, genes undergoing RNA editing were consistently enriched in genes involved in neurological disorders and cancer, confirming the relevant biological role of RNA editing in human.

Although investigations in bulk tissues are extremely useful, they do not capture the transcriptomic heterogeneity of multiple cell types constituting the ensemble tissue.

Methods

To characterize the complexity of RNA editing at single cell resolution, we investigated this phenomenon in single cells from adult human cortex obtained from living subjects in which transcriptome diversity was already surveyed by single cell RNA sequencing (scRNA-seq) [6]. Using a comprehensive collection of known RNA editing events, we explored inosinome profiles in 466 cortex cells. Individual scRNAseq data were quality checked by FASTQC and poor regions at 3' ends were trimmed by means of trim_galore tool. Cleaned read were then mapped onto the human reference genome by STAR aligner. RNA editing candidates were detected using our REDItools [7] and analyzed by custom scripts.

Results

We found that the identification of A-to-I RNA editing in single cells was strongly correlated with the amount of generated RNA reads. RNA editing profiles were quite heterogeneous also inside the same cell population. However, the observed RNA editing profile as well as the Alu editing index were sufficient to discriminate major cell types as neurons, astrocytes and oligodendrocytes, underlining the cell specific nature of RNA editing. Interestingly, recoding RNA editing were mainly detectable in neurons, remarking the primary role of A-to-I editing in modulating brain functions through key modifications in receptors for neurotransmitters.

References

1. Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM: Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science 2009, 324(5931):1210-1213.

2. Mehler MF, Mattick JS: Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease. Physiological reviews 2007, 87(3):799-823.
3. Maas S, Kawahara Y, Tamburro KM, Nishikura K: A-to-I RNA editing and human disease. RNA biology 2006, 3(1):1-9.
3. Khermesh K, D'Erchia AM, Barak M, Annese A, Wachtel C, Levanon EY, Picardi E, Eisenberg E: Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. Rna 2016, 22(2):290-302.

4. Picardi E, Manzari C, Mastropasqua F, Aiello I, D'Erchia AM, Pesole G: Profiling RNA editing in human tissues: towards the inosinome Atlas. Sci Rep 2015, 5:14941.
6. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Hayden Gephart MG, Barres BA, Quake SR: A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences of the United States of America 2015, 112(23):7285-7290.
5. Picardi E, Pesole G: REDItools: high-throughput RNA editing detection made easy. Bioinformatics 2013, 29(14):1813-1814.

A tool for genotyping variable length tandem repeats in high throughput sequencing data

Geraci F(1), Manzini G(1,2), Genovese LM(1), D'Aurizio R(3), Pellegrini M(1,3)

(1) Istituto di Informatica e Telematica del CNR, Pisa.
(2) Istituto di Informatica, DiSIT Universita' del Piemonte Orientale, Alessandria.
(3) LISM, Istituto di Informatica e Telematica e Istituto di Fisiologia Clinica del CNR, Pisa

Motivation

The human genome is rich in tandem repetitions, whose study is important for a wide range of applications, such as: forensics, medical genetics, and population studies. Individual variability in TR is directly linked to dozens of diseases, mostly neurodegenerative and neuromuscolar disorders, including Huntington disease (HD), Kennedy disease (SBMA), and several types of Spinocerebral Ataxias (SCA) [Orr et al 2007]. Variability in TR may also influence predisposition to cancer [Boland et al. 1998]. With the advent of high throughput sequencing technology it is now possible to sequence large cohorts of patients and search for highly variable TR in the sequenced data within genotype/phenotype association studies. There is, however, a pressing need for computational tools supporting for this type of studies, and this research area has been burgeoning in the last few years. Currently the available tools like lobSTR [Gymrek et al 2012] or VNTRseek [Gelfand 2014] suffer from some limitations: they aim at analyzing just short tandem repeats, and are very computationally demanding. There is thus space for improvements.

Methods

We have developed a new method for genotyping variable length tandem repeats in high throughput sequencing data, attaining better precision than existing state of the art methods. Our algorithm works in four phases: flanking regions alignment, TR alignment, copy number prediction, and statistical evaluation of alleles length. In each of these phases we use innovative algorithms and data structures to reach high precision and speed.

Flanking regions alignment accepts as input the upstream and downstream sequences for each tandem repeat of interest and align them with the set of reads. To speed up the alignment we preliminary build a data structure for the set of upstream and downstream sequences. The data structure allows the efficient search of overlaps between sequences and reads even in the presence of mismatches. Being based on the SDSL library [Gog et al. 2014], the memory footprint of the data structure is small even for large collections of sequences.

Locating the flanking regions is not enough to ensure that a read covers a tandem repeat. The TR alignment phase uses a semi-global sequence alignment algorithm to verify the presence of the TR between the flanking regions. Subsequently, we use a greedy approach to split the TR into non-overlapping intervals such that the overall hamming distance between each element and the TR motif sequence is minimized. We use the number of segments with length equals to the TR motif as an estimation of the copy number of the TR in the read.

In the last phase we build for each TR a histogram with the distribution of the number of reads on the predicted copy number. We than use a peak detection algorithm to determine the zigosity of the TR and alleles lengths.

Results

For our experiments we used the dataset suggested in [Gymrek et al 2012] consisting in the whole

genome NGS sequencing of human HapMap trio NA12877, NA12878, NA12882 using the Illumina HiSeq 2000 sequencer. Each sample has coverage higher than 30x and consists of about 820M of paired-end 100bp reads. We profiled a collection of 485 highly polymorphic tandem repeats extracted from the Marshfield panel [Ghebranious et al 2003].

We measured the Mendelian Inheritance (MI) coherence between the genotyping calls among the trio. We succeed when the genotype predicted for the son can be tightly traced to the genotype of the father and mother. Moreover, we measured a slightly less restrictive case where we admit a divergence of one unit of copy number from the predicted genotype of the son and the corresponding from one of the two parents.

According to our experiments, our tool has reported a consistent MI in 344 cases while with less stringent setting the number of consistent predictions increased to 405 cases. We compared this result with that obtained running lobSTR. This software has reported a tightly consistent MI in 307 cases and 338 in the more permissive case.

We further investigated the accuracy of prediction of the zygosity. To this end, we profiled the two males of the trio (NA12877 and NA12882) on the Y-STR: a collection of 85 homozygous tandem repeats lying on the Y chromosome. Our experiments show that we were able to exactly predict the homozygosity state in 62 (for NA12877) and 63 (for NA12882) cases, while in further 6 cases the error is bounded within one unit of copy number.

As for running time evaluation we report that our tool has been able to complete all three steps of the analysis of each sample within one hour using a moderate parallelism (8 threads) on the aligning phase while lobSTR exceeded the three hours of computation even with a higher degree of parallelism.

- *References:*

Harry T Orr and Huda Y Zoghbi. Trinucleotide repeat disorders. Annu. Rev. Neurosci., 30:575-621, 2007.
- Melissa Gymrek, David Golan, Saharon Rosset, and Yaniv Erlich. lobSTR: a short tandem repeat profiler for personal genomes. Genome research, 22(6):1154-1162, 2012.
- Thomas Willems, Melissa Gymrek, Gareth Highnam, David Mittelman, Yaniv Erlich, 1000 Genomes Project Consortium, et al. The landscape of human str variation. Genome research, 24(11):1894-1904, 2014.
- Yevgeniy Gelfand, Yozen Hernandez, Joshua Loving, and Gary Benson. VNTRseek: a computational tool to detect tandem repeat variants in high-throughput sequencing data. Nucleic acids research, page gku642, 2014.
- Nader Ghebranious, David Vaske, Adong Yu, Chengfeng Zhao, Gabor Marth, James L. Weber. STRP screening sets for the human genome at 5 cM density. BMC Genomics 4:6, 2003
- Simon Gog, Timo Beller, Alistar Moffat, Matthias Petri. From Theory to Practice: Plug and Play with Succinct Data Structures. Proc. 13th International Symposium on Experimental Algorithms. Springer Verlag Lecture Notes in Computer Science, vol. 8054:326-337, 2014.
- Richard Boland, Stephen Thibodeau, Stanley Hamilton, David Sidransky, James Eshleman, Randall Burt, Stephen Meltzer, Miguel Rodriguez-Bigas, Riccardo Fodde, Nadia Ranzani, et al. A national cancer institute workshop on microsatellite instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer research, 58(22):5248-5257, 1998.

Var2GO: a web-based tool for gene variants annotation and selection

Granata I*, Sangiovanni M*, Guarracino MR

High Performance Computing and Networking Institute, National Research Council of Italy-CNR, Napoli, Italy
** Equal contributors*

Motivation

Next Generation Sequencing (NGS) data analysis is a wide, cost-effective approach to identify and study genetic variants across the genome. Genetic variants, often the cause of complex and rare diseases, are usually investigated by whole genome or exome sequencing approaches. These techniques produce large amounts of data, which represent at the same time a powerful knowledge

source and a big challenge.

The output of the variant calling pipeline is a huge VCF (Variant Calling Format) file containing hundreds of thousand rows, which correspond to called single-nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs).

Inferring the biological meaning of the genetic variations is a crucial and laborious step. The variants are usually defined exploiting several annotation and effect prediction tools, such as ANNOVAR, snpEFF or AnnTools.

One of the most problematic issue in the whole process is handling the data generated from the annotation pipeline, and filtering the obtained genes to retain only the ones strictly related to the topic of interest.

Several tools permit to gather annotations at different levels of complexity for the detected genes and to group them according to the pathways and/or processes they belong to. However, it might be a time consuming and frustrating task due to the size of the resulting file, that might contain many thousands of genes, and to the search of associated variants, that usually requires a gene-by-gene investigation and annotation approach.

As a consequence, the initial gene list is often reduced exploiting some a-priori knowledge of variants effect, novelty and genotype, with the potential risk of loosing meaningful pieces of information.

Here we present Var2GO, a new web-based tool to support the annotation and filtering process of variants and genes coming from variant calling of high-throughput sequencing data.

Methods

Var2GO allows to upload and store an unprocessed VCF file into an on-the-fly generated database, through a simple and user-friendly interface. The variants are then submitted to the annotation step, realized by the integration of SnpEff and Annovar tools, to obtain the information needed for their characterization (gene name, functional class, minor allele frequency, etc.).

The genes associated to the variants are automatically annotated with the corresponding Gene Ontology terms covering the three GO domains: Molecular function, Cellular component, and Biological process.

Using the web interface is then possible to filter and extract, from the whole list, variants and genes having annotations in the domain of interest, by simply specifying filtering parameters and one or more keywords.

At each step the user can look at the resulting table and, in case, go back to change or add new filters. The resulting data can be downloaded, either as a complete table containing both the input variants and the gathered annotations, or as a simple gene list. Also the on-the-fly created and populated database can be downloaded, to be locally imported.

The relevance of this tool is demonstrated on a NGS exome sequencing dataset coming from a variant discovery study on a family with polyglucosan body myopathy (PGBM).

Results

Var2GO is a user-friendly and flexible tool that implements a topic-based approach, expressly designed to help biologists in narrowing the search of relevant genes coming from variant calling analyses.

When used on the PGBM data, Var2GO permitted to easily identify the variants that are likely involved in the disease. The list refinement was performed through the definition of several filters, involving annotations such as the ESP 6500 allele frequency, the VQSLOD filter (derived by the variant quality score recalibration walker of GATK to assign a well-calibrated probability to each variant call), the snpEFF functional class, the genotype, and the depth of coverage. We were able to reduce the number of genes to analyze from 19453 (size of the input list) to 352.

The resulting gene list allows to retrieve additional information from pathway and/or gene-disease association studies, such as the Database for Annotation, Visualization and Integrated Discovery (DAVID).

Var2Go is available online at http://www-labgtp.na.icar.cnr.it/VAR2GO.

INVITED LECTURE

Structure, affinity and specificity riddles in biomolecular interactions

Alexandre M. J. J. Bonvin

Computational Structural Biology Group, Department of Chemistry, Faculty of Science, Utrecht University, 3584CH, Utrecht, The Netherlands. a.m.j.j.bonvin@uu.nl

Motivation

The prediction of the quaternary structure of biomolecular macromolecules is of paramount importance for fundamental understanding of cellular processes and drug design. In the era of integrative structural biology, one way of increasing the accuracy of modelling methods used to predict the structure of biomolecular complexes is to include as much experimental or predictive information as possible in the process.

Methods

We have developed for this purpose a versatile information-driven docking approach HADDOCK (http://www.bonvinlab.org/software/haddock2.2) [1-3]. HADDOCK can integrate information derived from biochemical, biophysical or bioinformatics methods to enhance sampling, scoring, or both [4]. The information that can be integrated is quite diverse: interface restraints from NMR, mutagenesis experiments, or bioinformatics predictions; shape data from small-angle X-ray scattering [5] and, recently, cryo-electron microscopy experiments [6].

Results

In my talk I will illustrate HADDOCK's capabilities with various examples, including results from our participation to CAPRI [7,8]. HADDOCK has demonstrated sustained prediction and scoring performance since the start of its participation to CAPRI. This is due, in part, to its ability to integrate experimental data and/or bioinformatics information into the modelling process, and also to the overall robustness of the scoring function used to assess and rank the predictions. Thirteen years after the original publication, the HADDOCK scoring function remains a simple linear combination of OPLS [9] intermolecular van der Waals and Coulomb electrostatics energies, an empirically-derived desolvation energy term [10], and one or more restraints energy terms reflecting the agreement between models and experimental/prediction information. Our simple scoring scheme successfully selected acceptable/medium quality models for 18/14 of the 25 targets during the combined CASP-CAPRI prediction round, making us rank at the top. Considering that for only 20 targets acceptable models were generated our effective success rate reaches as high as 90% (18/20)! These results underline the success of our simple but sensible prediction and scoring scheme.

I will end by discussing the problem of binding affinity prediction, showing that current scoring functions in macromolecular docking fail at predicting the affinity of protein-protein complexes and introduce a simple, contact-based predictor that outperforms complex, energy-based predictors [11].

References

1. C.P van Zundert, J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, E. Karaca, A.S.J. Melquiond, M. van Dijk, S.J. de Vries and A.M.J.J. Bonvin. The HADDOCK2.2 webserver: User-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.*, **428**, 720-725 (2015).
2. J. de Vries, M. van Dijk and A.M.J.J. Bonvin The HADDOCK web server for data-driven biomolecular docking. *Nature Protocols*, **5**, 883-897 (2010).
3. Dominguez, R. Boelens and A.M.J.J. Bonvin HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **125**, 1731-1737 (2003).
4. P.G.L.M Rodrigues and A.M.J.J. Bonvin Integrative computational modeling of protein interactions. *FEBS J.*, **281**, 1988-2003 (2014).
5. Karaca and A.M.J.J. Bonvin. On the usefulness of Ion Mobility Mass Spectrometry and SAXS data in scoring docking decoys. *Acta Cryst. D.*, **D69**, 683-694 (2013).
6. C.P. van Zundert, A.S.J. Melquiond and A.M.J.J. Bonvin. Integrative modeling of biomolecular complexes: HADDOCKing with Cryo-EM data. *Structure*. **23**, 949-960 (2015).
7. P.G.L.M. Rodrigues, A.S.J. Melquiond, E. Karaca, M. Trellet, M. Van Dijk, G.C.P. Van Zundert, C. Schmitz, S.J. de Vries, A. Bordogna, L. Bonati, P.L. Kastritis and A.M.J.J. Bonvin Defining the limits of homology modelling in information-driven protein docking *Proteins: Struc. Funct. & Bioinformatics*, **81**, 2119-2128 (2013).
8. J. de Vries, A.S.J. Melquiond, P.L. Kastritis, E. Karaca, A. Bordogna, M. van Dijk, J.P.G.L.M. Rodrigues and A.M.J.J. Bonvin Strengths and weaknesses of data-driven docking in CAPRI *Proteins: Struc. Funct. & Bioinformatic*, **78**, 3242-3249 (2010).

9. Jorgensen, W. L. & Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Am. Chem. Soc.* **110**, 1657–1666 (1988).
10. Fernández-Recio, J., Totrov, M. & Abagyan, R. Identification of protein-protein interaction sites from docking energy landscapes. *Mol. Biol.* 335, 843–65 (2004).
11. A Vangone and A.M.J.J. Bonvin. Contacts-based prediction of binding affinity in protein-protein complexes. *eLife*, 4, e07454 (2015).

Computational screening and bioinformatics functional analysis for the investigation of apple polyphenols chemopreventive effects in cancer

Scafuri B(1,2), Marabotti A(1,2), Carbone V(1), Minasi P(1), Dotolo S(1),

Facchiano A(1)

(1)CNR-ISA, National Research Council, Institute of Food Science, Avellino, Italy
(2) Department of Chemistry and Biology "A. Zambelli", University of Salerno, Fisciano (SA), Italy

Motivation

Apple fruits are particularly rich in antioxidant compounds, and are object of investigation for their known chemo-preventive effects against colorectal cancer [Ribeiro et al., 2014; Teller et al., 2013; Koch et al., 2009]. Apples varieties are characterized by a different content of phenolic compounds, and in any case by a high content of antioxidants, considered of potential benefits for human health. It is known that certain polyphenols inhibit the growth of colon carcinoma cell, through the inhibition of p38/CREB signaling, the stimulation of a G2/M phase cell cycle block, the decrease of COX-2 expression by signalling through a number of pathways, including the mitogen-actived protein kinase (MAPK) pathway [Corona et al., 2007].

Although the search for protein targets of antioxidants revealed that these compounds might interfere directly on the protein activities, only few studies in the literature have been focused to the binding between antioxidant molecules and specific protein targets. Therefore, we decided to investigate in detail the potential protein targets of apple antioxidant compounds, and their potential role in chemoprevention.

Methods

The phenolic compounds in apple extracts from three varieties, i.e. Annurca, Red Delicious, and Golden Delicious, were experimentally determined by HPLC–UV/Vis and Electrospray Ionization multistage Ion Trap Mass Spectrometry (ESI-ITMSn) analyses.

Molecular structure of the phenolic compounds were retrieved from PubChem database. Reverse docking search for protein targets was performed by IdTarget web server. AutoDock4.2 was used to investigate in more detail by direct docking analysis the interaction between each compound and the potential protein targets. The lists of potential protein targets and of the related genes were investigated for functional analysis by means of Cytoscape, GeneMANIA, BioGPS.

Results

Starting by the analytic characterization of phenolic compounds in three apple varieties, i.e. Annurca, Red Delicious, and Golden Delicious, we applied reverse docking approach to search for protein targets of the identified compounds, obtaining a long list of potential targets.

After a first level of selection, direct docking validation of the potential protein-ligand interactions has generated a short list of human proteins, potential targets of the phenolic compounds. A functional analysis by comparison with experimental gene expression data and interaction networks, obtained from public repositories, suggests that chemo-preventive effects of apple extracts in human pathologies, in particular for colorectal cancer, may be caused by the interference with the activity of nucleotide metabolism and methylation enzymes, similarly to known classes of anticancer drugs.

Modelling the intertwined network of PPIs along the AhR:ARNT dimer

Corrada D(1), Bonati L(1)

(1) Dept. of Earth and Environmental Sciences, University of Milano-Bicocca, Milan

Motivation

The Aryl hydrocarbon Receptor (AhR) is a trancription factor activated by binding to a wide range of exogenous and endogenous chemicals, including many xenobiotics such as polycyclic- and halogenated-aromatic hydrocarbons. Upon ligand binding, AhR dimerizes with the AhR Nuclear Translocator (ARNT), then the complex binds its specific DNA recognition site. This event triggers the expression of a large battery of genes, some of them involved in detoxification pathways. Comprehension of these molecular mechanisms would require structural information on the protein domains responsible for the different steps, in particular the bHLH motif (involved in DNA binding and AhR:ARNT dimerization) and the PAS domains (involved in both ligand binding and dimerization). Only two crystallographic structures of homologous bHLH-PAS protein dimers including the full length N-terminal region are available: the HIF2a:ARNT and the CLOCK:BMAL1 complexes. Even though the inter-domain interfaces of the two complexes share an impressive similarity, the overall architectures show noteworthy differences, thus motivating investigation of the most reliable AhR:ARNT dimerization mode.

Methods

Two homology models were built for the AhR:ARNT dimer, according to the alternative dimerization modes showed by the available X-ray structures of homologous complexes. The Protein-Protein Interaction (PPI) interfaces were evaluated through calculation of the Solvent Accessible Surface Area variation (dSASA). The binding free energy (dGbinding) of each dimer model was calculated by the MM-GBSA method and further decomposed to establish the main inter-residue energetic couplings. A variant of the “energy decomposition analysis” method was developed to obtain dGbinding-based interaction energy matrices. While structural information on the core fold and the spatial arrangements of the domains in the AhR:ARNT complex could be derived by the template structures, no knowledge based data are available for the flexible loops and inter-domain linkers (more than 100 residues). The ab initio Rosetta approach for loop-modeling, including a MonteCarlo-Simulated Annealing strategy for energy refinement, was employed for these regions.

Results

The preliminary analysis of the homomeric PPIs in the crystallographic CLOCK:BMAL1 and HIF2a:ARNT dimers further confirmed the models of the individual AhR:ARNT PAS domain dimers we developed in a previous work [1]. On the other hand, these experimental structures show that heteromeric interactions (between bHLH and PAS-A or PAS-A and PAS-B) play an important role in the dimerization of the bHLH-PAS proteins. In this work we propose, for the first time, the homology model of the full-length N-terminal region of the AhR:ARNT complex. On the basis of the templates available, we developed two alternative dimer models in which the AhR domains give mutual contacts to form a contiguous surface and the ARNT domains wrap around it (according to the HIF2a:ARNT complex) or viceversa (according to the CLOCK:BMAL1 complex). Unexpectedly the models share nearly identical values of dGbinding and the energy decomposition analysis reveals a common mapping of the main energetic determinants in the 3D structures. The analysis of dSASA highlights a core interface in which homomeric bHLH and PAS-A PPIs are intertwined, and sheds light on the roles played by specific secondary structure elements. Finally, the ab initio prediction of the long PAS-A loops suggests an additional extensions of the dimerization interface. Beside structural characterization, Molecular Dynamics based studies will be needed to evaluate the dimer stability and to establish putative functional roles of both the loops and the inter-domain linkers. These studies offer an unprecedented starting point for the molecular investigation of the AhR dimerization and transformation into its functional DNA binding form.

Info

References
[1] Corrada D, Soshilov AA, Denison MS, Bonati L Deciphering dimerization modes of PAS domain: computational and experimental analyses of the AhR:ARNT complex reveal new insights into the

mechanisms of AhR transformation, PLoS Comput Biol under review.

Figure

Structural characterization of the Hepatitis C Virus E2 protein: computational and experimental approaches

Balasco N(1,2), Barone D(1,2), Sandomenico A(1), Iaccarino E(1,2), Ruvo M(1), Vitagliano L(1)

(1) Institute of Biostructures and Bioimaging, C.N.R., Naples 80134, Italy
(2) DiSTABiF, Second University of Naples, Caserta 81100, Italy

Motivation

Hepatitis C virus (HCV) infection is a major cause of chronic liver disease worldwide. Although effective therapeutic approaches, based on specific inhibitors of the viral proteins NS3/4A and NS5B, have been recently discovered, their use is limited by the elevated costs. Currently, there is neither an effective immune globulin for prophylaxis nor a vaccine for the prevention of hepatitis C. It is commonly believed that a full structural/functional characterization of the immunogenic E2 protein, a key factor for HCV entry in host cells, represents an important step for the development of effective vaccines. Although the crystallographic structure of the E2 protein core has been recently solved by two independent groups in complex with antibodies (Kong et al. 2013, Khan et al. 2014), no structural data have been derived for the region encompassing residues 412-423 (E2_412-423), the most highly conserved antigenic site of the protein. Several independent crystallographic structures of the isolated peptide with the corresponding antibody solved in recent years have indicated that this portion of the protein is likely endowed with a significant level of structural versatility. Indeed, it is able to adopt distinct, often unrelated, conformations (β-hairpin and extended structures) when bound to different antibodies (Sautto et al.2013 and references therein). In this framework, we have recently developed and structurally characterized novel antibodies specifically targeted against a restrained variant of E2_412-423 (Sandomenico et al. 2016). We here extended these investigations through the study of the dynamic properties of the E2 protein core freed from the antibody and the intrinsic structural preferences of E2_412-423 epitope by combining computational and experimental approaches.

Methods

Classical Molecular Dynamics (MD) simulations were performed on the HCV E2 protein core (PDB ID: 4MWF) and on the different structures available for the E2_412-423 peptide (PDB IDs: 4GAG, 4DGV, 4HS8, 4HS6, 4WHT, 4XVJ) upon mAb removal. Moreover, in order to enhance the sampling of the epitope, Replica Exchange (REMD) was performed on the different conformations adopted by E2_412-423 (PDB IDs: 4HS8, 4WHT, and 4XVJ). GROMACS software package 4.5.5 was used to perform the MD/REMD simulations in explicit solvent (TIP4P water model) under AMBER99SB force field. The simulations were run with a time step of 0.002 ps applying periodic boundary conditions. In the REMD simulations the sampling for all systems was composed of 24 replicas ranging from 298 to 363K.

Different variants of the E2_412-423 peptide were prepared by solid synthesis following standard Fmoc chemistry protocols. Peptide secondary structure was analyzed by far-UV Circular Dichroism. The intrinsic Fluorescence of the peptide was measured in the emission range 310-450 nm upon excitation at 290 nm. Binding of Thioflavin T to peptide aggregates was assayed by scanning the fluorescence emission from 300 to 600 nm, with excitation at 440 nm.

Results

The MD simulation performed on the E2 protein core provided interesting information on both global dynamics of the protein and local features of the two most important antigenic regions. Our data indicate that E2 combines a flexible structure with a network of covalent bonds. We found that a fluctuating β-hairpin represents a populated state by E2_412-423. Interestingly, we observed that the conformations adopted by the epitope region E2_427-446, that undergoes a remarkable

rearrangement in the simulation, have significant similarities with the structure that this fragment adopts in complex with a neutralizing antibody. Moreover, the analysis of E2_412–423 flexibility in the context of the whole protein provides insights into the mechanisms that some antibodies adopt to anchor Trp437 that is fully buried in E2 (Barone et al. 2016). The computational and experimental analyses performed on the isolated peptide E2_412-423 provide interesting insights on the conformational preferences of this key portion of the protein. In line with solution studies which highlight the presence of a very limited content of secondary structure, MD/REMD data indicate that this region is endowed with a remarkable structural versatility. Interestingly, our simulations show that the peptide populates all the different structural states detected in its complexes with antibodies. These findings clearly indicate that antibodies targeting this region operate through a conformation selection mechanism which does not require any induced fit in the recognition process. Moreover, the peptide explores novel conformations that may be potentially recognized by other antibodies. These observations could be effectively exploited for the design of anti-hepatitis vaccines. Unexpectedly, the experimental characterization of the peptide has also shown that, in physiologic-like conditions, it forms β-structured aggregates able to bind Thioflavin T (Balasco et al. In preparation).

Tandem Repeat proteins at a glance: function, disease and role in protein-protein interaction networks

Paladin L(1), Richard F(2), Kajava AV(2), Tosatto SCE(1,3)

(1) Dept. of Biomedical Sciences, University of Padua, viale G. Colombo 3, 35121 Padova, Italy

(2) Centre de Recherches de Biochimie Macromoléculaire, CNRS, Université Montpellier 1 et 2, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France

(3) CNR Institute of Neuroscience, viale G. Colombo 3, 35121 Padova, Italy

Motivation

Tandem repeat (TR) protein structures are abundant in nature and widespread across all types of organisms. Their periodic sequence folds into a modular and elongated architecture. They play a role in a number of different pathways, suggesting their association to a large number of diseases. A possible explanation of the evolutionary framework that lead to the wide distribution of TR proteins and their importance is given by the properties of their modular structure, ideal for highly specialized and rapidly evolving binding functions. The present study shows an assessment of TR proteins function and role in the protein-protein interaction (PPI) network. In addition, it presents the first analysis of the relationship between repeat proteins and Pfam repeat domains and diseases from OMIM (Online Mendelian Inheritance in Man).

Methods

The dataset of 417 proteins containing repeat regions is collected from RepeatsDB. RepeatsDB provides a resource for structurally validated TR proteins, extracted from the PDB and classified. The background for enrichment calculation is given by SwissProt and PDB data banks, and three additional sets of disease-related classes were identified to compare against human TR proteins, i.e. kinases, homeobox proteins and ion channels. The GO annotation of all proteins was collected and processed to define enriched functions in the TR dataset through a Fisher test. A non-redundant set of protein-protein interactions is retrieved from the IMEx consortium, the general features of the network (degree, connectivity, neighbor degree) were analyzed and compared to the TR dataset features. The significance of the differences was assesses performing a T-test. The disease annotation is extracted by the field “diseases” of the Uniprot description. The number of proteins associated to at least one OMIM disease ID was divided by the total number of proteins, obtaining the fraction of disease-associated proteins in each dataset. The associated diseases were characterized through a systematic method that takes advantage of the OMIM Clinical Synopsis information. Data was plotted in a matrix enriched with data from the whole human Swissprot sequences.

Results

The results show that TR regions are ubiquitous in organism, cellular location, biological role and functional pathways. In addition, they are characterized by a higher number of interactors than the

protein-protein interaction network average. They are also significantly associated to diseases. The disease annotation is consistent with protein function and can be used to characterize RepeatsDB subclasses. In addition, Pfam repeat domains are frequent among the top ranking results. This is in agreement with the observation that typical disease genes are intermediate between hubs of core biological functions, whose disruption causes lethality, and genes for which high haplotype diversity and mutation rates are advantageous. The results highlight the importance of TR protein recognition, classification and study for a better understanding of the cell machinery and disease insurgency, as well as opening up a promising field in biomolecular engineering.

Session 5: Special Session: Clinical Bioinformatics

HIGHLIGHT LECTURE

Multilayered transcriptional crosstalks are sustained by cooperating micro-societies in human colorectal cancer

Tommaso Mazza

Casa Sollievo della Sofferenza (FG) e CSS-Mendel (Roma)

Alterations in the balance of mRNA and microRNA (miRNA) expression profiles contribute to the onset and development of colorectal cancer. The regulatory functions of individual miRNA-gene pairs are widely acknowledged, even if group effects are largely unexplored. Interactions between mRNA-miRNA and miRNA-miRNA expression profiles, measured from matched specimens of human colorectal cancer tissues and adjacent non-tumorous mucosa, were summarized by a hypernetwork-based model, which highlighted the propensity of several miRNAs to aggretate into tight micro-assemblies. Some of these miRNAs resulted to modulate several genes, which in turn participate to fulfill a set of significantly enriched cancer-enhancer and cancer-protection biological processes, still being under the control of miR-145, a cell cycle and MAPK signaling cascade master regulator. Thus, miR-145 came up as a potent upstream regulator of a complex RNA-RNA crosstalk, and to mechanistically coordinate several signaling pathways and regulatory circuits that - when deranged - contribute to the colorectal carcinogenesis.

Confirming and Investigating the Role of Breast Cancer PIK3CA-ERBB2 Genes in Anti-Cancer Drug-Resistance with a new Framework for the Inference of Cancer Progression Graphs using Vector Integration Sites Data

Spinozzi G(1,2), Calabria A (1), Caravagna G(3), Graudenzi A(2), Ramazzotti D(2), Antoniotti M(2), Mauri G(2), Montini E(1)

(1) San Raffaele Telethon Institute for Gene Therapy
(2) University of Milano-Bicocca; Department of Informatics, Systems and Communication
(3) University of Edinburgh, Laboratory of Machine Learning for Computational Biology and Bioinformatics

Motivation

Evolution plays a key role in Cancer as the result of the accumulation of genetic alterations, which provide selective advantages to a tumor cell, allowing resistance to anti-cancer drugs. Unfortunately, however, the identification of the driver mutations and thus the mechanisms underlying anti-cancer drug resistance (ACDR) still remains a challenge. We previously

demonstrated that lentiviral vectors (LVs), when properly modified, might integrate near specific genes, alter their expression and induce cancer or ACDR in vivo and in vitro [Ranzani et al. 2013; Ranzani et al. 2014]. The analysis of vector-cellular genomic junctions in tumor or ACDR cells allowed identifying causative genes of HER2+ breast cancer cell line using a statistical approach defined Common Insertion Sites (CIS) that highlight genomic regions targeted at significantly higher frequency than expected by a random distribution. The reconstruction of cumulative cancer progression from CISs genes has not been yet addressed and may produce causative gene networks. The aim of this project is studying anti-cancer drug resistance from exclusive and co-occurring genes using cumulative cancer progression from our cell line CISs genes and investigating the relation between them.

Methods

Bioinformatics tools aimed at inferring cancer progression models, in terms of selective advantage relation among relevant genomic alteration from cross-sectional data (Next Generation Sequencing platforms), would allow identifying specific combinations of targeted drugs to overcome the occurrence of resistance. In a new context of vector integration sites (ISs), we developed an integrated bioinformatics workflow composed of: (i) an updated and more accurate version of VISPA (Vector Integration Site Parallel Analysis) [Calabria et al. 2014], a pipeline for automated ISs identification and annotation based on a distributed environment with a simple web based interface; (ii) identification of the CISs with a sliding window approach developed in [Abel et al. 2011] and [Presson et al. 2011]; (iii) a new statistical tool, CAncer PRogression Inference (CAPRI) - [Ramazzotti et al. 2015], to infer selective advantage relations among various mutational events in cancer cell genomes, mostly in relation with drug-resistance. The model is based on probabilistic causation and is able to reconstruct our cancer progression Direct Acyclic Graphs (DAGs), involving the CIS genes. With the use of GENEMANIA (http://www.genemania.org), Enrichr (http://amp.pharm.mssm.edu /Enrichr) and Cytoscape (http://www.cytoscape.org), we studied the protein-protein interaction, Gene Ontology and Pathway relations between selected genes, collecting and visualizing results in gene networks.

Results

By applying our new method to the published ISs dataset from the two cell lines, we were able to generate progression models involving relevant genes (confirming that these are not mutually exclusive genes, by Mutex [Babur et al. 2015]), which are consistent with previously validated results, confirming the role of PIK3CA-ERBB2 genes in ACDR. Unfortunately, one of the two cell line has a low quality samples. For this reason, CAPRI was not able to generate the progression DAG. In Figure 1 the progression for a cell line, BT474 (nodes are CISs, in blue; merged genes - same progression - are in green), pre-treatment and post-treatment with lapatinib respectively. Now we are investigating the relations between genes, produced by the model, trying to find some useful new interactions and confirmations for ACDR studies (i.e. SUMO1-ERBB2-PIK3CA-CSMD3). New insertional mutagenesis data from lung cancer cell lines aimed to induce ACDR in vivo and in vitro are ongoing and will allow to validate and/or identify novel cancer progression models, as well as possible combinatorial therapies.

Figure

XTENS: a neuroblastoma copy number variation repository at the BIT-Gaslini biobank

Izzo M(1,2), Cangelosi D(1), Pezzolo A(3), Morini M(1), Varesio L(1)

(1) Laboratorio di Biologia Molecolare, Istituto Giannina Gaslini, Genova
(2) Oxford e-Research Centre, University of Oxford, Oxford
(3) Laboratorio di Oncologia, Istituto Giannina Gaslini, Genova

Motivation

Neuroblastoma is the major paediatric solid tumour. Unfortunately, about 50% of high risk patients are refractory to treatment and die, demanding new prognostic indicators for improving and personalising therapy. Biomarkers discovery depends on mining molecular, biological, and clinical data, thus making integrated biobanks the ideal collectors of this heterogeneous information and the essential structure for this task.

The Biobank Integrating Tissue-genomics of Gaslini Institute, Genova, Italy (BIT-Gaslini) has adopted XTENS [1,2], a web-based data management platform to handle the sample management workflow and integrate it with the patients' clinical records and molecular data. As of December 2015 the biobank collected over 3700 primary samples (2140 tissue and 1600 fluid) and 1650 derivatives (1030 DNA and 620 RNA). Besides sample management information - such as specimens characterisation, aliquot deliveries to external laboratories and/or centres, and quality control reports - XTENS stored clinical and molecular details for over 900 neuroblastoma patients as retrieved from the National Neuroblastoma Registry and 175 microarray profiles from primary tumour tissues. We have decided to test the capability of the platform as a repository for Copy Number Variations (CNVs, gain/amplification and loss of DNA) data obtained from oligonucleotide array-Comparative Genomic Hybridisation (aCGH) analyses routinely executed at Gaslini Institute. aCGH data are essential for patients' assignment to specific treatments' protocols and they are a valuable source of information for mining the biology of neuroblastoma and the identification of new prognostic indicators.

Methods

To identify CNVs (numerical and segmental chromosomal alterations) in neuroblastoma samples, we used aCGH, that was performed using the Agilent Human Genome CGH microarrays 180K following the manufacturer's protocol (Agilent Technologies, Santa Clara, California, USA). Slides were scanned using a G2565BA scanner, and analysed using Agilent CGH Analytics software Genomic Workbench 7.0 (Agilent Technologies Inc.) with the statistical algorithm ADM-1 and a sensitivity threshold of 6.0. Probes with a log ratio value greater than 2 were considered as amplified. To automate the CNV upload to XTENS, we have designed a novel page on the website where the operator can upload the aCGH processed files containing all the analysis metadata and the full list of found aberrations (gain and loss analysis of DNA). A Node.js server-side script parses the uploaded file, and stores all the relevant information in the database. The raw data files are then uploaded using the standard XTENS data management interface and are stored on a distributed file system managed by the iRODS data grid middleware, transparently integrated with XTENS through a REST interface.

Results

As of March 2016, we have uploaded 345 aCGH analyses of neuroblastoma tissue samples run between 2007 to 2015 on XTENS 2, for a total of 6511 aberration calls (3947 gain/amplifications, 2564 losses). Each aberration is characterised by type (amplification or deletion), location (chromosome, cytogenetic band, start and end position in the reference genome), and the list of affected genes and miRNAs. Overall, there are 2.18 million gene annotations from 21,919 different protein-coding genes, and 107,400 miRNAs annotations from 1,040 different miRNA genes. All the metadata concerning the copy number variations are stored using the binary JSON format (JSONB) of PostgreSQL that is natively supported by the XTENS server environment, running on a Node.js server.

JSONB provides a more efficient schemaless solution than traditional metadata storage in relational Entity-Attribute-Value (EAV) catalogues. Queries composed from the website query builder tool to retrieve subjects based on gene annotations for the aberration calls are executed in 260.9 ± 2.1 ms without indexing and 35.9+2.1 ms if a General Inverted (GIN) Index is created for the JSONB metadata field. The execution times when GIN is adopted, below the 50 ms limit usually required by data-intensive web applications, evidence that XTENS is a suitable web-based solution to manage clinical, biological, and genomic metadata for medical research collaborations. The data uploader we have designed can be extended to other data sources. We plan to support automatic upload of circulating miRNA expression profiles obtained by blood plasma samples, in order to provide a broader database for identifying novel prognostic factors in neuroblastoma. Initial data mining applications will be presented.

Info

[1] Izzo, Massimiliano, et al. "XTENS-A JSON-Based Digital Repository for Biomedical Data Management." Bioinformatics and Biomedical Engineering. Springer International Publishing, 2015. 123-130.

[2] Izzo, Massimiliano. "Results: XTENS 2, A JSON-Compliant Repository." Biomedical Research and Integrated Biobanking: An Innovative Paradigm for Heterogeneous Data Management. Springer International Publishing, 2016. 61-88.

Session 6: Algorithms for Bioinformatics and Systems Biology

INVITED LECTURE

Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet

Michael M. Hoffman

Department of Computer Science, University of Toronto, Toronto, ON, Canada

Princess Margaret Cancer Centre, Toronto, ON, Canada

Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

michael.hoffman@utoronto.ca

Motivation

Many transcription factors (TFs) initiate transcription only in specific sequence contexts, providing the means for sequence specificity of transcriptional control. A four-letter DNA alphabet only partially describes the possible diversity of nucleobases a TF might encounter. Cytosine is often present in the modified forms: 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5hmC). TFs have been shown to distinguish unmodified from modified bases. Modification-sensitive TFs provide a mechanism by

which widespread changes in DNA methylation and hydroxymethylation can dramatically shift active gene expression programs.

Methods

To understand the effect of modified nucleobases on gene regulation, we developed methods to discover motifs and identify TF binding sites (TFBSs) in DNA with covalent modifications. Our models expand the standard A/C/G/T alphabet, adding

m (5mC) and h (5hmC). We additionally add symbols to encode guanine complementary to these modified cytosine nucleobases and represent states of ambiguous modification. We adapted the position weight matrix model of TFBS affinity to an expanded alphabet. We developed a program, Cytomod, to create a modified sequence. We also enhanced the MEME Suite to be able to handle custom alphabets. We created an expanded-alphabet sequence using whole-genome maps

of 5mC and 5hmC in naive ex-vivo mouse T cells.

Results

Using this sequence and ChIP-seq data from Mouse ENCODE and others, we identified

modification-sensitive cis-regulatory modules. We elucidated various known methylation binding preferences, including the preference of ZFP57 and C/EBP for methylated motifs and the preference of c-Myc for unmethylated E-box motifs. We demonstrated that our method is robust to parameter perturbations, with TF sensitivities for methylated and hydroxymethylated DNA broadly conserved across a range of mod-

ified base calling thresholds. Hypothesis testing across different threshold values was used to determine cutoffs most suitable for further analyses. Using these known binding preferences to tune model parameters enables discovery of novel modified motifs.

Discussion

Hypothesis testing of motif central enrichment provides a natural means of differentially

assessing modified versus unmodified binding affinity. This approach can be readily extended to other DNA modifications. As more high-resolution epigenomic data becomes available, we expect this method to continue to yield insights into altered TFBS affinities across a variety of modifications

---

Data Fusion for cleavage target prediction

Marini S(1), Demartini A(2), Vitali F(2), Bellazzi R(2), Akutsu T(1)

(1) Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan.
(2) Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy.

Motivation

Protein cleavage is a pivotal process in cell metabolism. It is involved, among other processes, in cell differentiation and cycle control, stress and immune response, removal of abnormally folded proteins and cell death. Proteases (i.e. protein responsible for cleavage) account for ~2% of all gene products. As consequence, wrongly regulated proteolytic activity may result in diseases. The problem of predicting cleavage targets have been addressed by a number of algorithms [1]. Traditional prediction models tackle the cleavage target machinery encoding directly related information to the outcome class (e.g. by extracting sequence patterns or frequency matrices). We are aware, however, that a huge amount of indirectly-related information is available in public data sets. Peptidases and targets are both proteins, and share similarities as well as non-cleavage interactions in knowledge bases; they are both encoded by genes, and gene interactions are also in databases. Our proposed Data Fusion algorithm leverages on these secondary information sources to infer novel peptidase targets.

Methods

Our approach is based on tri-factorization [2]. The multiplicity of data are fused by inferring a joint model, and without altering their original structure, i.e. data are explicitly represented in the form of a relational block matrix R. Diagonal blocks of R are set to 0, while other blocks are sparse matrices, populated with the relations harvested from the various data sources. R elements are constrained into the range [0, 1], where 0s represent negative or unknown relationships, while 1s are interpreted as certain relationships. We considered three elements in our matrix, namely peptidases, targets and genes. From MEROPS we obtained 657 human peptidases affecting 3460 targets and forming 8931 pairs. From their mapping on Uniprot, 3833 genes coding for pepidases or targets were retained. This information was used to populate the peptidase-target, peptidase-gene and target-gene R blocks.

During the data fusion process, each R block is decomposed into three sub-matrices, characterized by low dimensions (if compared to the original R block size). There is no clear consensus about a technique to define these dimensions [2], and we proceeded by choosing a rank for a given block based on the number of known interactions. Once the dimensions are set, the three sub-matrices are used to reconstruct a user-defined target block Rt. R decomposition is obtained through an iterative process, where constraint matrices play an important role. Constraint matrices are populated with the associations relating objects of the same type. In our application we utilized five constraints: one gene-gene interaction matrix from BIOGrid; two target-target and protease-protease interaction matrices from STIRNG (0.7 as combined score threshold); two target-target and protease-protease BLAST similarity matrices (10-10 as e-value threshold). Once the convergence is reached, the target block Rt is examined to infer novel relationships. Our objective is to find protease-target putative interactions, therefore our target block is the protease-target one. To detect a new interaction, we applied the row-centric rule [2]. Note that since the iterative process starts from a random initialization, we repeated the whole data fusion process 15 times with different random initializations and retained only the interactions that satisfied the row-centric rule in all the runs.

Results

1787 new protease-targets were predicted by our approach, involving 139 proteases and 716 targets. To validate our results, we utilized an independent algorithm, CasCleave [1]. CasCleave is based on traditional Machine Learning, therefore it is complementary to our data fusion approach. Though our approach pinpointed targets for all possible peptidases, we could validate only the 73 Caspase-interacting subset of our targets, since CasCleave has a limited scope. By comparing the cleavage

CasCleave probability distributions of our predicted targets with the ones over the whole human proteome, we found 6 new targets predicted for Caspase-1 (p-value 1.23×10-5), 37 for Caspase-3 (p-value 2.2×10-16); 4 for Caspase-6 (p-value 6.8×10-4); 5 for Caspase-7 (p-value 9.14×10-3); 4 for Caspase-8 (p-value 1.1×-3); and 17 for Granzyme B (p-value 6.84×10-3). P-values were computed with KS test. The average interaction probability of our targets predicted by Cascleave is 0.82. In future research we will expand the list of considered objects in the data fusion (e.g. domains) and validate our results with wet lab experiments.

References

[1] Wang, Mingjun, et al. "Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets." Bioinformatics (2013): btt603.

[2] Zitnik, Marinka, and Blaz Zupan. "Data fusion by matrix factorization." Pattern Analysis and Machine Intelligence, IEEE Transactions on 37.1 (2015): 41-53.

TCGAbiolinks: An R/Bioconductor package for integrative analysis with TCGA data

Colaprico A(1,2), Silva TC(3,4), Olsen C(1,2), Garofano L(5,6), Cava C(7), Garolini D(8), Sabedot T(3,4), Malta TM(3,4), Pagnotta SM(5,9), Castiglioni I(7), Ceccarelli M(10), Bontempi G(1,2), Noushmehr H(3,4)

(1) Interuniversity Institute of Bioinformatics in Brussels (IB)2, Brussels, Belgium
(2) Machine Learning Group (MLG), Department d'Informatique, Université libre de Bruxelles (ULB), Brussels, Belgium
(3) Department of Genetics Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, São Paulo, Brazil
(4) Center for Integrative Systems Biology - CISBi, NAP/USP, Ribeirão Preto, São Paulo, Brazil
(5) Department of Science and Technology, University of Sannio, Benevento, Italy
(6) Unlimited Software srl, Naples, Italy
(7) Institute of Molecular Bioimaging and Physiology of the National Research Council (IBFM-CNR), Milan, Italy
(8) Physics for Complex Systems, Department of Physics, University of Turin, Italy
(9) Bioinformatics Laboratory, BIOGEM, Ariano Irpino, Avellino, Italy
(10) Qatar Computing Research Institute (QCRI), HBKU, Doha, Qatar

Motivation

The Cancer Genome Atlas (TCGA) research network has made public a large collection of clinical and molecular phenotypes of more than 10 000 tumor patients across 33 different tumor types. Using this cohort, TCGA has published over 20 marker papers detailing the genomic and epigenomic alterations associated with these tumor types. Although many important discoveries have been made by TCGA's research network, opportunities still exist to implement novel methods, thereby elucidating new biological pathways and diagnostic markers. However, mining the TCGA data presents several bioinformatics challenges, such as data retrieval and integration with clinical data and other molecular data types (e.g. RNA and DNA methylation). We developed an R/Bioconductor package called TCGAbiolinks to address these challenges and offer bioinformatics solutions by using a guided workflow to allow users to query, download and perform integrative analyses of TCGA data. We combined methods from computer science and statistics into the pipeline and incorporated methodologies developed in previous TCGA marker studies and in our own group. Using four different TCGA tumor types (Kidney, Brain, Breast and Colon) as examples, we provide case studies to illustrate examples of reproducibility, integrative analysis and utilization of different Bioconductor packages to advance and accelerate novel discoveries.

Methods

Here, we describe a new software tool called TCGAbiolinks that aids in querying, downloading, analyzing and integrating TCGA data within a single collective Bioconductor package. TCGAbiolinks was developed exclusively in R and features many of the Bioconductor-specified package and object designs, which are necessary for integration with other Bioconductor packages. The Bioconductor project ensures high-quality, well-documented and interoperable software and the possibility of integration with hundreds of available packages within R. The Bioconductor project was also endorsed by the editors at Nature Genetics as a bioinformatics resource.
The aim of TCGAbiolinks is four-fold: (i) to facilitate data retrieval via TCGA's DCCWS; (ii) to prepare

the data using the appropriate preprocessing strategies; (iii) to provide a means to conduct different standard analyses and advanced integrative analyses and (iv) to allow the user to download a specific version of the data and thus easily reproduce earlier research results. We introduce public methods used in several marker papers to integrate DNA methylation and gene expression data. In addition, our tool extracts published molecular subtype information for each TCGA sample within a tumor type (generally embedded in supplementary tables, PDFs or external websites). Because our tool was developed in the language of R specifically for integration within the Bioconductor project, we have provided most of the TCGA data objects as the Bioconductor-specified 'SummarizedExperiment' class, thereby allowing easy integration with other data types and statistical methods that are common in the Bioconductor repository.

Results

To introduce and describe the utility and application of TCGAbiolinks, we used four different TCGA cancer types (Brain, Kidney, Breast and Colon) as examples. For each tumor type, we describe methods to extract the different experimental types and integrate the information into a cohesive, biologically specific and hypothesis-driven approach. We also describe how to generate a starburst plot (16). The starburst plot was introduced to illustrate the results of integrating DNA methylation and gene expression data. In addition, we describe how TCGAbiolinks prepares data for integration with other recently published packages, such as ELMER (12), a new Bioconductor package designed to identify candidate regulatory elements in the non-coding regions of the genome associated with cancer, and DNET (24), a new R package designed to uncover the existence of an underlying gene network that is defined by somatic mutations and that at least partially controls cancer survival independently of tumor origin and type. Our package is freely available within the Bioconductor project at http://bioconductor.org/packages/TCGAbiolinks/.

For detailed results see NAR's online paper about four case studies and for reproducible R codes in supplementary informations and related vignette.

http://nar.oxfordjournals.org/content/suppl/2015/12/23/gkv1507.DC1/nar-03136-met-n-2015-File009.pdf

Info

TCGAbiolinks it was recently used for section (4) mRNA Expression and (5) DNA methylation profiling in last TCGA's marker paper published in early 2016.

See citation in supplementary informations.

Ceccarelli et al, Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma, Cell 164 Issue 3: p550-563, 2016

https://tcga-data.nci.nih.gov/docs/publications/
http://www.sciencedirect.com/science/article/pii/S009286741501692X
http://www.sciencedirect.com/science/MiamiMultiMediaURL/1-s2.0-S009286741501692X/1-s2.0-S009286741501692X-mmc1.pdf/272196/html/S009286741501692X/d3371d49bef3810100e52cf6ec732f74/mmc1.pdf
Bioconductor's links

https://www.bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html
https://www.bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/tcgaBiolinks.html
http://bioconductor.org/packages/stats/bioc/TCGAbiolinks.html

---

A graph-based method to evaluate clonally related Immunoglobulins

Tomasi F(1), Squillario M(1), Bagnara D(2), Verri A(1), Barla A(1)

(1) DIBRIS – Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi – Università degli Studi di Genova
(2) The Feinstein Institute for Medical Research, Manhasset, NY 11030, USA

Motivation

Igs repertoire analysis is a hot research topic in molecular biology because the understanding of diversity in the immune response is key to determine the response to antigens (e.g., bacteria).

In this context, the aim is to predict the clonally related Igs, i.e. those who originate from the same ancestral cells, usually through clustering methods. This is a tough problem because of Igs complex structure and high mutation level.

Ig primary structure is formed by an ordered combination of three types of genes (V, D and J). When building a new Ig, those genes are selected almost randomly and, while putting together V, D and J genes, some nucleotides between V-D and D-J are inserted or deleted (indels). Another issue is the multiple assignment of probable V and J genes for those Igs in which their attribution can be ambiguous.

Furthermore, when an Ig is first formed and before encountering an antigen (i.e., "naive"), it has low (or zero) mutation levels. After encountering antigens, Igs undergo a series of mutations ("memory"). Since mutations are the normal behaviour of Igs, the identification of clones has an intrinsic difficulty. Moreover, the presence of the junction region, which binds to the antigens and therefore is the most hypervariable portion of an Ig, makes the clustering task even more challenging.

The most popular framework for the identification of Igs clones is a two-step tool named Change-O. The first step groups Igs that have at least one V or J in common and the same junction length; the second step performs a hierarchical clustering for each group defined before.

Methods

We propose a new method that finds clonally related Igs by exploiting a graph structure together with an ad-hoc similarity measure between the junctions.

First, we build a special case of a tripartite undirected graph that links each Ig to its corresponding V and J genes (Figure 1). The nodes of the graph consist in Igs, V genes and J genes. Edges link Igs with their correspondent V genes and J genes. In this graph the vertices can be partitioned into 3 different independent sets in which the elements in a set are only allowed to be linked to elements belonging to different sets. With this representation, the pairwise similarity between Igs can be computed with an extension of standard associativity index such as Jaccard, Simpson and geometric indexes for tripartite graphs with the possibility of assigning different weights, in particular, to common V genes with respect to common J genes. Then, a clustering procedure is applied to the similarity matrix obtaining the clonal families.

Results

The method has been validated using both hierarchical and spectral clustering, which extends to the tripartite representation of the data, for an increasing number of clusters k. The quality of clustering has been evaluated using an average silhouette value for each k.

The analysed dataset is formed of 4379 naive and 10631 memory Igs and derived from an NGS experiment using Illumina MiSeq sequencing system considering one healthy donor.
We considered naive and memory Igs separately. Since naive Igs have no mutations, we expect that the best average silhouette value would be obtained considering a number of clusters equal to the unique naive Igs.

Instead, considering the memory Igs, we expect to obtain a lower number of clusters with respect to the unique memory Igs due to their high mutation level.

Our results show that the best silhouette value is obtained at around 4200 clusters for naive and 5000 for memory Igs, as expected.

Differently from Change-O, the method allows a customisation of the distance function. Not only this is done by choosing one of the standard distances used in graph theory, but also by assigning a different weight to common V or J genes. Furthermore, each distance is adjusted according to the mutation levels of the compared Igs. In this way the method is capable of considering as similar two Igs even when at least one of them is highly mutated.

Info

Reference:

Gupta, Namita T and Vander Heiden, Jason A and Uduman, Mohamed and Gadala-Maria, Daniel and Yaari, Gur and Kleinstein, Steven H. "Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data". Bioinformatics, 2015, 31(20).

Figure

A Computational Systems-Level Approach to Decipher Inborn Errors of Metabolism

Pagliarini R(1), Castello R(1), Napolitano F(1), Borzone R(1), Annunziata P(1), Mandrile G(2,3), De Marchi M(2,3), Brunetti-Pierri N(1,4), di Bernardo D(1,5)

(1) Telethon Institute of Genetics and Medicine, Naples, Italy
(2) Medical Genetics, San Luigi University Hospital, Orbassano, Italy
(3) Department of Clinical & Biological Sciences, University of Torino, Torino, Italy
(4) Department of Translational Medicine, Federico II University, Naples, Italy
(5) Department of Chemical, Materials and Industrial Engineering, Federico II University, Naples, Italy

Motivation

Inborn errors of metabolism (IEMs) are a group of Mendelian disorders resulting from genetic disruption of single enzymes carrying out metabolic reactions. Our understanding of the consequences of single enzyme deficiencies on the whole metabolism are largely underappreciated because most studies have been narrowly focused on the affected metabolic reactions, thus neglecting alterations of more distant metabolites. To overcome this problem, we propose an innovative computational approach to model the alteration of metabolism caused by IEMs. We applied it to predict and analyse, in-silico, the metabolic alterations occurring in hepatocytes with loss-of-function (LoF) of the peroxisomal enzyme alanine:glyoxylate aminotransferase (AGT) encoded by the AGXT gene, mutated in Primary Hyperoxaluria type 1 (PH1). PH1 is an autosomal recessive disease presenting with hyperoxaluria, progressive renal involvement, and systemic deposition of calcium oxalate in multiple organs and tissues. Although the enzyme is only expressed in hepatocytes, the lack of AGT results in excessive production of oxalate by the liver leading to oxalate-induced damage in several tissues, and particularly in kidneys. PH1 is a severe disease that results in high morbidity, pain, disability, poor quality of life, and early death if treated late or untreated. Effective treatments for PH1 are still lacking and combined liver-kidney transplantation is the only available therapeutic option for patients with severe forms.

Methods

We first extended HepatoNet1 (Gille et al., 2010), a genome-scale metabolic network model of human hepatocytes, by including additional enzymatic and transport reactions and metabolites related to glyoxylate metabolism that were needed to model PH1. We next applied an algorithm that we developed, Differential Flux-balance Analysis (DFA), to predict differences in metabolic fluxes and metabolite levels between WT and single-enzyme defective hepatocytes across 442 different metabolic objectives. DFA is based on Flux Balance Analysis (FBA) (Orth et al., 2010), a mathematical procedure to estimate the metabolic flux of each metabolic reaction at steady-state when satisfying a given metabolic objective. For each metabolic objective, DFA computes the difference between the metabolic fluxes in the wild-type genome-scale model (WT) versus the perturbed model (LoF, or GoF, of a single enzyme or of a set of them). The end result is a ranked list of metabolic fluxes sorted by their difference in the LoF (or GoF) model versus the WT model. This difference is computed as the average across the 442 metabolic objectives. Hence, the reactions that are most affected by the LoF (or GoF) will be found at the top of the ranked list. Once these differential metabolic fluxes have been identified, the metabolites involved in the associated reactions are ranked to predict the most affected ones. It is done, for each metabolite, by summing the contribution of all the differential fluxes involving the same metabolite. This value is then used for the ranking, so that the metabolites whose fluxes are most affected by the LoF (or GoF) will be found at the top of the ranked list. Finally, we developed a modified version of Flux Variability Analysis (FVA) (Duarte et al., 2007) to estimate whether metabolites tend to increase, or decrease, in the LoF (or GoF) model compared to the WT model. The significance of model predications was assessed by Metabolite Set Enrichment Analysis (MSEA) (Xia and Wishart, 2010). MSEA is a statistical procedure for metabolomic studies that takes in input a list of altered metabolites in a patient and automatically predicts the most likely metabolic disorder. This is done by checking whether the metabolites in input are statistically enriched for known biomarkers of the disease.

Results

We applied our systems-levels approach to simulate the effect of AGT enzyme LoF, causative of PH1. The in-silico model correctly reproduced accumulation of known PH1-related metabolites. Unexpectedly, the model also predicted that an alteration of histidine and histamine should occur in PH1. We confirmed in-vitro, in-vivo, and in PH1 patients a significant reduction in histidine and histamine levels. Moreover, AGT deficient mice showed decreased vascular permeability, a read-out of in-vivo histamine activity. In-depth analysis of the in-silico model revealed that histamine reduction is caused by increased catabolism of the histamine-precursor histidine, caused by a redirection of alanine from AGT to the glutamic-pyruvate transaminase (GPT). We predicted in-silico and confirmed in-vivo that alanine administration reduces histamine levels in WT mice, while overexpression of GPT in PH1 mice increases plasma histidine, normalizes histamine levels, restores vascular permeability, and decreases urinary oxalate levels. Our work demonstrates that genome-scale metabolic models are clinically relevant and can link genotype to phenotype in metabolic disorders.

General contact info

amarabotti@unisa.it

robtag@unisa.it

Università degli Studi di Salerno
Via Giovanni Paolo II, 132
84084 Fisciano SA, Italy