SEQUENCE ANALYSIS, ALIGNMENT AND NEXT GENERATION SEQUENCING

Chairs: Geoff Barton and Des Higgins

A-1. Resolving the genomic breakpoints of a microdeletion using next-generation sequencing in a family

de Ligt J (1,), Vissers LELM (1), Kloosterman W (2), Hehir-Kwa JY (1), Bruijn E (3), Gilissen C (1), Guryev V (3), Cuppen E (2,3) , Brunner HG (1), Veltman JA (1)*

Some genomic regions are particularly difficult to map and analyse due to repeat rich content, inversion prone regions and polynucleotide stretches. One such region has a commonly occurring alternative locus which is located on chromosome 17q21 and is associated to different disorders. The common haplotype is H1, the H2 haplotype occurs in approximately 20% of the Caucasian population and is characterized by an 900kb inversion. The H2 haplotype predisposes to a microdeletion that causes a frequent mental retardation syndrome.

Materials and Methods

In this study different strategies are used to fine map the breakpoints of a deletion within the complex region of 17q21. We use a combination of genomic microarrays and Next Generation Sequencing (NGS) strategies; mate-pair sequencing and coverage depth comparison. Data from one pedigree was analysed; parents (♀ homozygous H2, ♂ homozygous H1), child (heterozygous H1/H2) with a deletion within the inversion region of H2 which resulted in the above mentioned syndrome. NGS should be able to provide the data to pinpoint the exact position and orientation of the inversion and the deletion.

Results

While microarray data provided a breakpoint in the patient, the mate-pair data was inconclusive in detecting these breakpoints for both the inversion and the deletion. The coverage values did show a deletion region when compared with a general reference, but the signal was too noisy to determine the breakpoints. Pedigree data provided a more reliable reference sequence and enabled better coverage comparison, which allowed us to capture the elusive breakpoints of the causative deletion. NGS predicted a more proximal start position (~40kb apart), while the range for the distal breakpoint is 2kb.

Discussion

The identification of the deletion breakpoints would not have been possible if the currently available reference genome would have been used. We conclude that many structural variants in variable or complex genomic regions will go un-noticed in high throughput approaches due to the fact that mapping errors will cause too much noise or result in un-mapped regions.

Presenting Author

Joep de Ligt (j.ligt@antrg.umcn.nl)
Department of Human Genetics UMC St Radboud, Nijmegen, The Netherlands

Author Affiliations

1 Department of Human Genetics UMC St Radboud, Nijmegen, The Netherlands 2 Department of Medical Genetics, UMC Utrecht, The Netherlands 3 Hubrecht Institute for Developmental Biology and Stem Cell Research, Utrecht, The Netherlands

Acknowledgements

TechGene

A-2. Identification of unknown environmental sequences by biased distribution of signature oligonucleotides

Labuschange P (1), Emmett W (1), Davenport CF (2), Reva ON (1,)*

The next generation sequencing technologies make it possible to sequence all the DNA in given samples, but data mining remains a challenge. While k-mer classification of unknown sequences is widely discussed in the literature, the sensitivity of the approaches based on short k-mers is still low. Further, the analysis of frequencies of longer words is computer intensive. Frequencies of 8 to 14-mer oligonucleotides are loaded with significant taxonomic information, but there is a lack of computer algorithms to exploit this information in large-scale studies.

Materials and Methods

The distribution of oligonucleotides was analysed in DNA sequences of bacterial chromosomes available in the NCBI database. A template of 172,636 signature words was created. Each oligonucleotide and its reverse complement were considered as the same word so that the two different stands of the DNA molecule will be assigned identical scores. To keep the size of the database small enough for easy use, the numeric frequencies of word occurrence were replaced by percentiles. For case studies several metagenome datasets were simulated by MetaSim and clustered by LikelyBin.

Results

Here we present new genome linguistics algorithms and a portable program with a database of signature oligonucleotides. Frequencies of signature oligonucleotides were calculated in 768 bacterial chromosomes and this database can be downloaded from the project web-site. The program includes several algorithms that help one to find the words that discriminate best between certain bacterial genomes or taxonomic units. Once such words have been found, another algorithm implemented in the program allows one to make inferences about the phylogeny of unknown DNA sequences.

Discussion

The case studies showed how the program and signature words perform in practice; however, the quality depends on sequence length. Our current work focuses on the development of a Web-based interface that will make it possible to perform a semi-automated identification in several hierarchical steps. First, a cluster of reads will be associated with a high level taxonomic groups sharing compositional sequence similarity based on pre-selected signature words. Then, the cluster will be assigned to a smaller group of organisms until the confidence in the assignment drop below a certain threshold.

URL

http://www.bi.up.ac.za/SeqWord/oligodb/index.html

Presenting Author

Oleg N. Reva (oleg.reva@up.ac.za)
University of Pretoria

Author Affiliations

(1) Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria,Pretoria 0002, Republic of South Africa; (2) Klinische Forschergruppe, OE 6711, Hannover Medical School, Carl-Neuberg-Strasse 1, D-30625 Hanover, Germany

Acknowledgements

Funding for this research was provided by the National Research Foundation of South Africa (NRF) grant for the National Bioinformatics and Functional Genomics Programme.

A-3. Identification of conserved repeats with advanced background models

Łabaj P-P (1,), Sykacek P (1), Kreil D-P (1)*

The functional interpretation of sequencing data remains a key challenge, especially the identification of biologically relevant patterns against a backdrop of functionally meaningless variation. Considering the unexpected abundance of single amino acid repeats, a detection of repeats with potential biological functions is of interest. As their occurrence cannot be assessed by established models, an exploration of more complex models is necessary. We demonstrate the efficacy of an application specific background model for the identification of novel conserved sequence patterns.

Materials and Methods

An empirical model was developed for the study of repeat abundance in signal peptides compared to mature proteins. It was trained on a comprehensive unbiased reference sample from UniProt. Input parameters comprised protein length, amino acid composition, and repeat length. A two-stage approach was found to be efficient. First, a logistic regression is used to model the probability that a given protein has at least one repeat. Then a relevance vector machine predicts the conditionally expected number of repeats.

Results

Considering that the abundance of single amino acid repeats cannot be assessed by standard models (including higher order HMMs) we have developed an application specific background model for their analysis. In a survey of these repeats in signal peptides of eukaryotes, a significant enrichment was only detected for leucine repeats. It explains their unexpected abundance in general. We could then show that these repeats were conserved in tetrapods, indicating a functional role. This was particularly clear in mammals.

Discussion

While standard models perform well for globular protein domains, they are known to break down in regions of stronger compositional bias or low complexity. As demonstrated, to meaningfully study such regions, application specific background models may be required. As repeats in general seem not to be conserved but rather tolerated as neutral, these features are filtered and excluded in regular sequence analysis such as similarity searches. We identify leucine repeats in signal peptides that are more strongly conserved than their host sequence, indicating novel functional roles.

Presenting Author

Paweł P. Łabaj (eccb2010@boku.ac.at)
Chair of Bioinformatics, BOKU University Vienna, Austria

Author Affiliations

1 Chair of Bioinformatics, BOKU University Vienna, Austria

Acknowledgements

This work was supported by the Vienna Science and Technology Fund (WWTF), Baxter AG, Austrian Research Centres Seibersdorf, and the Austrian Centre of Biopharmaceutical Technology.

A-4. Searching for structural homologues of LeuT-like secondary transporters beyond sequence similarity

Khafizov K (1,), Staritzbichler R (1), Stamm M (1), Forrest LR (1)*

X-ray structures have recently revealed that numerous secondary transporter proteins belonging to different sequence families unexpectedly share the so-called LeuT fold. The core of this fold consists of two units of five transmembrane helices (TM). That these two units share similar structures implies that LeuT-like transporters arose from gene duplication and fusion events. The evolutionary origins of this fold may be relevant to their function. However, the repeat units have no sequence similarity, so very sensitive methods are required to search for relationships between such 5TM proteins.

Materials and Methods

We developed AlignMe, which can use various types of input information to perform pair-wise alignments of (membrane) protein sequences. Aside from conventional similarity matrices, AlignMe can also use specific amino-acid properties (e.g., hydrophobicity); predictions of secondary structure and/or membrane-spanning segments; as well as combinations of these data. We used AlignMe to search for 5TM homologues of LeuT-fold transporters by screening hydropathy profiles of 5TM repeats from known structures against hydropathy profiles constructed for a large dataset of membrane protein sequences.

Results

We first demonstrate that AlignMe hydropathy profile alignments accurately match up the TM segments of LeuT-like secondary transporter proteins, which share similar structures but lack significant sequence similarity. We then report the identification of 5TM proteins that possess similar hydropathy profiles to 5TM repeats of LeuT-like transporters and, therefore, may share a common ancestor to the 5TM repeat units in the LeuT fold.

Discussion

The comparison of the LeuT-fold transporters shows that AlignMe is a useful tool for alignment of membrane proteins that are remote homologues. By identification of 5TM proteins that may share a common ancestor with the LeuT-fold transporters using AlignMe we hope to provide some insight into the evolution of these remarkable structural features, whose symmetry is thought to allow the formation of two alternate states during the alternating access mechanism of secondary transport.

URL

http://www.forrestlab.org/alignme

Presenting Author

Kamil Khafizov (kamil.khafizov@biophys.mpg.de)
Max Planck Institute of Biophysics

Author Affiliations

(1) Computational Structural Biology group, Max Planck Institute of Biophysics, 60438 Frankfurt am Main, Germany

Acknowledgements

SFB807

A-5. Why inverse proteins are relatively abundant

*Nebel J-C**

Studies of peptide chains created by inversing protein sequences have established there is no correlation between the structures of a protein and its inverse. However, inverse sequences are more common in nature than one would expect by chance. They must display some characteristics that are not present in random sequences. However, to date, the relative abundance of inverse proteins remains unexplained. Here, we investigate the proposition that inverse peptide chains are more common than random ones mainly because they display periodicity and repeat patterns present in protein sequences.

Materials and Methods

We designed a new artificial peptide dataset, i.e. 'opprotein', sharing repeat patterns with reference proteins, but displaying very different amino acid (aa) environments. An 'opprotein' is the peptide chain which is the most unlike a given protein sequence. More specifically, the opposite of a protein sequence is defined as the aa sequence where each aa is replaced by an aa with the most opposite physico-chemical properties. 3 artificial datasets, i.e. inverse, opprotein and random, containing 5489 sequences each were analysed using BLAST to establish similarity to existing proteins.

Results

Inverse, opprotein and random sets had, respectively, 16, 5 and 0 significant hits. Consequently, opproteins share some properties with inverse and real sequences, which are not found in random sequences. Remarkably, 4 protein sequences produced both inverse and opprotein hits. In most cases (15), matching proteins display repeats or repeating structural elements, i.e. beta-barrel, beta propeller and duplicated folds.

Discussion

Our study suggests that repeats are the main contributor of the abundance of inverse proteins. However, since inverse proteins are more common than opproteins that share the same repeat patterns, other factors are involved. Since, unlike opproteins, inverse peptide sequences have the same residue propensity as known proteins, amino acid distribution must also play a part in the similarity between a peptide chain and a protein. A consequence of this work is that the use of inverse sequences as a negative set in experiments should be done with caution as they cannot be considered as random.

URL

http://staffnet.kingston.ac.uk/~ku33185/ProteinSequenceConverter/OpProtein.html

Presenting Author

Jean-Christophe Nebel (j.nebel@kingston.ac.uk)
Kingston University

Author Affiliations

Kingston University, London

Acknowledgements

Why inverse proteins are relatively abundant, J.-C. Nebel and C. Walawage, Protein & Peptide Letters, 17(7), 2010

A-6. HERD: the highest expected reward decoding for HMMs with application to recombination detection

Nánási M (1,), Vinař T (1), Brejová B (1)*

Hidden Markov models (HMMs) are an important tool for modeling biological sequences and their annotations. By sequence annotation we mean assignment of labels to each symbol according to its function. The Viterbi algorithm, commonly used for HMM decoding, finds the most probable annotation some HMMs. In general, the sequence annotation is NP-hard (Brejová et al. 2007) and the Viterbi algorithm is used as a heuristic. Recently it has been shown that other decoding methods lead to more accurate results in specific applications (Kall et al 2005, Gross et al. 2007, Brown and Truszkowski 2010).

Materials and Methods

We propose a new efficient HMM decoding algorithm called the highest expected reward decoding (HERD). It is appropriate, when we want to partition sequence into features, but allow some tolerance in the placement of each feature boundary. We define our objective function in the terminology of gain functions (Hamada et al. 2009), where the gain characterizes the similarity between a predicted and the correct annotation. Then we seek the annotation with the highest expected gain where the expectation is taken over all annotations in the HMM, since we do not know the true annotation.

Results

We evaluate our approach on the problem of detecting recombination in the HIV genome and compare it with an existing tool called jumping HMM which uses the Viterbi algorithm (Schultz 2009). On artificial recombinants created from real HIV sequences, HERD has significantly higher feature specificity and sensitivity than the Viterbi algorithm using the same HMM. Here, a feature is a predicted block between two adjacent recombination points, and we count it as correctly predicted if its boundaries are misplaced by at most 10 symbols.

Discussion

The results show that the HERD predicts viral recombination with higher accuracy than the Viterbi algorithm. Here, the Viterbi algorithm finds the highest scoring alignment of a query to a profile HMM. In contrast, we marginalize over all alignments and over nearby placements of recombination points. Our algorithm can be easily extended to other similar gain functions, and therefore it can be used in other application domains. Our method has several parameters that have significant impact on the prediction accuracy. It remains an open question how to systematically choose their values.

URL

http://www.compbio.fmph.uniba.sk/herd/

Presenting Author

Michal Nánási (mic@eccb.ksp.sk)
Comenius University in Bratislava

Author Affiliations

(1) Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

Acknowledgements

This research is funded by European Community FP7 grants IRG-224885 and IRG-231025 and VEGA grant 1/0210/10.

A-7. Visualizing the next generation of sequencing data with GenomeView

Abeel T (1,2,), Van Parys T (1), Galagan J (2), Van de Peer Y (1)*

High-throughput sequencing instruments are widely available. As costs decrease and volumes increase, sequencing is rapidly moving to become a general purpose assay as a read-out for a specific biological question. Thus, applications including RNA-seq, Chip-Seq and SNP discovery have become common. Moreover, as sequencing a single genome has become more affordable, it is no longer unusual to have sequences for several closely related genomes. Yet, although it is relatively easy to acquire HTS data, there is still a critical need for tools that allow scientists to effectively use these data.

Materials and Methods

Data types that can be visualized include reference sequences, annotation, multiple alignments, high-throughput sequencing mappings and synteny. Multiple file format are supported for each data type and we cover most of the common file formats. Furthermore it is very easy to link GenomeView with any existing data sources,because GenomeView supports a broad range of widely used data formats. Integrating our tool with a website consists of two steps: (i) put the data on a webserver, (ii) make a URL pointing to GenomeView and the data. Creating this URL is easy and documented in the manual.

Results

To address this need for interactive and dynamic visualization of gigantic datasets coming from HTS, we have developed a user-friendly and intuitive tool, called GenomeView. GenomeView can handle a broad range of the emerging information that results from NGS. In particular, GenomeView is the only tool currently available that allows biologists to rapidly and interactively browse, edit, and analyze genome wide datasets of mapped short read sequences and comparative data. Website: http://genomeview.org/ (free download, tutorials, manual, screencasts, etc.)

Discussion

GenomeView provides biologists with a unique tool to quickly explore their data in an intuitive and visually appealing way. It has value throughout any genomics project. At the start of a project it can act as a portal to get an overview of the data and to generate hypotheses. Later on, it provides valuable visual sanity checks on any preliminary results. And finally, it is also valuable at the end of the project, to present the results and data in an appealing way to the scientific community.

URL

http://genomeview.org

Presenting Author

Thomas Abeel (thomas@abeel.be)
VIB-UGent

Author Affiliations

(1) VIB Department of Plant systems biology, Ghent University, Gent, Belgium (2) Broad Institute of MIT and Harvard, Cambridge, MA, USA

A-8. IsUnstruct: a method based on Ising's model for prediction of disordered residues from protein sequence alone

Lobanov MY (1), Galzitskaya OV (1,)*

Intrinsically disordered regions serve as molecular recognition elements, play an important role in the control of many cellular processes and signaling pathways. It is useful to be able to predict positions of disordered residues and disordered regions in protein chains.

Materials and Methods

In this work, a statistical analysis of disordered residues was made by considering 28727 unique protein chains taken from the PDB database. In this database, 4.65% of residues are disor-dered. The statistics was obtained separately for the N- and C-termini as well as for the central part of the protein chain. The optimal parameters have been sorted out for predictions using prepared databases.

Results

A new method based on the dynamic programming was developed for searching not only disordered regions but also individual disordered residues in protein chain. This method correctly finds 77% of disordered residues as well as 87% of ordered residues in the CASP8 database.

Discussion

Comparison of our method with other methods has shown that our method is one of the best.

URL

http://antares.protres.ru/IsUnstruct/

Presenting Author

Oxana V. Galzitskaya (ogalzit@vega.protres.ru)
Institute of protein research

Author Affiliations

(1) Institute of Protein Research, Russian Academy of Sciences

Acknowledgements

This work was supported by the Russian Academy of Sciences ("Molecular and Cell Biology" and "Fundamental Sci-ence to Medicine" programs), by the Russian Foundation for Basic Research (08-04-00561). We are grateful to Prof. J.L. Sussman for providing us the materials from CASP8.

A-9. MASiVE: Mapping and Analysis of SireVirus Elements in plant genome sequences

Darzentas N (1,), Bousios A (1), Apostolidou V (1), Tsaftaris AS (1,2)*

The availability of large plant genome sequences has necessitated the development of tools for high throughput discovery and annotation of transposable elements (TE), which comprise the majority of these genomes. Such algorithms are available, however they are based on general TE structural characteristics that lead to misannotations or incomplete identification. Our discovery that the abundant (20% of the maize genome) Sireviruses contain highly conserved motifs in their genome, provided the opportunity to develop MASiVE, a tool able to identify with high precision intact Sirevirus elements.

Materials and Methods

MASiVE is written in Perl and makes use of external algorithms (LTRharvest, Vmatch and GeneWise) and of the highly conserved Sirevirus multiple polypurine tract (PPT) signature and primer binding site (PBS) to build a pipeline of filters applied sequentially. Initially, LTRharvest scans the genomic sequence for LTR retrotransposons (LTR-RTN) and only elements that overlap with a multiple PPT signature are retained. Next, a cascade of distance requirements, LTR-RTN specific gene hits and within-element homology-based searches result in the accurate annotation of intact Sirevirus elements.

Results

We demonstrate the specificity of the multiple PPT signature and its capacity to be used by MASiVE. The stepwise application of the filters rejected 2086 elements from a total of 3870 LTRharvest-predicted Sireviruses that overlapped with a multiple PPT signature. The final set totals 1784 high quality intact Sireviruses in maize chromosome one. Comparison with the annotation from the Maize Transposable Element Consortium (MTEC) shows that MASiVE not only recovered 76% of the MTEC set, but also added another 50% intact elements to the pool.

Discussion

Based on the highly conserved genome structure of Sireviruses, we developed MASiVE, a novel method able to discover with high sensitivity and precision intact Sireviruses in plant genomic sequences. MASiVE also provides detailed information on elements that failed the filters, thus expanding the view on the genome distribution of Sireviruses. Since Sireviruses are among the most abundant LTR-RTN genera of plant genomes, we believe MASiVE to be a valuable addition to the toolbox of scientists trying to untangle the intriguingly complex genomic landscape of plants.

Presenting Author

Nikos Darzentas (ndarz@certh.gr)
Institute of Agrobiotechnology, Centre for Research and Technology Hellas (CERTH)

Author Affiliations

1 Institute of Agrobiotechnology, Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece 2 Department of Genetics and Plant Breeding, Aristotle University of Thessaloniki (AUTh), Thessaloniki, Greece

Acknowledgements

This work was supported by the FP6 Biosapiens Network of Excellence (contract number LSHG-CT-2003-503265).

A-10. Signals in human promoter sequences

*Putta P**

It is believed that there are no universal or conserved core promoter sequence in eukaryotes as that of clear signals like TATAAT and TTGACA around -10 and -35 from TSS in prokaryotes.Eukaryotic promoters are structurally complex and so it is believed that transcription in more complex than currently believed.This has motivated us to understand the complexity in sequence-specific binding nature of TFs in promoter recognition during initial stages of transcription based on the 6-nts that are present within the promoter sequences.

Materials and Methods

We have looked at functional promoter sequences in Homo sapiens from Eukaryotic Promoter Database (EPD). We have located relatively conserved GC-rich 6-nt sequences that are common in the promoter sequences. Next looked for their distribution with in the promoter sequences and also in human miRNA datasets (from miRBase database) and transcription factor binding sites from JASPAR database.

Results

Sigmoidal distribution of 6-nt sequences in upstream and downstream of TSS suggests existence of some internal co-operativity in their distribution.We also noticed that more than half of the promoter sequences contain multiple signals that favor strong binding signals for TFs.Many of these 6-nts were identified in miRNA datasets suggesting that miRNAs play significant role in recognition of promoters.We have identified 6-nts sequences in transcription factor binding sites of TFs.We correlated the promoters, miRNA and TFs based on common 6-nts.

Discussion

We have located GC-rich 6-nt signals around TSS in human promoter sequences.Sigmoidal distributions of 6-nts suggest that at least many of them are likely to have important biochemical roles.TFs must straddle on both strands at or near the TSS though they must be relived for continuation of transcription.Correlation between promoters and miRNAs suggest that a group of promoters or genes could be regulated by few miRNAs.We postulate that miRNAs play significant role in promoter recognition.We aim to classify TFs based on 6-nt sequences and also studying their biological significance.

Presenting Author

Padmavathi Putta (padma_bioinfo@yahoo.co.in)
Dept of Biochemistry, University of Hyderabad

Author Affiliations

Department of Biochemistry, University of Hyderabad, Hyderabad - 500046 Andhra Pradesh, India.

A-11. Iterative read mapping and assembly allows the use of more distant reference in metagenome assembly

Dutilh BE (1,), Huynen MA (1), Gloerich J (2), Strous M (3,4,5)*

Most microbial species can not be cultured in the lab. Metagenomic sequencing may still yield a complete genome if the sequenced community is enriched and the sequencing coverage is high. However, a natural population may contain multiple related strains. Moreover, it is not uncommon that these strains represent a quasispecies that is only distantly related to the closest available reference genome. This can confound strict assembly programs and lead to a fragmented assembly.

Materials and Methods

The full methods are explained in [Dutilh et al. 2009, Bioinformatics 25:2878]. Briefly, we used a permissive (BlastN) and a strict (Maq) mapping algorithm to assemble a majority consensus. As the initial assembly better represents the sequenced genomes than the reference genome does, we iterated the mapping and assembly several times.

Results

We show that by iteratively mapping short metagenomic sequencing reads from a population of strains to a related reference genome, we can create a genome that captures the consensus of the population's sequences. Iteration allows us to map more of the reads, leading to a higher coverage and depth of the assembled consensus genome. At the same time, the similarity with the reference genome decreases, while the number of metaproteomic peptides that can be explained by the assembly increases.

Discussion

This indicates that the assembly becomes less dependent on the reference genome and approaches the consensus genome of the multi-strain population. Thus, by exploiting the homology offered by a reference genome in combination with permissive, iterative read mapping, we get a better view of both the consensus genome sequence of the quasispecies present in the sample and of the sequence diversity between the strains.

URL

http://www.cmbi.ru.nl/~dutilh/

Presenting Author

Bas E. Dutilh (dutilh@cmbi.ru.nl)
CMBI / NCMLS / Radboud University Nijmegen Medical Centre

Author Affiliations

(1) Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands (2) Nijmegen Proteomics Facility, Radboud University Nijmegen Medical Centre, Geert Grooteplein-Zuid 10, 6525 GA, Nijmegen, The Netherlands (3) Department of Microbiology, Radboud University Nijmegen, Heyendaalsweg 135, 6525 AJ, Nijmegen, The Netherlands (4) MPI for Marine Microbiology, Celsiusstrasse 1, D-28359, Bremen, Germany (5) Centre for Biotechnology, University of Bielefeld, Bielefeld, Germany

Acknowledgements

This research was funded in part by Dutch Science Foundation (NWO) Horizon Project 050-71-058.

A-12. APPRIS: a database of annotated principal splice variants for the human genome

Rodríguez J-M (1,2,), Ezkurdia I (1), Lopez G (1), Pietrelli A (1), Maietta P (1), Wesselink J-J (1), Valencia A (1,2), Tress M (1)*

The role played by alternative splicing in the modulation of cellular function is some way from being clarified. Recent works have suggested that many alternative isoforms are likely to have altered structure, localisations and cellular functions. Given the likely ubiquity in the cell of alternative isoforms, the role of these alternatively spliced gene products is becoming an increasingly important question. Determining principal functional variants is a critical first step in the study of the implications of alternative splicing.

Materials and Methods

We have developed a database to aid in the annotation of any set of well-annotated genes. The database deploys a range of computational methods. For example, variants are mapped to known structures using the module MATADOR and Firestar predicts individual functionally important residues. The principal isoform for each gene is selected based on the annotations from these methods. The database is currently being used in collaboration with the GENCODE consortium to annotate the human genome and is part of the scale up of the ENCODE project to annotate 100% of the human genome.

Results

We have annotated the 22,304 genes in the current HAVANA release of the human genome (release 3C). Based on this information we have been able to select a principal isoform for over 75% of the genes in the human genome. Many of the alternative variants are likely to have changed structure, localisation or function. Over 35% of the annotated alternative splice variants have damaged Pfam functional domains, while at least 30% of the alternative variants are likely to have substantially altered protein structure.

Discussion

Alternative splicing has the potential to expand the cellular protein repertoire by altering the biological function of the expressed proteins. Many of the alternative splice variants that we see here are predicted to have changed structure, localisation or function in relation to the selected principal variant. These are results are an interesting first step. What is required now is experimental evidence of the expression of these isoforms as proteins and the characterization of the structure and in vivo function and localisation of the alternative isoforms.

URL

http://appris.bioinfo.cnio.es

Presenting Author

Jose M. Rodriguez Carrasco (jmrodriguez@cnio.es)
Spanish National Bioinformatics Institute (INB)

Author Affiliations

(1)-Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO). Madrid, Spain.
(2)-Spanish National Bioinformatics Institute (INB).

Acknowledgements

This is a NIH-funded ENCODE project and was supported in part by Spanish National Bioinformatics Institute (INB).

A-13. Accurate long read mapping using enhanced suffix arrays

Vyverman M (1,), Fack V (1), De Schrijver J (2), Dawyndt P (1)*

With the rise of high throughput sequencing, new programs have been developed for dealing with the alignment of a huge amount of short read data to reference genomes. Recent developments in sequencing technology allow longer reads, but the mappers for short reads are not suited for reads of several hundreds of base pairs. We propose an algorithm that can handle large reads with mutations and long insertions and deletions.

Materials and Methods

Our algorithm is based on chaining together maximal exact matches (MEMs) and uses heuristics to speed up the search. To compute the MEMs we use a specialized index structure, called enhanced suffix array. The MEMs are used as anchors in the alignment and are combined to a full local alignment using a combination of the Needleman-Wunsch algorithm and some heuristics.

Results

The proposed algorithm has been tested on simulated BRCA1 data. BRCA1 is a gene known to be involved in the development of breast cancer when mutated. A reference sequence of 80000bp was used, together with about 80000 queries of length between 8bp and 1021bp. We compared our results with results obtained by Bowtie. Our algorithm was able to find an optimal alignment for 95% of the queries, whereas Bowtie found only 31% of them. Moreover, our algorithm is very fast, taking only a few seconds.

Discussion

We presented an algorithm for mapping large reads to a reference genome which is fast and accurate in finding an optimal local alignment. The algorithm is easy to understand and has few parameters. Our first results show that the algorithm is able to map reads with insertions, deletions and mutations and with a length of several hundreds of base pairs successfully to a reference sequence.

Presenting Author

Michaël Vyverman (Michael.Vyverman@UGent.be)
Ghent University, Dept. Applied Math. & Comp. Sci.

Author Affiliations

(1) Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281-S9, 9000 Ghent, Belgium
(2) Department of Molecular Biotechnology, Ghent University, Coupure Links 653, 9000 Ghent, Belgium

A-14. Detection and architecture of small heat shock protein monomers

Poulain P (1), Gelly J-C (1), Flatters D (1,)*

sHSPs are chaperone-like proteins able to prevent aggregation of misfolded proteins. Found in most organisms, these proteins form large oligomeric complexes. Their chaperone function is tightly associated to their highly dynamics quaternary structures. The features that regulate the sequence - structure - function relationships remain still unclear. Here, we propose a better understanding of the architecture, organization and properties of the sHSP family through structural and functional annotations of monomers. We focused on ACD, a beta-sandwich fold, that is the hallmark of the sHSP family.

Materials and Methods

sHSP proteins are characterized by low sequence identity, few available structures but with a strong structural pattern. We developed a new approach for detecting protein sequences of sHSPs and delineating ACDs. This approach is based on an iterative Hidden Markov Model algorithm using a multiple alignment profile generated from structural data on ACD. Protein sequences identified as sHSPs were found in the UniProt databank that contains more than 10 million protein sequences. sHSPs were grouped based on their taxonomic origin.

Results

We found 4478 sequences identified as sHSP showing a very good coverage with the corresponding PROSITE and Pfam profiles. We showed that taxonomic-based groups of sHSPs have unique features regarding their ACD. We detailed highly conserved residues and patterns specific to the whole family or to some groups of sHSPs. For 96% of studied sHSPs, we identified in the C-terminal region a conserved I/V/L-X-I/V/L motif that acts as an anchor in the oligomerization process. The fragment defined from the end of ACD to the end of this motif is named the C-terminal Anchoring Module (CAM).

Discussion

In this work, we annotated structural components of ACD and quantifies properties of thousands sHSPs. Structural elements identified as essential to the oligomerization process are associated with specific sequence properties in each group. We showed the relevance of the iterative HMM approach based on structural data, learned on sequences but aimed toward elucidating structural/functional properties of sHSPs. As a perspective, we intend to make a web server dedicated to this family.

Presenting Author

Delphine Flatters (delphine.flatters@univ-paris-diderot.fr)
INSERM UMR-S 665 and Université Paris Diderot

Author Affiliations

(1) DSIMB, Inserm UMR-S 665 and Université Paris Diderot – Paris 7, INTS, Paris, France

Acknowledgements

Financial support was provided by grants from the French Ministry of Research, the Paris Diderot - Paris 7 University, the National Institute for Blood Transfusion (INTS) and the Institute for Health and Medical Research (INSERM).

A-15. Assembly of mitochondrial genomes using next-generation sequencing data

Felder M (1,), Taudien S (1), Groth M (1), Münsterkötter M (2), Navarro-Quezada A (3), Knogge W (3), Platzer M (1)*

Classical approaches to sequence and analyse mitochondrial DNA are based on the separate isolation and/or amplification of mitochondrial DNA. The aim of our approach was to identify and analyse mitochondrial DNA using next-generation data derived from whole-genome shotgun sequencing of Rhynchosporium secalis UK7 and 6a.1. These two isolates are of particular interest as pathogens of barley and rye, respectively.

Materials and Methods

For both isolates whole-genome 454 libraries were generated according to the manufacturer's protocol and sequenced using a GS FLX (Roche). In addition, whole-genome paired-end DNA libraries with insert sizes of about 330 bp were constructed and sequenced on a Solexa/Illumina Genome Analyser II using standard procedures. In total, for each of the isolates more than 2 Gb of usable sequence were generated. The nucleotide sequences were assembled using the Newbler software (Roche).

Results

De novo assemblies using datasets with reduced read counts resulted in putative mitochondrial sequences of the R. secalis isolates. By remapping reads to these sequences the complete circular mitochondrial genomes of R. secalis UK7 (69,581 bp) and 6a.1 were obtained (68,729 bp). A comparison of the R. secalis mitochondrial DNAs with fungal mitochondrial reference sequences confirmed the high degree of conservation of sequence and structure within these fungi.

Discussion

Despite the extremely high mitochondrial and the low nuclear sequencing depths the mitochondrial genomes of two strains of the plant pathogenic fungus R. secalis were identified and analyzed from whole-genome shotgun sequencing data.

Presenting Author

Marius Felder (mfelder@fli-leibniz.de)
Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI)

Author Affiliations

(1) Genome Analysis, Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI), Beutenbergstr. 11, D-07745 Jena, Germany (2) MIPS - Institute of Bioinformatics and Systems Biology, Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany (3) IPB - Leibniz Institute of Plant Biochemistry, Weinberg 3, D-06120 Halle, Germany

Acknowledgements

Collaborative PAKT project of IPB Halle, Germany

A-16. AnsNGS: Annotation system of structural variation for Next Generation Sequencing data

Na Y-J (1,2,), Park CH (1,2), Cho Y (1), Kim JH (1,3)*

The ongoing revolution in sequencing technology has led to the production of sequencing machines with dramatically low costs and high throughput. While the promise of next-generation sequencing technologies has become a reality, they also present substantial informatics challenges. Most current informatics methods have focused on preprocessing such as alignment and assembly or visualization of mapping read sequences. However, in order to satisfy the impending need for deciphering the large-scale data generated from next-generation sequencing, it is essential to the development of an integrated

Materials and Methods

AnsNGS can import data from MAQ and SOAP which are the most popular read mapping methods. It supports annotated sequence information by integrating genome feature annotation databases such as genes, non-coding RNAs and structural variations. To speed up query, R-Tree with quadratic splitting to index spatial columns was used to retrieve annotation information from the integrated database. The use of a parallel computing system based on multi-core processor architectures will scale its performance. AnsNGS can export the annotation results for post-analysis.

Results

To determine the effectiveness of AnsNGS, we have tested AnsNGS on several Roche 454, Illumina Solexa and ABI SOLiD data sets with data size ranging from 10 to 50 million reads. All mapped reads are annotated, using dual-core CPU processors on a Linux machine with 16G RAM.

Discussion

AnsNGS supports correlations between aligned read sequences and transcriptome features. Moreover, it provides known structural variation as well as sequence annotation information. Elucidation and utilization of individual genetic differences including various structural variations would be a major hallmark of personalized medicine.

URL

http://clara.snubi.org:8080/shortRead/

Presenting Author

Young-Ji Na (yjna01@snu.ac.kr)
Seoul National University Biomedical Informatics (SNUBI) Seoul National University College of Medicine

Author Affiliations

1Seoul National University Biomedical Informatics (SNUBI) Seoul National University College of Medicine 28 Yongon-dong Chongno-gu Seoul 110-799, Korea 2Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Korea 3Division of Biomedical Informatics Seoul National University College of Medicine 28 Yongon-dong Chongno-gu Seoul 110-799, Korea

A-17. Assessing the accuracy and completeness of the bonobo genome sequence

Prüfer K (1,), Ptak SE (1), Fischer A (1), Good JM (1), Mullikin JC (2), Miller J (3), Kodira CD (4), Knight JR (4), The Bonobo Genome Consortium, Kelso J (1), Pääbo S (1)*

We analyze the quality of the bonobo genome assembly, generated from 25-fold coverage of 454 SequencingTM data, encompassing both shotgun and paired end reads. This assembly has a N50 scaffolds size of 9.6 megabases and a N50 contig size of 67 kilobases.The bonobo genome sequence is particularly suited for the analysis of assembly quality since the genome sequences of two closely related species (human and chimpanzee) are available for comparison.

Materials and Methods

We use whole genome alignments between the genomes of human, chimpanzee and bonobo to estimate the completeness and sequence accuracy of the Bonobo genome in comparison with the chimpanzee genome.

Results

Our analysis shows that the bonobo assembly achieves comparable quality to the chimpanzee assembly in terms of sequence accuracy and genome completeness. Using the high quality of the finished chimpanzee chromosome 21 we are able to estimate a rate of approximately two errors in 10,000 base pairs for the bonobo genome.

Discussion

The quality of the Bonobo X chromosome assembly is as good as the other autosomes since a female individual was selected for sequencing. This places this bonobo's X and all her autosomes in an excellent position for comparative analysis to other primates. We conclude that large and complex genomes can be de novo assembled from next generation sequencing data.

Presenting Author

Kay Prüfer (pruefer@eva.mpg.de)
Max-Planck Institute for Evolutionary Anthropology

Author Affiliations

1 Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany 2 National Institute of Health, Bethesda, MD, USA 3 J. Craig Venter Institute, Rockville, MD, USA 4 454 Life Sciences, Branford, CT, USA

A-18. Estimating effective DNA database size via compression

Visnovska M (1,), Nanasi M (1), Vinar T (1), Brejova B (1)*

Recent progress in genome sequencing technologies led to a rapid increase in the size of DNA sequence databases. In large databases, many query sequences will have high scoring matches purely by chance. To distinguish real matches from spurious ones, we need to assess their statistical significance. One of the main parameters used in P-value computation is the database size n. Instead of the real database size, we propose to use an effective database size which will account for redundancies in the database.

Materials and Methods

One way of describing the information content of a database D is its Kolmogorov complexity (Li and Vitanyi, 2008). The Kolmogorov complexity is not computable, but the compressed size of D obtained by a fixed compression algorithm can be used as an upper bound. We use the compressed size of D as its effective size. We have implemented an algorithm that computes exact P-values for a given database D and all of its prefixes in the simple scenario of finding the closest match to a short query under Hamming distance. It can be used on real or randomly generated databases.

Results

We tested our approach on databases of three different types: randomly generated databases with GC-contents 75% and 90%, artificial databases that are concatenation of many mutually similar sequences of the same length, and real genomic data from human, chimpanzee, and rhesus macaque. Experimental results suggest that P-value estimates based on effective size are appropriate for large databases, but not conservative for the small. However, conditional entropy as a correction factor for small databases seems promising.

Discussion

We have considered methods for more accurate estimation of P-values for sequence homology search. In particular, we propose to adjust the size of the database to compensate for the structure present in the database due to the fact that individual sequences are related by evolution. Our method is flexible and easy to implement. In future, we would like to extend our work to more complex scenarios of homology search. Longer query sequences will require handling of insertions and deletions. More complex scoring schemes on both nucleotide and protein sequences also need to be examined.

Presenting Author

Martina Visnovska (visnovska@ii.fmph.uniba.sk)
Department of Applied Informatics, Faculty of Mathematics, Physics, and Informatics, Comenius University

Author Affiliations

1 Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava

Acknowledgements

This research is funded by European Community FP7 grants IRG-224885 and IRG-231025, VEGA grant 1/0210/10, and Comenius University grant UK/151/2010.

A-19. ChIP-seq analysis: from peaks to motifs

Thomas-Chollier M (1,), Defrance M (2), Sand O (3), Herrmann C (4), Thieffry D (4), van Helden J (4,5)*

The treatment of the raw ChIP-seq data proceeds in three steps :(i) read mapping; (ii) peak calling; (iii) detecting regulatory motifs in peak sequences. Whereas many software tools are being developed to address the two first steps, most classical methods for analyzing regulatory motifs are not able to cope with the unprecedented amounts of sequences to be processed: computing time and memory requirements constitute the main bottleneck for peak sequence analysis. Consequently, a common strategy is to restrict the analysis to a subset of the top-scoring peak sequences.

Materials and Methods

We present ChIP-motifs, a pipe-line combining several motif discovery approaches that are scalable to genome-wide datasets. Those approaches rely on distinct statistical criteria to detect exceptional words: global over-representation, local over-representation and positional bias. The analysis covers oligonucleotides as well as spaced pairs (dyads), which are particularly suited for dimeric transcription factors. Discovered motifs are compared to collections of previously known motifs. The pipe-line also analyses mono- and di-nucleotide compositions of the peaks.

Results

When applied to 13 ChIP-seq peak sets from mouse transcription factors controlling stem cell pluripotency [Chen et al. (2008) Cell 133:1106-17], ChIP-motifs successfully recovers the reference motifs, and reveals potential co-factors. In most cases, several algorithms of the pipe-line independently discover the same motifs, revealing both over-representation and bias towards the center of the peaks.

Discussion

The approach proposed here answers a crucial need for the treatment of full datasets resulting from ChIP-seq and ChIP-chip experiments. The program chip-motifs is integrated in the Regulatory Sequence Analysis Tools (RSAT), and can be accessed via a Web form, SOAP/WSDL Web services, or as a stand-alone application. Its low time and memory requirements enable to treat datasets comprising tens of thousands peaks on a personal computer.

URL

http://rsat.ulb.ac.be/rsat/

Presenting Author

Jacques van Helden (Jacques.van.Helden@ulb.ac.be)
Université Libre de Bruxelles

Author Affiliations

1. Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany. Email: morgane@bigre.ulb.ac.be 2. Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México. Av. Universidad, Cuernavaca, Morelos 62210, Mexico. Email: defrance@ccg.unam.mx 3. CNRS-UMR8199 Institut de Biologie de Lille. Génomique et maladies métaboliques. 1, rue du Pr Calmette, 59000 Lille, France. Email: sand@good.ibl.fr 4. Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Université de la Méditerranée. Campus de Luminy, F - 13288 Marseille, France. Email: {thieffry,herrmann}@tagc.univ-mrs.fr 5. Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium. Email: Jacques.van.Helden@ulb.ac.be

Acknowledgements

This work was supported by the Belgian Program on Interuniversity Attraction Poles (IAP), initiated by the Belgian Federal Science Policy Office, project P6/25 (BioMaGNet). The BiGRe laboratory is supported by the MICROME Collaborative Project funded by the European Commission within its FP7 Programme, under the thematic area "BIO-INFORMATICS - Microbial genomics and bio-informatics", contract number 222886-2." MTC is supported by the Alexander von Humbold Stiftung.

A-20. Applying mass spectrometry data to improve annotation of the reference mouse genome

Saunders GI (*), Brosch M, Frankish A, Collins MO, Yu L, Harrow J, Choudhary JS, Hubbard T

With the advent of the next-generation DNA sequencing technologies, our understanding of the transcriptome has rapidly improved very recently; however, the same cannot be said of the proteome. Recent advances in protein Mass Spectrometry (MS) are of great excitement to genome annotators, offering the chance to marry high-throughput peptide sequencing to predicted coding sequences, and to identify new protein-coding loci. For the first time, we have systematically validated and improved mouse genome annotation with tandem MS data.

Materials and Methods

This study is based on 10,465,149 tandem MS spectra, of which 9,735,566 were obtained from the Peptide Atlas project and 729,583 from in-house experiments on nuclear protein extracts of murine embryonic stem cells and murine brain membrane fractions. All MS peaklist data were searched with Mascot and post processed with Mascot Percolator. The processing parameters used in each programme were identical to those employed by Brosch 2009 (Brosch et al. Journal of Proteome Research 8: 3176-3181). We used the Ensembl Perl API to unambiguously map the resulting peptides to the reference mouse genome.

Results

We have validated 32% of all mouse protein-coding genes, 17% of all protein-coding exons and 7% of all protein-coding splice boundaries. In addition we present evidence for the identification of 53 high confidence genes with multiple alternative translations. Moreover, we have uncovered 10 novel protein-coding loci that are not present in any mouse annotation data sources. Our study also provides 24 high confidence peptide spectrum matches that map to and support the translation of 10 processed pseudogenes, indicating their resurrection into coding forms.

Discussion

In spite of the mouse proteome being far from saturated by MS based peptide identifications, we have applied our novel annotation pipeline to enable the first proteogenomic study to both validate and improve the annotation of this genome. We believe that this method to identify protein-coding loci will become a mainstay of genome annotation pipelines in the near future and we provide results to support our view that, in order to avoid erroneous annotation, it should be our high stringency approach that is adopted.

URL

http://das.sanger.ac.uk/das/ms_das

Presenting Author

Gary I. Saunders (gs6@sanger.ac.uk)
Wellcome Trust Sanger Institute

Author Affiliations

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK, CB10 ISA

Acknowledgements

This work was funded by the Wellcome Trust.

A-21. Prediction of splice-modifying SNPs in human genes using a new analysis pipeline called AASsites

Hotz-Wagenblatt A (1,), Faber K (1), Risch A (2), Glatting K-H (1)*

Single nucleotide polymorphisms (SNPs) become more and more important in disease research. The aim of most existing SNP tools is the annotation of SNPs mainly by analyzing variances in the coding region of the affected gene and possible effects on the transcriptome and proteome levels. In contrast, our AASsites pipeline (Automatic Analysis of SNP sites) indicates if a change in the splice pattern due to a SNP is likely to occur.

Materials and Methods

Implemented in the W3H-pipeline system, AASsites uses a combination of different gene prediction methods (GenScan, GeneID, HMMgene, GlimmerHMM and GrailExp), and an ESE (Exonic Splice Enhancer) detection to predict changes in the elements which are relevant for splicing. Additionally, Genewise analysis determines changes in the ORF (open reading frame). The rule based classification system ranks the SNPs and their effects into different categories.

Results

Modifications in the exon-intron structure and variances in ESE and ORF are shown in the final result of the pipeline including new, disappeared and modified exons and introns. 37 SNPs with known effects on splicing and 23 negative SNPs were used for testing AASsites. 73% of the positive SNPs and 100% of the negative SNPs were classified correctly. We checked 80,000 SNPs from the human genome which are located near splice sites for their ability to change the splicing pattern of the gene. We identified 301 “likely” and 985 “probable” classified SNPs with such characteristics.

Discussion

We have shown with a set of SNPs with known changes, that the pipeline can predict the change in splicing caused by the SNP in 83% of 60 cases correctly. The problem of testing and improving the rule system for combining the results is caused by the small number of experimentally proven SNP-derived modifications in splicing. With more experimental data available we could replace the rule system by a knowledge system based on machine learning algorithms.

Presenting Author

Agnes Hotz-Wagenblatt (hotz-wagenblatt@dkfz.de)
German Cancer Research Center (DKFZ)

Author Affiliations

1 Bioinformatik (HUSAR), Core Facility Genomics Proteomics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany 2 Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

A-22. Structator: a tool for fast index-based matching of RNA structural motifs

Meyer F (1,), Will S (2), Beckstette M (1)*

The exponential growth rate of biological sequence databases asks for tools that allow for fast search of biologically relevant motifs. While several sequence-pattern matching problems can be vastly accelerated by index data structures like suffix trees and suffix arrays, matching RNA structural motifs is more efficient with affix arrays. Despite not being a new data structure, its potential was neither recognized by the scientific community, nor was there a public implementation of it.

Materials and Methods

Structator matches sets of patterns employing the affix array index data structure introduced by [Strothmann, 2007]. The index is the combination of the enhanced suffix array of [Abouelhoda et al., 2004] and the enhanced reverse prefix array of a given sequence. Patterns are matched in a bidirectional way, i.e. possibly beginning with a character in the middle with consecutive left and right extensions. In a second step, Structator performs RNA family classification by building chains of matches where the order of the patterns is preserved using the algorithm of [Abouelhoda et al., 2005].

Results

We measured the running time to search with 392 motifs describing families of Rfam rel. 10.0 where the consensus structure contained at least 5 stem-loop motifs. On the precomputed affix array of the database, Structator performed all searches in 57 sec., whereas a simple online algorithm required 105 min. 12 sec. In a second experiment we searched with eight motifs from family RF00193 of Rfam rel. 9.1. Via chaining of the thousands of matches, Structator removed all spurious hits, thus identifying the 26 members of the family in the full alignment.

Discussion

Structator's main features are persistent affix array construction, flexible alphabet handling (DNA, RNA, proteines) including alphabet transformation, matching on forward and reverse strand, support of various pattern types including IUPAC wildcard symbols, support of standard and user-defined base pairing rules, integrated chaining algorithms, output of results in different formats e.g. BED for visualization in the UCSC Genome Browser.

Presenting Author

Fernando Meyer (meyer@zbh.uni-hamburg.de)
Center for Bioinformatics, University of Hamburg

Author Affiliations

(1) Center for Bioinformatics, University of Hamburg (2) Computation and Biology Group, Massachusetts Institute of Technology

A-23. Project HOPE: Providing the last piece of the puzzle...

Venselaar H (1)*

Recent developments in next-generation sequencing increased the detection of disease-related mutations, a significant part affecting the protein's 3D-structure. Explaining their structural effect requires knowledge of the protein, obtained from a series of sources including the 3D-structure and several databases such as the uniprot database. Existing online prediction-servers often do not focus on the structural effect of the mutation. Therefore, we developed project HOPE, a web-server for automatic point-mutant analysis that is easy-to-understand for anyone in the (bio)medical field.

Materials and Methods

The core of Project HOPE is a Java-implemented module that collects information from a series of sources and stores this in a database. This information consists of calculations on the real structure or a homology model by WHAT IF web-services, sequence based predictions by DAS-servers, and structural features annotated in Uniprot. All information is stored per residue in a PostGreSQL database. A decision scheme is used to reach a conclusion about the effect of the mutation that focuses on the structural changes. The conclusion is shown on a website and illustrated with figures and animations.

Results

Our approach was tested in numerous collaborative projects with researchers from several (bio)medical departments. Our contribution consisted mainly of mutation studies (non-sense and missense mutations) and suggestions for experimental design. Often, our findings provided insight that was necessary to understand the mutation and can therefore be seen as the last piece that completed the molecular puzzle.

Discussion

Our examples show that a structural analysis provides essential insight in the effect of a mutation. We developed Project HOPE to provide easy accessible and understandable structural information that can be beneficial for research in the (bio)medical field.

URL

http://www.cmbi.ru.nl/hope

Presenting Author

Hanka Venselaar (h.venselaar@cmbi.ru.nl)
CMBI, NCMLS, Radboud University Nijmegen Medical Centre

Author Affiliations

Centre for Molecular and Biomolecular Informatics (CMBI) NCMLS, Radboud University Nijmegen Medical Centre

A-24. Simple not-PWM-based approach for transcription factor binding site prediction

Fazius E (1), Shelest V (1), Shelest E (1)*

The prediction of transcription factor binding sites (TFBS) is crucial for promoter modeling and consequent network inference. Despite the huge number of tools for the TFBS prediction, no satisfactory solution for the high number of false predictions is found so far. All available tools are PWM-based and differ only in the ways of calculating the cut-offs. Recent new approaches tend to use more mathematically sofisticated methods, without breakthrough results. We decided to develop a straightforward non-heuristic metod based on direct comparison of training set motifs with the query sequence.

Materials and Methods

The tool is implemented in PHP.

Results

In our approach, the query sequence is scanned for motifs with a sliding window. Each motif is assigned a weight that takes into account the probabilty of the motif occurrence by chance, the number of matching motifs of the training set, and the number of mismatches. The method does not require any preliminary alignment of motifs of the training set. The comparison of our method with Jaspar database searching tool and Match (TRANSFAC) reveals its better performance in both sensitivity and specificity. Most important, it allows to significantly reduce the number of false positive predictions.

Discussion

We suggest a simple and transparent not-PWM-based approach that is implemented in a tool called SiteTracker. Not requiring an alignment of the training set motifs, we avoid the uncertainties introduced by the alignment procedure, which is especially problematic for very degenerate motifs, and the following trimming, which can lead to a loss of potentially important sequence information.

Presenting Author

Ekaterina Shelest (ekaterina.shelest@hki-jena.de)
Hans Knoell Institute, HKI, Jena

Author Affiliations

Leibniz institute for Natural Product Research and Infection Biology, - Hans Knoell Institute, Jena

A-25. PICMI: mapping point variations on genomes

Le Pera L (1,), Marcatili P (1), Tramontano A (1)*

Several international collaborations and local projects are producing extensive catalogues of genomic variations which are supplementing existing collections such as the OMIM catalogue. This flood of data will keep increasing and, especially, it will be used by a wider user base, including also clinical researchers. Mapping the observed variations on a genome, identifying whether it affects a gene and - if so - whether it also affects different isoforms of the same gene, is not a straightforward task. To face this issue we developed a web server called PICMI.

Materials and Methods

Multiple nucleotidic and amino acidic variations can be used as input of the PICMI server, which finds for each of them (querying different releases of the Ensembl database) whether it lies in a gene and if it is in a non-coding region (upstream or downstream, in the 5' or 3' untranslated region, in a stop-codon or in a skipped-exon) or in a coding region of each transcript. In this last case, the mutation is mapped and classified for all the isoforms. Notably, when the input is an amino acidic mutation, the tool verifies if a corresponding single base mutation can be unambiguously inferred.

Results

The PICMI web server provides a user-friendly, extremely easy to use tool for mapping single site variations on a genome and its products, including alternative spliced isoforms. The server can map nucleotide variations, but also amino acidic mutations when the corresponding nucleotide substitution can be unambiguously identified.

Discussion

We believe that this easy-to-use tool can reveal to be very useful both to simplify the mapping of nucleotidic variations, for example it is set to get the SNPs from the 1000 genomes project as input, and, especially, to analyse a number of pathological and physiological variations at the nucleotide level when they are only available at the protein level, as it is often the case for reports recorded in OMIM and in SwissProt variations.

Presenting Author

Loredana Le Pera (loredanalepera@gmail.com)
Department of Biochemical Sciences, Sapienza University of Rome

Author Affiliations

1. Department of Biochemical Sciences, Sapienza University of Rome, P.le A. Moro, 5 - 00185 Rome

Acknowledgements

FIRB ITALBIONET

A-26. Detection of alternative splice isoforms in the human genome

Ezcurdia I (1,), del Pozo A (1), Rodriguez JM (2), Ashman K (3), Valencia A (1,2), Tress ML (1)*

Recent studies have estimated that almost all multi-exon human genes can produce at least two differently spliced mRNA transcripts by alternative splicing. Although multiple transcripts are strongly supported by mRNA and EST sequence evidence, the expression of multiple protein isoforms is less well studied. Proteomics technologies ought to provide conclusive evidence for alternative protein variants, but MS/MS experiments can only identify a fraction of the peptide ions present in protease digests and it remains a technical challenge to detect proteins that are expressed at low levels.

Materials and Methods

We have carried out a comprehensive re-mapping of experimental spectra to the annotated sequences from the Human proteome. We used as data all the relevant experiments stored in two huge proteomics data repositories, the GPM and PeptideAtlas. We reanalysed the mass spectra using X!Tandem and mapped the results to the HAVANA release 3C annotations.

Results

We confirmed that many human genes do indeed express multiple alternative protein isoforms that are stable enough and expressed in sufficient quantities to be detected in proteomics experiments. In addition to demonstrating the expression of alternatively spliced isoforms in the human proteome we have also been able to provide some insight into the functional role of alternative splicing in the cell.

Discussion

This work highlights the growing importance of proteomics in the validation of predicted proteins, especially as a complement to large-scale annotation efforts such as the ENCODE project. However, we also show that careful statistical testing is a necessary part of validating any proteomics results.

Presenting Author

Iakes Ezcurdia (iezkurdia@cnio.es)
Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO)

Author Affiliations

(1) Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C./ Melchor Fernandez Almagro, Madrid, Spain. (2) Spanish National Bioinformatics Institute (INB), Spanish National Cancer Research Centre (CNIO), C./ Melchor Fernandez Almagro, Madrid, Spain. (3) Proteomics Division, Spanish National Cancer Research Centre (CNIO), C./ Melchor Fernandez Almagro, 3, Madrid 28029, Spain.

A-27. Boundaries for short-read, low-coverage de novo assembly and resequencing

Schatz F (1,), Schimmler M (1)*

For resequencing and deNovo assembly the amount of input data is crucial. On the one hand, too little input data results in missing information about parts of the DNA sequence; and on the other hand, too much information makes the computations take too much time. This conflict is why this work is important for the sequencing community, because the results of this work can be used prior to the sequencing process to estimate the amount of data needed, which are input parameters for sequencing.

Materials and Methods

It is assumed that the positions of reads within the genome of a sequencing process are distributed equally over the genome. Although this assumption might not fit to all sequencing techniques, this simplification allows us to deduce manageable formulas for calculating the contig length distribution using only the parameters of sequence length, read count, and read length. From this general formula reflecting general knowledge about a possibly optimal assembly, a formula calculating N50 is deduced.

Results

Theoretical results have been shown to be the same as random data for fixed input parameters. Furthermore, these results have been compared to real data from the human genome. From this it was deduced that real data are much worse in terms of having a higher number of sections with no coverage. Finally, it is shown that the given formulas can estimate the expected contig length distribution very precisely. Even the calculated N50 value was shown to give good estimations of real data or, depending on the read data's quality, give an upper bound for a best-case assembly.

Discussion

A mathematical model to precalculate the contig length distribution for equally distributed reads has been proposed and been proven to be correct by simulation. It has been shown that good predictions even for real biological data can be made. Because read positions generally are not distributed equally, another approach is to take different distributions to fit real data even more exactly than the proposed model. Besides standard distribution functions an interesting point of research would be to try to define a distribution function by the inverse transformation method.

Presenting Author

Florian Schatz (mail@florianschatz.de)
Christian-Albrechts-Universität zu Kiel

Author Affiliations

(1) Christian-Albrechts-Universität zu Kiel, Department of Computer Science

A-28. TAPIR: a web server for the prediction of plant microRNA targets, including target mimics

Bonnet E (1,2), He Y (1,2,), Billiau K (1,2), Van de Peer Y (1,2)*

MicroRNAs (miRNAs) constitute a prominent class of small non-coding RNAs that regulate gene expression at the post-transcriptional level. MiRNA targets can be identified through base pair complementarity of the miRNA sequence to mRNA sequences. Recent reports have also revealed the existence in plants of a phenomenon called miRNA target mimicry, where duplexes with a large bulge around the cleavage site actually sequester miRNAs, inhibiting the miRNA activity. We have designed and implemented TAPIR, a novel web server dedicated to the prediction of plant miRNAs, including miRNA target mimics.

Materials and Methods

For the detection of miRNA:mRNA duplexes, we use two different previously published algorithms. The first is the classical FASTA local alignment program, which is very fast but cannot detect the duplexes having a lot of bulges and/or mismatches. The second algorithm is RNAhybrid, an algorithm for a precise detection of the miRNA:mRNA duplexes.

Results

We tested the sensitivity and specificity of the TAPIR web server using a refernce set of 102 experimentally verified plant miRNA-target pairs. We also compared the results to existing web servers performing the same task. The TAPIR web server is performing equally well or even slightly better than the other tools, and seems to have a more flexible and user-friendly interface. The other tools don't offer the ability to predict target mimics.

Discussion

The TAPIR web server is bringing new features compared to existing solutions. The ability to use two different search engines, the rich output results featuring a precise calculation of the free energy and free energy ratio, the possibility to look for target mimics and to use pre-computed results should make TAPIR a useful resource for the plant research community.

URL

http://bioinformatics.psb.ugent.be/webtools/tapir

Presenting Author

Eric Bonnet (eric.bonnet@psb.vib-ugent.be)
Ghent University

Author Affiliations

(1) Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium. (2) Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Gent, Belgium.

Acknowledgements

This work is supported by an Innovatie door Wetenschap en Technologie (IWT) grant for the Bioframe project and by an IUAP grant for the BioMaGNet project (ref. p6/25).

A-29. GRID/CPCA analysis of the SULTs superfamily: multivariate characterization of the most relevant structural and physicochemical differences

Rebollido-Rios R (1,2), Zamora-Rico I (3)*

Sulfation plays an important role in the second phase metabolism. In this work provide a multivariate characterization of the SULTs superfamily, in order to identify the most relevant structural differences that could be responsible for the selective biding affinity and could be used in the design of new potential ligands determinants for drug metabolism. To achieve this goal, we employed the GRID/CPCA computational procedure; which has shown to be an able tool to distinguish selectivity differences from the point of view of the receptor. Consequently, it is independent of the availability of appropriate ligands for a based QSAR analysis.

Materials and Methods

The methodology involves the following steps:(1)Obtain 3D structures of the target proteins.(2)Produce a superposition of the well known conserved regions, in this particular case the (TYPKSTTW) sequence motif, and also of the binding sites.(3) Get a multivariate characterization of these regions derived from the GRID force field.(4)Use CPCA to handle the multivariate data.(5)Graphical analysis and interpretations of the results of the CPCA.

Results

CPCA analysis of the SULT superfamily: In general, the CPCA plot reveals the different subfamilies of SULTs that were included in this work: SULT1, SULT2, SULT4. CPCA analysis interspecies (mSULTs and hSULTs): The replacement of a few amino acids that are involved in the ligand binding site can make a large difference between mouse and human SULTs. The aromatic residues Phe 81 y Phe 142 have a major role in the phenol recognition. CPCA analysis of hSULT1 and hSULT2 families: The CPCA procedure was able to distinguish between the isoforms of both families.

Discussion

The overall aim of this study was to compare the SULTs superfamily structurally. This was done in order to identify the most relevant structural differences that could be responsible for the selective binding affinity. Despite of mSULT and hSULT sharing some sequence identity, CPCA shows important structural differences regarding to some key amino acids that are involved in the ligand recognition. The presence of the substrate in the active site plays an important role in the determination of some structural differences between SULTs isoforms. We can conclude that it is a computational tool that allows distinguishing structural differences between these enzymes.

Presenting Author

Rocio Rebollido-Rios (rocio.rebollido-rios@stud.uni-due.de)
Dept. of Bioinformatics, Center of Medical Biotechnology.University of Duisburg-Essen

Author Affiliations

1.Department of Molecular Design and Synthesis. Higher Institute of Technologies and Applied Sciences. Ciudad de La Habana. CP 10600,AP 6163,Cuba. 2.Dept. of Bioinformatics.Center of Medical Biotechnology.University of Duisburg-Essen. 45117. Essen. Germany 3.Lead Molecular Design, S.L., Vallés 96-102 (27). 08190. San Cugat del Vallés. Municipal Institute for Medical Research. IMIM. Pompeu Fabra University. 08003. Barcelona. Spain.

A-30. Theoretical and empirical quality assessment of transcription factor binding motifs

Medina-Rivera A (1,3), Abreu-Goodger C (2), Thomas-Chollier M (4), Salgado-Osorio H (1), Collado-Vides J (1), van Helden J (1,3)*

Position-specific scoring matrices are routinely used to predict transcription factor (TF) binding sites in genome sequences. However, their reliability to predict novel binding sites can be far from optimum, due to the use of a small number of training sites or the inappropriate choice of parameters when building the matrix or when scanning sequences with it. Measures of matrix quality such as E-value and information content rely on theoretical models, and may fail in the context of full genome sequences.

Materials and Methods

The program matrix-quality combines theoretical and empirical score distributions to assess the predictive capability of position-specific matrices. The theoretical distribution provides an estimate of the false prediction rate. Empirical distributions indicate the enrichment of binding sites in various collections of sequences: known binding sites (positive control), all upstream regions of a genome, microarray clusters, ChIP-seq peaks. Negative controls are performed by analyzing the same sequence collections with column-permuted matrices.

Results

We applied the method to estimate the predictive capacity of matrices for bacterial, yeast and mouse transcription factors. The evaluation of 60 matrices from RegulonDB revealed some poorly predictive motifs, and allowed us to quantify the improvements obtained by applying multi-genome motif discovery. Interestingly, the results reveal differences between global and specific regulators. It also highlights the enrichment of binding sites in sequence sets obtained from high-throughput ChIP-chip (bacterial and yeast TFs), and ChIP-seq and experiments (mouse TFs).

Discussion

Users are often restricted by the available databases, which contain motifs of variable quality. In this context, the method presented here has many applications: (i) selecting reliable motifs before scanning sequences; (ii) improving motif collections in transcription factor databases; (iii) evaluating motifs discovered in massive datasets resulting from new sequencing technologies (ChIP-seq, ChIP-chip), to cite a few.

URL

http://rsat.ulb.ac.be/rsat/

Presenting Author

Alejandra E Medina Rivera (amedina@lcg.unam.mx)
Center for Genomic Sciences, Universidad Nacional Autonoma de Mexico

Author Affiliations

1. Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México. Av. Universidad s/n. Cuernavaca, Col. Chamilpa, Morelos 62210; Mexico. Email: amedina@lcg.unam.mx, heladia@ccg.unam.mx, collado@ccg.unam.mx 2. EMBL - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. Email: cei@ebi.ac.uk 3. Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe). Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe. B-1050 Bruxelles, Belgium. Email: Jacques.van.Helden@ulb.ac.be 4. Department of Computational Molecular Biology. Max Planck Institute for Molecular Genetics. Ihnestrasse 73. 14195 Berlin. Germany. Email: thomas-c@molgen.mpg.de

Acknowledgements

We acknowledge the members of the BiGRe laboratory for useful comments on the manuscript. Funding: A.M-R. was supported during her Ph.D. studies (Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México) by a fellowship from the Consejo Nacional de Ciencia y Tecnología (Mexico). C.A-G. was supported by a Sanger Institute Postdoctoral Fellowship. The BiGRe laboratory is supported by the BioSapiens Network of Excellence funded under the sixth Framework program of the European Communities (LSHG-CT-2003-503265), by the Belgian Program on

Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office, project P6/25 (BioMaGNet) and by the MICROME Collaborative Project funded by the European Commission within its FP7 Programme, under the thematic area "BIO-INFORMATICS - Microbial genomics and bio-informatics", contract number 222886-2.". Travel costs for A.M.-R. were partly supported by the Actions de Recherches Concertées de la Communauté Française de Belgique (ARC grant number 04/09-307), the Bureau des Relations Internationales et de Coopération (BRIC, Université Libre de Bruxelles) and UNAM. J.C.-V. acknowledges support by the National Institutes of Health, grant number R01 GM071962-05. MTC is supported by the Alexander von Humbold Stiftung.

A-31. Characterization and prediction of protein nucleolar localization sequences

*Scott MS (1, *), Boisvert F-M (2), McDowall MD (1), Lamond AI (2), Barton GJ (1)*

The nucleolus is the cellular site of ribosomal subunit assembly but is also known to be a key player in numerous other fundamental cellular activities. Although the nucleolar localization of proteins is often believed to be mediated primarily by non-specific retention to core nucleolar components, many examples of short nucleolar targeting sequences have been reported in recent years. These motifs are referred to as NoLSs (Nucleolar Localization Sequences).

Materials and Methods

To characterize NoLSs, we performed an extensive curation of the literature, uncovering 46 human NoLSs, and performed statistical analysis on these motifs. Then the sequence and predicted protein secondary structure of the curated NoLSs and a carefully generated negative set which included nuclear localization signals (NLSs) were used to train an artificial neural network for the prediction of NoLSs. The predictor was applied to the complete human proteome and ten of the highest scoring previously unknown NoLSs were experimentally confirmed by fusion to GFP and visualization by microscopy.

Results

Statistical analysis of the NoLSs revealed that 48% of their residues are basic, while 99% of their residues are predicted as solvent-accessible. These and other features were used to train an artificial neural network NoLS predictor. At a true positive rate of 54%, the predictor's overall false positive rate (FPR) was estimated to be 1.52%, which can be broken down to FPRs of 0.26%, 0.80% and 12% for randomly chosen cytoplasmic, nucleoplasmic and NLS sequences respectively. All previous unknown predicted NoLSs chosen for validation were experimentally confirmed as localizing to the nucleolus.

Discussion

NoLSs are a prevalent type of targeting motif that can be computationally predicted. We predict 19% of human proteins to encode NoLSs. Although recognized as different in the cell, the NoLS is a targeting signal that has a similar composition to NLSs and both signals are often confused or used interchangeably. This study should help define the differences between the two types signals. The proteome-wide prediction of NoLSs may be searched and downloaded from http://www.compbio.dundee.ac.uk/www-nod/.

URL

http://www.compbio.dundee.ac.uk/www-nod/

Presenting Author

Michelle S. Scott (michelle@compbio.dundee.ac.uk)
Division of Biological Chemistry and Drug Discovery , College of Life Sciences, University of Dundee

Author Affiliations

(1) Division of Biological Chemistry and Drug Discovery and (2) Wellcome Trust Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, United Kingdom

Acknowledgements

MSS is a recipient of a post-doctoral fellowship from the Caledonian Research Foundation.

A-32. Jalview: past, present, future

Procter JB (1,), Troshin PV (1), Martin MAM (1), Barton GJ (1)*

Sequence alignments are essential to comparative sequence analysis, and they provide a scaffold for highlighting variation. Jalview provides a light weight and flexible sequence alignment visualization environment, either as a web based or stand-alone system. However, new annotation capabilities are needed to better visualize quantitative annotation - such as amino acid profiles and positional scores derived from computational prediction and experimental observations. Furthermore, a new web service infrastructure is needed to enable users to exploit their own computational resources.

Materials and Methods

The Jalview visualization pipeline and datamodel were extended to support the display of alignments overlaid with quantitative sequence annotation. A new program execution engine and parameter specification model was developed in Java 1.6, which is able to manage local and distributed job execution. The engine provides a web service interface layer and its client was embedded in the Jalview desktop. Jalview's user interface was also extended to support the configuration of multiple web service installations.

Results

Jalview's new annotation visualization capabilities were evaluated with data retrieved from public servers via the Distributed Annotation System (DAS), and General Feature Format (GFF) files. The new interface enables regions to be shaded according to score or any associated text to better distinguish annotation of the same type. Several multiple sequence alignment programs were configured for execution using the new web service execution engine. Jalview users are able to access a public installation of the engine, or download them for local installation.

Discussion

The extensions to Jalview's alignment, visualization, and analysis capabilities that we described enable it to be a more useful system for biologists, and expert bioinformaticians. We wish to extend the repertoire of analysis types accessible via the new Jalview web services system, to include sequence database searches and automated alignment analysis and annotation tools. The new capabilities also introduce new levels of complexity for the user. Therefore, we are now developing new training materials to help biologists and bioinformaticians effectively exploit Jalview's new capabilities.

URL

http://www.jalview.org

Presenting Author

James B. Procter (j.procter@dundee.ac.uk)
University of Dundee

Author Affiliations

1. School of Life Sciences Research, College of Life Sciences, University of Dundee, UK.

Acknowledgements

Jalview’s core development is supported by the UK’s Biotechnology and Biological Sciences Research Council (BBSRC).

A-33. Enrichment of eukaryotic linear motifs in viral proteins

Pushker R (1,), Edwards R-J (2), Shields D-C (1)*

Eukaryotic Linear Motifs (ELMs) in proteins are short and linear motifs spanning 3 to 10 amino acids that may play an important role in virus-host interaction and subversion of host cell signaling. These ELMs have been found to be more prevalent in the disordered regions (lacking secondary or tertiary structure) of proteins. Since many viral genomes display a propensity for intrinsic disorder, we expect that there would be an enrichment of these ELMs in viruses. We were also interested to see whether disordered regions are more enriched with ELMs than ordered regions of viral proteins.

Materials and Methods

119 ELMs were searched in the ordered and disordered regions of 1,646 viruses consisting of 33,269 proteins. For each ELM, the number of occurrences in all proteins was recorded and compared with that of control ELMs; typically the reversed sequences. Chi-square test with Bonferroni correction ($p < 4.2E-4$; this threshold allows for multiple testing) was performed to test whether certain ELMs are over-represented or under-represented in viruses. Since the statistical test relies on non-independence of the data; we also searched for these ELMs in evolutionary Unrelated Protein Clusters (UPCs).

Results

We found that 23 ELMs were over-represented in the ordered region of viral proteins. Proprotein convertase subtilisin/kexin (PCSK) cleavage site motif, TRAF2 binding site motif and SH2 ligand binding motif were the most significantly over-represented ELMs. Integrin binding site motif was the most significant one among the 12 under-represented ELMs. After searching for ELMs in the UPCs, we found that the ordered regions of viral proteins were enriched with an ELM that binds to Proliferating Cell Nuclear Antigen (PCNA) protein and forms an alpha helix on binding.

Discussion

There were 14 ELMs significantly over-represented; coincidently, PCSK cleavage site motif, being the most significant one in the disordered regions of viral proteins. Analyzing the disordered regions of UPCs, we found that two PDZ ligand binding motifs were still significantly over-represented in these UPCs. Interestingly, the N-myristoylation site motif was found to be under-represented in the disordered region of UPCs. We conclude that there is an evidence for enrichment of PCNA binding motif in the ordered region and PDZ ligand binding motifs in the disorder region of viral proteins.

Presenting Author

Ravindra Pushker (pushker@ucd.ie)
University College Dublin, Ireland

Author Affiliations

1. UCD Complex and Adaptive Systems Laboratory; UCD Conway Institute of Biomolecular & Biomedical Research; School of Medicine and Medical Science, University College Dublin, Belfield, Dublin 4, Ireland. 2. School of Biological Sciences, University of Southampton, Southampton, UK.

Acknowledgements

This work was funded by Science Foundation Ireland and University College Dublin, Ireland. We also wish to acknowledge the SFI/HEA Irish Centre for High-End Computing (ICHEC) for providing computational facilities and support.

A-34. Analyzing ChIP-Seq data using qips

Gogol-Döring A (1,), Chen W (1)*

Chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-Seq) is a common method for the genome-wide analysis of protein-DNA interactions. Specific antibodies select the proteins of interest together with any piece of randomly fragmented DNA bound to them. The origin of the selected DNA fragments is then determined by sequencing and mapping to a reference genome; genomic regions binding to proteins feature an increased number of mapped sequencing reads.

Materials and Methods

qips computes a set I of nonoverlapping and high-scoring intervals which is optimal, i.e. for each alternative interval j exists an overlapping interval in I having at least the same score as j. The scores are derived from a statistical model that implies the mappability of genomic regions and, if available, the read distribution of a control data set. Our software also provides a new and accurate way for estimating the length distribution of single-end sequenced DNA-fragments which does not require distinctive peaks and is robust with respect to noise and mapping biases.

Results

Calculating p-Values for intervals of arbitrary length and without fixed starting positions makes our method more precise than alternative approaches.

Discussion

Our software generates several figures and statistical output valuable for evaluating the overall quality of ChIP-Seq data. It will be published free and open source.

Presenting Author

Andreas Gogol-Döring (andreas.doering@mdc-berlin.de)
Berlin Institute for Medical Systems Biology, MDC-Berlin

Author Affiliations

(1) Berlin Institute for Medical Systems Biology, MDC-Berlin

A-35. COMA server for protein distant homology search

Margelevičius M (*), Laganeckas M, Venclovas Č

The concept of homology (common evolutionary origin) is at the heart of most studies dealing with protein sequence, structure and function. In the absence of protein structure, inference of homology usually has to rely exclusively on sequence data. At present, most sensitive sequence-based methods use comparison of multiple sequence alignments represented as either Hidden Markov Models or sequence profiles.

Materials and Methods

With the goal of improving distant homology detection, we have recently developed a new method COMA, based on sequence profile-profile comparison (Margelevičius and Venclovas, 2010, BMC Bioinformatics 11, 89). The new method has at least two major features distinguishing it from other profile-profile comparison methods. They comprise a position-specific variable gap penalties, and a global score system leading to improved estimation of statistical significance of detected similarities. To make this method accessible for larger biological community, we have developed a web server.

Results

The server has a simple and intuitive user interface for setting up homology search jobs, and uses the latest version of the COMA method. The actual homology search is done by comparing query-based sequence profile against the selected profile database (SCOP, PDB, and PFAM). The structured output of the homology search is enriched with useful functionality. The user may extract the combined alignment between the query and any number of hits for further analysis. The server also provides a possibility to generate a 3D model of the query based on the corresponding alignment.

Discussion

To illustrate the utility of the COMA server for distant homology detection, we present newly discovered evolutionary links to the highly diverse PD-(D/E)XK nuclease superfamily. We also provide some examples where COMA performs better than a structure comparison method. The provided illustrations demonstrate the COMA server's utility in inferring structure and function of specific protein families, and non-trivial evolutionary relationships.

URL

http://bioinformatics.ibt.lt/coma/

Presenting Author

Mindaugas Margelevičius (minmar@ibt.lt)
Institute of Biotechnology

Author Affiliations

Institute of Biotechnology, Graiciuno 8, LT-02241 Vilnius, Lithuania

Acknowledgements

This work was supported by Lithuanian State Science and Studies Foundation and Howard Hughes Medical Institute.

A-36. Protein disorder and short motifs in disordered regions are enriched near the cytoplasmic side of single-pass transmembrane proteins

Stavropoulos I (1, 2), Khaldi N (1, 2), Davey NE (3), Finian M (2), Shields DC (1, 2)*

We investigated the distribution of predicted intrinsic disorder in human single-pass transmembrane proteins in relation to the plasma membrane. The proteins were also studied to identify potentially functional short linear motifs, by searching for regions in the protein sequences that are both intrinsically disordered and relatively conserved compared to adjacent residues in orthologous proteins. The distribution of the predicted motifs was also investigated in relation to the plasma membrane.

Materials and Methods

All 803 single-pass human transmembrane proteins were analyzed using IUPred to predict disorder. Gopher algorithm calculated the relative local conservation for every residue in the proteins and by combining information from disorder prediction identified regions of relative conservation in disordered regions. As disordered regions are devoid of structural constraints, conservation of residues relative to a background of random mutation is generally due to functional relevance. Using this method, the protein sequences were searched for statistically over conserved groupings of residues.

Results

Human single-pass proteins have higher mean disorder in their cytoplasmic segments than their extracellular parts. Since the region immediately adjacent to the membrane is less disordered, the disorder peaks at around 30 residues from the membrane. We noted a marked increase in the incidence of conserved motifs within the disordered regions at the same location, suggesting that this area may be extensively used in protein-protein interactions. We noted that this excess of conserved motifs is seen even after adjusting for the disorder level.

Discussion

We conclude that many transmembrane proteins have highly disordered cytoplasmic tails, and that these may play roles in cellular signalling.

Presenting Author

Ilias Stavropoulos (ilias.stavropoulos@ucd.ie)
University College Dublin

Author Affiliations

1-UCD Conway Institute of Biomolecular and Biomedical Research, School of Medicine and Medical Sciences, University College Dublin, Dublin 4, Republic of Ireland. 2-UCD Complex and Adaptive Systems Laboratory, University College Dublin, Dublin 4, Republic of Ireland. 3-European Molecular Biology Laboratory, Heidelberg, Germany.

Acknowledgements

This work was funded by the Irish Research Council for Science, Engineering & Technology (IRCSET) Graduate Education Research Programme (http://bioinformatics.ucd.ie/PhD).

A-37. Variation in positional preferences by 20 amino acids in alpha helices with different size

Fallahi H (1,), Dokanehiifard S-A(1), Hosseini- Nasab S-M-E (2)*

Many studies indicated that different amino acids have different preferences for taking part in alpha helix conformation. As a result of these studies, substitution matrices have been prepared to compare the proteins secondary and tertiary structures. Unfortunately, in many cases these matrixes are unable to produce reliable results due to the complexity of the conformation. In order to improve prediction of the secondary structures based on the peptide sequences, a more detailed look at these helices is inevitable. Here we present the result of dissecting different alpha helixes by position.

Materials and Methods

The amino acid sequences of proteins containing alpha helixes were obtained from PDB. Alpha helixes in these proteins were then categorized based on the number of amino acids into classes with 3 to 25 amino acids. The amino acid contents of each class and each position within the classes were then determined. Amino acid contents of the same position in different classes were also compared. Using PERL language programming all the scripts created and tested locally on a Mac operating system (Mac OSX ver.10.5.8).

Results

The results indicate differences in the preferences of different amino acids for each position. For example, while alanine did not show significant tendency toward any position, histidine were mostly present in positions 18 and afterwards. Besides, the same positions in different classes of helixes were occupied by different amino acids. As an example, position two in the 4-aa classes of helixes was mostly occupied by proline, followed by alanine. Interestingly, in the 5-aa helixes the same position was contained mostly glutamic acid followed by alanine and proline was the third.

Discussion

The result of the current study might be an indication of greater differences among helices with different number of amino acids. Therefore, one might need separate substitution matrices for each class for alignment and comparison studies. Currently used matrices are not effective, because AA substitutions are regarded to be the same in all helixes. These results might also imply alternative evolutionary route for different sized helixes. The longer helixes could be a result of merging the smaller one together in different combinations, rather than simple expansion of the smaller helices.

Presenting Author

Hossein Fallahi (fallahi.hossein@gmail.com)
Razi Unversity, Kermanshah, IRAN (I.R. of))

Author Affiliations

1. Dep. of Biology, School of Science, Razi University, Kermanshah, I.R. of Iran 2. Dep. of Statistics, Shahid Beheshti University, Evin, Tehran, I.R. of Iran * corresponding author

Acknowledgements

This work was partially supported by Razi University, through a research fund provided to Dr. Hossein Fallahi.

A-38. Comparison of mapping and variant calling accuracies for next-generation sequence data in relation to coverage depth: a case study in leukemia cell lines

Kalender Z (1,2,), Geerdens E (2), Cuppens H (3), Cools J (2), Aerts S (1)*

We found that combinations of different mapping and variant calling algorithms lead to significantly different variant predictions. Moreover, we have observed that most of the variant caller algorithms have a coverage depth threshold of 10 reads, yet there might be high quality reads with low coverages reflecting true variations, especially in targeted sequencing studies. Thus, we aimed to identify the most accurate combination of mapping and variation calling methods and to study how the performance of different methods varies with coverage depth.

Materials and Methods

We re-sequenced the exons of 97 genes in eight leukemia cell lines on the Roche 454 platform. Nine analysis pipelines are constructed using different combinations of existing mapping and variant algorithms. Sequence reads are aligned to the human reference genome using one of the four mapping algorithms (BLAT,BWA-SW, SSAHA2, and GsMapper). Then, single nucleotide variations are predicted using one of the four variant calling algorithms (SAMTools,VarScan,Atlas-SNP2, and GsMapper). Furthermore, a set of 192 predicted variations are sequenced with capillary sequencing to create a validation set.

Results

To assess how the coverage is affecting variant calling, separate variant calling is performed for low coverage (3 to 10 reads per base, LC) and high coverage (more than 10 reads per base, HC) regions. By comparing the sensitivity (SN) and specificity (SP) between pipelines, we found that different pipelines perform good in different coverages (HC:BWA-SW+SAMTools;SN=99%,SP=87% and LC:BLAT+Atlas-SNP2 and SSAHA2+Atlas-SNP2; SN=71%, SP=100%). Moreover, higher performance can be achieved using combinations of high specificity pipelines in both LC and HC (HC:SN=92%, SP=98% and LC:SN=79%, SP=100%).

Discussion

A new computational pipeline that combines different combinations of mapping and variant calling algorithms leads to more accurate and specific identification of genomic variants in cancer genomes. We show that it is possible to extract valuable information even in LC regions. We believe these results will contribute to the identification of high-confidence mutations in cancer genomes, and we are currently using the SNV predictions from our best pipeline to identify driver mutations in a cohort of re-sequenced leukemia cell lines and patient samples.

Presenting Author

Zeynep Kalender (zeynep.kalender@med.kuleuven.be)
Laboratory of Computational Biology, Center for Human Genetics, KULeuven

Author Affiliations

1 Laboratory of Computational Biology, Center for Human Genetics, KULeuven 2 Laboratory for Molecular Pathogenesis of Leukemia, Center for Human Genetics, KULeuven 3 Laboratory for Cytogenetics and Genome Research, Center for Human Genetics, KULeuven

Acknowledgements

This project was supported by the National Taskforce on Cancer from the Belgian government.

A-39. Monitoring strain diversity in metagenomics data using meta-MLST SNP profiling

De Jager VCL (2,4,5), Van der Sijde MR (1,2,3) Hazelwood LA (1,3) Kutahya O (1,3), Kleerebezem M (1,3,5) Smid EJ (1,3), Siezen RJ (1,2,3,4), Van Hijum SAFT (1,2,3,4)*

The diversity in many industrial and naturally occurring metagenomes (e.g. a complex starter culture) with a limited set of species is mostly represented by strain variants. Properties of such metagenomes may be explained by the presence of combinations of such strains. It is therefore it is important in these metagenomes to profile the diversity at the bacterial strain level.

Materials and Methods

The method allows following bacterial diversity in metagenomes at the strain level. It involves selecting "core" genes expected to be present in all strains of interest. Next, single nucleotide polymorphisms (SNPs) are determined that support strain classification. These SNPs are subsequently used to assign strain classes to indvidual reads or contigs obtained from metagenomic samples, similar to multi locus sequence typing (MLST). We apply our method to follow strain level diversity of Lactococcus lactis in multiple-timepoint metagenomics data obtained during the cheese making process.

Results

We obtained core clusters of genes from sequenced species expected to be present in the metagenome of interest and from sequenced individual isolated strains. SNPs were detected on these core genes and separating SNPs were selected for read classification. We were able to assign strain classes to individual reads and contigs from a multiple-timepoint metagenomics dataset obtaining known and new varieties.

Discussion

The ability to follow the presence of individual but closely related bacterial strains in metagenomes allows to gain insight in the dynamics of such populations. The combination of metagenome data and sequences of selected individual isolates allows detection of known classes and new strain variants.

Presenting Author

Victor C.L. de Jager (victor.de.jager@nbic.nl)
NBIC / CMBI / TIFN / WUR

Author Affiliations

De Jager V-C-L (2,4,*), Van der Sijde M-R (1,2,3) Hazelwood L-A (1,3) Kutahya O (1,3), Kleerebezem M (1,3,5) Smid E-J (1,3), Siezen R-J (1,2,3,4), Van Hijum S-A-F-T (1,2,3,4) 1) NIZO food research, P.O. Box 20, 6710 BA Ede, the Netherlands 2) Center for Molecular and Biomolecular Informatics, Nijmegen Center for Molecular Life Sciences, Radboud University Medical Centre, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands 3) TI Food and Nutrition, P.O. Box 557, 6700 AN Wageningen, the Netherlands 4) Netherlands Bioinformatics Centre, 260 NBIC, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands 5) Wageningen University, Laboratory of Microbiology, Dreijenplein 10, Buildingnumber 316, 6703 HB, Wageningen, The Netherlands

Acknowledgements

Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands, programme: BioRange 2.4

A-40. A variance stabilizing parametrization to detect differentially expressed genes in RNA-Seq data

Sales G (1,), Risso D (2), Chiogna M (2), Romualdi C (1)*

Several studies have shown that deep sequencing provides a new level of detail for the analysis of the transcriptome. Current protocols for the processing of RNA-Seq data, however, are affected by a transcript length bias in the calling of differentially expressed genes. This effect introduces a distortion in the ranking of differentially expressed genes and may affect adversely studies comparing the behavior of multiple genes (for example, in the context of pathway analysis).

Materials and Methods

The Poisson regression offers a simple model to deal with count data, such as that obtained from deep sequencing technologies. In the context of the Generalized Linear Model framework, the choice of the link function represents a crucial issue. Although the canonical link is often the default choice, other links can present favorable properties. Here, we explore the use of the square root as the variance stabilizing link for the Poisson regression. We propose a Wald type test (equivalent to the classical t-test within the Poisson model) to test for differentially expressed genes.

Results

By means of simulations and real data (from the MAQC-III project), we show that, in the Poisson model with the canonical link function, the variance of the estimators of the parameters strongly depends on the mean count level. By exploiting a variance stabilizing transformation of the parameters, this dependency is removed without affecting the biological interpretation.

Discussion

We propose a novel approach for the detection of differentially expressed genes in RNA-Seq data. In particular, we have observed that the transcript length dependency reported in other studies can be explained almost completely by the more fundamental link between the intercept and the variance of the estimators in the Poisson regression used to model short reads data. Building on this conclusion, we have developed a variance stabilizing parametrization compensating the bias and allowing the use of the standard t-test for the identification of differentially expressed genes.

Presenting Author

Gabriele Sales (gbrsales@gmail.com)
Department of Biology, University of Padua

Author Affiliations

1) Department of Biology, University of Padua, via U. Bassi 58/B, 35121 Padova, Italy. 2) Department of Statistical Sciences, University of Padua, via C. Battisti 241, 35121 Padova, Italy.

Acknowledgements

University of Padua.

A-41. DNA annotation induction: from RefGene on human chr. 22 to genome-wide CAGE for human and mouse

Bedo J (1,2,), Steininger A (1,3), Izhak Haviv (4,5), Adam Kowalczyk (1,2)*

Transcriptions start site (TSS) prediction on the human genome (using RefGene) has recently been extensively studied as a benchmark problem for such applications, and supervised learning techniques were demonstrated as a very robust approach that outperformed other competing approaches. We show that supervised techniques are capable of even more: they can extract novel knowledge from a small sample of annotated DNA and build predictive models capable of accurate annotation of the same or even another species, generalising beyond the apparent scope of the initial annotation.

Materials and Methods

Our approach is to segment the genome into small tiles assuming that some corresponding sets of binary labels are available for the tiles. A small subset of tiles (e.g., a single chromosome) is used to train a predictive model. The model is then applied to the rest of the genome. We used different genome annotations: RefGene genes, CAGE tags and also broad, unedited ChIP-Seq for Pol-II binding from the ENCODE project. Five different predictive model were used -- three developed for human and two for mouse -- with each tested on both mouse and human.

Results

We have shown that the knowledge of DNA sequences and functional annotation on the smallest autosomal human ch. (22) is sufficient to build robust genome-wide models for CAGE tag prediction on human and mouse. For human we detect the top 10^4 tags with precision ~95%. Similar performance was observed for four additional models: two human models from ChIP-Seq peaks for Pol-II binding and CAGE tags, and two mouse models from RefGene and CAGE annotations. We conclude that a significant fraction of TSS and CAGE tag locations are associated with very similar local DNA properties.

Discussion

We believe our algorithm provides a good baseline in-silico tool for extending empirical data obtained during phase I of ENCODE through to the rest of the genome, further to the TSS task we explored. Furthermore, our predicted TSS annotations merits consideration by the human ENCODE Genome Annotation Assessment Project (EGASP), and could improve our annotation of functional elements in the context of interpretation of genetic studies, such as genome wide disease-allelic associations.

Presenting Author

Justin Bedo (justin.bedo@nicta.com.au)
NICTA

Author Affiliations

1: NICTA, Victoria Research Laboratory 2: Dept. of Electrical and Electronic Engineering and 4: Dept. of Biochemistry and Molecular Biology of The University of Melbourne VIC 3010, Australia 3: University of Applied Sciences Technikum Wien, Vienna, Austria 5: The Alfred Medical Research and Education Precinct, Baker Medical Research

Acknowledgements

JB and AK acknowledge the support of NICTA in conducting this research. NICTA is funded by the Australian Government's Department of Communications, Information Technology and the Arts, the Australian Research Council through Backing Australia's Ability, and the ICT Centre of Excellence programs.

A-42. A new resampling method

*Bakis Y (1, *), Sezerman OU (2)*

Increase in molecular data has caused diversity of methods in analysing this data. Among them, some recently introduced approaches does not require an alignment procedure. However, it wouldn't be possible to test the robustness of the resulting phylogenies. In this work, we propose a new resampling method to use with those programs.

Materials and Methods

We have chosen a biological way, mutation, to create resamples from the original sequence. A mutation can be performed by three different ways; insertion, deletion and substitution. Mutations were performed at random positions and with random rates. A resampling is result of random mutations on certain percent of original sequences. We have used Relative Complexity Measure method to construct the phylogenies and TreeDist routine of PHYLIP 3.65 to calculate tree distances. We have used molecular sequences from ITS region of DNA to test for.

Results

We have performed our method on different set of molecular sequences. Results showed that proposed resampling method produced reasonable resamples at 10% mutation level for ITS region of DNA. The resulting resamples were produced very diverse phylogenies, some have very distant topology while some others were very close to the original topology.

Discussion

Proposed method has produced resamples successfully. Since all other methods are based on aligned sequences, presence of a method which does not require multiple alignment is crucial. The method is also unique by its way of creating resamples by means of biological mutations.

Presenting Author

Yasin Bakis (bakis_y@ibu.edu.tr)
Abant Izzet Baysal University

Author Affiliations

1 Department of Biology, Abant Izzet Baysal University, Bolu, 14280 Turkey 2 Biological Sciences and Bioengineering, Sabanci University, Tuzla, Istanbul, 34956 Turkey

A-43. Clustering massive sequence databases

Vu D (1,), Robert V (1,2)*

The amount of genomic data grows significantly. Currently there are more than 6 millions fungal sequences in GENBANK databases. Unfortunately, many sequences have been submitted wrongly or without any indication on the species name. Thus there is a need for tools to classify sequences. Existing tools such as UPGMA require heavy computations as they aim to construct phylogenetic trees. Thus, it is not feasible to use them for clustering very large databases. We aim to develop a clustering tool that can group sequences from the same species which is is able to handle large numbers of sequences.

Materials and Methods

We use a greedy algorithm as the starting point of our approach. For each sequence, we find its cluster by searching for the nearest matching sequence in a sequence database obtained from the existing clusters. If the similarity between the sequence and its nearest match is sufficient, the sequence is placed in the cluster of its match. Otherwise a new cluster is created for the sequence. To determine a sufficient similarity, we created a database of identified sequences to learn this number. It is the similarity that gives the closest clustering to the species of the given sequences.

Results

We have implemented the algorithm in BioloMICS, a software solution for biological data management. We have experimented with more than 5 millions fungal sequences from GENBANK databases. The sufficient similarity of sequences in the same cluster for the most used regions such as ITS, 26S has been predicted. Preliminary experiments of clustering show promising results as a number of ITS and 26S fungal sequences have been detected as incorrectly identified.

Discussion

The advantage of our approach is that it is time efficient since we do not need to take into account the relationship between organisms. Another approach, which is similar to ours, is the tool called UCLUST by C. Edgar. However, in our algorithm, a cluster can have more than one representative sequences, and therefore, it respects a transitive closure of sequences with respect to similarity. Moreover, we use BLAST searching for local matches of a sequence, while UCLUST searches for global matches. BLAST may find less accurate matches, but it is more time efficient than other global alignments.

Presenting Author

Duong Vu (d.vu@cbs.knaw.nl)
Bioinformatics group, CBS-KNAW fungal biodiversity centre

Author Affiliations

(1) Bioinformatics group, CBS-KNAW fungal biodiversity centre, Utrecht, The Netherlands (2) BioAware, www.bio-aware.com

A-44. WordCluster: detecting clusters of DNA words and genomic elements

Barturen G, Oliver JL, Hackenberg M*

Many k-mers (DNA words) and genomic elements are known to be clustered in the genome. Well established examples are the genes, TFBSs, CpG dinucleotides and ultra-conserved non-coding regions. Currently, no algorithm exists to find these clusters in a statistically comprehensible way. The detection of clustering often relies on densities and sliding window approaches or arbitrarily chosen distance thresholds.

Materials and Methods

The algorithm is based on entity distances and an assigned p value. Two different types of input data can be supplied: 1) a group of k-mers and a genomic sequence to be scanned by the program (user supplied or chosen among the 19 genome assemblies available in our database); and 2) a file in BED format with the coordinates of the genomic elements whose clustering properties should be analyzed. The output includes a basic statistics of the clusters, the co-localization with gene annotation, and GO enrichment/depletion analysis for the genes overlapped by the predicted clusters.

Results

WordCluster detects clusters of DNA words (k-mers) or genomic elements, based on the distance between neighboring copies and an assigned p-value. The method was implemented into a web server which also determines the co-localization with gene annotations. We demonstrate the usefulness of this approach by detecting the clusters of CAG/CTG (cytosine contexts that can be methylated), showing that the degree of methylation varies drastically inside/outside of the clusters. As a second example, we search for statistically significant clusters of olfactory receptor (OR) genes in the human genome.

Discussion

WordCluster generalizes the previous CpGcluster algorithm to any word or genomic element in the genome, at the same time associating a statistical significance to the detected clusters. It outperforms current methods relying on densities and sliding window approaches or arbitrarily chosen distance thresholds. The implementation as a web server connected to online databases allows for co-localization studies with different gene regions, as well as for genome wide enrichment/depletion analysis in GO terms of the predicted clusters, which may facilitate relating them to biological function.

URL

http://bioinfo2.ugr.es/wordCluster/wordCluster.php

Presenting Author

Guillermo G.B. Barturen (bartg01@gmail.com)
University of Granada

Author Affiliations

Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, 18071-Granada & Lab. de Bioinformática, Centro de Investigación Biomédica, PTS, 18100-Granada, Spain

Acknowledgements

Spanish Government Grant No. BIO2008-01353 to JLO. Spanish 'Juan de la Cierva' grant to MH and Basque Country 'Programa de formación de investigadores del Departamento de Educación, Universidades e Investigación' grant to GB.

A-45. Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs

Seroussi E (1,), Glick G (1), Shirak A (1), Yakobson E (1), Weller JI (1), Ezra E (2), Zeron Y (3)*

Copy number variation (CNV) has been recently identified in human and other mammalian genomes and there is a growing awareness of CNV's potential as a major source for heritable variation in complex traits. Genomic selection is a newly developed tool for the estimation of breeding values for quantitative traits through the use of genome-wide genotyping of SNPs. The Illumina BovineSNP50 BeadChip has been used worldwide to genotype over 30,000 Holstein cattle. On this chip, 54,001 SNPs are represented (~SNP/50,000 bp), and some of them fall within CNV regions.

Materials and Methods

We used the BeadChip data obtained for 912 Israeli bulls to investigate the effects of CNV on SNP calls.For each of the SNPs, we estimated the frequencies of occurrence of loss of heterozygosity (LOH) and of gain based either on deviation from the expected Hardy-Weinberg equilibrium (HWE) or on signal intensity (SI) using the PennCNV “detect” option.

Results

Correlations between LOH/CNV frequencies predicted by the two methods were low (up to r = 0.08). Nevertheless, 418 locations displayed significantly high frequencies by both methods. Efficiency of designating large genomic clusters of olfactory receptors as CNVs was 29%. Frequency values for copy loss were distinguishable in non-autosomal regions, indicating misplacement of a region in the current BTA7 map. Analysis of BTA18 placed important net merit QTLs in regions rich in segmental duplications and CNVs. Enrichment of transporters in CNV loci suggested their potential effect on milk-production traits.

Discussion

Expansion of HWE and PennCNV analyses allowed estimating LOH/CNV frequencies, and combining the two methods yielded more sensitive detection of inherited CNVs and better estimation of their possible effects on cattle genetics. Although this approach was more effective than methodologies previously applied in cattle, it has severe limitations. Thus the number of CNVs reported here for the Holstein breed may represent as little as one-tenth of inherited common structural variation.

Presenting Author

Eyal Seroussi (seroussi@agri.huji.ac.il)
The Agricultural Research Organization (ARO), Volcani Center

Author Affiliations

1 Institute of Animal Sciences, ARO, The Volcani Center, Bet Dagan 50250, Israel 2 Israel Cattle Breeders Association, Caesaria Industrial Park 38900, Israel 3 Sion, AI Institute, Shikmim 79800, Israel

Acknowledgements

This research was supported by grants from the Israel Milk Marketing Board; the European Sixth Research and Technological Development Framework Programme, Proposal No. 016250-2 SABRE, and Research Grant Award No. IS-4201-09 from BARD, The United States - Israel Binational Agricultural Research and Development Fund.

A-46. The role of the Occludin MARVEL domain in intracellular targeting and clustering

Yeheskel A (1,), Yaffe Y (2), Hirschberg K (2), Pasmanik-Chor M (1)*

Tight junctions (TJs) are complex proteinatious integral membrane assemblies indispensable for epithelial function. Occludin is a MARVEL domain containing protein, found in all TJs and is associated with regulating it functions. The MARVEL domain contains 4 transmembrane (TM) segments and its function is largely unclear. We propose that this lipid membrane interacting motif is associated with a number of central functions such as targeting to the tight junctions and clustering. The purpose of this study is to gain insight into Ocln MARVEL function using protein sequence analysis.

Materials and Methods

94 SwissProt MARVEL homologues were obtained from PFAM. Alignment was performed by MAFFT and phylogeny tree was built by RaxML. The TM boundaries of Ocln were determined using secondary structure prediction, conservation and hydrophobicity profile. Consensus prediction of TM helices was used to produce helical wheel presentation of each helix, using Textopo. Bi-molecular fluorescent complementation analysis was applied to study Ocln oligomerization. Site directed mutagenesis of conserved aromatic residues was applied. For cell microscopy, a fluorescent protein-tagged Ocln-MARVEL motif was used

Results

Phylogenetic tree of the MARVEL family illustrates distinct subgroups, with Ocln as one of them having unique properties. The first extracellular loop contains a glycine-tyrosine rich motif. The second extracellular loop has a distinctive secondary structure. Live cell confocal microscopy shows that Ocln-MARVEL is correctly sorted to the basolateral membrane and TJs. Alanine substitutions had little effect on the surface expression of Ocln-MARVEL. However 4 residue mutants resulted in reduced oligomerization, increased tendency to aggregate and slowed down secretory transport.

Discussion

We propose a mechanism for the binding of 2 Ocln molecules from opposing membranes of adjacent cells using the first extracellular loop, previously suggested to contribute to protein flexibility and organization. The second extracellular loop was predicted to have a distinctive secondary structure which may contribute to paracellular transport and cell-cell adhesion. Site directed mutagenesis showed a role for conserved aromatic amino acids in cluster formation. These data shed light on the potential role of protein–lipid interactions of the Ocln MARVEL domain and membrane in TJ functions.

URL

http://www.tau.ac.il/lifesci/bioinformatics.html

Presenting Author

Adva Yeheskel (suezadva@tauex.tau.ac.il)
Bioinformatics Unit, G.S.W. Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Author Affiliations

1. Bioinformatics Unit, G.S.W. Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel 2.Department of Pathology, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

A-47. Multi-Harmony: detecting functional specificity from sequence alignment

Feenstra KA (*), Brandt BW, Heringa J

Many protein families contain sub-families with functional specialization, such as binding different ligands or being involved in different protein-protein interactions. A small number of amino acids generally determine functional specificity. The identification of these residues can aid the understanding of protein function and help finding targets for experimental analysis.

Materials and Methods

Here, we present multi-Harmony, an interactive web sever for detecting sub-type-specific sites in proteins starting from a multiple sequence alignment. Combining our Sequence Harmony (SH) and multi-Relief (mR) methods in one web server allows simultaneous analysis and comparison of specificity residues; furthermore, both methods have been significantly improved and extended. SH has been extended to cope with more than two sub-groups. mR has been changed from a sampling implementation to a deterministic one, making it more consistent and user friendly. For both methods Z-scores are reported.

Results

The Sequence Harmony and multi-Relief methods are benchmarked on a test-set comprising 23 protein families with known specificity sites and functional sub-grouping. Overall performance is shown to be comparable to or better than the best of alternative available methods, such as SDPpred, ProteinKeys, PROUST-II and Xdet. Performance at highest precision continues up to higher recall compared to all other methods tested.

Discussion

The multi-Harmony web server provides state-of-the art specificity detection, and is easily accessible to non-expert users. An example script for remote access for expert users is available on request. The webserver produces a dynamic output page, which includes interactive connections to the Jalview and Jmol applets, thereby allowing interactive analysis of the results. Multi-Harmony is available at http://www.ibi.vu.nl/programs/shmrwww.

URL

http://www.ibi.vu.nl/programs/shmrwww

Presenting Author

K. Anton Feenstra (feenstra@few.vu.nl)
IBIVU / Free University

Author Affiliations

Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, De Boelelaan 1081A, 1081HV Amsterdam, The Netherlands

Acknowledgements

ENFIN, a Network of Excellence funded by the European Commission within its FP6 Programme, under the thematic area 'Life sciences, genomics and biotechnology for health' (LSHG-CT-2005-518254).

A-48. Fast approximate statistical ranking for sequence motif discovery under higher order Markov background model

Shida K (*)

Sequence motif discovery is still a difficult problem in bioinformatics. The motif candidates found by motif discovery programs are routinely ranked by their p-values to ensure that the most statistically significant answer is presented. However, accurate p-value evaluation is a complicated and time consuming computation and there is considerable difficulty in taking higher order Markov background models into account precisely.

Materials and Methods

For a motif discovery problem under k-th order Markov background, we propose (1) a novel polynomial approximation scheme for the motif p-value (2) a MCMC method to optimize the coefficients of the polynomial to maximize the accuracy of the resultant p-value and (3) a system to validate the obtained approximation using a motif discovery problem on random binary sequences as an example. The optimization of polynomial parameters by means of MCMC sampling is surely time consuming but reusable: such a sampling is necessary only once for a given condition.

Results

First, the numerical precision of the p-value estimated by this method is merely acceptable. However, unlike any previously proposed method, this method is applicable for higher order (k=>2) Markov background without excessive loss of precision nor catastrophic increase of computational complexity. Second, because this method is just a calculation of polynomial, the approximate evaluation of p-value is almost comparable to some of typical motif score functions, such as the relative entropy, in speed. Third, a prototype of this method is successfully tested on a small Human dataset.

Discussion

The previous methods of motif p-value estimation are focused on near-exact precision for limited level of correlation in the background requiring large computational complexity. On the contrary, this new method can handle increased level of background correlation with minimum sacrifice of p-value precision. The computational burden is increased only in the pre-processing stage. Some other problems in bioinformatics can be heavily affected by background correlations that are hidden or neglected. This method can be extended to analyze and cancel the effect of such correlations.

Presenting Author

Kazuhito Shida (shida@imr.edu)
Institute for materials research, Tohoku university

Author Affiliations

Institute for materials research, Tohoku university

A-49. Networking interacting residues, evolutionary conservation and energetics in protein structures

Vidotto M (1), Martin AJM (1), Boscariol F (1), Walsh I (1), Tosatto SCE (1,)*

Recently, there has been a growing interest in representing protein structures as so-called residue interaction networks (RINs). RINs consider single amino acids in the protein structure as nodes and connections as physico-chemical interactions, such as covalent bonds and non-covalent contacts. The intuitive idea is to analyze a structure with the same approach as a protein interaction network in order to investigate whether the same rules apply. RINs have already been used to analyze protein stability and folding, allosteric communication, enzyme catalysis or mutation effect prediction.

Materials and Methods

Here, we present RING (URL: http://protein.bio.unipd.it/ring/) as a novel tool to generate RINs for use in Cytoscape. The tools was conceived to yield a simple, intuitive representation that is also physico-chemically meaningful, i.e. distinguish salt bridges, hydrogen bonds and aromatic interactions from generic van-der-Waals contacts. Structural features are generated for each node. Protein sequence conservation, determined by PSI-BLAST, and local conformational energy preferences to the best of our knowledge represent two novel features for RINs.

Results

The web server provides two user interfaces. A simple user interface with meaningful default values for most parameters and a more complex interface where the user can specify how the RIN parameters. The output file contains the network definition files for Cytoscape. Online help pages are provided with extensive documentation. The generated RINs contain information regarding different interaction types as well as secondary structure and energetics of each residue. Once generated, the RIN can be easily visualized in Cytoscape and network properties analyzed with available plugins.

Discussion

A special feature of RING is to generate meaningful sub-networks. RING allows the user to limit the generated network to buried residues, conserved residues or both. Intuitively, limiting RIN analysis to a network of conserved residues will help to focus on the essential interactions for a protein fold. In summary, RING is a novel web tool for use with Cytoscape designed for the analysis of protein structures in terms of physico-chemical interactions, evolutionary conservation and energetics while taking advantage of the powerful network paradigm.

URL

http://protein.bio.unipd.it/ring/

Presenting Author

Silvio C.E. Tosatto (silvio.tosatto@unipd.it)
Dept. of Biology, University of Padua

Author Affiliations

1 - Dept. of Biology, University of Padua, Italy

A-50. An iterative scheme for making long structural alignments of ncRNAs using FOLDALIGN

Havgaard J H (*), Gorodkin J

Structured non-coding RNAs (ncRNAs) are often more conserved at the structural level than in the primary sequence. Sequence based alignments of ncRNAs are therefore often inaccurate. Specialized approaches, such as FOLDALIGN, which also takes the structures of the molecules into account are needed to obtain the proper alignment. However, due to the high run times of such methods heuristics are needed to lower the time and memory requirements to make the algorithms practical.

Materials and Methods

The FOLDALIGN algorithm (Sankoff based) makes pairwise alignments of RNA sequences using both sequence and structure information. Pre-aligned positions or whole subalignments are introduced as seed constraints to FOLDALIGN. The algorithm builds longer alignments from the shorter seed constraints. At each step in the iteration a subset of the generated (length constrained) alignments are used as seeds which are extended into longer alignments. A subset of these alignments are then used as seeds in the next step of the iteration.

Results

We present putative results of test runs. These preliminary results indicates that the use of seed constraints makes it possible to make global structural alignments of ribosomal sequences on a modern computer, and thus promise to make de novo search for similar size ncRNAs. The seed constraints also result in lower run time for local alignments. We are currently working a more systematic benchmark. An effort has also been made to improve the statistical evaluation of the local alignments.

Discussion

Tools which simultaneously can fold and align RNAs have been shown to be an essential component in do novo searches for ncRNAs. However, a main limitation of these methods are their time and memory requirements. With the implementation of seed constraints FOLDALIGN can now make alignments of size 1500 - 2000 nt. compared to the previous 300 - 500 nt. This makes it possible to find not only the small but also some of the larger ncRNAs. With the improvement in the evaluation of the significans of the alignments the selection of candidates for experimental verification have become more feasible.

URL

http://foldalign.ku.dk

Presenting Author

Jakob H. Havgaard (hull@genome.ku.dk)
University of Copenhagen

Author Affiliations

Center for non-coding RNA in Technology and Health, Genetics and Bioinformatics IBHV, Faculty of Life Sciences
University of Copenhagen

Acknowledgements

The Danish Council for Independent Research (Technology and Production Sciences) The Danish Council for Strategic Research (Strategic growth Technologies) The Danish Center for Scientific Computing

A-51. A multi-objective approach to the prediction of sRNAs in bacteria

Arnedo J (1), Romero-Zaliz R (1), del Val C (1)*

Bacterial small non-coding RNAs (sRNAs) are recognized as novel widespread regulators of gene expression. There are several algorithms available for their ab initio prediction, but all of them present intrinsic limitations. We present in this work a methodology, which uses a multi-objective approach approach to extract the best methods' aggregations by maximizing the specificity and sensitivity of the program´s individual predictions

Materials and Methods

All potential aggregations, methods', form a space of potential hypotheses, which can be represented as a lattice structure. We search for the best methods' aggregations, moving from hypothesis to hypothesis towards the most general, the union of all methods, and the most specific, their intersection, which are located at the top and the bottom of the lattice, respectively. Salmonella typhimurium LT2 was used reference genome, and the algorithms: zMFold, eQRNA, RNAz, Alifoldz, Dynalign and MSARi.

Results

Results show a major improvement in specificity and sensitivity when our methodology is compared to the performance of individual methods. The use of methods' aggregations increase either prediction's sensitivity or specificity, and a few aggregations increase both objectives when compared with the individual methods. The best results are obtained in aggregations containing 2 or 3 methods via either union or intersection.

Discussion

The here proposed methodology is an automatic method generator, and a step forward to exploit all already existing methods, by providing optimal methods' aggregations to answer concrete queries for a certain biological problem with a maximized accuracy of the prediction. As more approaches are integrated for each of the presented problems, de novo accuracy can be expected to improve further.

Presenting Author

Coral del Val (delval@decsai.ugr.es)
University of Granada

Author Affiliations

Dpt. of Computer Science and Artificial Inteligence. University of Granada, Spain

Acknowledgements

This work was supported by the Consejeria de Innovacion, Investigacion y Ciencia de la Junta de Andalucia under project TIC-02788

A-52. Scalability of large-scale protein domain family inference

Rezvoy C (1,), Vivien F (2), Kahn D (3)*

The ProDom database is a repository of protein domain families inferred automatically from homologies between protein sequences of the Uniprot database. Since 1999, ProDom has been built using MkDom2, a sequential algorithm of quadratic complexity. Because of its sequential nature, MkDom2 cannot keep up with the exponential increase of Uniprot over the years, to the point that it is no longer possible to envision a new release of ProDom with this sequential program. given past records, running MkDom2 on a recent release of Uniprot would have taken over 10 years.

Materials and Methods

Mpi_MkDom2 is a master-worker greedy algorithm. At each iteration, the shortest sequences are considered potential domains and used as queries to recruit homologous domains using PSIBLAST. If two sequences share an homology in an all-against-all BLAST, only the shortest is used as query. New families are validated by checking if they do not overlap with each other. Valid families are removed from the database before the next iteration. To assess the performance of the algorithm, MPI_MkDom2 was tested with nested databases (27MB and 210MB) and different numbers of computing nodes.

Results

Experiments show that MPI_MkDom2 is able to achieve a reasonable speedup on a small database (27 MB), but starts loosing efficiency above 40 workers. A larger database, however, allows a better load balancing. MPI_MkDom2 is thus able to efficiently use more workers when processing larger databases. While allowing to substantially decrease the construction time of ProDom, the new algorithm still creates results that are consistent with those of the original algorithm: on average domain families created by MkDom2 are more than 90% similar to those created by MPI_MkDom2.

Discussion

The distributed algorithm presented here is able to provide reasonable speed-ups while retaining the structure of the protein domain families built buy the sequential algorithm. Even when achieving maximal speedup the clustering stays consistent with the sequential result. The speedup being dependent of the size of the database processed, this new algorithm will provide a mean to efficiently construct new releases of ProDom while staying consistent with the original sequential heuristic.

URL

http://prodom.prabi.fr

Presenting Author

Clément Rezvoy (Clement.Rezvoy@ens-lyon.fr)
Éns Lyon

Author Affiliations

´ 1 - ENS Lyon, Universite de Lyon, LIP, UMR 5668, ENS Lyon - CNRS - INRIA - UCBL, Lyon, FRANCE 2 - Université de Lyon, Laboratoire de Biometrie et Biologie Evolutive, UMR 5558, UCBL - CNRS, INRA, Villeurbanne, FRANCE 3 - INRIA, Universite de Lyon, LIP, UMR 5668, ENS Lyon - CNRS - INRIA - UCBL, Lyon, FRANCE

Acknowledgements

The ProDom project was supported by the FP6 EMBRACE Network of Excellence, the FP7 IMPACT Research Infrastructure programme and the France-Israel Research Network Program in Bioinformatics. Experiments presented in this paper were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr). This work was granted access to the HPC resources of CINES under the allocation 2010-c2010076425 made by GENCI (Grand Equipement National de Calcul Intensif).

A-53. Evolutionary study of eye-developmental gene expression across multiple Drosophila species by high-throughput tag sequencing

Naval-Sanchez M (*), Aerts S

Retinal differentiation is initiated in the eye imaginal disc during third instar larval development. During this process the genes eyeless and atonal activate a cascade of signaling events that end in retina formation. Drosophila eye-development is a well-studied system but information about gene expression and gene regulatory interactions remains sparse. Here, we aim to gain further insight into the gene regulatory network underlying retinal differentiation by determining conserved and divergent patterns of gene expression in various Drosophila species using next-generation sequencing.

Materials and Methods

RNA was extracted from eye-antennal imaginal discs and wing imaginal discs (controls) during third instar larval stage from three species, namely D.melanogaster, D.yakuba and D.virilis. Tag libraries for digital gene expression were created using the NlaIII restriction enzyme and were sequenced on the Illumina GAII platform in six separate lanes. Reads were mapped against their reference genome using bowtie, considering only uniquely mapped tags. Gene expression levels were normalized as tags per million (TPM). Gene set and functional enrichment analyses were performed using GSEA.

Results

First, we show that DGE yields meaningful expression values in D.melanogaster through a comparison with microarray data on the same tissue. Next, we confirm that gene expression divergence increases over evolutionary time, being larger for Dmel-Dvir than for Dmel-Dyak. Finally, we identify genes with conserved expression in the eye disc. Surprisingly, only a small fraction of eye specific genes in Dmel is also eye specific in the other species (30% and 17% overlap with Dyak and Dvir respectively). These conserved genes are mostly involved in photoreceptor cell fate specification.

Discussion

To our knowledge this is the first time that digital gene expression is used to measure tissue-specific gene expression in multiple Drosophila species. We conclude that DGE-based gene expression measurements are accurate and appropriate for determining gene expression levels for species without a microarray platform. Finally, we are able to find expression conservation and divergence patterns across Drosophila species allowing us to investigate further the evolutionary differences in the mechanisms governing gene expression.

Presenting Author

Marina Naval Sanchez (Marina.NavalSanchez@cme.vib-kuleuven.be)
Laboratory of Computational Biology, Center for Human Genetics, KULeuven

Author Affiliations

Laboratory of Computational Biology, Center for Human Genetics, KULeuven.

A-54. Higher order repeat with largest primary repeat unit (~2.4 kb) and located within a gene in human genome

Glunčić M (1,), Paar V (1), Paar P (2), Rosandić M (1), Vlahović I (1)*

Using our novel computational method Global Repeat Map (GRM), based on Key String Algorithm (KSA), which is particularly convenient for identification and analysis of repeats and higher order repeats (HORs) characterized by very long repeat units, we performed a case study of Build 37.1 genomic ensemble of Y chromosome, addressing the following questions: (i) Is there a HOR in chromosome Y in addition to the known HORs based on 5-bp and alphoid 171-bp primary repeat units? (ii) Is there a HOR with very large basic repeat unit far beyond the size of all those found so far in human genome? (iii) Is there a HOR fully contained within a gene for which no evidence was reported so far?

Materials and Methods

The Key String Algorithm (KSA) framework is based on the use of a freely chosen short sequence of nucleotides, a key string, which cuts a genomic sequence at each location of the key string. The lengths of ensuing KSA fragments form KSA length array. Any periodicity appearing in the KSA length array enables identification and location of repeat in a given sequence. Analysis of repeat sequences at position of any periodicity gives consensus repeat unit. Any presence of higher order periodicity in the KSA length array reveals the presence of HOR at that position. The GRM algorithm is a graphical extension of KSA to an ensemble of all key strings of the same length and superposition of their segmentation results.

Results

We discovered in Y chromosome a new Higher Order Repeat (HOR), based on ~2.4 kb primary repeat units. We classified an ensemble of ~2.4 kb monomers into five highly homologous monomer families denoted m01, m02, m03, m04, and m05. Three monomer families, m02, m03, and m04 form the basic repeat units of 2mer HOR copies and 3mer HOR copies which have the largest primary repeat units discovered so far in the whole human genome. We found that each of four tandem sequences involving HOR copies is contained within one of the genes DAZ1 – DAZ4. Highly homologous simple repeats of m05 monomer family, are organized into two distinct tandem arrays; they are wholly contained within the DAZ2 and DAZ4 gene, respectively.

Discussion

Using our novel computational method GRM we discovered in Y chromosome a new HOR with largest known basic repeat unit, and wholly contained within genes. That HOR may shed a new light at a unique role of Y chromosome. Here we open the question on a possible role of HOR structure within genes, and in particular if the basic repeat unit is very large. It is therefore of interest to search for another examples of HOR structure within genes. The use of Global Repeat Map is especially suited for bioinformatics identification and analysis of repeats with very large monomeric repeat units and for the case of periodic sequencing of several families of basic monomeric repeat units as in the case of higher order repeats.

URL

http://www.hazu.hr/KSA/app/tools.html

Presenting Author

Matko Glunčić (matko@phy.hr)
University of Zagreb

Author Affiliations

(1) Faculty of Science, University of Zagreb, Zagreb, Croatia (2) Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

A-55. Fast and accurate digging for binding motifs in ChIP-Seq data using ChIPMunk software

Kulakovskiy I V (1,), Boeva VA (1,2,3,4,5), Favorov AV (1,6), Makeev VJ (1,7)*

The task of identification of transcription factor binding motifs in a limited number of short DNA sequences has a long history. Recently upcoming ChIP-Seq data provided a new challenge for motif discovery. Such data consist of thousands of sequences where a short overrepresented motif is to be found. Fortunately, in the case of ChIP-Seq data one has additional information, which helps to select the correct signal. This information is the coverage profile constructed for DNA fragments obtained from ChIP-Seq experiments.

Materials and Methods

We present a ChIP-Seq 'peak mode' of our efficient motif discovery tool "ChIPMunk". It is an expectation-maximization-with-bootstrapping algorithm which uses coverage profile information by setting a specific weight profile over each sequence to represent a preference for the sequence position to overlap with a binding site.

Results

We have tested our algorithm on different ChIP-Seq data sets, including those for NRSF, GABP and for the oncogenic protein EWS-FLI1. We successfully identified the correct motifs without traditional strict truncation of large enriched regions. This was possible because of the coverage profile information used as the motif positional preference prior. We show that ChIPMunk motif recognition quality is the same or better than that of the traditional (MEME, Multiple EM for Motif Elicitation) or ChIP-Seq-oriented (HMS, Hybrid Motif Sampler) tools while the speed is dramatically better.

Discussion

ChIPMunk can be effectively used to analyze large sequence sets and therefore is a helpful tool for prediction of transcription factor binding motifs in ChIP-Seq data. ChIPMunk is based on Java. The source code is freely available on the web.

URL

http://line.imb.ac.ru/ChIPMunk/

Presenting Author

Ivan V. Kulakovskiy (ivan.kulakovskiy@gmail.com)
Research Institute for Genetics and Selection of Industrial Microorganisms

Author Affiliations

(1) Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow, 117545 Russia (2) Institut Curie, 26 rue d'Ulm, Paris, F-75248 France (3) INSERM, U900, Paris, F-75248 France (4) INSERM, U830, Paris, F-75248 France (5) Mines ParisTech, Fontainebleau, F-77300 France (6) Johns Hopkins University, Baltimore MD 21205, US (7) Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991 Russia

Acknowledgements

This work was supported by RFBR#07-04-01623 and #07-04-01584 and RSASI#02.740.11.5008.

A-56. Mebitoo: a sequence analysis toolbox framework

Spaniol C (1,), Helms V*

Since we want to deploy our methods for public use and instead of implementing a variety of independent services, like online services, we want to speed up both the analysis and development workflow providing an incorporated framework that allows for: * module development independent from the core framework * storage of arbitrary data, based on module-individual XML concepts * a task processing schedule that each concurrently processes any set of data with any set of methods incremental analysis workflows, i.e. one method can make use of previously obtained results, gained by other methods.

Materials and Methods

Mebitoo is a platform-independent desktop application written in Java an based on the Netbeans Rich Client Platform. The software allows to import sequences persistently into an incorporated database engine, namely HSQL. A combined storage concept based on XML-files enables software extensions – plugins – to archive information in customized data structures. Automated data processing using those extensions can be invoked using a task execution interface, which enables to queue multiple operations and process multiple datasets in parallel.

Results

We started to incorporate our own existing prediction methods TMX and MINS as well as upcoming tools to predict helix cooperativitiy, and the HMM-based Beta-Barrel TMX (BTMX). Moreover, we started to adapt the MEME framework for motif search as a plugin. We aim to release the software including these tools combined with utilities like PFAM-based sequence alignments and visualization with TopoView and JMol.

Discussion

Mebitoo is intended to be framework software which offers researchers to develop new methods easily without getting lost in details of data input and output or GUI representation. Novice users may implement based on the interface we provide while advanced users are able to use the entire Netbeans RCP and customize their own data storage. Because of the task concept, researchers can import and automatically process large datasets easily with different methods by queuing several jobs.

Presenting Author

Christian Spaniol (christian.spaniol@bioinformatik.uni-saarland.de)
Center for Bioinformatics, Saarland University

Author Affiliations

(1) Center for Bioinformatics, Saarland University

A-57. Practical aspects of RNA-Seq data analysis that you should know and they don't tell you

Garcia F (1,), Tarazona S (1), Zuñiga S (2), Santoyo J (1), Dopazo J (1), Conesa A (1)*

RNA-Seq has recently emerged as a powerful tool for the analysis of the transcriptome as it is able to uncover novel transcriptionally active genome regions, and can characterize with an unprecendeted detail the complexity of the alternatively spliced transcriptional landscape. Unfortunately, the analysis of these data is complex and includes many "tricky" points. In this poster we report and comment some practical aspects concerningr RNA-seq data analysis that are not usually found in manuals/documentations, but should be taken into account when anlysing high-throughput count data.

Materials and Methods

We made use of published RNA-Seq data generated with each of the three current NGS platforms (Roche, Illumina, and SOLiD). We considered sequencing data of different properties (read length, pair-ends, strand-specificity, etc.), and applied state-of-the-art software (Tophat, Cufflinks, BFAST, AB WT, etc.), and in-house scripts for the analysis of the data.

Results

In general, we found that available documentation was inssuficient to clearly understand the effect of parameters in the analysis results. Moreover, provided examples are nearly impossible to reproduce. Also data quality is an issue. Some algorithms ignore it at all or are accompained by example data with strong biases. Finally we have detected problems with the scalability of the algorithms and different bugs. These include failure in processing parameter values, miss-match position misplacing resulting in false discoveries and significance divergences when using or not reference genome data.

Discussion

RNA-Seq technologies are extremely powerful for transcriptome studies. However, analysis technologies for this kind of data are still at their infancy and users should use them critically. As more datasets become available, algorithms will gain robustness and quality standards for RNA-Seq analysis will be established.

Presenting Author

Fernando Garcia (fdgarcia@cipf.es)
Bioinformatics Department, Centro de Investigacion Principe Felipe

Author Affiliations

(1) Bioinformatics Department, Centro de Investigacion Principe Felipe, Valencia, Spain (2) Bioinformatics Department, Sistemas Genomicos, Valencia, Spain

Acknowledgements

MICINN Projects BIO2009-10799 y BIO2008-05266-E

A-58. MutaBase: a framework and web-interface for the archival, management and interpretation of single nucleotide variants obtained by next generation sequencing

Sifrim A(1,), Van Houdt J (2), Vermeesch J (2), Moreau Y (1)*

Next generation sequencing allows the detection of novel variant sites which are elusive in array-based methods of genotyping using known common variants. As technology develops the throughput increases and the management and interpretation of tens of thousands of variants per sample becomes the new experimental bottleneck. We developed a flexible framework with an easy-to-use interface which allows the archival of variants for every sample and which integrates existing knowledge and prediction software for functional impacts of variants.This allows for the rapid interpretation of NGS results

Materials and Methods

Our framework is based on a Ruby on Rails system which includes parsing capabilities of most common single nucleotide variant callers. This system is linked to a local Ensembl database to gather exisiting knowledge about the given variants and genes. We use Polyphen2, FoldX, Waltz, Tango and other software to computationally predict the impact of variants on the phenotype. In order to make the task feasible in a reasonable time frame we provide an easy-to-setup parallelized and distributed task management system. The web-interface provides a multi-level complex query system for the archive.

Results

We present the application of MutaBase in a case-study of full-exome sequencing of 4 patients with Nicolaides-Baraitser syndrome. Doing this we predict the genetic consequences and known information of each variant and predict their possible functional impacts. By doing gene-centered cross-patients queries and taking into account existing knowledge, automatically retrieved by MutaBase, we can postulate candidate genes for downstream validation.

Discussion

Due to the sheer magnitude of the datasets produced by NGS experiments, managing and interpreting variant lists becomes a cumbersome and time-consuming task. As more and more data is generated, interesting complex cross-sample questions become more relevant but are hard to achieve. With MutaBase, we aimed to deliver a tool which handles such data conveniently and easily and which provides a front-end for the biologist or physician to quickly gain new insights from the data. In the future we plan to broaden the scope of MutaBase further by integrating more knowledge and prediction software.

URL

http://homes.esat.kuleuven.be/~asifrim/mutabase

Presenting Author

Alejandro Sifrim (alejandro.sifrim@esat.kuleuven.be)
Katholieke Universiteit Leuven

Author Affiliations

(1) Bioinformatics Group, Department of Electrical Engineering (ESAT-SCD), University of Leuven, Belgium. (2) Center for Human Genetics, University Hospital Gasthuisberg Leuven, Belgium.

A-59. Mining cleavage sites of the mouse peptidome

De Grave K (1,), Guns T (1), Vandekerckhove TTM (2), Landuyt B (3), Menschaert G (2), Van Criekinge W (2), Schoofs L (3), Luyten W (4)*

Recently, advances in peptidomics technology [1] have permitted identification from MALDI- or ESI-TOF/TOF mass spectra of over 100 novel peptides in tissues or body fluids of mice. The precursor proteins of all observed peptides could be identified. This enables data analysis of the cleavage sites, with the prospect of obtaining clues about novel cleavage enzymes or mechanisms. [1] K. Boonen, B. Landuyt, G. Baggerman, S.J. Husson, J. Huybrechts, L. Schoofs. Peptidomics: the integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis. J Sep Sci. 2008.

Materials and Methods

The problem of finding patterns in the cleavage locations is formulated as an item set mining task and solved using the CP4IM constraint-based itemset miner [2]. Ranking and pattern redundancy is controlled using the statistical methods described in Mining Informative Non-redundant Itemsets (MINI) [3]. We also apply established bioinformatics tools such as sequence logos. [2] L. De Raedt, T. Guns and S. Nijssen. Constraint programming for itemset mining. In ACM SIGKDD, 2008. [3] A. Gallo, T. De Bie and N. Cristianini. MINI: Mining informative non-redundant itemsets. In PKDD, 2007.

Results

[Preliminary analysis.] The cleavage sites are significantly different from the background model (independent and identically distributed amino acids according to the abundance in the mouse proteome). There is also a clear difference between previously known peptides and the ones observed by mass spectrometry during the project. Some patterns found confirm secondary cleavage documented in the literature, such as C-terminal dibasic KR-elimination. We are in the process of evaluating a conjectured leucine pattern.

Discussion

While distant (>2AA from the cleavage point) positions contain hardly any information individually, longer combinatorial patterns can be found that remain significant after conservative Bonferroni correction, revealing potential recognition sites for endopeptidases.

URL

http://www.peptidomics.be and *http://dtai.cs.kuleuven.be/CP4IM/*

Presenting Author

Kurt De Grave (kurt.degrave@cs.kuleuven.be)
K.U.Leuven, DTAI

Author Affiliations

(1) DTAI, Department of Computer Science, Faculty of Engineering, K.U.Leuven, Leuven, Belgium (2) Department of Molecular Biotechnology, Faculty of Bioscience Engineering, Laboratory for Bioinformatics and Computational Genomics, Ghent University, Ghent, Belgium (3) Biology Department, K.U.Leuven, Leuven, Belgium (4) Biomedical Sciences Group, K.U.Leuven, Leuven, Belgium

Acknowledgements

Kurt De Grave is funded by GOA/08/008 "Probabilistic Logic Learning". Tias Guns is funded by the Agency for Innovation by Science and Technology in Flanders (IWT). The project was supported by IWT grant SBO 50164 to L. Schoofs.

A-60. Properties of plant microRNAs affect miRDeep statistical scoring as well as accuracy

Thakur V (1,), Wanchana S (1), Xu M (2, Bruskiewich R (1), Mosig A (2), Zhu X (2)*

There are several tools which can identify miRNAs from deep-sequencing, however only a few of them like miRDeep, which can identify novel ones and also being available as standalone application. Given the difference between plant and animal miRNAs, particularly in terms of distribution of precursor length and complementarity, it's likely that the underlying prediction measures get affected. We studied the key prediction measures which get affected for plant specific miRNA prediction, in particular, those employed by miRDeep, and also attempted to estimate the new set of parameters.

Materials and Methods

The said effect on measures like minimum free energy, stability of secondary structures, excision length was examined, and the parameters of those, which displayed sizable changes, were estimated for plant specific miRNAs. Further, the fine-tuned miRDeep was tested on a Illumina dataset from maize leaves and compared the prediction with those obtained using the default parameters. The targets of these candidates was also predicted followed by their functional enrichment analysis.

Results

We found majority of the measures to take new set of values/distributions for plant specific miRNAs. While nucleus or seed region was relatively longer, the difference between distribution of minimum free energy was marginal. We identified 43 miRNA candidates; 13 of them were apparently novel, not showing any homology with any of the known plant miRNAs. The target prediction and its enrichment analysis, however, showed mixed results

Discussion

The present study characterizes the effect of precursor length on MFE, and several other measures, which provide insight into their use in plant miRNA prediction. Further, researchers applying miRDeep in plants, can improve their results by using the new set of parameters.

Presenting Author

Vivek Thakur (v.thakur@irri.org)
International Rice Research Institute

Author Affiliations

1. International Rice Research Institute (IRRI), Los Banos, Philippines. 2. MPG-CAS Partner Institute of Computational Biology (PICB), 320 Yue Yang Road, Shanghai, China.

Acknowledgements

This work has been done under C4 Rice project funded by Bill and Melinda Gates Foundation.

A-61. Using the Amazon elastic cloud computing resource for the analysis of next-generation sequencing data in Ensembl

Vogel J (1,), Aken B(1), Chiang G-T(1), Clapham P(1), Coates G(1), Fairley S(1), Hourlier T(1), Ruffier M(1), Tang A(1), White S(1), Zadissa A(1), Searle S(1), Hubbard T(1)*

Ensembl is one of the leading sources of genome sequence annotation data for a variety of different genomes. Ensembl genome annotation is currently available for more than 50 species, ranging from high-coverage reference genomes to genomes with highly fragmented, low-coverage assemblies. With the arrival of the Next-Generation transcriptome sequencing, a novel, valuable data source has become available to further extend our ongoing effort in annotating genome sequences. This new data type provides an unprecedented view of the whole transcriptome for coding- and non-coding transcripts.

Materials and Methods

Such data is available for different tissues, cell compartments and different developmental stages. The analysis of transcript expression levels and alternative splice forms can now be integrated into our gene annotations. The large amount of available data produced by NGS platforms sets new challenges to the development of bioinformatic tools and requires the access to a powerful compute infrastructure for a short time to process, map and analyze the sequence data. A cloud computing environment like the Amazon EC2 offers an affordable way to analyze large scale data sets.

Results

With the development of our Ensembl RNA sequence analysis pipeline (publication in preparation), we are now to able to handle and analyze large NGS based transcriptome data sets. This enables us to quantify short-read support for exons and introns of existing Ensembl annotation and, depending on the NGS input data, to produce complete de-novo gene sets for different tissues. Here, we give a detailed overview of how various Ensembl analysis modules can be configured and run in the Amazon EC2 environment using Sun GridEngine as a Job Submission System.

Discussion

We provide a practical example of how to use our pre-configured Amazon machine images to process and analyze NGS data with the Ensembl RNA sequence pipeline. With this infrastructure in place, the adaptation of further standard Ensembl analyses to the EC2 environment is facilitated. Additionally, we show how the Distributed Annotation System DAS can be used to display the obtained results in Ensembl. The analysis of NGS transcriptome data in the Amazon cloud with the Ensembl analysis pipeline offers a modular solution to analyse large-scale data sets without in-house compute infrastructures.

URL

http://www.ensembl.org

Presenting Author

Jan H Vogel (jan.vogel@gmail.com)
Wellcome Trust Sanger Institute

Author Affiliations

(1) Wellcome Trust Sanger Institute (2) European Bioinformatics Institute

A-62. Automated sequence extraction of relevant genomic features for targeted resequencing

De Wilde B (1,), D'Hont B (1), Lefever S (1), Hellemans J (1,2), Speleman F (1), Pattyn F (1), Vandesompele J (1,2)*

Massively parallel sequencing technologies enable researchers to determine the nucleotide sequence of DNA at ever-increasing throughput and decreasing cost. Focusing this sequencing power to a genomic region of interest to perform targeted resequencing is an important challenge in genetics. We created a web tool to assist in extracting sequences from relevant genomic features (e.g. coding exons) in large regions typically defined by by linkage analysis, homozygosity mapping or arrayCGH.

Materials and Methods

The tool is based on locally stored annotation tables originating from the genome databases of NCBI, UCSC, Ensembl, and other specialized sources of annotation information. The database content is managed by automated update scripts to allow version tracking and re-formatting of the data into a common pre-defined annotation format. This enables easy comparison and merging of the tables resulting in an integrated resource of high quality genome annotation information. Depending on the downstream application a reference table can be built containing only the required annotation information.

Results

The reference tables form the basis for an easy to use interface for genomic target selection. This tool allows a scientist without bioinformatics skills to select annotation information for a large genomic region based upon its characteristic features (e.g. protein coding, pseudogene or a non-coding element; a transcribed, untranscribed or translated part of an element; etc.). The annotated information can be visualized in a genome browser and outputted to a format ready for DNA chip or PCR based target selection.

Discussion

We present an easy to use web tool for genomic target selection based upon the most complete and uniform source of annotation information currently available. The tool itself has a focus on usability and should allow biologists to easily prepare their next generation sequencing experiments. We present results on the practical application of this tool for the selection of highly dispersed short non coding RNA genes for PCR based target selection as well as Nimblegen chip based capture of regions of several megabases in size for some mendelean human disorders.

URL

http://www.mellfire.ugent.be/NGSF

Presenting Author

Bram De Wilde (bram.dewilde@ugent.be)
Center for Medical Genetics

Author Affiliations

(1) Center for Medical Genetics, Ghent University, Ghent, Belgium (2) NXTGNT, Ghent University, Ghent, Belgium

Acknowledgements

Research Foundation - Flanders (FWO) Special Research Fund Ghent University (BOF-UGent)

AUTHOR INDEX