COMPARATIVE GENOMICS, PHYLOGENY AND EVOLUTION

Chairs: Martijn Huynen and Yves Van de Peer

Comparative Genomics, Phylogeny and Evolution 1

B-1. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates 3

B-2. A tool for comparison of complete proteomes between pairs of organisms 4

B-3. Toward an unified measure of intraspecific selective pressure 5

B-4. Assessment of genetic structure within and among Normandie (Northern France) populations of wild beet (Beta vulgaris sps. maritima Arcang.) 6

B-5. Trees inside trees for genome wide association studies 7

B-6. Validation of single cell arrayCGH on a 60-mer oligo microarray platform for preimplantation genetic diagnosis 8

B-7. Polymorphisms associated with mtDNA of elite Kenyan endurance athletes 10

B-8. Bayesian inference of community average gene copy numbers and its application for the metagenomic characterization of bacterioplankton community types in the Sargasso Sea 11

B-9. Phylogenetic mapping of non-model organism RNA-seq reads using a graph algorithm 12

B-10. GEI-DB: a database of atypical genomic elements in bacteria 13

B-11. The CNS twilight zone: limitations in comparing upstream regions in plants 14

B-12. TurboOrtho: a high performance alternative to OrthoMCL 15

B-13. i-ADHoRe 3.0 Detection of collinearity in large scale datasets 16

B-14. Comparative expression analysis of orthologous genes between Arabidopsis and rice 17

B-15. Phylogenomics and robust construction of prokaryotic evolutionary trees 18

B-16. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin 19

B-17. Towards molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data 20

B-18. Orthologous proteins associated with yeast telomeric complex identified by synteny and sequence similarity 21

B-19. Estimating average heterozygosity using multilocus dominant marker data: a maximum likelihood approach via the EM algorithm. 22

B-20. PLAZA 2.0 : a resource for plant comparative genomics 23

B-21. Branch testing in phylogenetic trees: comparison of computationally efficient methods 24

B-22. Parsimonious higher-order hidden Markov models for enhanced comparative genomics 25

B-23. Correction of bootstrap confidence levels using an iterated bootstrap procedure with computational efficient methods 26
B-24. Molecular evolution, structure and functional divergence of Lipoxygenase gene family in vertebrate 27
B-25. Cassis: detection of genomic rearrangement breakpoints 28
B-26. Methylated cytosines are less likely to mutate within CpG islands 29
B-27. Entropy approach reveals new features of genes and genomes 30
B-28. The hidden duplication past of the plant pathogen Phytophthora and its consequences for infection 31
B-29. Protein linkages between African Trypanosoma and 8 pathogenetic eukaryotic organisms 32
B-30. Reconstructing phylogenetic trees from clustering trees 33
B-31. Predictive modeling of psychrophilic adaptation on the proteome sequence level 34
B-32. Microbial phenotype prediction based on protein domain profiles 35
B-33. New insights into the metazoan evolution of cadherins: from basal to modern 36
B-34. Comparative mapping of transcription factor binding sites in plant genomes 37
B-35. Protein model selection with ProtTest increases phylogenetic performance 38
B-36. cn.FARMS: a probabilistic latent variable model to detect copy number variations 39
B-37. Genome-wide heterogeneity of the substitution process 40
B-38. Comparative microarray analysis to elucidate TF-networks in activated T-cells 41
B-39. Gene-trait matching analysis of Lactobacillus plantarum strains 42
B-40. ProtTest-HPC: fast selection of best-fit models of protein evolution 43
B-41. NTRFinder: an algorithm to find nested tandem repeats 44
B-42. Computational epigenomics of plant SNF2 genes 45
B-43. Comprehensive analysis of splice site evolution in primates using whole genome alignments and RNA-Seq data 46
Author Index 47

B-1. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates

Sela N (1,), Kim E (2), Ast G (2)*

Transposable elements (TEs) have played an important role in the diversification and enrichment of mammalian transcriptomes through various mechanisms such as exonization and intronization (the birth of new exons/introns from previously intronic/exonic sequences, respectively), and insertion into first and last exons. However, no extensive analysis has compared the effects of TEs on the transcriptomes of mammalian, non-mammalian vertebrates and invertebrates.

Materials and Methods

We determined the TE fraction within intronic sequences using the UCSC genome browser and GALAXY. Introns of chicken (G. gallus, Build 1.1), zebrafish (D. rerio, release Zv4), C. elegans (Release 2003) and D. melanogaster (Build 4.1) were extracted from the Exon-Intron Database (http://hsc.utoledo.edu/depts/bioinfo/database.html). When alternatively spliced isoforms of the same gene were present, only the first annotated isoform was extracted; all other isoforms were excluded in order to avoid redundancy. The analysis of the TE content was done using RepeatMasker software and rep

Results

We analyzed the influence of TEs on the transcriptomes of five species, three invertebrates and two non-mammalian vertebrates. Compared to previously analyzed mammals, there were lower levels of TE introduction into introns, significantly lower numbers of exonizations originating from TEs and a lower percentage of TE insertion within the first and last exons. Although the transcriptomes of vertebrates exhibit a significant level of exonizations of TEs, only anecdotal cases were found in invertebrates. In vertebrates, as in mammals, the exonized TEs are mostly alternatively spliced, indicating

Discussion

Exonization of TEs is wide-spread in mammals, less so in non-mammalian vertebrates, and very low in invertebrates. We assume that the exonization process depends on the length of introns. Vertebrates, unlike invertebrates, are characterized by long introns and short internal exons. Our results suggest that there is a direct link between the length of introns and exonization of TEs and that this process became more prevalent following the appearance of mammals.

Presenting Author

Noa Sela (noa.sela5@gmail.com)
Ludwig-maximilans university Munich

Author Affiliations

1. Department Biology I, Ludwig-Maximilians-University Munich (LMU) Großhaderner Str. 2, D-82152, Planegg-Martinsried, Germany 2. 1Department of Human Molecular Genetics, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

Acknowledgements

This work was supported by the Cooperation Program in Cancer Research of the Deutsches Krebsforschungszentrum (DKFZ) and Israel's Ministry of Science and Technology (MOST) and by a grant from the Israel Science Foundation (40/05), ICRF, DIP and EURASNET. NS is supported by the LMU excellence fellowship.

B-2. A tool for comparison of complete proteomes between pairs of organisms

*Karathia H (1, *), Alves R (1)*

Assignment of protein function to newly discovered or sequenced proteins based on the similarity of their sequence to that of other proteins of well-known function is a critical and fundamental problem in molecular system biology, comparative genomics and molecular biology. Although there are many tools that permit doing such an assignment on a protein by protein basis, very few tools are available to compare full proteomes between organisms.

Materials and Methods

We developed a tool that has been developed to perform: (a) a full comparison of the proteomes between two organisms (b) discrimination between orthologs and in-paralogs in each of the organisms being compared (c) full comparison of the functional pathways and processes in the two organisms to a third model organism in order to assess how similar or different specific physiological responses for the same biological processes may be between the organisms. The tool is implemented in PERL and graphical representations are generated in Mathematica.

Results

As a case study, we apply the tool to compare the proteome of 59 eukaryotic organisms with fully sequenced genomes to that of Saccharomyces cerevisiae. We identify different gene duplication and deletion events between each organism and the yeast. We use the results to predict how good a model organism is S. cerevisiae for the study of different biological processes and pathways in other eukaryotes. We rank both organisms and processes/pathways as to their probable closeness to those in S. cerevisiae.

Discussion

Our analysis and results are supported by the fact that fungi organisms are predicted to be the closest to S. cerevisiae. Candida albicans in particular is an exceptional species in fungi, because it has the highest proportion of duplicated proteins relative to S. cerevisiae. Our tool accurately classifies the duplications and deletions of genes between pairs of organisms. It also appears to be a good predictor of qualitative differences in the adaptive responses of corresponding biological processes in different organisms.

Presenting Author

Hiren Karathia (hiren@cmb.udl.cat)
Departament Ciències Mèdiques Bàsiques, Universitat de Lleida and IRBLleida

Author Affiliations

Departament Ciències Mèdiques Bàsiques, Universitat de Lleida and IRBLleida, Montserrat Roig 2, 25008 Lleida, Spain (1)

Acknowledgements

This work was partially funded by a Ramon y Cajal Research Award and by grant BFU2007-62772/BMC from the Spanish Ministry of Science and Innovation to RA and by a Generalitat de Catalunya Ph. D. fellowship to HK.

B-3. Toward an unified measure of intraspecific selective pressure

Amato R (1,2,), Miele G (1,2), Pinelli M (1,3), Cocozza S (1,3)*

During recent decades the study of human evolution has been of increasing interest, also due to the large amount of data now available. In the mean time, new applications of evolutionary biology to medical problems are being discovered at an accelerating rate. Several estimators for the selective pressure have been proposed. Being introduced to face different aspects of selective pressure, each measure has its own pros and cons.

Materials and Methods

We focused on the most widely used intraspecific estimators of selective pressure. In particular, to cope with both inter- and intra-group phenomena, we analysed the fixation index (FST) and some measures based on the extended haplotype homozygosity (namely REHH, iHS, XP-EHH). We assessed the performances of each measure on simulated data produced using Fregene, a tool developed and calibrated to reproduce, in a biologically sound manner, our evolutionary history, also taking into account for complex demographic, selection and recombination scenarios.

Results

By using GRID facilities to produce a large enough amount of data, we were able to exhaustively assess the performances of each estimator with regard to several biological parameters of the sites under selection (e.g. selection coefficient, dominance model, age of selection).

Discussion

The study of human evolution is obtaining increasing interest also for its potential medical applications and until now several measure of selective pressure have been proposed. Our analysis highlighted a complementary behaviour of these estimators suggesting a possible strategy of merging them in an unified and versatile measure of selective pressure.

Presenting Author

Roberto Amato (roamato@na.infn.it)
Dipartimento di Scienze Fisiche - Universita' degli Studi di Napoli "Federico II"

Author Affiliations

(1) Gruppo Interdipartimentale di Bioinformatica e Biologia Computazionale, Università di Napoli "Federico II" & Università di Salerno (2) Dipartimento di Scienze Fisiche, Università di Napoli "Federico II" - INFN Sezione di Napoli (3) Dipartimento di Biologia e Patologia Cellulare e Molecolare "L. Califano", Università di Napoli "Federico II"

B-4. Assessment of genetic structure within and among Normandie (Northern France) populations of wild beet (Beta vulgaris sps. maritima Arcang.)

Tran HT (1,), Cuguen J (2), Touzet P (2), Saumitou-Maprade P (2)*

The examination of coastal northern France (Normandie) beet populations is of particular concern for competent authorities for the regulation of transgenic organisms (GMO) due to the close proximity of wild beet to sugar beet fields. Gene flow from cultivated beets in this area is theoretically possible, if fields near wild beet habitats contain vernalized sugar beets or annual weed types. For purpose of in situ conservation of this genetic resource . For crop improvement, it is very important to examine genetic diversity of sea beet population in Normandie where still is a little information about this.

Materials and Methods

Young leafs gathered from from 396 plants in 9 populations coastal northern France. Nuclear microsatellite loci (codominant marker)used to analysis genetic variations.

Results

The global inbreeding coefficient Fit is 0.3336 , meaning that overall there is evidence for quite substantial inbreeding. We see that 17% of genetic variation is found among populations and 83% within popupulations. The Fst/(1-Fst) ratio for pairs of populations increased no linearly with the natural logarithm of the geographical distance. (Mantel test: P = 0.0005, R 2 = 0.0099), showing a pattern of no isolation by distance (Rousset 1997). There are two groups of populations widely separated: one group of three populations, fec4, veu5 and val6; the other with all 6 other populations. Within the latter group, mars-08 and pou9 are very closely related to each other.

Discussion

Information of biological characteristics from this study is usseful for consevation management such as the level of inbreeding, the population bottlenecks and the variance of gene dispersal distances.

URL

http://gepv.univ-lille1.fr/english/perso_pages_en

Presenting Author

Hoa Thi Tran (tranthihoa@agi.vaas.vn)
Institute of Agricultural Genetics

Author Affiliations

1*.Institute of Agricultural Genetics, Laboratory for Forest Genetics and Conservation, Hanoi 10000, Vietnam 2.Université des Sciences et Technologies de Lille I, Laboratoire de Génétique et Evolution des Populations Végétales, UPRESA 8016 du CNRS, Bâtiment SN2, 59655 Villeneuve d'Ascq Cedex, France

Acknowledgements

AUPEL-UREF- Bourses d’excellence

B-5. Trees inside trees for genome wide association studies

Botta V (1,), Geurts P (1), Wehenkel L (1)*

In the field of genome wide association studies, the individuals are described by hundreds of thousands of SNPs. Due to linkage disequilibrium, these descriptors are correlated. We propose here to exploit these correlations inside Random Forest by treating attribute groups instead of single variables within test-nodes. We expect that grouped variables can be exploited to construct more efficient tests and thus better predictors. Indeed, when a SNP is associated with the disease, its neighbors are expected to be disease associated too.

Materials and Methods

More precisely, each test-node becomes a node limited totally randomized tree learned from a subgroup of variables, which, in this case, correspond to a group of adjacent SNPs. During the learning process, at each node, K decision trees are constructed on K randomly picked groups, the most informative one is kept to split the learning set given a threshold on the probabilities generated by the resulting tree. Inside these tree test-nodes, the best nodes are expanded first, then tree complexity (called here the internal complexity) can be limited to use only some of the SNPs from a group.

Results

This approach has been applied on simulated data and on a real dataset related to Crohn's disease. The results are compared to Random-Forest testing individual SNPs. The resulting models are far less complex due to the dimension reduction implied by arbitrarily grouping adjacent SNPs. Also, we observe a predictive power increase with bigger values of the internal complexity. Furthermore, we extended variable importances to block importances which allow us to directly highlight chro- mosomal regions of interests.

Discussion

More generally, facing the increasing amount of data in GWAS, what we propose here is to use weak learners as test-nodes into decision trees in order to reduce the problem dimensionnality given the data structure. Further works will focus our efforts on exploring the internal complexity parameter space and on testing variant weak learners (like ensemble of stumps, ensemble of trees or HMM models) as test- nodes.

Presenting Author

Vincent Botta (vincent.botta@ulg.ac.be)
University of Liege : Department of Electrical Engineering and Computer Science & GIGA-Research

Author Affiliations

(1) University of Liege - Department of Electrical Engineering and Computer Science and GIGA-Research

Acknowledgements

This paper presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. VB is recipient of a F.R.I.A. fellowship and PG is a Research Associate of the F.R.S.-FNRS.

B-6. Validation of single cell arrayCGH on a 60-mer oligo microarray platform for preimplantation genetic diagnosis

Cheng J (1,), Vanneste E (2), Konings P (1), Van Eyndhoven W (4), Voet T (2), Ampe M (3), Verbeke G (3), Vermeesch J (2), Moreau Y (1)*

Preimplantation Genetic Diagnosis with Aneuploidy Screening (PGD-AS) is used for the selection of genetically normal embryos before implantation, thus increasing both baby take home rate and health of the fetuses. Array CGH technologies have been employed to detect chromosomal anerploidies in single blastomeres on a genome-wide level. The aim of this study was to explore whether the use of high resolution Agilent 244K Human CGH arrays would improve the detection of segmental aneuploidies in single cells compared to the previously used BAC and SNP arrays (Vanneste et al., Nat Med, 2009).

Materials and Methods

In total, 8 Epstein Barr Virus (EBV) transformed lymphoblastoids and 43 blastomeres derived from human embryos were analyzed. Data were normalized by the new developed channel based method. Subsequently, mean-median calculation, Circular binary segmentation (CBS) and Haarseg algorithms have been optimized and employed to analyze the accuracy of the data to call for a priori known and/or de novo whole chromosome and segmental aberrations in single blastomeres.

Results

We succeeded in developing a new normalization method to preprocess Agilent 244K human mircroarray data, enabling single cell data analysis on oligo platforms. We proved that this channel based method can effectively correct for artifacts/biases introduced by single cell amplification while maintaining a similar dynamic range of the log-ratios as compared to unamplified genomic DNA (Cheng et al., BMC Genomics, submitted). The mean-median calculation method can identify a priori known aberrations while the CBS and the Haarseg algorithms can detect both a priori known and de novo aberrations.

Discussion

In conclusion, whole chromosome and segmental imbalances can be detected at the single cell level using Agilent 244K human CGH microarrays. However, the aberration calling accuracy of the three detection algorithms optimized for Agilent is in the same order of magnitude as compared to BAC and SNP based arrays. Further validation studies are needed before any array CGH technology can be introduced in the clinic as a standardized assay for PGD.

Presenting Author

Jiqiu Cheng (jiqiu.cheng@esat.kuleuven.be)
Katholieke Universiteit Leuven

Author Affiliations

1. Katholieke Universiteit Leuven, ESAT-SCD: SISTA/COSIC/DOCARCH, Leuven. 2. Katholieke Universiteit Leuven, Center for Human Genetics, Leuven, Belgium. 3. Katholieke Universiteit Leuven, Center for Biostatistics, Leuven, Belgium. 4. Agilent Technologies.

Acknowledgements

This work was supported by Research Council K.U.Leuven: ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys, START 1, several PhD/postdoc & fellow grants; Flemish government: FWO-PhD/postdoc grants, FWO-projects, G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), FWO- research communities (ICCoS, ANMMM, MLDM), G.0733.09 (3UTR), G.082409 (EGFR), IWT- PhD Grants, Silicos, SBO-BioFrame, SBO-MoKa, TBM-IOTA3, FOD-Cancer plans; Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011); EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED. We would like to thank Agilent for providing us with the Agilent 244K arrays. We would like to thank Pascal Yazbeck and Sigrun Jackmaert for preparing and performing the array experiments. We appreciated technical discussions with Kristof Engelen, Inge Thijs, Marijke Bauters, Guy Froyen and Ernesto Iacucci. E.V.

was supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

B-7. Polymorphisms associated with mtDNA of elite Kenyan endurance athletes

Seiler M (1,), Fuku N (4), Diao L (1), Solovyov A (2), Seiler S (3), Tanaka M (4), Bhanot G (1,5,6,7)*

Kenyan dominance in endurance running events in recent decades is now well-established. Of the top 500 outdoor marathon official running times, 42% belong to Kenyans, including the top 4. It is natural to suspect that dominance by a single ethnic group may have a partial genetic basis. It has also been demonstrated that perhaps the most important contributor to endurance running success, aerobic capacity, can be partially attributed to matrilineal inheritance. Since mitochondria are strictly maternally inherited, mtDNA may harbor polymorphisms associated with endurance performance.

Materials and Methods

5023 mitochondrial sequences were obtained from MITOMAP and aligned to rCRS using the stretcher algorithm as implemented in the EMBOSS package. An additional 74 mtDNA samples from elite Kenyan endurance athletes were sequenced and added to the dataset. All sequences were organized into mitochondrial haplogroups using known sequence markers and recursive consensus clustering as implemented in the ConsensusCluster package. Polymorphisms enriched in the 74 medalists were identified within haplogroups using a strict permutation test for significance.

Results

The consensus clustering procedure identified 7 robust clades (L0/1/2/3/5/6/7) in the data as well as polymorphisms distinguishing these clades with 90% or higher accuracy under sample bootstrap. Our assignments agreed with the mtDNA tree in Behar et al, 2008 in common samples. We found 4 SNPs enriched in athletes of the L0 clade, two of which result in non-synonymous mtDNA mutations. We did not find any SNPs to be significantly enriched in African samples of the same haplogroup yet absent in runners, suggesting that SNPs which reduce endurance performance are not under selection.

Discussion

A previous study of elite Kenyan endurance athletes found a significant representation of L0 samples and a dearth of L3 samples (Scott et al., 2008), though no study was made of the polymorphisms which might contribute to enhanced performance. The SNPs we have identified in the L0 elite endurance athlete samples studied are virtually absent in the L0 clade members of the general population, suggesting these polymorphisms are very recent. Experiments on cell lines derived from samples with these SNPs can be used to show whether these cells show increased mitochondrial efficiency.

Presenting Author

Michael W Seiler (miseiler@gmail.com)
Rutgers University

Author Affiliations

(1) BioMaPs Institute, Rutgers University, Piscataway, NJ, 08854, USA (2) Physics Department, Princeton University, Princeton, NJ, 08544, USA (3) Sarah W. Stedman Nutrition and Metabolism Center, Durham, North Carolina 27710, USA (4) Department of Genomics for Longevity and Health, Tokyo Metropolitan Institute of Gerontology, 35-2 Sakae-cho, Itabashi-ku, Tokyo 173-0015, Japan (5) Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ, 08540, USA (6) Department of Physics; Department of Molecular Biology & Biochemistry, Rutgers University, Piscataway, NJ, 08854, USA (7) Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ, 08903, USA

B-8. Bayesian inference of community average gene copy numbers and its application for the metagenomic characterization of bacterioplankton community types in the Sargasso Sea

Beszteri B (1,2,), Frickenhaus S (2), Giovannoni SJ (1)*

We recently drew attention to the fact that average genome sizes (GS) of communities are expected to influence apparent relative gene abundances in comparative metagenomics (doi: 10.1038/ismej.2010.29). We suggested a method to account for this possible bias by estimating average GSs and incorporating them into gene count normalization in a generalized linear modelling framework. This approach, however, did not account for the uncertainty in estimated GSs, and led to an arbitrary scaling of inferred relative gene abundances. The methods presented here remedy these weaknesses.

Materials and Methods

We developed a fully Bayesian framework for incorporating gene length, sampling effort and community-averaged genome size into estimating gene abundances from metagenomic counts. We implemented an adaptive normal jump Metropolis-Hastings scheme to fit the model which scales well to real life metagenomic data set sizes, and used an ANOVA variant of this model to identify genes characteristic of previously described (doi:10.1038/ismej.2009.60), seasonally and vertically differentiated planktonic microbial community types at the Bermuda Atlantic Time Series Site.

Results

Besides identifying some biological processes potentially relevant for adaptive differentiation across the communities compared, our results also highlight some inherent limitations of gene centric community -omics approaches. In several cases, average copy numbers of different gene families with similar fuctions showed complementary changes among community types. The inferred differences in gene copy numbers often do not directly reflect functional adaptation but rather differences in the taxonomic composition between communities.

Discussion

The method introduced represents a next step towards more fully accounting for biases impacting gene centric comparative metagenomics. The Bayesian framework we propose is flexible and can be extended to incorporate further biasing factors. However, our results also highlight that gene abundance differences among communities do not necessarily reflect differential environmental adaptation of community functional repertoire as often implicitly assumed in similar studies.

Presenting Author

Bank M Beszteri (bank.beszteri@awi.de)
Alfred Wegener Institute for Polar and Marine Research

Author Affiliations

1: Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany 2: Department of Microbiology, Oregon State University, Corvallis, OR, USA

Acknowledgements

This work was funded by the Gordon and Betty Moore Foundation

B-9. Phylogenetic mapping of non-model organism RNA-seq reads using a graph algorithm

Vilella A (1,), Massingham T (1), Loytynoja A (1)*

RNA-seq analyses of non-model organisms suffer from the lack of a reference gene set to map the reads against. We propose to tackle this problem by aligning the reads to multiple alignments of gene phylogenies from related species using the PAGAN sequence graph aligner. Our simulation study and analyses of real data indicate better performance than simply using current de novo assembly approaches. Our PAGAN graph aligner can take graphs from de novo assemblers and phylogenetically map them to reference gene family alignments, increasing the bridging between contigs and separating paralogs.

Materials and Methods

Our approach builds profile-HMM models from EnsemblCompara GeneTrees codon alignments, scans the RNA-seq reads using HMMER3 and clusters reads as putatively homologous to the highest scoring gene tree with evalue

Results

The aligned reads from each species reflect the underlying paralogous haplotypes in duplicating genetrees and indicates the inferred missing site info. We performed simulations to study the expected coverage of a given RNA-seq run for a non-model species with no close sequenced genome Our method performs well even for species 300 MYA to the closest reference genome. We show that there is a benefit in using multiple references even when a single closest reference genome is available, due to the incompleteness of genomic annotation and the wider scope of gene families in representing evolution.

Discussion

We applied our approach to reconstruct the brain transcriptome of several publicly available non-model transcriptomes in the NCBI SRA against the available gene families in EnsemblCompara. Even for low-coverage runs (1x-4x of 454 Ti sequencing) we already obtained an average coverage of 50% transcripts at 60% median of full length. Considering current NGS technologies, we predict that our phylogenetic approach is be able to deliver comprehensive transcript models for a vertebrate species in a sampled subgroup for less than $10k in sequencing costs.

URL

https://www.ebi.ac.uk/~avilella/talks/20100901_ECCB

Presenting Author

Albert J. Vilella Bertran (avilella@ebi.ac.uk)
EBI-EMBL

Author Affiliations

1 EMBL, EBI, WTGC, Hinxton, CB10 1SD, United Kingdom

B-10. GEI-DB: a database of atypical genomic elements in bacteria

Bezuidt O (1,), Lima-Mendez G (2), Reva O (1)*

Comparative genome analysis revealed lateral transfer as one of the dominant factors in the evolution of bacteria. Analysis of these organisms showed that mobile genomic elements (MGE) appear as atypical genomic entities that influence the dissemination of genes that contribute to virulence and adaptation to environmental selection pressures. Developments of tools are desperately needed to aid the identification of such elements to fully understand their role and influence in bacterial evolution. This work focuses on the analysis of pathways of distribution of MGE in bacteria.

Materials and Methods

SeqWord genomic islands sniffer (SWGIS) was used in the study to search for genomic islands (GI) in bacteria. SWGIS detects GI by the analysis of intragenomic variations and compositonal biases in the genome-wide distribution of tetranucleotide usage patterns. It evaluates variances of oligonucleotide frequencies (OUV) and calculates distances (D) between local pattern deviations from the patterns calculated for the whole chromosomal DNA.

Results

The current version of GEI-DB offers an array of 3518 precalculated genomic islands (GI) identified in 637 prokaryotic genomes by SWGIS. Annotation of genes that are contained within each islands is provided. The resource also provides the tRNAs that each island is associated with identified by tRNA-scan tool; the list of homologous genomic regions which were acquired upon a comparison of all the genomic islands against one another using BLASTN and classifications of functional related genomic islands in the form of protein families identified by a Markov clustering algorithm.

Discussion

The GEI-DB serves as a resource for MGE that could further be used for more analysis by the scientific community. It has available in it a collection of MGE from prokaryotes of various backgrounds that are of medical and environmental importance. It allows browsing of genomic elements that were classified according to DNA and protein compositional features that they have in common and also gene entities that are associated with these elements, which also allow the study of evolutionary profiles that are shared among MGE.

URL

`http://anjie.bi.up.ac.za/geidb/geidb-home.php` *(firefox, konqueror and safari only)*

Presenting Author

Oliver Bezuidt (bezuidt@gmail.com)
University of pretoria, Department of biochemistry

Author Affiliations

University of Pretoria, Department of Biochemistry, Bioinformatics and Computational Biology Unit, Lynnwood Rd, Hillcrest, Pretoria, South Africa (1*), Laboratoire de Bioinformatique des Genome et des Reseaux, Universite de Bruxelles, 1050 Bruxelles, Belgium (2)

Acknowledgements

The work was funded by the National Bioinformatics Network of South Africa.

B-11. The CNS twilight zone: limitations in comparing upstream regions in plants

Reineke AR (1,), Bornberg-Bauer E (1), Gu J (1)*

Two main problems challenge the investigation of upstream regions in plants through comparative genomic analysis. First, the search for clear orthologous gene pairs between plants is difficult due to many in- and outparalogs arisen from frequent duplication events. Second, the search for conserved non-coding sequences (CNS) is difficult because they are less frequent and much shorter in upstream regions of plants compared to animals.

Materials and Methods

With Inparanoid (Remm et al.,2001), orthologous genes were found in plant pairs with divergence times between 5 and 108 mya. Upstream regions of genes were extracted in 500 bp segments up to 3 kb upstream of the transcription start site. With DIALIGN (Morgenstern et al., 1998), a tool combining global and local alignments, similarities and CNS between orthologous upstream regions were analyzed. The missing plant divergence times were estimated using BEAST (Drummond et al., 2007).

Results

Upstream region similarities of monocots are significantly lower compared to dicots. In both plant groups, CNS frequency decreases with increasing plant divergence time. The decay rate of UTR similarity is estimated using the exponential function. CNSs are found to be mostly embedded in the first +500 bp upstream, followed by -500 bp downstream. The similarity between UTR decreases with distance from the transcription start site (TSS). In monocots, the signal for significant similarity is not detectable for a distance of >+1kb in monocots and +1.5kb in dicots from the TSS.

Discussion

Our findings identifies the twilight zone, where significant similarity of upstream regions becomes difficult to distinguish, is estimated to be around 70 mya divergence time for plants. Furthermore, independent duplication events reduce this time of useful divergence. In regions with a distance over +1.5 kb from transcription starts site, CNS detection is difficult because the weak CNS signal cannot be distinguished from the background noise. Upstream region comparisons between monocots and dicots have to be done with caution, due to different CNS properties.

Presenting Author

Anna R. Reineke (a.reineke@uni-muenster.de)
Institute for Evolution and Biodiversity, University of Münster

Author Affiliations

Institute for Evolution and Biodiversity, University of Münster

Acknowledgements

This work was supported by HFSP R6P0033/2006-C and the Alexander von Humboldt Foundation.

B-12. TurboOrtho: a high performance alternative to OrthoMCL

Ekseth O (1), Lindi B (1), Kuiper M (1), Mironov V (1,)*

Establishing genuine orthology relations among species is a hard problem, therefore in practice some heuristic approaches are used. In most cases BLAST, TerraBLAST or mpiBLAST are used in the all-against-all mode to generate a list of gene/protein similarities which serves as the starting point for elucidating putative orthology relationships. A number of software packages are available for this purpose, OrthoMCL being one of the most popular. However, its performance is limited by its implementation based on Perl. Therefore, we developed a High Performance alternative to OrthoMCL.

Materials and Methods

TurboOrtho was implemented in C++. For benchmarking TurboOrtho we used a set of 144829 protein sequences from four species: H. sapiens, A. thaliana, S. cerevisiae and S. pombe. The BLAST output file contained 30841618 hits and was 2.2 G large. The analysis was performed on a Hewlett-Packard machine with 125 G memory and 2.66 GHz speed.

Results

The execution time for the benchmarking set was 1080 min for OrthoMCL v.1, 70 min for OrthoMCL v.2 and below 5 min for TurboOrtho.

Discussion

We have re-implemented the OrthoMCL algorithm (with minor amendments) in C++ and demonstrate a many-fold increase in the performance. In particular, TurboOrtho outperforms OrthoMCL v.1 and OrthoMCL v.2 by the factor of 216 and 14 respectively. TurboOrtho is also easier to use than OrthoMCL v.2 because it combines four steps of the OrthoMCL procedure in a single program that does not require a relational database.

Presenting Author

Vladimir N. Mironov (mironov@bio.ntnu.no)
Norwegian University for Science and Technology (NTNU) Trondheim, Norway

Author Affiliations

Norwegian University for Science and Technology (NTNU),Trondheim, Norway

B-13. i-ADHoRe 3.0 Detection of collinearity in large scale datasets

Proost S (1,2,+,), Fostier J (3,+,*), Dhoedt B (3), Demeester P (3), Van de Peer Y (1,2), Vandepoele K (1,2)*

Though several tools to detect collinearity exist, few are able to go beyond a pairwise comparison and harvest the information from multiple genomes to detect additional diverged regions. Current tools supporting multiple genomes have memory and run-time issues when scaled up to a dozen or more genomes. As recently several new genomes have been sequenced, and many more can be expected in the near future, there is a clear need for a novel tool able to cope with this abundance of data in an efficient and accurate way.

Materials and Methods

Several algorithmic improvements are featured in this version, the most important one being a novel, graph-based alignment of gene lists. Also, support for multi-core CPUs and computer clusters is implemented using multi-threading and MPI support for parts of the algorithm that could run in parallel. Statistical improvements have been thoroughly benchmarked using permutation tests and different datasets derived from PLAZA (http://bioinformatics.psb.ugent.be/plaza/) and Ensembl release 57 (http://www.ensembl.org/index.html).

Results

The new version of i-ADHoRe made it possible to analyze an unprecedented large dataset containing 49 genomes in less then 5 hours (using 32 CPUs). We show the merits of the algorithmic improvements and the technical adjustments that greatly reduce the runtime and memory requirements, compared to previous versions and other tools. Furthermore, a few preliminary results from the analysis of this large dataset are shown, including a large alignment of several Hox-custers found in a variety of species.

Discussion

As new sequencing technologies greatly reduce the time and costs to sequence new genomes, the downstream analysis of this data will become the bottleneck in the absence of fast and accurate tools to analyze them. In this study, it was demonstrated that the combination of fast heuristics, efficient implementation and support for modern hardware can be a significant improvement over the current state-of-the-art.

Presenting Author

Sebastian J.J. Proost (sebastian.proost@psb.vib-ugent.be)
Department of Plant Systems Biology, Bioinformatics and Systems Biology Division, VIB

Author Affiliations

1 Department of Plant Systems Biology, Bioinformatics and Systems Biology Division, VIB, Technologiepark 927, 9052 Ghent, Belgium 2 Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium 3 Ghent University - IBBT, Department of Information Technology (INTEC), Gaston Crommenlaan 8, Bus 201, 9050 Ghent, Belgium +contributed equally

Acknowledgements

S.P. thanks the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship. K.V. is a Postdoctoral Fellow of the Research Foundation–Flanders. This project is funded by the Research Foundation–Flanders and the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet).

B-14. Comparative expression analysis of orthologous genes between Arabidopsis and rice

Movahedi S (1,2,), Van de Peer Y (1,2), Vandepoele K (1,2)*

Recently microarray experiments have yielded massive amounts of expression information for many genes under various conditions or in different tissues for different model species. Expression compendia grouping multiple microarray experiments performed in similar (or different) experimental condition make it possible to define correlated expression patterns between genes or go a bit further and see how tissue-specific or constitutive expression is linked with expression context conservation and protein evolution.

Materials and Methods

In this study we use three main data bases including microarray expression data, GO functional annotations and orthologous gene families to calculate Expression Coherence (EC) and Expression Context Conservation (ECC) measures. We also calculate protein evolution rate (Ka) in order to study correlations between tissue specificity, expression context conservation and protein evolution.

Results

Among 4,630 1:1 orthologous Arabidopsis-rice gene pairs, 60% are ECC conserved, 21% ECC diverged and 19% have non-significant ECC . ECC conserved genes are evolving slightly slower at the protein level compared to genes with non-conserved coexpression contexts. ECC conserved genes overall show a higher expression breadth compared to non-conserved ECC genes. Colinear genes do not show a different ECC conservation pattern but Core genes are slightly more conserved. Finally genes with increased rates of protein evolution show reduced expression breadth.

Discussion

It is not easy to find a clear correlation pattern between tissue specificity, expression context conservation and protein evolution, most of the previous attempts gave conflicting results. In this study we also found tissue specificity and protein evolution are weakly linked with the conservation of gene expression. ECC conservation pattern of colinear genes and core genes is not striking as well but we have strong evidences showing tissue specific genes are fast-evolving at the protein level.

Presenting Author

Sara Movahedi (sara.movahedi@psb.vib-ugent.be)
Department of Plant Systems Biology, Bioinformatics and Systems Biology Division, VIB

Author Affiliations

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

B-15. Phylogenomics and robust construction of prokaryotic evolutionary trees

Korenblat K (1), Volkovich Z (1), Bolshoy A (2,3,)*

Usually, the phylogeny produced by means of molecular evolution is based on the comparison of the highly conserved small subunit rRNA sequences (ssu-rRNA) as originally proposed by Woese and colleagues. We wanted to demonstrate the power of the phylogenomics' approach. For this purpose we constructed an evolutionary tree of a few hundreds of prokaryotes. We wanted to show that our method of construction of prokaryotic evolutionary trees produces very robust results. We wanted to show that more genes are considered for tree construction better trees are obtained.

Materials and Methods

Data consist of median protein lengths related to several thousands of COGs from more than 600 genomes. To demonstrate robustness we selected 60 genomes. To evaluate robustness of trees we used two approaches. The first way was to compare distances between trees constructed on real data with those obtained for randomized data. The second method was to compare distances among trees based on different subsets of all COGs set. Partition distance between trees was chosen to measure distances between obtained evolutionary trees.

Results

1) An average distance between trees obtained on different COG-subsets (randomly selected 300 COGs out of 900 COGs) is significantly smaller than an average distance between trees based on partially randomized data. 2) An average distance between trees obtained on different COG-subsets is negatively correlated with a number of randomly selected N COGs out of 900 COGs. 3) Robustness of the method allows obtaining reliable consensus phylogenomic trees of hundreds of prokaryotes.

Discussion

Our results make a positive contribution to resolve an ongoing discussion whether phylogenomics can successfully assist in construction of a "species tree" as opposite to a "gene tree".

Presenting Author

Alexander Bolshoy (bolshoy@research.haifa.ac.il)
University of Haifa

Author Affiliations

1 Software Engineering Department, ORT Braude Academic College, Karmiel, Israel 2 Department of Evolutionary and Environmental Biology, University of Haifa, Haifa, Israel 3 Genome Diversity Center of the Institute of Evolution, University of Haifa, Haifa, Israel

B-16. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin

Almén MS (1,), Nordström KJ (1), Fredriksson R (1), Schiöth HB (1)*

The transmembrane proteins form a link between the cell and the surroundings, allowing it to interact and respond to environmental changes and communicate with neighboring cells. Large research efforts have concerned these proteins over the years as they include the majority of our drug targets. Although this numerous group of proteins has received a lot of attention it is largely unexplored and a comprehensive overview is missing.

Materials and Methods

We mined the human proteome and identified the membrane proteome subset using three prediction tools for alpha-helices: Phobius, TMHMM, and SOSUI. This dataset was reduced to a non-redundant set by aligning it to the human genome and then clustered with our own implementation of the ISODATA algorithm. The genes were classified and each protein group was manually curated, virtually evaluating each sequence of the clusters, applying systematic comparisons with a range of databases and other resources.

Results

We identified 6,718 human membrane proteins and classified the majority of them into 234 families of which 151 belong to the three major functional groups: receptors (63 groups, 1,352 members), transporters (89 groups, 817 members) or enzymes (7 groups, 533 members). Also, 74 miscellaneous groups with 697 members were determined. Interestingly, we find that 41% of the membrane proteins are singlets with no apparent affiliation or identity to any human protein family. Further, we have extended our study with a thorough evolutionary investigation of the understudied human 4TM proteins.

Discussion

In conclusion, we estimate that 27% of the total human proteome are alpha-helical transmembrane proteins and have created a complete overview for all major families of transmembrane proteins and functionally classified them. This roadmap will aid future identification and classification of novel transmembrane proteins by providing a functional context of the known human transmembrane repertoire.

Presenting Author

Markus Sällman Almén (markus.sallman-almen@neuro.uu.se)
Dept. of Neuroscience, Uppsala Universitet

Author Affiliations

1. Department of Neuroscience, Functional Pharmacology, Uppsala University, Uppsala, Sweden

Acknowledgements

The studies were supported by the Swedish Research Council, The Novo Nordisk Foundation, Swedish Royal Academy of Sciences, and Magnus Bergvall Foundation. RF was supported by the Göran Gustafssons foundation.

B-17. Towards molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data

Raes J (1,2,), Letunic I (1), Yamada T (1), Jensen LJ (1,3), Bork P (1,4)*

Meta-omics (metagenomics, metatranscriptomics, metaproteomics) are powerful tools for the analysis of complete microbial communities in both environmental (e.g. ocean, soil) and medical (e.g. the human microbiome) context. Because of its complexity, meta-omics data has required the development of novel computational approaches to move from 'parts lists' to ecosystem functioning and and understanding of the ecology of these communities.

Materials and Methods

To deduce important ecological indicators such as environmental adaptation, molecular trait dispersal, diversity variation and primary production from the gene pool of an ecosystem, we integrated 25 ocean metagenomes with geographical, meteorological and geo-physicochemical data.

Results

We find that climate (temperature, sunlight) is the major determinant of the biomolecular repertoire of each sample and the main limiting factor on functional trait dispersal (absence of biogeographic provincialism). Molecular functional richness and diversity show a distinct latitudinal gradient peaking at 20°N and correlate with primary production. The latter can also be predicted from the molecular functional composition of an environmental sample.

Discussion

Together, our results show that the functional community composition derived from metagenomes can be used as quantitative predictor for molecular trait-based biogeography and ecology. We see this study as proof-of-principle that molecular functional composition can be used in various other environmental settings such as the human microbiome where it could be integrated with clinical data to study the molecular ecology and temporo-spatial variation of the ‘human’ ecosystem.

Presenting Author

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

Author Affiliations

1 Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany 2 Molecular and Cellular Interactions Department, VIB – Vrije Universiteit Brussel, 1040 Brussels, Belgium 3 NNF Center for Protein Research, 2200 Copenhagen, Denmark 4 Max Delbrück Center for Molecular Medicine, 13125 Berlin-Buch, Germany

Acknowledgements

This research is supported by the European Union FP7 Program Contract no. HEALTH-F4–2007-201052. JR is supported by the FWO Oddyseus programme.

B-18. Orthologous proteins associated with yeast telomeric complex identified by synteny and sequence similarity

Macko M (1,), Tomáška L (2), Vinař T (3)*

Proteins of telomeric complex protect integrity of chromosome and thus they have an important function during the replication process. These proteins have been previously identified in yeast Saccharomyces cerevisiae, but not in other yeast species e.g. Candida parapsilosis. Search for orthologous proteins in other species is complicated by faster evolution of these proteins and species specific sequences on chromosome ends where these proteins bind. Identification of proper orthologs in several yeast species will enable further analysis of evolution of this complex.

Materials and Methods

We searched for the putative orthologs by a combination of sequence comparison and synteny. The first batch of orthologs have been identified by the reciprocal best BLAST hits between proteins of Saccharomyces cerevisiae and several other yeast species. Next, we used orthologs located by our search and orthologs already identified in KEGG database, to create pHMM and motifs by software HMMER and MEME. These models and motifs were used to identify additional orthologs. Confidence for new orthologs is evaluated by the naive Bayes classifier with use of the sequence and synteny information.

Results

We have searched for orthologous proteins of 121 Saccharomyces cerevisiae proteins associated with telomeres in genomes of 6 related yeast species. Higher sensitivity of search was accomplished by combining information from known orthologs and creating profile hidden markov models for each group of orthologous proteins. For example, significant matches for EST3 protein were not found by BLAST search but with motifs and pHMM searches. On the other hand protein CDC13 had no significant matches even though most proteins that interact with it according to database UniPROT were found.

Discussion

After obtaining putative orthologs for most proteins of the telomeric complex, protein interaction networks could be reconstructed, by assuming conserved protein interactions between orthologs of proteins which have interaction identified in UNIProt database. This could help locate “missing” nodes in these putative networks and point out cases where further analysis is needed.

Presenting Author

Martin Macko (martin.macko1@gmail.com)
doctoral student, Department of Applied Informatics, Faculty of Mathematics, Physics, and Informatics Comenius University in Bratislava

Author Affiliations

(1)doctoral student,Department of Applied Informatics, Faculty of Mathematics, Physics, and Informatics Comenius University (2) head of the Department of Genetics, Faculty of Natural Sciences, Comenius University (3) Assistant Professor,Department of Applied Informatics, Faculty of Mathematics, Physics, and Informatics Comenius University

Acknowledgements

This work has been supported by a grant VEGA 1/0210/10, Marie Curie Fellowship to Tomas Vinar IRG-224885, and Comenius University young researcher grant G-10-229-00

B-19. Estimating average heterozygosity using multilocus dominant marker data: a maximum likelihood approach via the EM algorithm.

Khang TF (1,), Yap VB (2)*

Many preliminary studies of natural plant and animal populations use molecular methods such as amplified fragment length polymorphism (AFLP) and random amplified polymorphic DNA (RAPD) to assess average heterozygosity in a population. For small sample sizes, ascertainment bias can lead to inaccurate estimates when the distribution of null homozygote proportions for dominant loci is J-shaped. We show how an EM algorithm can be used effectively to correct for ascertainment bias when estimating average heterozygosity under Hardy-Weinberg equilibrium assumption for all dominant loci.

Materials and Methods

From the full likelihood function, which is a product of beta-binomial functions, we derived the Expectation and Maximization steps of the EM algorithm. We used simulation to check the error distribution of maximum likelihood estimates of average heterozygosity via the EM algorithm. We did this for several different beta distribution profiles for the null homozygote proportions. Subsequently, We compared its performance with two other maximum likelihood approaches: one that does not correct for ascertainment bias, and another that corrects for it using a truncated beta-binomial likelihood.

Results

Our results show that ML estimates via the EM algorithm as well as the truncated beta-binomial likelihood lead to subtantial improvement in accuracy and root mean square error for J-shaped beta profiles. For U-shaped profiles, the improvement is nontrivial; and for inverse-J shaped profiles, there is negligible improvement as expected.

Discussion

Many published results do not incorporate correction for ascertain bias when estimating average heterozygosity. As we have shown, this can lead to estimates that are biased upwards, with magnitude dependent on the beta profile of the null homozygote proportions. This finding is of concern because it may mislead researchers into thinking that a particular population is genetically diverse. Therefore, to properly interpret the results of published work that do not correct for ascertanment bias, we propose to do a simple graphical check of the beta profiles.

Presenting Author

Tsung Fei Khang (tfkhang@um.edu.my)
Institute of Biological Sciences, University of Malaya

Author Affiliations

1. Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia. 2. Department of Statistics & Applied Probability, Faculty of Science, National University of Singapore, Singapore 117546.

Acknowledgements

The National University of Singapore supported the first author with a research scholarship while the work was carried out as part of his PhD thesis.

B-20. PLAZA 2.0 : a resource for plant comparative genomics

Van Bel M (1,2,), Proost S (1,2), Wischnitzki E (1,2), Van de Peer Y (1,2), Vandepoele K (1,2)*

Comparative sequence analysis has significantly altered our view on the complexity of genome organization and gene function. To explore all this genome information, a centralized infrastructure is required where all data generated by different sequencing initiatives is integrated and combined with advanced methods for data mining. Here we describe PLAZA2.0, an update to the PLAZA platform. This resource integrates structural and functional annotation of published plant genomes together with a large set of interactive tools to study gene function and gene and genome evolution.

Materials and Methods

The update PLAZA version contains the nuclear and organelle genomes of twenty three species within the Viridiplantae kingdom: eleven eudicots, five monocots, two (club-)mosses and five green algae. The integration of all gene annotations provided by the different sequencing centers yielded a data set of 843,854 gene models, of which 88.3% represent protein-coding genes. The remaining genes are classified as transposable elements, RNA, and pseudogenes (11.3%, 0.1%, and 0.3%, respectively).

Results

Gene families were automatically generated using MCL clustering, resulting in far fewer single species gene families compared with PLAZA 1.0, thus providing evidence that the growing number of species leads to a more balanced view of species-specific gene families. Automated whole-genome statistical comparisons give further insights in variations in codon usage between species, and the new and improved tools allow users to gain knowledge faster and more efficiently.

Discussion

We provide an update to the PLAZA platform, where users can easily browse genes, gene families and the associated functional annotation.

URL

http://bioinformatics.psb.ugent.be/plaza

Presenting Author

Michiel Van Bel (Michiel.vanbel@psb.ugent.be)
Universiteit Gent

Author Affiliations

(1) Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium. (2) Department of Plant Systems Biology, Ghent University, B-9052 Ghent, Belgium.

Acknowledgements

This work was supported by European Union EUFP6 Food Safety and Quality Contract FOOD-CT-2006-016214. This project is funded by the Research Foundation?Flanders and the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet).

B-21. Branch testing in phylogenetic trees: comparison of computationally efficient methods

Czarna A (1,), Wróbel B (1,2)*

It has become standard procedure to assess uncertainty in phylogenetic reconstruction by calculating bootstrap support or Bayesian posterior probability for clades. However, when the tree is reconstructed using maximum likelihood, the computational cost of calculating bootstrap support may be prohibitive, especially for large datasets. The same goes for Bayesian analysis. Other approaches to branch testing were proposed, but they have not been compared thus far in a systematic manner.

Materials and Methods

We used nucleotide sequences of yeast species for which correct topology is known. The alignment was separated into 1000-nt subsets. We have first analyzed how the bootstrap support values are affected by the method used for the reconstruction (ML, NJ/BIONJ, PhyML). Then, the branches were tested using the quartet puzzling (QP) method, the Dopazo test, aLRT and WLS-LRT. Finally, we have investigated how the results are affected by simplifying the substitution model (using JC+Γ+I model instead of GTR+Γ+I) or by not addressing the rate heterogeneity among sites (GTR instead of GTR+Γ+I).

Results

The PhyML bootstrap support values were almost the same as the values obtained when a computationally expensive ML method was used for the reconstruction of pseudotrees. Using PhyML, NJ or BIONJ gave results similar to ML in terms of false positives. On the other hand, the Bayesian approach, aLRT and Dopazo test resulted in less false negatives and more false positives. The QP method and WLS-LRT gave intermediate results. Using GTR instead of JC+Γ+I resulted a higher support for true branches. Both simplified models caused an increase of false positives.

Discussion

Calculating bootstrap support or Bayesian posterior probability for clades may not be possible for large data sets. Our results show that more computationally efficient methods can be used to find support for branches, although some result in more false positives, especially when the model is too simple. Using PhyML for pseudotrees reconstruction was the most robust to model misspecification; the QP method the least robust. However, bootstrapping trees with PhyML took noticeably more time than using WLS-LRT, aLRT or Dopazo test.

URL

http://www.evosys.org

Presenting Author

Aleksandra Czarna (aczarna@iopan.gda.pl)
Computational Biology Group, Institute of Oceanology, Polish Academy of Sciences

Author Affiliations

1 – Computational Biology Group, Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland 2 – Evolutionary Systems Group, Laboratory of Bioinformatics, Adam Mickiewicz University in Poznań, Poland

Acknowledgements

Polish Ministry of Science and Education (project N303 291234).

B-22. Parsimonious higher-order hidden Markov models for enhanced comparative genomics

Seifert M (1,), Banaei A (1), Gohr A (2), Strickert M (1), Grosse I (3)*

Arabidopsis thaliana is an important model organism in plant biology with a broad geographic distribution. Precise knowledge of genomic differences between ecotypes is a fundamental component for a better understanding of their phenotypic variation. For measuring genomic differences, the method of Array-Comparative Genomic Hybridization (Array-CGH) is applied routinely nowadays. The analysis of the resulting huge data sets requires efficient bioinformatics tools. Long-range dependencies between measurements on the chromosomes stimulate an extension of standard Hidden Markov Models.

Materials and Methods

Array-CGH has been applied to compare the genomes of the two important ecotypes Col and C24 of the model plant Arabidopsis thaliana. Sequence polymorphisms have been identified by a newly developed parsimonious higher-order Hidden Markov Model (PHHMM) modeling dependencies between more than two directly neighboring probes on a chromosome. The PHHMM is compared to a standard HMM and to other methods for Array-CGH data analysis. This is done based on sequence polymorphisms identified in publicly available re-sequencing experiments of C24.

Results

The identification of sequence polymorphisms by the PHHMM is noticeably higher than for the standard HMM and other methods for Array-CGH data analysis. Sequence polymorphisms have been identified widespread over all chromosomes in good accordance with polymorphisms known from re-sequencing experiments. A significantly large proportion of the identified sequence polymorphisms in C24 is covered by transposable elements in the reference genome of Col. Genes and their functional components are significantly less affected.

Discussion

By modeling higher-order dependencies in a data-dependent manner, the newly developed PHHMM overcomes the limitation of the commonly used HMM that only models dependencies between two adjacent probes on a chromosome. This leads to an improved identification of sequence polymorphisms also in comparison to other methods. The biological relevance of the identified sequence polymorphisms is emphasized by the over-representation of transposable elements that are known to be involved in driving the variation among ecotypes. The PHHMM turns out to be a useful tool for the analysis of ACGH data sets.

Presenting Author

Michael Seifert (seifert@ipk-gatersleben.de)
IPK Gatersleben

Author Affiliations

(1) Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany (2) Leibniz Institute of Plant Biochemistry, Halle (Saale), Germany (3) Martin Luther University Halle, Halle (Saale), Germany

Acknowledgements

Development of the PHHMM is funded by the Ministry of Culture Saxony-Anhalt (XP3624HP/0606T), and the Array-CGH experiments were funded by the DFG (HO1779/7-2).

B-23. Correction of bootstrap confidence levels using an iterated bootstrap procedure with computational efficient methods

Czarna A (1), Wróbel B (1,2,)*

The uncertainty in phylogenetic analysis is affected by many factors which cause that the bootstrap support values are often underestimated or overestimated. Using the method proposed by Rodrigo (1993) allows for correction of bootstrap confidence levels. The method requires bootstrapping at three levels. Howver, even with only 20 pseudosamples at each level (8000 in total), the computational cost is prohibitive when maximum likelihod is used for the reconstruction of pseudotrees. Here we investigate if the NJ or PhyML methods can be used to approximate the correct bootstrap confidence levels.

Materials and Methods

The iterated bootstrap procedure was used for nucleotide sequences of yeast species for which correct topology was known. We separated the alignment of over 100000 nt into 100 datasets, 20 replicates were obtained for each. Pseudotrees were reconstructed using NJ and PhyML. 20 2nd-level bootstrap replicates were obtained for each 1st-level replicate and again 20 3rd-level bootstrap replicates for each 2nd-level replicate. Finally, 8000 3rd-level pseudotrees were reconstructed to allow for the correction of bootstrap confidence levels.

Results

We analyzed 5 true and 3 wrong branches for one hundred 1000-nt datasets. True branches got support in 47-99% of datasets when PhyML was used for pseudotree reconstruction (depending on the branch). When NJ was used, they got support in 22-92% of datasets. Wrong branches got support in 9-21% (PhyML) and 7-17% (NJ). The results were similar to using 70% threshold in the standard bootstrap procedure: in 32-100% of datasets (PhyML) and 17-85% (NJ) the bootstrap proportions (BP) were above 70% for true branches. In 2-13% (PhyML) and 3-12% (NJ) datasets wrong branches had BP above 70%.

Discussion

The use of iterated bootstrap resulted in a much lower number of false negatives than the standard bootstrap method. However, the correction of the bootstrap support values also caused undesirable increase of support for wrong branches. When PhyML was used for the reconstruction of pseudotrees, all the branches received higher support than when NJ was applied. However, the computations using PhyML took 750 times longer. Our results support the notion that iterated bootstrap using computationally efficient methods can be applied as an approximate method to correct bootstrap confidence levels

URL

http://www.evosys.org

Presenting Author

Borys Wróbel (bwrobel@iopan.gda.pl)
Computational Biology Group, Institute of Oceanology, Polish Academy of Sciences

Author Affiliations

1 – Computational Biology Group, Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland 2 – Evolutionary Systems Group, Laboratory of Bioinformatics, Adam Mickiewicz University in Poznań, Poland

Acknowledgements

Polish Ministry of Science and Education (project N303 291234).

B-24. Molecular evolution, structure and functional divergence of Lipoxygenase gene family in vertebrate

Padmanabhan R (1), Kuhn H (2), Reddanna P (1)*

Lipoxygenase are a group of non-heme, iron containing enzymes involved in oxygenation of PUFA. The products of LOX reactions are hydroperoxides (eicosanoids) which are powerful mediators of inflammation and other medically important states, like asthma, atherosclerosis, cancer etc.Although discoveries about LOX family in mammals and other species have been reported in some studies, the evolution and functional and structural divergence of LOX in mammals are not clearly understood.Hence to study the phylogenetic relationships, structural and functional divergence of lipoxygenas.

Materials and Methods

Sequence data sets of the LOX protein family were obtained by Blast-Explorer and genomic databases at Ensembl.Amino acid sequences were aligned with muscle and refined in Bioedit and Gblock. Phylogentic trees were constructed with ML, NJ programs implemented in the PHYLIP program package and PhyML.GeneTree was used to resolve the incongruence between the gene and species trees by predicting duplication and losses.Estimation of substitution rates and testing positive selection was done using PAML.Site-specific changes in evolution rate after gene duplication (type I) was estimated using DIVERGE

Results

Sequence similarity searches using BLAST-EXPLORER was used to identified the homologus sequences of LOX in all the kingdoms. A total of 152 sequeces from 46 species representing four major lineages (fishes,amphibeia, sauria, mammalia) were included in the study.Phylogenetic analysis was used to characterize the family evolutionary history by identifying two major duplications early in vertebrate evolution., to further analyse this GeneTree software was used.Estimates of synonymous and nonsynonymous substitution rates suggest that the duplication events were associated with positive selection.

Discussion

We present a phylogeny describing the evolutionary history of the Lipoygenase gene family and shows that the genes have evolved through duplications in animalia followed by positive selection., even though many sites are higly conserved in the gene familiy and preserve an over all structural feature of these enzymes. The critical amino acid residues likely relevant for the distinct functional properties of the prologues has been identified. Intron-Exon structure evolution of these genes showed well conservation of exons and intron expansion in mammals.

Presenting Author

Roshan Padmanabhan (rosaak@gmail.com)
The School of Life Sciences, University of Hyderabad, Hyderabad-500046, Andhra Pradesh, India

Author Affiliations

(1) The School of Life Sciences, University of Hyderabad, Hyderabad-500046, Andhra Pradesh, India (2)Institute of Biochemistry, University Medicine Berlin-Charité, Monbijoustrasse 2, D-10117 Berlin, Germany

B-25. Cassis: detection of genomic rearrangement breakpoints

Baudet C (1,3,), Lemaitre C (1,2), Dias Z (3), Gautier C (1), Tannier E (1), Sagot M-F (1)*

Genomic regions which have undergone a rearrangement are called breakpoints. Current methods for breakpoint identification are in fact strategies for detecting synteny blocks or conserved regions (that are orthologous regions which have not been rearranged): breakpoints are obtained only as a byproduct, simply by returning regions that are not found in a conserved synteny. Cassis aims to go one step further and to extend the synteny blocks by focusing on the breakpoints themselves in order to improve the precision of the breakpoints.

Materials and Methods

Cassis receives as input a list of one-2-one orthologous genes. It processes the given list to determine the breakpoint regions. For each breakpoint, its sequence, called SR (in the reference genome), is aligned against its orthologous sequences SoA and SoB (in the second genome). A sequence of discrete values is defined along the sequence SR: +1 (-1) for the position that has a hit only with the sequence SoA (SoB) and 0 for the position that has hits with both sequences. A segmentation algorithm analyses this sequence to narrow the breakpoint, and its result is then statistically assessed.

Results

We compared the genomes of human and mouse with Cassis. First, we used the orthologous genes between the two species available in Ensembl and we identified 369 breakpoints. A total of 340 breakpoints were narrowed with an average length reduction of 62%. The median size of the refined breakpoints is 54.9 kbp. Secondly, we ran Cassis with lists of synteny blocks obtained by other methods (Compara and Mauve). We show that Cassis is able to refine most of the breakpoints identified by the other methods.

Discussion

Cassis is a method that allows the identification and refinement of breakpoints by a comparison between two genomes. Our results showed that it is able to narrow a large number of breakpoints, and that it can work also with the synteny blocks identified by other methods. Cassis thus enables to better characterise breakpoint sequences and their distribution along the genomes, which could lead to a better understanding of the mechanisms and evolutionary properties of chromosomal rearrangements.

URL

http://pbil.univ-lyon1.fr/software/Cassis/

Presenting Author

Christian Baudet (christian.baudet@univ-lyon1.fr)
Equipe BAMBOO, INRIA Grenoble Rhône-Alpes et Laboratoire de Biométrie et Biologie (UMR 5558) CNRS, Université Lyon 1

Author Affiliations

(1) Equipe BAMBOO, INRIA Grenoble Rhône-Alpes et Laboratoire de Biométrie et Biologie Évolutive (UMR 5558) CNRS, Université Lyon 1, F-69100 Villeurbanne, France (2) Université de Bordeaux, Centre de Bioinformatique - Génomique Fonctionnelle Bordeaux, F-33000 Bordeaux, France 3 (3) Institute of Computing, University of Campinas (Unicamp), Av. Albert Einstein, 1251 - Cidade Universitária, Caixa Postal 6176 - CEP 13083-970, Campinas - São Paulo, Brazil

Acknowledgements

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior [4676/08-4] and partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico [472504/2007-0, 479207/2007-0 and 483177/2009-1]. This work was funded by the ANR [project BLAN08-1335497], ANR-BBSRC [project ANR-07-BSYS 003 02] and the ERC AdG (project Sisyphe).

B-26. Methylated cytosines are less likely to mutate within CpG islands

Medvedeva Y (1,), Panchin A (2,3) Mitrofanov S (2), Makeev V (1)*

CpG islands (CGIs) in mammals are regions with about 7-10 times higher CpG dinucleotide frequency comparing to the rest of the genome. The main cause of this effect is though to be the decreased level of cytosine methylation within the CG context inside CGIs. Methyl-cytosines followed by guanines (5mCpGs) are about 10 times more likely to mutate into TpGs than nonmethylated CpGs. However, recent studies show that 5mCpGs are not very rare in CGIs. We decided to compare the frequency of 5mCpG to TpG mutations inside CGIs and across the remaining parts of the human genome.

Materials and Methods

From a dataset of human SNPs with cytosine/thymine variants we selected those containing cytosine in orthologous positions of chimp and orangutan genomes. These SNPs represent C to T mutations in human population. We took into consideration only SNPs with double strand cytosine methylation in the human embryonic stem cell line. For each 5mCpG within CGIs a control 5mCpG was selected outside CGIs with the same degree of methylation and coverage depths. We compared the percentage of mutated 5mCpGs in both groups using Fisher's exact test. to check for statistical significance.

Results

5mCpGs are over 1.5 fold less likely to mutate into TpGs inside CGIs comparing to the rest of the genome. This effect is statistically significant ($p < 10e-20$).

Discussion

We propose two possible explanations for the observed decrease of 5mCpG mutation percentage in CGIs. First, deleterious variants are expected to be underrepresented among SNPs. CGIs have been shown to carry certain functions; therefore some SNPs within CGIs might have been eliminated from the population as a result of natural selection. Second, the mutation frequency of 5mCpG->TpG itself or the effectiveness of reparation systems could be subject to local GC content influence. Several studies show that adjacent methylated cytosines increase the effectiveness of mismatched thymine replacement.

Presenting Author

Yulia Medvedeva (ju.medvedeva@gmail.com)
Research Institute for Genetics and Selection of Industrial Microorganisms

Author Affiliations

(1) Research Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia (2) Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia (3) Institute for Information Transmission Problems RAS, Moscow, Russia

B-27. Entropy approach reveals new features of genes and genomes

Putintseva Y (1,), Sadovsky M (1,2)*

Investigation of statistical properties of the nucleotide sequence brings new facts, new knowledge, and inspires a researcher with new methodology. Our approach is mathematically oriented and aims to retrieve as much knowledge, as possible, from a sequence itself, avoiding an implementation of any additional, external information. Here we proposed three new indices of codon usage bias, as well as some other issues. All the indices are based on mutual entropy calculation. They differ in the codon frequency distribution supposed to be "quasi-equilibrium" one.

Materials and Methods

1031 bacterial, 89 archaeal and 8 yeast annotated genome sequences were retrieved from the European Bioinformatics Institute Database (http://www.ebi.ac.uk/genomes/). S. cerevisiae nucleotide sequences were taken from Saccharomyces Genome Database (http://www.yeastgenome.org/). Sequences containing letters other than A, C, G, T were excluded from the analysis. Frequency dictionary (including expected frequencies) and its entropy figures (including mutual entropy) are the key tools of the studies.

Results

Three indices of mutual entropy and absolute entropy were calculated at the level of genomes and genes. Correlations between the figures and the taxonomic position of the organism were observed with varying levels of success. Genome regions, which differ considerably in their entropy characteristics from the average of the genome, were identified (for example, in S. cerevisiae genome a group of genes involved into flocculation process has remarkably specific entropy characteristics). Numerous examples of statistically identified sites have been studied, as well.

Discussion

We developed a new index measuring codon usage bias. The index, designated as mutual entropy, aims to retrieve as much knowledge, as possible, from the genome sequence itself. Representation of all of the information for a gene and genome by a single statistic is essentially a reduction in information; therefore, no single measure reflects all aspects of codon usage. Although measures, proposed here, also suffers from some limitations we can conclude that it is a promising method for obtaining quantitative information about the degree of overall synonymous codon usage of a gene and genomes.

Presenting Author

Yuliya A. Putintseva (yuliya-putintseva@rambler.ru)
Siberian Federal University

Author Affiliations

1 - Siberian Federal University 2 - Institute of Computational Modelling of Russian Academy of Sciences

B-28. The hidden duplication past of the plant pathogen Phytophthora and its consequences for infection

Martens C (1,2,), Van de Peer Y (1,2)*

Oomycetes of the genus Phytophthora are pathogens that infect a wide range of plant species. For dicot hosts such as tomato, potato and soybean, Phytophthora is even the most important pathogen. Previous analyses of Phytophthora genomes uncovered many genes, large gene families and large genome sizes that can partially be explained by significant repeat expansion patterns.

Materials and Methods

Analysis of the complete genomes of three different Phytophthora species, using a newly developed approach, unveiled a large number of small duplicated blocks, mainly consisting of two or three consecutive genes. The detected number of duplicated genes was further compared with ten other eukaryotes including parasites, algae, plants, fungi, vertebrates and invertebrates, of which the gene and genome duplication history is known.

Results

The performed analyses suggest that the ancestor of P. infestans, P. sojae and P. ramorum most likely underwent a whole genome duplication (WGD). Genes that have survived in duplicate are mainly genes that are known to be preferentially retained following WGDs, but also genes important for pathogenicity and infection of the different hosts seem to have been retained in excess. As a result, the WGD might have contributed to the evolutionary and pathogenic success of Phytophthora.

Discussion

The fact that we find many small blocks of duplicated genes indicates that the genomes of Phytophthora species have been heavily rearranged following the WGD. Most likely, the high repeat content in these genomes have played an important role in this rearrangement process. As a consequence, the paucity of retained larger duplicated blocks has greatly complicated previous attempts to detect remnants of a large-scale duplication event in Phytophthora.

URL

http://bioinformatics.psb.ugent.be

Presenting Author

Cindy Martens (cindy.martens@psb.vib-ugent.be)
Ghent University

Author Affiliations

(1) Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium (2) Bioinformatics and Systems Biology, Department of Plant Biotechnology and Genetics, Technologiepark 927, Ghent University, B-9052 Ghent, Belgium

Acknowledgements

IWT

B-29. Protein linkages between African Trypanosoma and 8 pathogenetic eukaryotic organisms

Dimitriadis D (1,), Trimpalis P (1,2), Choli-Papadopoulou T (2), Anagnou NP (3), Kossida S (1)*

Biological cells are regulated and controlled by interacting proteins in metabolic and signaling pathways. Moreover, as the processing power of computers is constantly being improved, the implementation of computational methodologies studying biological systems and quantitatively understanding their functionality and complexity is feasible and necessary.

Materials and Methods

The methodology we followed for the domain fusion analysis is described in the below given steps: 1.1.The protein sequences of the African trypanosome, Trypanosoma brucei,were retrieved. 1.2.The translated sequences of 8 genomes (say which ones in this paranthesis) from NCBI & UniProt databases were also retrieved. 2.Trypanosoma brucei protein sequences were compared (blastp) against the 8 retrieved proteomes by the "in-house" software to predict functional associations based on fusion events. 3.Validation of the final predictions based on some criteria's (see figure)

Results

We identified 176 component of which 12 were excluded for further analysis: 1.Functional annotation prediction. 2.Biological pathways through the evolution, based on domain fusion analysis method.

Discussion

We compared the 5 fused events against all the family of eukaryotic organisms in each subkingdom, using BlastP. The scope of this step was to make an observation and prediction for the evolution of the fusion domains based on fusion or fission events within other eukaryotic organisms.

The in vivo confirmation of the fused predicted events have being undergone. Future efficient results may

a.surface more details about the biological function and the metabolic pathways of compounds.

b.Motivate the development of new drugs.

•we have supported the accuracy of the "in-house" software

Presenting Author

Dimitris D.D Dimitriadis (dimitrisdimitriadis@mac.com)
Bioinformatics and Medical Informatics Team of the Biomedical Research Foundation of the Academy of Athens.

Author Affiliations

1)Bioinformatics and Medical Informatics Team of the Biomedical Research Foundation of the Academy of Athens. 2)department of chemistry-Aristotle University of Thessaloniki 3)department of medicine-National and Kapodistrian University of Athens

B-30. Reconstructing phylogenetic trees from clustering trees

Costa E (1,), Vens C (1,2), Blockeel H (1,3)*

In the context of phylogenetic tree reconstruction, divisive clustering methods can be used to infer phylogenetic trees in a top-down way. These methods have the important advantage of providing an explanation for the resulting topology, since the splits are described by polymorphic locations in the sequences. However, the quality of the resulting trees is rather variable. In this work, we argue that trees induced by top-down methods can be viewed not just as phylogenetic trees, but also as identifying constraints that the real phylogenetic tree must satisfy.

Materials and Methods

We analyzed trees inferred by Clus-φ, a distance based method for phylogenetic tree reconstruction based on a conceptual clustering method that extends the well-known decision tree learning approach. Each split defines two subclusters, such that the total branch length of the tree is minimized. However, the split does not define how the subclusters have to be connected. We propose a post-processing method that processes the clustering tree bottom-up, at each split finding the internal branch that connects the two subclusters with a minimal number of mutations.

Results

To evaluate this method we used a number of synthetic datasets generated by an evolutionary process simulator. In general, the post-processed Clus-φ trees are more similar to the underlying target trees of the synthetic datasets than the original Clus-φ trees, which shows that the post-processing step yields a better approximation of the target tree. When we consider Neighbor Joining and Parsimony results in this comparative analysis, we observe that the post-processed Clus-φ trees tend to be better than the NJ trees and are comparable to the parsimony trees.

Discussion

The results show that trees resulting from top-down phylogenetic tree construction can be improved by post-processing them. This post-processing is based on the viewpoint that the methods do not necessarily return the correct tree, but return constraints that the correct tree must satisfy. These constraints allow to guide the search for the tree with a minimal number of mutations in a more exhaustive way than the greedy search performed by parsimony methods. In general, the quality of post-processed Clus-φ trees are comparable to that of parsimony trees.

Presenting Author

Eduardo Costa (eduardo.costa@cs.kuleuven.be)
Katholieke Universiteit Leuven

Author Affiliations

1 - Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Leuven, Belgium 2 - Institut National de Recherche Agronomique, UMR 1301, 400 Route des Chappes, 06903 Sophia-Antipolis, France 3 - Leiden Institute of Advanced Computer Science, Universiteit Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
{eduardo.costa,celine.vens,hendrik.blockeel}@cs.kuleuven.be

Acknowledgements

Eduardo Costa is supported by the Research Foundation – Flanders (FWO) and the GOA Probabilistic Logic Learning. Celine Vens is a Postdoctoral Fellow of the Research Foundation –Flanders (FWO).

B-31. Predictive modeling of psychrophilic adaptation on the proteome sequence level

Frickenhaus S (1,), Beszteri B (1,2)*

Psychrophilic sequence adaptation of cold-active sequences is nowdays commonly analysed under the hypothesis that protein-flexibility must be maintained in the cold. Based on structural models and evaluation of local free-energy contributions the existence of families of locally as well as globally cold-adapted protein sequences were found. The present approach evaluates amino-acid properties numerically, allowing for discriminative analyses of families of protein-homologues, negelcting structural details.

Materials and Methods

Protein sequence alignments of homologues derived from completed genome sequencing projects of six psychrophilic and six mesophilic species are computed. The aligned aminoacid-sequences are translated residue-wise into a matrix by a set of five biopysically meaningful factor-scores. In a windowing approach along the alignment matrix linear discriminant models are computed for each factor-score independently to explain psychrophilic origin.

Results

Statistical analyses of the resulting models demonstrate a slight advantage in the predictive power for the factor-score IV corresponding to amino-acid heat-capacities. However, it is not possible to derive a clear relation between successful discrimination models and supersets of protein function.

Discussion

The result is consistent with the interpretation of flexibility as amplitude of positional fluctuation, as experimentally shown by neutron-scattering. A similar trend of heat-capacity is found from proteome-average scores in a linear discriminant analysis.

URL

http://www.awi.de/en/go/bioinformatics

Presenting Author

Stephan Frickenhaus (stephan.frickenhaus@awi.de)
Alfred-Wegener-Institute for Polar and Marine Research Bremerhaven

Author Affiliations

1: Alfred-Wegener-Institute for Polar and Marine Research, 2: Oregon State University

Acknowledgements

The project is funded within the AWI-research programme PACES.

B-32. Microbial phenotype prediction based on protein domain profiles

Lingner T (1,), Mühlhausen S (1), Meinicke P (1)*

Accurately predicting microbial phenotypes solely based on genomic features will allow us to infer relevant phenotypic characteristics when the availability of a genome sequence precedes experimental characterization. Although this scenario is favored by the advent of novel high-throughput and single-cell sequencing techniques, no practical solution exists. A particular challenge arises from the growing need to establish genotype-phenotype associations without the application of an orthology search step that is usually required by current phylogenomics approaches.

Materials and Methods

Our approach for phenotype prediction is based on the complete genomic sequences and the NCBI phenotype annotation of more than 1000 prokaryotic organisms. Pfam domain family occurrence profiles are inferred from all organisms' genomic sequences and are then represented as high-dimensional vectors of counts reflecting the organism-specific domain abundance. The phenotype annotation is used to construct phenotype-specific binary classification problems, which are then evaluated using a discriminative machine learning technique.

Results

Our approach provides high prediction accuracy regarding the phenotype categories motility, Gram stain, oxygen requirement and spore formation without the need for any orthology assignments. Our method substantially outperforms an approach that is based on inferred metabolic pathways and yields an average area under ROC curve of 0.972. Furthermore, the set of discriminative domains provides biological insight into the underlying mechanisms for a given phenotype and enables deriving hypotheses on the possible functions of uncharacterized domains.

Discussion

Fast and accurate orthology-free prediction of microbial phenotypes based on genomic protein domain content is feasible and has the potential to provide novel biological insights. First results of a systematic check for annotation errors indicate that our approach may also be applied to semi-automatic correction and completion of the existing phenotype annotation. Furthermore, the analysis of a hierarchical clustering of discriminative domain families indicate that the method can be used to suggest phenotype-specific functions for uncharacterized protein families.

URL

http://www.gobics.de

Presenting Author

Thomas Lingner (thomas@gobics.de)
Department of Bioinformatics, Göttingen, Germany

Author Affiliations

Department of Bioinformatics, Georg-August-University Göttingen, Göttingen, Germany

Acknowledgements

DAAD PostDoc Fellowship to Thomas Lingner FUGE Vest fellowship to Thomas Lingner

B-33. New insights into the metazoan evolution of cadherins: from basal to modern

Hulpiau P (1,), van Roy F (1,2)*

Specific cell-cell adhesion and intracellular communication are key processes in multicellular animals. Members of the cadherin superfamily are essential players in these processes. The details of premetazoan and metazoan evolution of the cadherin superfamily are largely unknown, but the many newly sequenced genomes of key metazoan organisms provide a rich resource for unraveling evolution of such complex protein families.

Materials and Methods

We used a combination of tBLASTn and profile HMM to identify the cadherin repertoires in amphioxus, sea anemone and the placozoan Trichoplax. To determine the evolution within the cadherin superfamily, we extensively compared domain organization and sequences of cadherin and cadherin-related proteins in more than 10 organisms from different metazoan lineages. Using bl2seq, we compared extracellular cadherin (EC) domain blocks, individual EC repeats and also non-EC domains. To complete the phylogenetic study we constructed a Neighbor-Joining and Bayesian Inference tree of cadherins all across the metazoan kingdom.

Results

Classical cadherins, such as E-cadherin, arose from an Urmetazoan cadherin, which progressively lost N-terminal extracellular cadherin repeats while its cytoplasmic domain, which binds p120ctn and β-catenin, remained quite conserved from placozoa to man. Protocadherins, until now considered a chordate innovation, predate the Bilateria and are likely rooted in an ancestral FAT cadherin. Flamingo,FAT-like and dachsous cadherins have remained essentially unchanged. The remarkable ancient origin of several cadherin types suggests that each of them has fulfilled separate and essential needs throughout metazoan evolution.

Discussion

The last common ancestor of animals expressed at least five types of cadherins found in nearly all its descendants: a ‘classical’ cadherin, a Flamingo cadherin and three cadherin-related members FAT, FAT-like and dachsous. Their similar domain composition suggests that they were originally paralogs. Both modern ‘classical’ cadherins and protocadherins have short ectodomains suggesting that the role of particular cadherins has changed from intercellular signaling via loose contacts to tight interactions. This might reflect the evolution from the limited requirements of a colonial unicellular eukaryote to the strict morphogenetic programs essential for metazoan organs.

Presenting Author

Paco Hulpiau (paco.hulpiau@dmbr.vib-ugent.be)
VIB & Ghent Uiniversity

Author Affiliations

(1) Department for Molecular Biomedical Research, VIB, Ghent, Belgium (2) Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium

Acknowledgements

Supported by the Research Foundation Flanders (FWO) and the Geconcerteerde Onderzoeksacties of Ghent University.

B-34. Comparative mapping of transcription factor binding sites in plant genomes

Wischnitzki E (1,2,), Van de Peer Y (1,2), Vandepoele K (1,2)*

The growing number of sequenced genomes provides an unique opportunity to study the conservation of transcription factor binding sites in plant species. The detection and evaluation of functional transcription binding sites is essential but unfortunately not trivial. It is especially hindered in plants by several small and large scale duplication events, leading to a very diverse number of orthologous or in many cases inparalogous genes. Our comparative mapping approach incorporates the individual information of each related set of genes and is applicable to any combination of genomes.

Materials and Methods

To estimate the evolutionary conservation the candidate motifs are evaluated using regulatory regions from orthologous genes. For each gene set and motif the conservation value is calculated as a probability for this event (p-value). The calculation is performed individually for each combination of gene set and motif and takes the individual composition and size of the set into consideration. Using this approach the p-value can be assigned and the significance of the result is defined. The results are evaluated using published TF target sets and enrichment for functional categories.

Results

Genes with significantly conserved binding sites show higher enrichment of functional categories than a simple genome wide mapping or a naive conservation filtering. Additionally known binding sites lacking functional enrichment using a naive mapping approach can now be linked with the underlying biological process they regulate. The comparison with experimental data shows a clear reduction of false positive instances and increased enrichment for the target sets. Our comparative mapping performed essentially better than the other two methods which sometimes were equal to random.

Discussion

The results demonstrate that comparative mapping is a powerful approach to identify functional binding sites and to distinguish them from false positives instances. Our approach can also be applied to study gene regulatory network inference and promoter evolution after speciation and/or duplication. This makes it a very effective tool to detect conserved and potentially functional transcription factor binding sites. Especially with the growing number of sequenced genomes the applications of our approach become more distinct and more detailed evolutionary conservation patterns can be studied.

Presenting Author

Elisabeth Wischnitzki (elwis@psb.vib-ugent.be)
VIB/UGent

Author Affiliations

1 Department of Plant Systems Biology, Bioinformatics and Systems Biology Division, VIB, Technologiepark 927, 9052 Ghent, Belgium 2 Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, 9052 Ghent, Belgium

B-35. Protein model selection with ProtTest increases phylogenetic performance

Patricio M (1,), Abascal F (1,2), Zardoya R (2, Posada D (1)*

Maximum likelihood and Bayesian phylogenetics rely on the use of explict probabilistic models of evolution. Best-fit models of evolution can be selected for the data at hand using statistical techniques like hierarchical likelihood ratio tests, or information criteria. While the connection between model selection and phylogenetic performance has been systematically studied for DNA sequence alignments, we do not much about this relationship in the case of protein alignments.

Materials and Methods

We simulated 5000 protein alignments (40 sequences with 1000 amino acids) using 5000 non-clock trees. For each alignment, we estimated the ML tree for every competing models (112 in total), and identified the best-fit model using the AIC and BIC criteria as implemented in ProtTest. To compare the phylogenetic performance of every competing model against the best-fit model we measured the differences between the generating trees and the estimated trees using the Robinson-Foulds symmetric distance (RF), the Branch-Score (BS) and the K-tree score (K) with its associated scale factor (SF).

Results

ProtTest correctly identified the true empirical model of amino acid replacement, or a very close one, most of the time. In general, the best-fit model was selected with little uncertainty and therefore receiving large weights, while model-averaged parameter estimates were also very precise. Phylogenetic trees reconstructed under the best-fit model selected by ProtTest were most accurate, together with the trees estimated under the true model used in the simulations. This was so regardless of the criteria used for model selection.

Discussion

Protein model selection using ProtTest increased phylogenetic accuracy independently of the criteria used. Not surprisingly, the most influential parameter was the shape of the gamma distribution, since the performance of the +G models was always closer to that of the generating model. A few particular models seem to work only slightly worse, like WAG and LG, which were derived from comprehensive datasets using ML estimationFurther simulations should explore whether these results hold under different generating models, including more time-consuming mechanistic models not explored here.

URL

http://darwin.uvigo.es

Presenting Author

Mateus Patricio (mateus@uvigo.es)
University of Vigo

Author Affiliations

(1) Departament of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain. (2) Departament of Biodiversity and Evolutionary Biology, National Museum of Natural Sciences (CSIC), Madrid, Spain.

Acknowledgements

This study was partially funded by the Spanish Ministry of Science, Xunta de Galicia and European Research Council.

B-36. cn.FARMS: a probabilistic latent variable model to detect copy number variations

Clevert D-A (1,2+,), Mitterecker A (1), Mayr A (1), Klambauer G (1), Tuefferd M (3), De Bondt A (3), Talloen W (3), Goehlmann HWH (3), Hochreiter S (1)*

Existing pre-processing methods for DNA microarrays designed to detect copy-number variations (CNVs) lead to high false discovery rates (FDRs). High FDRs misguide researchers especially in the medical context where CNVs are wrongly associated with diseases. Existing pre-processing methods for DNA microarrays designed to detect copy-number variations (CNVs) lead to high false discovery rates (FDRs). High FDRs misguide researchers especially in the medical context where CNVs are wrongly associated with diseases.

Materials and Methods

We suggest modeling the DNA copy numbers as a latent variable in our factor analysis model, which is optimized by Bayesian maximum a posteriori estimation. We originally developed this model for gene expression arrays but we extend our approach for copy number analysis. In comparison to other approaches we don't need to correct for sequence effects to remove the typical chromosome specific wave pattern correlated with GC content. cn.FARMS can be used for single-locus and multi-loci CNV analysis. The latter constructs so-called "meta-probe sets" combining probes from adjacent probe sets to one

Results

We compared cn.FARMS on HapMap Mapping250K_Nsp and SNP6.0 benchmark data sets to CRMA, dChip, CNAG and CNAT v4. The aim, described in this presentation, is to distinguish males from females based on the X chromosome copy numbers, where males possess one copy and females two. The ROC curve serves to compare the FDR for different true positive rates. cn.FARMS clearly outperforms its competitors. In addition we have assessed the CNVs detection performance under the objective of the area under the precision-recall curve. We found that, cn.FARMS performs superior to all other compared methods.

Discussion

Of course cn.FARMS is not limited to the Affymetrix platform and can be applied to other platforms like Illumina bead arrays. The concept remains the same: do genomic adjacent measurements tell the same story about copy numbers?

URL

http://www.bioinf.jku.at/software/farms/farms.html

Presenting Author

Djork-Arne Clevert (okko@clevert.de)
Institute of Bioinformatics, Johannes Kepler University Linz

Author Affiliations

(1) Institute of Bioinformatics, Johannes Kepler University Linz, Linz, Austria (2) Department of Nephrology and Internal Intensive Care, Charit e University Medicine, Berlin, Germany (3) Johnson & Johnson Pharmaceutical Research & Development, a Division of Janssen Pharmaceutica, Beerse, Belgium

Acknowledgements

This work has been funded by Janssen Pharmaceutica N.V. and the Belgium IWT project 80536.

B-37. Genome-wide heterogeneity of the substitution process

Arbiza L (1), Dopazo H (2), Posada D (1,)*

At genomic scales, the patterns that have shaped molecular evolution are largely heterogeneous. In particular, modern phylogenetic approaches should use appropriate probabilistic substitution models that capture the main features under which different genomic regions have evolved. While efforts have concentrated in the development and understanding model selection techniques, no descriptions of overall model fit at a complete genomic scale have been reported. Because NGS technologies are providing vast arrays of biological data, the understanding of model heterogeneity is fundamental.

Materials and Methods

The longest transcripts of orthologous coding genes from the complete genomes of 5 mammals and 15 vertebrates were obtained from Ensembl v54. Sequences were aligned using Muscle and filtered with Gblocks. In addition, filtered alignments for each of the 12 Drosophila genomes were obtained from the Drosophila 12 Genomes Consortium. The jModelTest and Phyml programs were used to estimate best-fit models of nucleotide substitution and ML trees. Putative functional association with model parameters was studied using the GO database, the program FatiScan and logit regressions using R.

Results

Out of the 88 models considered, 82 were selected as best-fit models at least in one occasion, although with very different frequencies. Most parameter estimates also varied broadly among genes. The patterns found for vertebrates and Drosophila were quite similar, and usually more complex than those found in mammals. Phylogenetic trees derived from models in the 95% CI set showed much less variance and were significantly closer to the tree estimated under the best-fit model than trees derived from models outside this interval. BIC selected simpler models, but suggested similar patterns.

Discussion

The analysis of ~20,000 alignments from three different genomic sets clearly shows that different genes are best explained by different models of nucleotide substitution, suggesting that the reasonably large variety of substitution models available is justified in order to maximize statistical model fit. The relevance of taking model selection uncertainty into account in phylogenetic analysis is clear from our results. We conclude that the use of model selection techniques is necessary to obtain accurate phylogenetic estimates from real data at a genomic scale.

URL

http://darwin.uvigo.es

Presenting Author

David Posada (dposada@uvigo.es)
University of Vigo

Author Affiliations

(1) Department of Biochemistry, Genetics and Immunology, University of Vigo, Spain (2) Comparative Genomics Unit, Bioinformatics Dept., CIPF, Valencia, Spain.

Acknowledgements

This work was partially supported by the European Research Council, Spanish Ministry of Science and Education and Xunta de Galicia

B-38. Comparative microarray analysis to elucidate TF-networks in activated T-cells

Kröger S (1,2,), Scheel T (1), Leser U (2) , Baumgrass R (1)*

A better knowledge of the transcriptional regulation during T cell activation is essential to understand the physiology and pathophysiology of this important hub of the adaptive immune system. Therefore, we created a basic gene regulatory network of T cell activation using public available data sets of gene expression arrays. The major challenges therein were the selection and heterogeneity of the available data sets. Eventually, we had to combine data sets across experimental techniques, species, stimulation time, and further experimental context.

Materials and Methods

From GEO [Geo2009] a pre-selected set of of gene expression sets were retrieved and curated by a immune biologists. All data sets were subjected to local (data set specific) and global (over all sets) RMA. The subsequent mapping of probe sets to Ensembl Gene IDs enables the comparison across microarray platforms. Mapping was improved by ANOVA sibling consolidation [Li2008]. We then used Network Component Analysis (NCA) [Liao2003] to construct a regulatory network of T cell activation.

Results

We present a large data set of T cell activation experiments consisting of more than 300 microarrays. Using NCA we constructed a basic gene regulatory network of differentially expressed transcription factors during T cell activation. Thereby, we confirm that the joint analysis of different microarray data sets is possible and helps to re-engineer regulatory networks.

Discussion

Further efforts are necessary to improve the process of data selection. In addition, validation and expansion of the basic transcription factor regulatory network is planned using analysis of global transcription factor binding [Lee2008], DNA methylation and ChIP-sequencing.

URL

http://informatik.hu-berlin.de/~kroeger/eccb2010/

Presenting Author

Stefan Kroeger (s.kroeger@drfz.de)
German Rheumatism Research Center Berlin

Author Affiliations

1 German Rheumatism Research Center Berlin 2 Humboldt Universität zu Berlin

B-39. Gene-trait matching analysis of Lactobacillus plantarum strains

Bayjanov JR (1,2,), Siezen RJ (1,2,3,4,5), van Hijum SAFT (1,2,3,4,5)*

Identifying genotype-phenotype (gene-trait) relations serves multiple purposes: (i) screen microorganisms for desired traits using gene content, (ii) gain insight in gene function, and (iii) gain knowledge on the effects of environmental factors that lead to gain or loss of a certain trait in an organism. A very cost-effective way for inferring gene-trait relations is to use high-throughput techniques such as comparative genome hybridization, transcriptomics, metabolomics, and phenotypic microarrays. A major bottleneck lies in the integration of these complex multivariate data.

Materials and Methods

We are developing a method and software that allows for non-parametric integration of multivariate data to identify many-to-many relationships. Genes linking to phenotypes are presented to the user, allowing for in-depth mining of such complex datasets. The software will be made available as a web-tool.

Results

The method is being tested on available genotype data (presence/absence of genes based on comparative genome hybridization results) and phenotype data (growth of strains on different sugars) for different Lactobacillus plantarum strains. Our method allowed determining (i) gene clusters associated to a single phenotype and (ii) gene clusters that were associated to multiple phenotypes. Additionally, there were leads to probable re-annotation of some genes in number of gene-phenotype relations.

Discussion

Current methods in multivariate data analysis have in general the following shortcomings: (i) they correlate only one observed variable (e.g., phenotypic measurement) with predictors (e.g., presence or absence of genes), and (ii) the actual relation between gene presence and phenotype is difficult to extract.

Presenting Author

Jumamurat R. Bayjanov (jumamurat@gmail.com)
Centre for Molecular and Biomolecular Informatics

Author Affiliations

1) Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Center, the Netherlands. 2) NBIC, Toernooiveld 1, 6525 ED, Nijmegen, the Netherlands 3) TI Food and Nutrition, Wageningen, the Netherlands 4) NIZO food research BV, Ede, the Netherlands 5) Kluyver Centre for Genomics of Industrial Fermentation, Delft, the Netherlands.

Acknowledgements

BSIK grant [through the Netherlands Genomics Initiative (NGI)]; BioRange programme [as part of, the Netherlands Bioinformatics Centre (NBIC)]; NGI (as part of the Kluyver Centre for Genomics of Industrial Fermentation).

B-40. ProtTest-HPC: fast selection of best-fit models of protein evolution

Darriba D (1,2), Taboada G-L (2), Doallo R (2), Posada D (1,)*

Using appropriate models of amino acid replacement is key for the study of protein evolution and can increase phylogenetic performance. ProtTest is a very popular tool for the selection of best-fit empirical models of amino acid replacement. The maximum likelihood calculations required for model selection can take a considerable time and can be unfeasible for large alignments. Because the latter are becoming increasingly common, it is necessary to develop computationally efficient solutions for protein model selection.

Materials and Methods

We developed a new version called ProtTest-HPC that is able to distribute the workload (i.e., ML optimization) into the available computational resources. We implemented and compared different strategies using shared (available cores in a machine), distributed (message-passing version on a whole cluster) and hybrid shared/distributed (two-level parallelism, relying on message-passing for inter-node communications and on a thread-based approach to exploit the available cores within each node) memory using MPI and OpenMP.

Results

We have redesigned ProtTest for high performance computing. ProtTest-HPC showed almost linear scalability with large input data in shared memory architectures with a low number of cores. For the distributed memory benchmarking, a two-step workload distribution was most efficient, scattering tasks among cluster nodes and taking advantage of the available processor cores per node. Moreover, we included new functionalities like the ability of building consensus trees and fault tolerance.

Discussion

ProtTest-HPC provides a user-friendly graphic interface for model selection tasks on multicore commodity/desktop systems, for current multicore processors, and also an interface for queueing systems (script-based) for distributed memory clusters. The speed up obtained with this version can be considerable. For instance, a large alignment analysis could be reduced from about 7 days using ProtTest to around 1 hour with ProtTest-HPC.

URL

http://darwin.uvigo.es/software/prottesthpc/

Presenting Author

David Posada (dposada@uvigo.es)
University of Vigo

Author Affiliations

(1) Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain. (2) Department of Electronics and Systems, University of A Coruna, 15071 A Coruna, Spain.

Acknowledgements

This work was partially supported by the European Research Council, Spanish Ministry of Science and Education and Xunta de Galicia.

B-41. NTRFinder: an algorithm to find nested tandem repeats

Matroud A(1,2,), Hendy M(1,2), Tuffley C(2)*

The ITS region of rRNA from Colocasia esculenta (taro), a plant of ethnobotanical interest, has been found to contain a 1600 bp complex repetitive structure consisting of two distinct tandem repeat motifs interspersed with one another. We propose that such "nested tandem repeats" are significant population markers and have developed the algorithm NTRFinder to search for NTRs in other sequences. A major issue is parsing, the determination of the optimal boundaries of the motifs, which is a significant problem in the analysis of NTRs.

Materials and Methods

NTRFinder adapts Wexler et al's heuristic for finding tandem repeats to the problem of finding NTRs, and aligns and analyses them using an extension of Fischetti et al's wrap-around dynamic programming. The algorithm has been implemented in java. We propose a model of motif duplication, mutation and deletion to explain the observed patterns. Under this model the parsing problem is solved using the most parsimonious tree linking the observed variants approximating each motif, minimising the number of duplicate substitutions in the development of the NTR.

Results

We have implemented and verified the algorithm with extensive testing on simulated NTRs. We have begun searching real sequence data. To date two other significant NTR regions of interest have been found. Variations in the NTRs of taro varieties from different regions are indicating its potential as a marker, and the rate of duplication appears to be approximately twice the rate of nucleotide substitution.

Discussion

We have begun the search for NTRs from other genomes using next generation sequencing. As the ITS region of rRNA is repeated hundreds of times throughout the taro genome, the consistency of the NTR is a result of concerted evolution, and further NTRs offer another window to understand this process. Comparisons of these additional NTRs will both give a more detailed pointer to the global spread of this prehistoric crop.

URL

http://arxiv.org/abs/1006.1730

Presenting Author

Atheer Matroud (a.a.matroud@massey.ac.nz)
Institute of Fundamental Sciences , Massey University

Author Affiliations

1 Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Private Bag 11 222 Palmerston North, New Zealand. 2 Institute of Fundamental Sciences , Massey University, Private Bag 11 222 Palmerston North, New Zealand.

Acknowledgements

This project was funded by the Allan Wilson Centre for Molecular Ecology and Evolution.

B-42. Computational epigenomics of plant SNF2 genes

Bargsten JW (,1,2,4), Mlynárová L (3,5), Nap JPH (1,5)*

Chromatin remodeling controls the accessibility of DNA for transcription via DNA methylation and histone modification, mediated by ATPases of the so-called Snf2 family. Arabidopsis contains at least 41 members of this family, some of which have already been associated with desirable traits such as tolerance to biotic or abiotic stress. The Snf2 family is at the same time conserved and highly diversified, posing computational challenges for identification, characterization and classification of such genes in other plant genomes as means to infer the function of individual family members.

Materials and Methods

Public domain genome resources of various plant genomes (arabidopsis, poplar, rice, cucumber, tomato and more) and data in different data warehouses (phytozome, UniRef, GenBank, ChromDB and more) were used for the comparative analyses. As soon as allowed, whole genome analyses of the soon-to-be-released tomato and potato genomes will be included in the comparisons. Analysis and modeling involved the appropriate software available for identification and classification, such as Blast, HMMER, MUSCLE, MEME and PHYLIP, in addition to Perl-based custom approaches for sequence masking and clustering.

Results

The Snf2 family consists of genes with conserved ATPase and helicase domains, interspersed with more diverse sequences. No less than 19 subfamilies in Arabidopsis may indicate the diverse functions of subfamily members. We screened the publicly available genome data with the conserved domains to identify family members and clustered genes on the basis of the remaining domain-spanning sequences. Results show that the total number of putative family members is surprisingly different in different plant genomes and that various subfamilies differ in the number of individual members.

Discussion

The distribution over subfamilies and the differences in absolute number of Snf2 family members both hint at diversified functions within plant species. These functions could go beyond the functions already identified in or suggested for Arabidopsis. Further analyses of the structural details of the subfamily members that differ between plant species, as well as comparative expression analyses of the family members in targeted cells and/or issues and/or mutational analyses for further functional characterization of this gene family, will be presented at the meeting.

Presenting Author

Joachim W. Bargsten (joachim.bargsten@wur.nl)
Applied Bioinformatics, Plant Research International, Plant Sciences Group, Wageningen University and Research Centre, The Netherlands

Author Affiliations

1. Applied Bioinformatics, Plant Research International, Plant Sciences Group, Wageningen University and Research Centre, The Netherlands 2. Laboratory for Plant Breeding, Plant Sciences Group, Wageningen University, The Netherlands 3. Laboratory for Molecular Biology, Plant Sciences Group, Wageningen University, The Netherlands 4. Netherlands Bioinformatics Centre (NBIC), Nijmegen, The Netherlands 5. Centre for BioSystems Genomics 2012 (CBSG2012), Wageningen, The Netherlands

Acknowledgements

This project was cofinanced by the Netherlands Bioinformatics Centre (NBIC), and was carried out within the research program of the Centre of BioSystems Genomics (CBSG), both of which are part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research.

B-43. Comprehensive analysis of splice site evolution in primates using whole genome alignments and RNA-Seq data

Ongyerth M (1,), Prüfer K (1), Kelso J (1)*

In recent years many new closely related genomes became available for the comparison with the human genome. Next generation sequencing is often used to supplement the genomic information with RNASeq data to quantify expression levels and discover differences in transcript structure. Splice sites, as one of the determinants of transcript structure, are encoded in the majority of cases through a well characterized sequence motif. We use this property to determine potential non-functional sites in non-human species and verify our findings with RNASeq data.

Materials and Methods

We use the Ensembl and Vega gene annotation of the human genome and a whole genome alignment between the human, chimpanzee, bonobo, orangutan and rhesus macaque genomes to identify orthologous positions of functional human splicing motifs in several primate species. To evaluate the functional impact of changes in splice sites, we score motifs using a profile weight matrix trained on human splice motifs. We use the score differences to the non-human species to analyze correlations to several features of the splice site and gene, and to identify potential non-functional sites.

Results

We show that high splice score differences between non-human and human species are more often observed with higher divergence between species. Sites in protein coding transcripts are generally more conserved than splice sites in pseudogenes or noncoding transcripts. Similarly, constituative splice sites are more conserved than alternative splice sites. When using RNASeq data, we observe that score differences are highly predictive of splice site usage. However this result applies only for high score differences. Minor score changes are not correlated with splice site usage.

Discussion

Our analysis shows that a simple model of splice sites can be used to identify non-functional splice sites in closely related primate species using solely human annotation. This result paves the way to a comprehensive analysis of splice site overturn over the cause of evolution in primate species. The analysis of individual candidates may uncover functionally important differences caused by a difference in transcriptome structure between human and other primates.

Presenting Author

Matthias Ongyerth (ongyerth@eva.mpg.de)
Max-Planck Institute for Evolutionary Anthropology

Author Affiliations

1 Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

AUTHOR INDEX

Abascal F 38
Albert J 12
Almén MS 19
Alves R 4
Amato R 5
Ampe M 8
Anagnou NP 32
Arbiza L 40
Ast G 3
Banaei A 25
Bargsten JW 45
Baudet C 28
Baumgrass R 41
Bayjanov JR 42
Beszteri B 11, 34
Bezuidt O 13
Bhanot G 10
Blockeel H 33
Bolshoy A 18
Bork P 20
Bornberg-Bauer E 14
Botta V 7
Cheng J 8
Choli-Papadopoulou T 32
Clevert D-A 39
Cocozza S 5
Costa E 33
Cuguen J 6
Czarna A 24, 26
Darriba D 43
De Bondt A 39
Demeester P 16
Dhoedt B 16
Diao L 10
Dias Z 28
Dimitriadis D 32
Doallo R 43
Dopazo H 40
Ekseth O 15
Fostier J 16
Fredriksson R 19
Frickenhaus S 11, 34
Fuku N 10
Gautier C 28
Geurts P 7
Giovannoni SJ 11
Goehlmann HWH 39
Gohr A 25
Grosse I 25
Gu J 14
Hendy M 44
Hochreiter S 39
Hulpiau P 36
Janssen P 39
Jensen LJ 20
Kelso J 46
Khang TF 22
Kim E 3
Klambauer G 39
Konings P 8
Korenblat K 18
Kossida S 32
Kröger S 41
Kuhn H 27
Kuiper M 15
Lemaitre C 28
Leser U 41
Letunic I 20
Lima-Mendez G 13
Lindi B 15
Lingner T 35
Loytynoja A 12
Macko M 21
Makeev V 29
Martens C 31
Massingham T 12
Matroud A 44
Mayr A 39
Medvedeva Y 29
Meinicke P 35
Miele G 5
Mironov V 15
Mitterecker A 39
Mlynárová L 45
Moreau Y 8
Movahedi S 17
Mühlhausen S 35
Nap JPH 45
Nordström KJ 19
Ongyerth M 46
Padmanabhan R 27
Patricio M 38
Pinelli M 5
Posada D 38, 40, 43
Proost S 16, 23
Prüfer K 46
Putintseva Y 30
Raes J 20
Reddanna P 27
Reineke AR 14
Reva O 13
Sadovsky M 30
Sagot M-F 28
Saumitou-Maprade P 6
Scheel T 41
Schiöth HB 19
Seifert M 25
Seiler M 10
Seiler S 10
Sela N 3
Siezen R 42
Siezen RJ 42
Solovyov A 10
Strickert M 25
Taboada G-L 43
Talloen W 39
Tanaka M 10
Tannier E 28
Tomáška L 21
Touzet P 6
Tran HT 6
Trimpalis P 32
Tuefferd M 39
Tuffley C 44
Van Bel M 23
Van de Peer Y 16, 17, 23, 31, 37
Van Eyndhoven W 8
van Roy F 36
Vandepoele K 16, 17, 23, 37
Vanneste E 8
Vens C 33
Verbeke G 8
Vermeesch J 8
Vilella A 12
Vinař T 21
Voet T 8
Volkovich Z 18
Wehenkel L 7
Wischnitzki E 23, 37
Wróbel B 24, 26
Yamada T 20
Yap VB 22
Zardoya R 38