BioVis 2015: The 5th Symposium on Biological Data Visualization

July 10-11th 2015

Co-located with ISMB 2015

Dublin, Ireland

Program for the 5th Symposium on Biological Data Visualization

Friday, July 10

8:30	8:45	BioVis Welcome, Kay Nieselt
8:45	9:45	Keynote Lecture: **Visual data science - Advancing science through visual reasoning** Speaker: ***Torsten Möller***, *University of Vienna, Austria*
9:45	10:30	Primer: **Four Levels of Visualization,** Session Chair: Sean O'Donoghue Speaker: ***Min Chen**, Oxford e-Research Centre, University of Oxford, UK*
10:30	11:00	*Coffee/Tea Break*
11:00	12:40	BioVis Papers: **Omics Data Visualization,** Session Chair: Michel Westenberg
11:00	11:25	PathwayMatrix: Visualizing Binary Relationships between Proteins in Biological Pathways, *T. N. Dang, P. Murray, A. Forbes*
11:25	11:50	Extended LineSets: A Visualization Technique for the Interactive Inspection of Biological Pathways, *F. Paduano, T. N. Dang, P. Murray, A. Forbes*
11:50	12:15	Pan-Tetris: an interactive visualisation for Pan-genomes, *A. Hennig, J. Bernhardt, K. Nieselt*
12:15	12:40	BactoGeNIE: A Large-Scale Comparative Genome Visualization for Big Displays, *J. Aurisano, K. Reda, J. Leigh, G.E. Marai, A. Johnson*
12:45	14:15	*Lunch Break with Lunch Box and Beverages*
14:15	15:00	Primer: **How to Visualize Data,** Session Chair: Sean O'Donoghue Speaker: ***Robert Kosara**, Tableau Software*
15:00	15:30	Short Poster Talks Session, Session Chairs: Alex Lex and Michael Westenberg
15:30	16:00	*Coffee/Tea Break*
16:00	17:15	BioVis Papers: **Visualizing Imaging Data,** Session Chair: Jos Roerdink
16:00	16:25	Physically-based In Silico Light Sheet Microscopy for Visualizing Fluorescent Brain Models, *M. Abdellah, A. Bilgili, S. Eilemann, H. Markram, F. Schürmann*
16:25	16:50	Visual parameter optimisation for biomedical image processing, *A.J. Pretorius, Y. Zhou, R.A. Ruddle*
16:50	17:15	GRAPHIE: Graph based Histology Image Explorer, *H. Ding, C. Wang, K. Huang, R. Machiraju*
17:15	17:30	Poster Fast Forward, Chairs: Alexander Lex, Michel Westenberg
17:30	19:30	**Poster Presentation** and **Reception** with Drinks and Finger Food

Saturday, July 11

8:30	9:30	Keynote: **Solving complicated problems in knowledge representation and visualization: cBioPortal, Pathway Commons, Precision Medicine,** Speaker: ***Chris Sander****, Sloan Kettering Genome Cancer Center, New York, USA*
9:30	10:45	BioVis Papers: **Proteins and Structures**, Session Chair: Jan Aerts
9:30	9:55	MoFlow: Visualizing Conformational Changes as Molecular Flow Improves Understanding*, W. Rumpf, S. Dabdoub, A. Shindhelm, W. Ray*
9:55	10:20	ReactionFlow: Visualizing Relationships between Proteins and Complexes in Biological Pathways, *T.N. Dang, P. Murray, A. Forbes, J. Aurisano*
10:20	10:45	Integrated visual analysis of protein structures, sequences, and feature data, *C. Stolte, K. S. Sabir, J. Heinrich, C. J. Hammang, A. Schafferhans, S. I. O'Donoghue*
10:45	11:15	*Coffee/Tea Break*
11:15	12:15	**Challenges Session**, Session Chair: Kay Nieselt, Sean O'Donoghue
11:15	11:45	**New Advanced Solutions for Genomic Big Data Analysis and Visualization**, Speaker: ***Nacho Medina***, Computational Biology Lab, University of Cambridge, UK
11:45	12:15	**Visualizing 3D Genomes,** Speaker: ***Marc Marti-Renom*****,** Genome Biology Group, Barcelona, Spain
12:15	14:00	*Lunch Break with Lunch Box and Beverages*
14:00	15:40	BioVis Papers: **Omics Data Visualization**, Session Chair: Nils Gehlenborg
14:00	14:25	miRTarVis: An Interactive Visual Analysis Tool for microRNA-mRNA Expression Profile Data, *D. Jung, B. Kim, R. Freishtat, M. Giri, E. Hoffman, J. Seo*
14:25	14:50	VisRseq: R-based visual framework for analysis of sequencing data, *H. Younesy, T. Möller, M. C. Lorincz, M. M. Karimi, S. J. M. Jones*
14:50	15:15	Epiviz: a view inside the design of an integrated visual analysis software for genomics, *F. Chelaru, H. Corrada Bravo*
15:15	15:40	XCluSim: A visual analytics tool for interactively comparing multiple clustering results of bioinformatics data, *S. L'Yi, B. Ko, D. Shin, Y.-J. Cho, J. Lee, B. Kim, J. Seo*
15:40	16:00	*Coffee/Tea Break*
16:00	17:00	Design Contest Presentations, Session Chairs: Eamonn Maguire and Ryo Sakai
17:00	17:30	Awards Ceremony and Closing Remarks, Liz Marai
17:30	18:30	Bring your own problem

Preface

The *Symposium on Biological Data Visualization (BioVis)*, established in 2011, is the premier international and interdisciplinary event for visualization in biology. This year BioVis is again a SIG at ISMB, the Intelligent Systems for Molecular Biology conference, held in July in Dublin, Ireland. The BioVis symposium is held on July 10-11, preceding the main conference.

Biology researchers face enormous challenges as they attempt to gain insight from large and highly complex data sets. Computational visualization and interaction techniques are essential to this process, and touch on all aspects of biology---from molecular to cell, tissue, organism, and population biology. The symposium brings together researchers from the visualization, bioinformatics, and biology communities with the purpose of educating, inspiring, and engaging visualization researchers in problems in biological data visualization. The prime motivation for collocating BioVis with ISMB is to allow bioinformaticians and biologists attending the main ISMB conference an in-depth introduction into state-of-the-art biological data visualization and foster communication and exchange of ideas between researchers from the different communities. The symposium also serves as a platform for researchers in biology and bioinformatics to share pressing visualization challenges and potential solutions in their fields; to initiate interdisciplinary collaborations; and to provide an outlet and training ground for junior visualization researchers with a keen interest in biological problems.

Given the goals of the meeting, the symposium solicited submissions in three categories: (i) *papers* (original, significant biodata visualization techniques); (ii) *posters*, describing work in progress and preliminary results, previously published work from journal venues, and visualization challenges relevant to the BioVis community; (iii) *data analysis and design contest entries*, consisting of short paper submissions and system demonstrations tackling a specific problem related to the domain of functional neuroimaging.

Papers

This year, the proceedings contain two different types of papers: (i) papers which are also accepted as journal papers in the thematic series on Biological Data Visualization of the journal BMC Bioinformatics; (ii) papers which were accepted for the conference only. From the 21 paper submissions, 14 were eventually accepted for the conference (67% acceptance rate) and 9 of these have also been accepted into the journal BMC Bioinformatics (43% acceptance rate).

This year the review process consisted of two rounds. In the first round, the submitted papers were assigned to Program Committee (PC) members and external reviewers, based on a match between paper topics and keywords and the declared expertise of the reviewers. Each paper was reviewed for novelty and contribution by at least two program committee members, and, on average, by two additional external reviewers. For each paper, experts from both the visualization and biology/bioinformatics communities were consulted during the review process.

When all reviews were completed, a discussion phase was initiated wherein reviewers of each paper could anonymously express their opinions with an opportunity to adjust their reviews and/or scores, and make an acceptance recommendation. The Program Chairs then finalized the decisions to either accept, conditionally accept, or reject the papers for the conference.

In the second review round, all accepted or conditionally accepted papers were judged for acceptance into BMC Bioinformatics or final acceptance into BMC Proceedings by the reviewers and the Paper Chairs, as well as by the respective BMC Editors.

Posters

Poster submissions were reviewed by the Poster Chairs for quality and relevance to the BioVis venue, with the goal of being as inclusive as possible and encouraging cross-domain interactions. 16 of the poster submissions were accepted for presentation at the symposium. The poster session includes 5 posters from the Data and Design Contest (see below).

Best paper and poster awards

To select the best paper, the Program Chairs have nominated four papers, based on the review scores and best paper recommendations of the reviewers. A best paper committee (BPC) consisting of three experts from the BioVis domain was assembled. Each member of the BPC independently ranked the selected papers with a short justification for each. Subsequently, the Program Chairs proposed a combined ranking of the selected papers that, along with the individual justifications, was circulated to the BPC. This process resulted in awarding the best paper prize to the highest-ranked paper, with an honorable mention for the paper that ranked second.

The best poster will be chosen during the BioVis 2015 symposium by a small panel of Program Committee members, to be recruited by the Posters Chairs; judging will be based on the quality of the presented posters and optional demos/videos on display during the poster session.

Data Analysis Contest

The BioVis Data Analysis contest brings the pressing needs of grand-challenge biological data analysis and visualization problems to the visualization community, and provides access to data and domain experts to support the visualization community in creating cutting-edge solutions to these important and pressing problems. Entries were judged by two panels, one composed of visualization experts and the other of biologists, with an overall winner chosen by consensus. Additional awards were given for entries deemed noteworthy by the judges.

Keynotes

BioVis 2015 features two Keynote talks.

The first BioVis 2015 Keynote Talk is titled “Visual data science -- Advancing science through visual reasoning” and will be given by **Torsten Moller**, professor at the **University of Vienna, Austria**. The talk discusses the latest visualization approaches and challenges in modeling and reasoning with uncertainty.

The second BioVis 2015 Keynote Talk is titled “Solving complicated problems in knowledge representation and visualization: cBioPortal, Pathway Commons, Precision Medicine” and will be given by **Chris Sander**, chair of the computational biology program at the **Memorial Sloan Kettering Cancer Center in New York**. The talk focuses on the challenges and advances in analyzing and simulating biological processes at different levels of organization.

Challenges

A challenges session will cover developments in a broad range of active research topics in modern biological data visualization.

Ignacio Medina (University of Cambridge) will talk about the challenges in genomic Big Data analysis and visualization.

Marc Martí-Renom (Centre Nacional d'Anàlis Genòmica, Spain) will talk about visualizing genome data in 3D, and will show how such visualizations can assist navigability, comprehension and discovery.

Primers

The program features two primer talks to introduce the visualization background and major visual design challenges that are relevant to biological data visualization. Since BioVis 2015 is a SIG at ISMB, the primer talks are aimed at making the BioVis paper presentations more accessible to researchers from biology and bioinformatics who do not have a background in visualization.

Min Chen (Oxford e-Research Centre, University of Oxford) will describe disseminative, observational, analytical and model-developmental visualization in the context of bioinformatics.

Robert Kosara (Tableau Software) will give an introduction to the basics of data visualization, show specific, common techniques and tools, and walk the audience through several examples.

Acknowledgments

Many people have contributed their time and energy to making the symposium a success. We thank the Paper Chairs, **Jan Aerts** and **Daniel Weiskopf**, as well as the Publication Chairs, **Cydney Nielsen** and **Marc Streit**, for their efforts and smooth collaboration. We thank the

Primer Chairs, **Miriah Meyer** and **Sean O'Donoghue**, as well as the Poster Chairs, **Alexander Lex** and **Michel Westenberg**, for their contribution. The energetic efforts of Contest Chairs **Eamonn Maguire, Ryo Sakai, Raghu Machiraju, William Ray,** and **Jason Bohland**, of the Challenges Chairs **Julian Heinrich** and **Greg Carter**, of the Industry and Fundraising Chairs **Cagatay Turkay, Harry Hochheiser,** and **Hanchuan Peng,** of the Publicity Chairs **Hendrik Strobelt** and **Andreas Hildebrandt**, and of the Website Chairs **Tengfei Yin** and **Tuan Dang**, are very much appreciated. We thank the Program Committee and the external reviewers for their timely, careful, considered, and balanced feedback. We all greatly benefited from the online submission and reviewing system provided by Precision Conference Solutions (PCS) and the steady support of James Stewart, as well as the help of the BMC section editor and editorial team for coordinating the publication in the journal BMC Bioinformatics and in the BMC Proceedings.

We thank our BioVis 2015 supporters **Autodesk** (platinum sponsor) and **Battelle** (bronze sponsor) for their generous contributions, and acknowledge the support of **BMC**. Finally, we thank the Steering Committee— **Larry Hunter, Jessie Kennedy, Nils Gehlenborg, Raghu Machiraju,** and **Jos Roerdink**—for their gentle guidance and feedback. It has been a great opportunity and privilege for all of us to work as a team in so many ways to make the Fifth BioVis symposium a success.

Kay Nieselt and G. Elisabeta Marai
General Chairs, 5th Symposium on Biological Data Visualization

Keynote Speakers

Torsten Möller

Head of research group of Visualization and Data Analysis
University of Vienna, Austria

Visual data science -- Advancing science through visual reasoning

Modern science is driven by computers (computational science) and data (data-driven science). While visual analysis has always been an integral part of science, in the context of computational science and data-driven science it has gained new importance. In this talk I will demonstrate novel approaches in visualization to support the process of modeling and simulations. Especially, I will report on some of the latest approaches and challenges in modeling and reasoning with uncertainty. Visual tools for ensemble analysis, sensitivity analysis, and the cognitive challenges during decision making build the basis of an emerging field of visual data science which is becoming an essential ingredient of computational thinking.

Biography

Torsten Möller is a professor at the University of Vienna, Austria, since 2013. Between 1999 and 2012 he served as a Computing Science faculty member at Simon Fraser University, Canada. He received his PhD in Computer and Information Science from Ohio State University in 1999 and a Vordiplom (BSc) in mathematical computer science from Humboldt University of Berlin, Germany. He is a senior member of IEEE and ACM, and a member of Eurographics. His research interests include algorithms and tools for analyzing and displaying data with principles rooted in computer graphics, image processing, visualization and human-computer interaction.

He heads the research group of Visualization and Data Analysis. He served as the appointed Vice Chair for Publications of the IEEE Visualization and Graphics Technical Committee (VGTC) between 2003 and 2012. He has served on a number of program committees and has been papers co-chair for IEEE Visualization, EuroVis, Graphics Interface, and the Workshop on Volume Graphics as well as the Visualization track of the 2007 International Symposium on Visual Computing. He has also co-organized the 2004 Workshop on Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration as well as the 2010 Workshop on Sampling and Reconstruction: Applications and Advances at the Banff International Research Station, Canada. He is a co-founding chair of the Symposium on Biological Data Visualization (BioVis). In 2010, he was the recipient of the NSERC DAS award. He received best paper awards from IEEE Conference on Visualization (1997), Symposium on Geometry Processing (2008), and EuroVis (2010), as well as two second best paper awards from EuroVis (2009, 2012).

Chris Sander

Computational Biosystems, MSKCC

Solving complicated problems in knowledge representation and visualization: cBioPortal, Pathway Commons, Precision Medicine

Biography

Chris Sander is the chair of the computational biology program at the Memorial Sloan Kettering Cancer Center in New York. His research focuses on analyzing and simulating biological processes at different levels of organization. This includes work on the identification of oncogenically altered pathways from genomic and molecular profiling in cancer, algorithms for the analysis of cancer genomics data, design of combinatorial cancer therapy, drug target identification, knowledge representation of biological pathways, protein evolution, specificity in protein networks, and the function of small RNAs.

Primers Speakers

Min Chen

BSc, PhD, FBCS, FEG, FLSW
Fellow of Pembroke College
Oxford e-Research Centre
University of Oxford

Four Levels of Visualization

In this talk, the speaker will discuss four levels of visualization in the context of bioinformatics. The four levels are disseminative, observational, analytical and model-developmental visualization, which reflect different visualization tasks as well as different levels of complexity. The speaker will also briefly discuss the latest development of an information-theoretic measure for optimizing the cost-benefit ratio of a data analysis and visualization process.

Biography

Min Chen developed his academic career in Wales between 1984 and 2011. He is currently the professor of scientific visualization at Oxford University and a fellow of Pembroke College. His research interests include visualization, computer graphics and human-computer interaction. He has co-authored some 160 publications, including his recent contributions in areas of volume graphics, video visualization, face modelling, automated visualization and theory of visualization. He has been awarded over £11M research grants from EPSRC, JISC (AHRC), TSB (NERC), Royal Academy, Welsh Assembly Government, HEFCW, Industry, and several UK and US Government Agencies. He is currently leading visualization activities at Oxford e-Research Centre, with a team of researchers working on a broad spectrum of interdisciplinary research topics, ranging from the sciences to sports, and from digital humanities to cybersecurity. His services to the research community include papers co-chair of IEEE Visualization 2007 and 2008, Eurographics 2011, IEEE VAST 2014 and 2015; co-chair of Volume Graphics 1999 and 2006, EuroVis 2014; associate editor-in-chief of IEEE Transactions on Visualization and Computer Graphics; and co-director of Wales Research Institute of Visual Computing. He is a fellow of British Computer Society, European Computer Graphics Association, and Learned Society of Wales.

Robert Kosara

Research Scientist
Tableau Software

How to Visualize Data

Visualization of data isn't just about creating pretty pictures for the cover of Nature, but really about being able to perceive and understand data. I will give a brief introduction to the basics of data visualization, show some common techniques and tools, and walk you through some examples. Presentation will not be ignored entirely, though it has to be preceded by exploration and analysis of the data.

Biography

Robert Kosara is a research scientist at Tableau Software. His focus is on the communication of data through visualization and visual storytelling. Robert is also working on furthering our understanding of visual perception and cognition, so we can make data easier to understand and develop tools to communicate it more effectively. His full list of publications can be found on his vanity website.

Before joining Tableau in 2012, Robert was Associate Professor of Computer Science at The University of North Carolina at Charlotte. Robert received his M.Sc. and Ph.D. Degrees in Computer Science from Vienna University of Technology (Austria). In his copious spare time, Robert likes to run long distances and writing articles for his website, eagereyes.org.

Challenges Speakers

Ignacio Medina

Head of Computational Biology Lab
University of Cambridge

New Advanced Solutions for Genomic Big Data Analysis and Visualization

Biography

Ignacio Medina received his BSc in Biochemistry and Molecular Biology from the University of Valencia, Spain, in 2001, and the BSc in Computer Science from the Polytechnic University of Valencia, in 2004. In 2009 he obtained the MSc in Genetics from University of Valencia. In 2006, he joined the Department of Bioinformatics and Genomics at the Prince Felipe Research Center, as a Bioinformatician and Researcher, and in 2010 became a Project Manager of several clinical and software development projects. In 2014, he joined EMBL-EBI as a Project Manager and Senior Software Architect at EBI Variation Team. In 2015, he joined University of Cambridge as a Head of Computational Biology Lab at HPC Service. His current research interests include HPC and big data software development for genomic-scale data analysis and visualization. He participates in several international projects such as EGA or Genomics England. He has authored more than 35 papers in international peer-reviewed journals.

Marc Martí-Renom

ICREA Research Professor
Genome Biology Group, Centre Nacional d'Anàlis Genòmica (CNAG, http://cnag.org), Barcelona, Spain.
Structural Genomics Group, Centre de Regulació Genòmica (CRG, http://crg.cat), Barcelona, Spain.

Visualizing 3D Genomes

The genome is conventionally represented and explored as a linear sequence, however chromatin is dynamic. The spatial organization and dynamics of the chromatin can be computationally modeled. The extent, detail and utility of these types of models present challenges in representing and interacting with such large, multi-scale models. Unfortunately, current genome browsers have limited overview 3D spatial data and do not provide an adequate end-user experience while browsing such generated models. New web-based, open-source, cross-platform technologies provide now the opportunity to address these issues with increasing ease, stability and security. Despite the limitations in three-dimensional (3D) representation due to occlusion, subjectivity and interpretation inherent in human system, 3D visualizations can assist navigability, comprehension and discovery by leveraging human visual processing, spatial reasoning and creativity. We will describe a new Web-based browser that aims at visualizing the genome from all its

dimensions (linear to 3D). Our approach, called TADkit, aims at integrating ‘1D’ genome sequence, its ‘2D’ aligned annotations and its ‘3D’ models to give a more complete vision of the forms and interactions.

Biography

Marc Marti-Renom obtained a Ph.D. in Biophysics from the UAB where he worked on protein folding under the supervision of Professors B. Oliva, F.X. Avilés and M. Karplus. He conducted postdoctoral training on protein structure modeling at the Sali Lab (Rockefeller University) as the recipient of the Burroughs, Wellcome Fund fellowship and was then appointed Assistant Adjunct Professor at UCSF. Between 2006 and 2011, Prof. Marti-Renom headed the Structural Genomics Group at the CIPF in Valencia (Spain). Currently, he is a ICREA research professor and leads the Genome Biology Group at the National Center for Genomic Analysis (CNAG) and the Structural Genomics Group at the Centre for Genomic Regulation (CRG), both in Barcelona. His group is broadly interested on how RNA, proteins and genomes organize and regulate cell fate. He is an Associate Editor of the PLoS Computational Biology and BMC Structural Biology journals and has published over 75 articles in international peer-reviewed journals.

BioVis 2015 Committees

Organizing committee

General chairs
Kay Nieselt, University of Tübingen, Germany
G. Elisabeta Marai, University of Illinois at Chicago, USA

Paper chairs
Daniel Weiskopf, Univ of Stuttgart, Germany
Jan Aerts, Leuven University, Belgium

Poster chairs
Alexander Lex, Harvard University, USA
Michel Westenberg, Eindhoven University of Technology, The Netherlands

Publication chairs
Cydney Nielsen, University of British Columbia, Canada
Marc Streit, Johannes Kepler University Linz, Austria

Primer/Tutorial chairs
Miriah Meyer, University of Utah, USA
Seán O'Donoghue, Garvan Institute/CSIRO, Australia

Challenges chairs
Julian Heinrich, CSIRO, Australia
Greg Carter, The Jackson Laboratory, USA

Data Contest chairs
Raghu Machiraju, The Ohio State University, USA
William Ray, The Ohio State University, USA
Jason Bohland, Boston University, USA

Design Contest chairs
Eamonn Maguire, CERN, Switzerland
Ryo Sakai, Leuven University, Belgium

Industry and fundraising chairs
Cagatay Turkay, City University, London, UK
Harry Hochheiser, University of Pittsburgh, USA
Hanchuan Peng, Allen Institute for Brain Science, Seattle, USA

Website chairs
Tengfei Yin, Seven Bridges Genomics, Cambridge, USA
Tuan Dang, University of Illinois at Chicago, USA

Publicity chairs
Hendrik Strobelt, Harvard University, USA
Andreas Hildebrandt, University of Mainz, Germany

Steering Committee

Nils Gehlenborg, Harvard Medical School, USA
Raghu Machiraju, The Ohio State University, USA (Chair)
Jos Roerdink, University of Groningen, The Netherlands
Jessie Kennedy, Edinburgh Napier University, UK
Larry Hunter, University of Colorado, USA

Program Committee

Danielle Albers Szafir , University of Wisconsin
Bilal Alsallakh, Vienna University of Technology
Chris Bartlett, Ohio State University
Bertjan Broeksema, Luxembourg Institute of Science and Technology
Stefan Bruckner, University of Bergen, Norway
Jian Chen, University of Maryland
Katja Bühler, VRVis Zentrum für Virtual Reality und Visualisierung
Çagatay Demiralp, Stanford University
Inna Dubchak, Lawrence Berkeley National Laboratory
David Duke, University of Leeds
Jim Faeder, University of Pittsburgh
Angus Forbes, University of Illinois at Chicago
Nils Gehlenborg, Harvard Medical School
Mohammad Ghoniem, Luxembourg Institute of Science and Technology
Nicholas Hamilton, University of Queensland
Helwig Hauser, University of Bergen
Julian Heinrich, CSIRO
Harry Hochheiser, University of Pittsburgh
Larry Hunter, University of Colorado
Radu Jianu, Florida International University
Fabrice Jossinet, University of Strasbourg
Andreas Kerren, Linnaeus University
Robert Kincaid, Agilent Laboratories
Karsten Klein, Monash University
Robert Kuhn, University of California Santa Cruz
Alexander Lex, Harvard University
Lars Linsen, Jacobs University Bremen
Lydia Müller, University of Leipzig
Raghu Machiraju, The Ohio State University
Eamonn Maguire, CERN Geneva
Han Mulder, Wageningen University
Tamara Munzner, University of British Columbia
Tim Nattkemper, Bielefeld University
Sergiy Nesterko, Harvard University
Benoît Otjacques, Luxembourg Institute of Science and Technology
Julius Parulek, University of Bergen
Jürgen Pleiss, University of Stuttgart
Yann Ponty, Ecole Polytechnique, University of Paris-Saclay
Bernhard Preim, University of Magdeburg
A. Johannes Pretorius, The University of Leeds
James Procter, University of Dundee
Jos Roerdink, University of Groningen
Francis Rowland, European Bioinformatics Institue
Ryo Sakai, KU Leuven
Reinhard Schneider, University of Luxembourg
Falk Schreiber, Monash University
Joost Schymkowitz, VIB-KULeuven
Hendrik Strobelt, Harvard University

Cagatay Turkay, City University London
Jim Vallandingham, Bocoup
Corinna Vehlow, University of Stuttgart
Toni Verbeiren, KU Leuven
Ivan Viola, Vienna University of Technology
Ting Wang, Washington University in St. Louis
Tengfei Yin, Seven Bridges Genomics
Dirk Zeckzer, University of Leipzig
Simone Zorzan, Luxembourg Institute of Science and Technology

Abstracts of Papers

VisRseq: R-based visual framework for analysis of sequencing data

Hamid Younesy, Torsten Möller, Matthew C. Lorincz, Mohammad M. Karimi, Steven J. M. Jones

Background: Several tools have been developed to enable biologists to perform initial browsing and exploration of sequencing data. However the computational tool set for further analyses often requires significant computational expertise to use and many of the biologists with the knowledge needed to interpret these data must rely on programming experts.

Results: We present VisRseq, a framework for analysis of sequencing datasets that provides a computationally rich and accessible framework for integrative and interactive analyses without requiring programming expertise. We achieve this aim by providing R apps, which offer a semi-auto generated and unified graphical user interface for computational packages in R and repositories such as Bioconductor. To address the interactivity limitation inherent in R libraries, our framework includes several native apps that provide exploration and brushing operations as well as an integrated genome browser. The apps can be chained together to create more powerful analysis workflows.

Conclusions: To validate the usability of VisRseq for analysis of sequencing data, we present two case studies performed by our collaborators and report their workflow and insights.

Pan-Tetris: an interactive visualisation for Pan-genomes

André Hennig, Jörg Bernhardt, Kay Nieselt

Background: Large-scale genome projects have paved the way to microbial pan-genome analyses. Pan-genomes describe the union of all genes shared by all members of the species or taxon under investigation. They offer a framework to assess the genomic diversity of a given collection of individual genomes and moreover they help to consolidate gene predictions and annotations. The computation of pan-genomes is often a challenge, and many techniques that use a global alignment-independent approach run the risk of not separating paralogs from orthologs. Also alignment-based approaches which take the gene neighbourhood into account often need additional manual curation of the results. This is quite time consuming and so far there is no visualisation tool available that offers an interactive GUI for the pan-genome to support curating pan-genomic computations or annotations of orthologous genes.

Results: We introduce Pan-Tetris, a Java based interactive software tool that provides a clearly structured and suitable way for the visual inspection of gene occurrences in a pan-genome table. The main features of Pan-Tetris are a standard coordinate based presentation of multiple genomes complemented by easy to use tools compensating for algorithmic weaknesses in the pan-genome generation workflow. We demonstrate an application of Pan-Tetris to the pan-genome of Staphylococcus aureus.

Conclusions: Pan-Tetris is currently the only interactive pan-genome visualisation tool. Pan-Tetris is available from http://bit.ly/1vVxYZT

Epiviz: a view inside the design of an integrated visual analysis software for genomics

Florin Chelaru, Héctor Corrada Bravo

Background: Computational and visual data analysis for genomics has traditionally involved a combination of tools and resources, of which the most ubiquitous consist of genome browsers, focused mainly on integrative visualization of large numbers of big datasets, and computational environments, focused on data modeling of a small number of moderately sized datasets. Workflows that involve the integration and exploration of multiple heterogeneous data sources, small and large, public and user specific have been poorly addressed by these tools. In our previous work, we introduced Epiviz, which bridges the gap between the two types of tools, simplifying these workflows.

Results: In this paper we expand on the design decisions behind Epiviz, and introduce a series of new advanced features that further support the type of interactive exploratory workflow we have targeted. We discuss three ways in which Epiviz advances the field of genomic data analysis: 1) it brings code to interactive visualizations at various different levels; 2) takes the first steps in the direction of collaborative data analysis by incorporating user plugins from source control providers, as well as by allowing analysis states to be shared among the scientific community; 3) combines established analysis features that have never before been available simultaneously in a genome browser. In our discussion section, we present security implications of the current design, as well as a series of limitations and future research steps.

Conclusions: Since many of the design choices of Epiviz are novel in genomics data analysis, this paper serves both as a document of our own approaches with lessons learned, as well as a start point for future efforts in the same direction for the genomics community.

XCluSim: A visual analytics tool for interactively comparing multiple clustering results of bioinformatics data

Sehi L'Yi, Bongkyung Ko, DongHwa Shin, Young-Joon Cho, Jaeyong Lee, Bohyoung Kim, Jinwook Seo

Background: Though cluster analysis has become a routine analytic task for bioinformatics research, it is still arduous for researchers to assess the quality of a clustering result. To select the best clustering method and its parameters for a dataset, researchers have to run multiple clustering algorithms and compare them. However, such a comparison task with multiple clustering results is cognitively demanding and laborious.

Results: In this paper, we present XCluSim, a visual analytics tool that enables users to interactively compare multiple clustering results based on the Visual Information Seeking Mantra. We build a taxonomy for categorizing existing techniques of clustering results visualization in terms of the Gestalt principles of grouping. Using the taxonomy, we choose the most appropriate interactive visualizations for presenting individual clustering results from different types of clustering algorithms. The efficacy of XCluSim is shown through case studies with a bioinformatician.

Conclusions: Compared to other relevant tools, XCluSim enables users to compare multiple clustering results in a more scalable manner. Moreover, XCluSim supports diverse clustering algorithms and dedicated visualizations and interactions for different types of clustering results, allowing more effective exploration of details on demand. Through case studies with a bioinformatics researcher, we received positive feedback on the functionalities of XCluSim, including its ability to help identify stably clustered items across multiple clustering results.

BactoGeNIE: a large-scale comparative genome visualization for big displays

Jillian Aurisano, Khairi Reda, Andrew Johnson, G.Elisabeta Marai, Jason Leigh

Background: The volume of complete bacterial genome sequence data available to comparative genomics researchers is rapidly increasing. However, visualizations in comparative genomics—which aim to enable analysis tasks across collections of genomes—suffer from visual scalability issues. While large, multi-tiled and high-resolution displays have the potential to address scalability issues, new approaches are needed to take advantage of such environments, in order to enable the effective visual analysis of large genomics datasets.

Results: In this paper, we present Bacterial Gene Neighborhood Investigation Environment, or BactoGeNIE, a novel and visually scalable design for comparative gene neighborhood analysis on large display environments. We evaluate BactoGeNIE through a case study on close to 700 draft Escherichia coli genomes, and present lessons learned from our design process.

Conclusions: BactoGeNIE accommodates comparative tasks over substantially larger collections of neighborhoods than existing tools and explicitly addresses visual scalability. Given current trends in data generation, scalable designs of this type may inform visualization design for large-scale comparative research problems in genomics.

Integrated visual analysis of protein structures, sequences, and feature data

Christian Stolte, Kenneth S. Sabir, Julian Heinrich, Christopher J. Hammang, Andrea Schafferhans, Séan I. O'Donoghue

Background: To understand the molecular mechanisms that give rise to a protein's function, biologists often need to (i) find and access all related atomic-resolution 3D structures, and (ii) map sequence-based features (e.g., domains, single-nucleotide polymorphisms, post-translational modifications) onto these structures.

Results: To streamline these processes we recently developed Aquaria, a resource offering unprecedented access to protein structure information based on an all-against-all comparison of SwissProt and PDB sequences. In this work, we provide a requirements analysis for several frequently occuring tasks in molecular biology and describe how design choices in Aquaria meet these requirements. Finally, we show how the interface can be used to explore features of a protein and gain biologically meaningful insights in two case studies conducted by domain experts.

Conclusions: The user interface design of Aquaria enables biologists to gain unprecedented access to molecular structures and simplifies the generation of insight. The tasks involved in mapping sequence features onto structures can be conducted easier and faster using Aquaria.

Physically-based in silico light sheet microscopy for visualizing fluorescent brain models

Marwan Abdellah, Ahmet Bilgili, Stefan Eilemann, Henry Markram, Felix Schürmann

Background: We present a physically-based computational model of the light sheet fluorescence microscope (LSFM). Based on Monte Carlo ray tracing and geometric optics, our method simulates the operational aspects and image formation process of the LSFM. This simulated, in silico LSFM creates synthetic images of digital fluorescent specimens that can resemble those generated by a real LSFM, as opposed to established visualization methods producing visually-plausible images. We also propose an accurate fluorescence rendering model which takes into account the intrinsic characteristics of fluorescent dyes to simulate the light interaction with fluorescent biological specimen.

Results: We demonstrate first results of our visualization pipeline to a simplified brain tissue model reconstructed from the somatosensory cortex of a young rat. The modeling aspects of the LSFM units are qualitatively analysed, and the results of the fluorescence model were quantitatively validated against the fluorescence brightness equation and characteristic emission spectra of different fluorescent dyes.

Visual parameter optimisation for biomedical image processing

A.J. Pretorius, Y. Zhou, R.A. Ruddle

Background: Biomedical image processing methods require users to optimise input parameters to ensure high-quality output. This presents two challenges. First, it is difficult to optimise multiple input parameters for multiple input images. Second, it is difficult to achieve an understanding of underlying algorithms, in particular, relationships between input and output.

Results: We present a visualisation method that transforms users' ability to understand algorithm behaviour by integrating input and output, and by supporting exploration of their relationships. We discuss its application to a colour deconvolution technique for stained histology images and show how it enabled a domain expert to identify suitable parameter values for the deconvolution of two types of images, and metrics to quantify deconvolution performance. It also enabled a breakthrough in understanding by invalidating an underlying assumption about the algorithm.

Conclusions: The visualisation method presented here provides analysis capability for multiple inputs and outputs in biomedical image processing that is not supported by previous analysis software. The analysis supported by our method is not feasible with conventional trial-and-error approaches.

GRAPHIE: graph based histology image explorer

Hao Ding, Chao Wang, Kun Huang, Raghu Machiraju

Background: Histology images comprise one of the important sources of knowledge for phenotyping studies in systems biology. However, the annotation and analyses of histological data have remained a manual, subjective and relatively low-throughput process.

Results: We introduce Graph based Histology Image Explorer (GRAPHIE)–a visual analytics tool to explore, annotate and discover potential relationships in histology image collections within a biologically relevant context. The design of GRAPHIE is guided by domain experts' requirements and well-known InfoVis mantras. By representing each image with informative features and then subsequently visualizing the image collection with a graph, GRAPHIE allows users to effectively explore the image collection. The features were designed to capture localized morphological properties in the given tissue specimen. More importantly, users can perform feature selection in an interactive way to improve the visualization of the image collection and the overall annotation process. Finally, the annotation allows for a better prospective examination of datasets as demonstrated in the users study. Thus, our design of GRAPHIE allows for the users to navigate and explore large collections of histology image datasets.

Conclusions: We demonstrated the usefulness of our visual analytics approach through two case studies. Both of the cases showed efficient annotation and analysis of histology image collection.

miRTarVis: An interactive visual analysis tool for microRNA-mRNA expression profile data

Daekyoung Jung, Bohyoung Kim, Robert J. Freishtat, Mamta Giri, Eric Hoffman, Jinwook Seo

Background: MicroRNAs (miRNA) are short nucleotides that down-regulate its target genes. Various miRNA target prediction algorithms have used sequence complementarity between miRNA and its targets. Recently, other algorithms tried to improve sequence-based miRNA target prediction by exploiting miRNA-mRNA expression profile data. Some web-based tools are also introduced to help researchers predict targets of miRNAs from miRNA-mRNA expression profile data. A demand for a miRNA-mRNA visual analysis tool that features novel miRNA prediction algorithms and more interactive visualization techniques exists.

Results: We designed and implemented miRTarVis, which is an interactive visual analysis tool that predicts targets of miRNAs from miRNA-mRNA expression profile data and visualizes the resulting miRNA-target interaction network. miRTarVis has intuitive interface design in accordance with the analysis procedure of load, filter, predict, and visualize. It predicts targets of miRNA by adopting Bayesian inference and MINE analyses, as well as conventional correlation and mutual information analyses. It visualizes a resulting miRNA-mRNA network in an interactive Treemap, as well as a conventional node-link diagram. miRTarVis is available at hcil.snu.ac.kr/~rati/miRTarVis/index.html.

Conclusions: We reported findings from miRNA-mRNA expression profile data of asthma patients using miRTarVis in a case study. miRTarVis helps to predict and understand targets of miRNA from miRNA-mRNA expression profile data.

PathwayMatrix: Visualizing binary relationships between proteins in biological pathways

Tuan Nhon Dang, Paul Murray, Angus Graeme Forbes

Background: Molecular activation pathways are inherently complex, and understanding relations across many biochemical reactions and reaction types is difficult. Visualizing and analyzing a pathway is a challenge due to the network size and the diversity of relations between proteins and molecules.

Results: In this paper, we introduce PathwayMatrix, a visualization tool that presents the binary relations between proteins in the pathway via the use of an interactive adjacency matrix. We provide filtering, lensing, clustering, and brushing and linking capabilities in order to present relevant details about proteins within a pathway.

Conclusions: We evaluated PathwayMatrix by conducting a series of in-depth interviews with domain experts who provided positive feedback, leading us to believe that our visualization technique could be helpful for the larger community of researchers utilizing pathway visualizations. PathwayMatrix is freely available at https://github.com/CreativeCodingLab/.

Extended LineSets: A visualization technique for the interactive inspection of biological pathways

Francesco Paduano, Angus Graeme Forbes

Background: Biologists make use of pathway visualization tools for a range of tasks, including investigating inter-pathway connectivity and retrieving details about biological entities and interactions. Some of these tasks require an understanding of the hierarchical nature of elements within the pathway or the ability to make comparisons between multiple pathways. We introduce a technique inspired by LineSets that enables biologists to fulfill these tasks more effectively.

Results: We introduce a novel technique, Extended LineSets, to facilitate new explorations of biological pathways. Our technique incorporates intuitive graphical representations of different levels of information and includes a well-designed set of user interactions for selecting, filtering, and organizing biological pathway data gathered from multiple databases.

Conclusions: Based on interviews with domain experts and an analysis of two use cases, we show that our technique provides functionality not currently enabled by current techniques, and moreover that it helps biologists to better understand both inter-pathway connectivity and the hierarchical structure of biological elements within the pathways.

MoFlow: Visualizing conformational changes in molecules as molecular flow improves understanding

Shareef M. Dabdoub, R. Wolfgang Rumpf, Amber D. Shindhelm, William C. Ray

Background: Current visualizations of molecular motion use a Timeline-analogous representation that conveys “first the molecule was shaped like this, then like this...”. This scheme is orthogonal to the Pathline-like human understanding of motion “this part of the molecule moved from here to here along this path”. We present MoFlow, a system for visualizing molecular motion using a Pathline-analogous representation.

Results: The MoFlow system produces high- quality renderings of molecular motion as atom pathlines, as well as interactive WebGL visualizations, and 3D printable models. In a preliminary user study, MoFlow representations are shown to be superior to canonical representations for conveying molecular motion.

Conclusions: Pathline-based representations of molecular motion are more easily understood than timeline representations. Pathline representations provide other advantages because they represent motion directly, rather than representing structure with inferred motion.

ReactionFlow: An interactive visualization tool for causality analysis in biological pathways

Tuan Nhon Dang, Paul Murray, Jillian Aurisano, Angus Graeme Forbes

Background: Molecular and systems biologists are tasked with the comprehension and analysis of incredibly complex networks of biochemical interactions, called pathways, that occur within a cell. Through interviews with domain experts, we identified four common tasks that require an understanding of the causality within pathways, that is, the downstream and upstream relationships between proteins and biochemical reactions, including: visualizing downstream consequences of perturbing a protein; finding the shortest path between two proteins; detecting feedback loops within the pathway; and identifying common downstream elements from two or more proteins.

Results: We introduce ReactionFlow, a visual analytics application for pathway analysis that emphasizes the structural and causal relationships amongst proteins, complexes, and biochemical reactions within a given pathway. To support the identified causality analysis tasks, user interactions allow an analyst to filter, cluster, and select pathway components across linked views. Animation is used to highlight the flow of activity through a pathway.

Conclusions: We evaluated ReactionFlow by providing our application to two domain experts who have significant experience with biomolecular pathways, after which we conducted a series of in-depth interviews focused on each of the four causality analysis tasks. Their feedback leads us to believe that our techniques could be useful to researchers who must be able to understand and analyze the complex nature of biological pathways. ReactionFlow is available at https://github.com/CreativeCodingLab/ReactionFlow.

Abstracts of Posters

A new hierarchical 3D random walk visualization for whole chromosome sequences

HWAN GUE CHO, DAE KEON KWON, HO YOUL JUNG, BYUNG CHUL KANG, Justin CHOI

Graphical techniques are a very powerful tool for the visualization and analysis of long DNA sequences by providing useful insights into local and global characteristics and the occurrences, variations and repetition of the genomic sequences. Within these methods, the random walk plot is a basic technique to visualize a huge genome on 2D or 3D geometric space. One problem in these geometric random plotting is the loss of genetic information due to the complementary nucleotide base pairs. In this work we propose a new 3D random walk plotting method to suppressing complementary base pair cancelation effect by allowing z-axis relaxation. Since the final “geometric genomes” constructed on 3D space is of similar complexity as fractal objects, it is necessary to simplify them into a manageable data structure. So we give one simplification algorithm for the geometrically visualized long sequences with alpha-hull polygon modeling. So we can compare the geometric similarity of these simplified polygons which are constructed from more than 100 mega base DNA sequences, which enables us to recognize the partial matching and the global similarity more intuitively without using any computation-intensive sequence alignment. Our experiment showed that our method is quite fast and effective in comparing whole chromosomes of more than 10 mammals.

Novel approach to modelling protein-protein interactions in biological space

Benedetta Frida Baldi

The main aim of this work was to develop an accurate and detailed method to simulate protein-protein interactions and the formation of large complexes, and use it to study the properties of Post Synaptic Density (PSD) formation in a dendritic spine. The methodology is based on simulating accurate protein geometries diffusing in space and interacting only via their binding sites. Coarse grained representation of the protein surfaces were produced starting from their PDB file structure, using Autodesk Maya. Subsequently, Unity, a game engine was used to simulate diffusion and reaction of the proteins models in the simulation environment. Diffusion in 2/3D was implemented as mass driven, with explicit calculations of translation and rotation, and the 2D diffusion was implemented to be independent of the shape of the container, its concave nature, and its tessellation. The protein binding strategy was implemented based on collision theory. The method was fully validated for the 2D and 3D diffusion and for complex formation, and applied to the study of the Post-Synaptic Density (a proteinaceous organelle present on the membrane of excitatory spines) assembly formation.

Interactive Visualization of Provenance Graphs for Reproducible Biomedical Research

Stefan Luger, Samuel Gratzl, Holger Stitz, Marc Streit, Nils Gehlenborg

A major challenge of data-intensive biomedical research is the collection and representation of provenance information to ensure reproducibility of the studies. The Refinery Platform (http://www.refinery-platform.org) is an integrated data management, analysis, and visualization system designed to support reproducible biomedical research. Refinery stores each data set as a directed acyclic experiment graph that associates every file in the data set with meta data annotations as well as the analyses and input files that were used to create it. In order to communicate and reproduce multi-step analyses on data sets that contain data for hundreds of samples, it is crucial to be able to visualize the provenance graph at different levels of detail. Most existing approaches for provenance visualization are based on node-link diagrams, however, they usually do not scale well to due to the limitations of the this visualization approach. Our proposed visualization technique dynamically reduces the complexity of subgraphs through hierarchical aggregation and application of a degree-of-interest (DOI) function to each node. Triggered by user interactions such as filtering for a subset of analyses or highlighting of a path in the graph, subgraphs are dynamically aggregated into a glyph representation. We further reduce the complexity of the provenance graph visualization by layering identical or similar sequences of parallel analysis steps into an aggregate sequence. We have implemented our approach in Refinery and our initial results are very promising. Future work will focus on the fine-tuning of the DOI function and comphrehensive user testing.

3D Visualization of Molecular Data using a Cloud Streaming Framework for Web-based Large Model Viewing

Merry Shiyu Wang, Michael Zyracki, Andrew Kimoto, Malte Tinnus, Florencio Mazzoldi

3D visualization of large and multi-scale biological data, from macro-molecular structures to whole organisms, is integral to building predictive models for biomedical research. However, existing industry standard desktop, plug-in, and web-based applications can easily push beyond the limits of advanced in-core processing and rendering. Autodesk Research Molecular Viewer is designed to overcome limitations of scalability, capability, accessibility, collaboration, and outreach for molecular datasets. Layered on an extensible cloud-based visualization framework optimized for large 3D models and data, the Viewer demonstrates common usage scenarios of molecule interaction by visualizing RCSB Protein Data Bank files directly in the web browser. The technology leverages a scalable, cloud-based translation and streaming pipeline, and utilizes three.js on webGL-enabled browsers to support a distributed data model. A root manifest streams in smaller component parts, which render as required, to absolve the need for entire model storage in memory. The Molecular Viewer visualization framework can be extended to other 3D data, enabling exploration of large-scale biomedical models.

Active Data Canvas: linking data to knowledge

Joon-Yong Lee, Ryan Wilson, Richard D. Smith, Nick Cramer, Samuel H Payne

Intuitive data visualization is critical for biomedical data analysis; especially essential is the ability for non-computational scientists to productively browse the data and apply their domain knowledge. We present Active Data Canvas, a web-based visual analytic tool suite that allows dynamic data interaction in familiar and intuitive contexts, like pathways and heatmaps. Users can track emerging hypotheses by pinning data to the Canvas, a Pinterest style board for aggregating and assimilating thoughts. As data are pinned, the Canvas proactively searches structured and unstructured knowledge bases (e.g. KEGG, Pubmed, Wikipedia etc.) to fetch relevant information. In this manner it becomes a productive digital assistant, lessening the burden on the researcher and freeing time for contemplative analysis. The Active Data Canvas promotes collaboration by using GitHub as the back-end data store. Via the ambitious task of versioning data, analyses and emergent hypotheses, users who collaborate on a project can share discoveries in real time. We used the Active Data Canvas to analyze ovarian cancer data for 174 tumors. After quality control and initial analysis, the R object was directly imported along with the clinical metadata. New hypotheses and conclusions were discovered via the interaction with the Active Data Canvas that had not been discovered in the months of analysis via statistical programming scripts. \

BioTapestry: New development supports integrated visualization of computational results

Kalle Leinonen, Suzanne Paquette, William Longabaugh

BioTapestry is a well-established tool for building, visualizing, and sharing models of gene regulatory networks (GRNs), with particular emphasis on the GRNs that drive development. BioTapestry features a hierarchical modeling framework that presents multiple views of the network at different levels of spatial and temporal resolution. This approach is well-suited to depicting the GRNs of an embryo that increases in spatial complexity over time. The most recent public release of BioTapestry, Version 7, can now be run as a web application which allows interactive, graphical GRN models to be viewed directly in a web browser. Researchers who wish to locally host an interactive GRN model can download and install the BioTapestry Web Application on a supported web server. Additionally, we are now hosting a set of published models of developmental GRNs on our new website, http://grns.BioTapestry.org/. For the upcoming Version 8, we are creating a new plugin framework that allows computational modeling tools to be integrated into BioTapestry to drive dynamic network visualizations. Simple computational models with a limited number of generic parameters (e.g. Boolean models) can help guide the process of choosing and refining the GRN model that best describes observed data. This new framework will allow the user to visualize the results of simple computational modeling of GRN behaviors in a network context, including direct visual comparisons between experimental data and model predictions while viewing the network model.

Polimero-bio - composable biological visualizations using web components

Daniel Alcaide, Ryo Sakai, Raf Winand, Toni Verbeiren, Thomas Moerman, Jan Aerts

As a visual data analysis lab, we often combine (brush/link) well-known data visualization techniques (scatterplots, barcharts, ...). Although this is possible in general purpose tools like Tableau, the specifics of the biological domain often require the development of custom visuals. This leads to the issue that we end up reimplementing the base visuals over and over if we want to build them into a specific analysis tool. Here, we present a proof-of-principle framework for creating composable linked data visualizations, including an initial collection of parsers and visuals with an emphasis on biology.

With polimero and polimero-bio, we want to create a scalable framework for building domain-specific visual data exploration tools using a collection of D3-based reusable components. Polimero is based on the emerging W3C-standard of web components (as enabled through Polymer; www.polymer-project.org), which allows for creating custom elements, HTML templates, shadow DOM and HTML imports. This makes it possible to create applications that are composable, encapsulated, and reusable. This is valuable both for the developer/designer who can easily create and plug-in custom visual encodings, and for the end-user who can create linked visualizations by dragging existing components onto a canvas using the polimero-designer.

Polimero and polimero-bio are available at http://bitbucket.org/vda-lab/polimero.

DeskGen | A novel platform for CRISPR gene editing

Kristian Kancleris, Victor Dillard, Riley Doyle, Edward Perello, Leigh Brody, Joseph Wolanski, Matt Couch, Neil Humphreys, Ben Corser

In this work we present a novel interactive approach to facilitate CRISPR genome editing. Our application combines extensive guide RNA (gRNA) visualisation and scoring tools with the information-rich environment of a genome browser. The application simultaneously displays data for: genes, transcripts, exons, introns, transcription start sites, gRNAs, PAM sites, on-target activity scores and off-target activity data. Our application was built through ethnographic study of gene editing scientists, and is the first tool that has been purpose-built to visualise all of the information necessary for a scientist to make an informed decision on a CRISPR gene editing strategy.

AnnoPeak: A visualization tool for Chip-Seq and Chip-Exo Analysis

Xing Tang, Arunima Srivastava, Gustavo Leone, Kun Huang, Raghu Machiraju

Chip-Seq and its improved resolution protocol Chip-Exo experiments, both describe location and strength of protein to DNA binding. These experiments are utilized in isolation as well as in a complementary fashion with microarray and RNA-Seq data to characterize binding events, and explore gene regulation and biological pathways. AnnoPeak aims to be a “one click”, simple to use, exploratory tool that efficiently generates sophisticated analyses of multiple experiments by annotating data with critical knowledge and utilizes a variety of different visualization techniques to characterize distinct experimental profiles. Current features include: binding peak’s genetic annotation, extraction of peaks from user selected regions, gene ontology enrichment of region specific binding, expedient linking to visualization of peaks on the UCSC genome browser, overlay of gene expression and alias, motif and ortholog based analyses. Primary visualizations constructed by AnnoPeak include histograms of peak density distribution, color coded and comparable representation of read depth around transcription start sites for multiple experiments, plots describing average peak shape, which can be readily normalized to compare binding profile in multiple experimental results, and graphs describing quality of replicates and data. Visual comparison of different experimental conditions by mapping colors to experiments and intensity gradients provides an easily interpretable summary of the data for downstream analysis. Since whole genome mapping using Chip-Seq is one of the most informative high throughput experiments in biology, a tool like AnnoPeak will provide a powerful utility for genetic exploration.

A Novel Tool for Isoform Visualization

Hendrik Strobelt, Bilal Alsallakh, Joseph Botros, Brant Peterson, Mark Borowsky, Hanspeter Pfister, Alexander Lex

Protein isoforms can be assembled from the same DNA sequence by means of alternative splicing, which selectively omits some of the coding regions (exons) associated with a gene. Detecting alternative splicing requires advanced data acquisition methods such as RNA-seq, as well as dedicated statistical inference methods such as MISO, TopHat, and RSEM. Analyzing the abundance of the corresponding isoforms under different conditions is important to understand and differentiate between normal processes and diseases, which helps in developing targeted therapies for these diseases. The data required for such analysis is multi-faceted, involving the abundance of exon reads, evidence for junctions between the exons, and predictions of isoform frequencies. Furthermore, conducting a comparative analysis involves multiple samples having varying values for the above information. State-of-the-art isoform visualizations such as Sashimi plots, SpliceGrapher, and SplicingViewer are limited both in terms of perceptual efficiency and in terms of comparing data of multiple samples.

To address these limitations, we developed a novel analysis tool for multi-faceted isoform data, in collaboration with biologists. Our open-source tool, called Vials, consists of coordinated multiple views to represent each of the data facets using the most effective visual encoding. It allows a simultaneous exploration of a large number of samples, and a comparative analysis between multiple groups of these samples. Our collaborators used Vials to analyze isoform data from The Cancer Genome Atlas (TCGA) and Illumina BodyMap 2. An online version of Vials along with a video demonstration and example analysis scenarios are available at http://vcglab.org/vials/

The Circular Secondary Structure Uncertainty Plot (CS2-UPlot) - Visualizing RNA Secondary Structure with Base Pair Binding Probabilities

Dan Tulpan

[Design Contest Entry] The Circular Secondary Structure Uncertainty Plot (CS2-UPlot) is an intuitive visual representation of an RNA secondary structure that includes the uncertainty of all possible base pairings. The CS2-UPlot uses a chord diagram layout and is comprised of 3 concentric graphical layers representing the three main information components of an RNA secondary structure required by the BioVis 2015 Design challenge: (i) the RNA sequence (outer layer), (ii) uncertainty and free energy (mfe) scatter plots for each base (middle layer) and, (iii) uncertainty and minimum free energy (mfe) base pairings (inner layer). The CS2-UPlot challenges the classical ways of representing RNA secondary structures and combines base pairings with dot-plot values in a single graphical representation capable of assisting biologists to quickly identify similarities and differences among a large number of secondary structures and their corresponding RNA sequences. Availability: Figures can be downloaded from: http://www.nrcbioinformatics.ca/cs2uplot/

Visualizing RNA Secondary Structure with Base Pair Binding Probabilities

Daniel P. Aalberts, William K. Jannen

[Design Contest Entry] RNAbow diagrams are a versatile tool for visualizing and comparing ensembles of RNA secondary structures. Previously, RNAbows have proved useful when investigating individual ensembles of folds, performing cluster analysis, and identifying conformational changes caused by single nucleotide polymorphisms. This contest submission highlights their usefulness in the understanding of the effects of mutations.

The RNAbows tool provides multiple modes of use. The AllPairs method is a generalization of the rainbow diagram and represents base pair probabilities with line thickness and darkness. Because our visual processing system naturally groups parallel lines, the AllPairs method makes compatible stems easy to identify. The Difference RNAbow facilitates the comparison of the folds of different ensembles in just a glance by coloring the regions of difference. Difference RNAbows are useful when analyzing the effects of mutations, as well as comparing of clusters of structures.

Visualizing Ensembles of Predicted RNA Structures and Their Base Pairing Probabilities

Peter Kerpedjiev, Ivo Hofacker

[Design Contest Entry] In this design contest submission we present an enhanced version of a traditional RNA dot plot containing a multitude of extra features and data, foremost among which is the inclusion of diagrams for the top Zuker sub-optimal RNA secondary structures. This new design facilitates and eases the interpretation of the dot plot by providing the viewer with an immediate representation of which structures the displayed base-pair probabilities belong to.

Visualizing ncRNA Structural Evolution with Arc Diagrams

Alyssa Tsiros, Lane Harrison

[Desing Contest Entry] The second challenge of this year's BioVis Symposium Design Contest was to create a visualization of noncoding RNA (ncRNA) structural evolution of the human accelerated region 1 (HAR1) gene in ancestral, denisovan, and human sequences. The commonly used chart types for this data, dot plots and node-link style graphs, do not support direct comparison of base pairings among the three structures. We propose a redesign that uses arc diagram visualization techniques to highlight conserved or otherwise evolved base pairings. These diagrams support the placement of unique and prevalent base pairings along a common scale and allow mapping task-specific features to color, providing viewers with a more direct means to identify differences in the structure predictions.

RNA-SequenLens for Visualizing RNA Secondary Structures

Florence Ying Wang, Arnaud Sallaberry, Mathieu Roche

[Design Contest Entry] Traditional RNA structure visualization methods (e.g. dot-plots, arc diagrams, graphs) have limitations such as: (1) with dot-plot and arc diagram, possible pairing nucleotides are not easily perceived, therefore it is difficult to compare RNA structures; (2) rainbow colors are often adopted, which can not reveal the characteristics of data that has implied ordering. Here, we introduce "RNA-SequenLens" to address above problems and the visualization tasks raised in the re-design contest.

In our visualization: (1) a RNA sequence is displayed circularly in the inner most circle, then color is used to encode the order of nucleotides on the RNA sequence. (2) A probability meter is plotted for each nucleotide of a sequence. On the probability meter of a nucleotide, all its possible pairing nucleotides are displayed sequentially based on probabilities. The closer the pairing nucleotides to the center, the higher the probability. Paring links are used to link possible binding nucleotide pairs. (3) A probability filter is introduced to allow the interactive visualization of pairing possibilities at the desired threshold. By interactively changing the size of the probability filter, paring links of different probabilities will be displayed in the center. (4) The visualization of predicted RNA secondary structures can be considered as a special case of uncertainty visualization where the threshold probability is fixed. Then, for comparison, we can highlight the differences between two RNA sequences and their predicted secondary structures.

[Acknowledgement: this work is sponsored by the Labex Numev (convention ANR-10- LABX-20)]

Visualizing RNA Secondary Structure Base Pair Binding Probabilities using Nested Concave Hulls

Joris Sansen, Romain Bourqui, Patricia Thebault, Julien Allali, David Auber

[Design Contest Entry] The challenge 1 of the BIOVIS 2015 design contest consists in designing an intuitive visual depiction of base pairs binding probabilities for secondary structure of ncRNA. Our representation depicts the potential nucleotide pairs binding using nested concave hulls over the computed MFE ncRNA secondary structure. Thus, it allows to identify regions with a high level of uncertainty in the MFE computation and the structures which seem to match to reality.

Abstracts of Contest Entries

Visualizing ncRNA Structural Evolution

Alyssa Tsiros
Worcester Polytechnic Institute

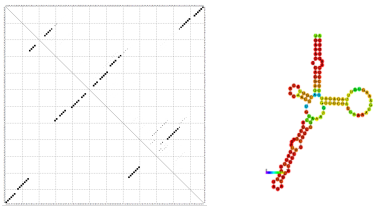
Abstract

This year's Biovis Symposium design challenge offered two challenges, the second of which is addressed here. The second design challenge is to create a visualization of noncoding RNA (ncRNA) structural evolution of the human accelerated region 1 (HAR1) gene in ancestral, denisovan, and human sequences. The Biovis Symposium design challenge #2 figure does not display the direct comparison base pairings among the three structures, whereas the new design "Relative distance vs. MFE base pairing for ancestral, denisovan, and human HAR1 genes" clearly highlights conserved or otherwise evolved base pairings. The direct comparisons of unique and common base pairings allows viewers to clearly identify differences in the structure predictions.

Keywords: Biovis Symposium, design challenge, structure prediction, noncoding RNA, structural evolution.

1. Introduction

Today, computational biologists are able to predict secondary structures from primary RNA sequences. These secondary structures are determined by a collection of predicted base pairs that minimize the free energy of the fold (Zuker, 1995), which is also known as the minimum free energy (MFE) structure. The visualization of this secondary structure typically comprises a "dot-plot", or a grid of possible base pairs (or non-pairs) at each nucleotide residue position (The W.C. Ray Lab, 2015), and a secondary structure graph of the MFE structure (Figure 1).



1. Dot-plot, or probabilities of base pair bindings, and secondary structure graph ("Design Contest", 2015).

There are several RNAs transcribed by the human genome known as non-coding RNAs, or ncRNAs, that do not code for proteins but still impact cellular processes. Until recent years, it has been understood that that most pertinent genetic information was obtained from proteins ("Non-coding RNA", 2006). Evidence suggests that the majority of mammal genomes is transcribed into ncRNAs (Makunin and Mattick, 2006). Such evidence has created a need to classify ncRNAs, where each class has been shown to have a characteristic secondary structure (Arora et al., 2014). Moreover, small changes in a primary RNA sequence can impact its structure and overall function, motivating molecular and structural biologists to compare secondary structures of evolved sequences.

atsiros@wpi.edu
774-275-0458

2. Methods

Several bioinformatics graduate students, a biovisualization professor, an evolutionary biologist professor, and a genetics professor were asked to comment on the Biovis Symposium design. The general consensus was that the new design should still clearly identify the different stems of the secondary structure prediction, but also provide a direct comparison of how the base pairings evolve from ancestral to human.

The presented redesign on the following page is titled "Relative distance versus minimum free energy base pairing for ancestral, denisovan, and human human accelerated region 1 genes". Each of the provided sequences have unique and common MFE structure base pairings. To highlight these pairings, a color was assigned to unique pairings, common pairings among ancestral and denisovan, common pairings among ancestral and human, common pairings among denisovan and human, and common among all three.

The first, middle, and last positions of the sequence are indicated in the figure. Moreover, relative distance was calculated in the following way, where k_1 and k_2 are the first and second positions of a MFE binding, respectively and L is the length of the sequence:

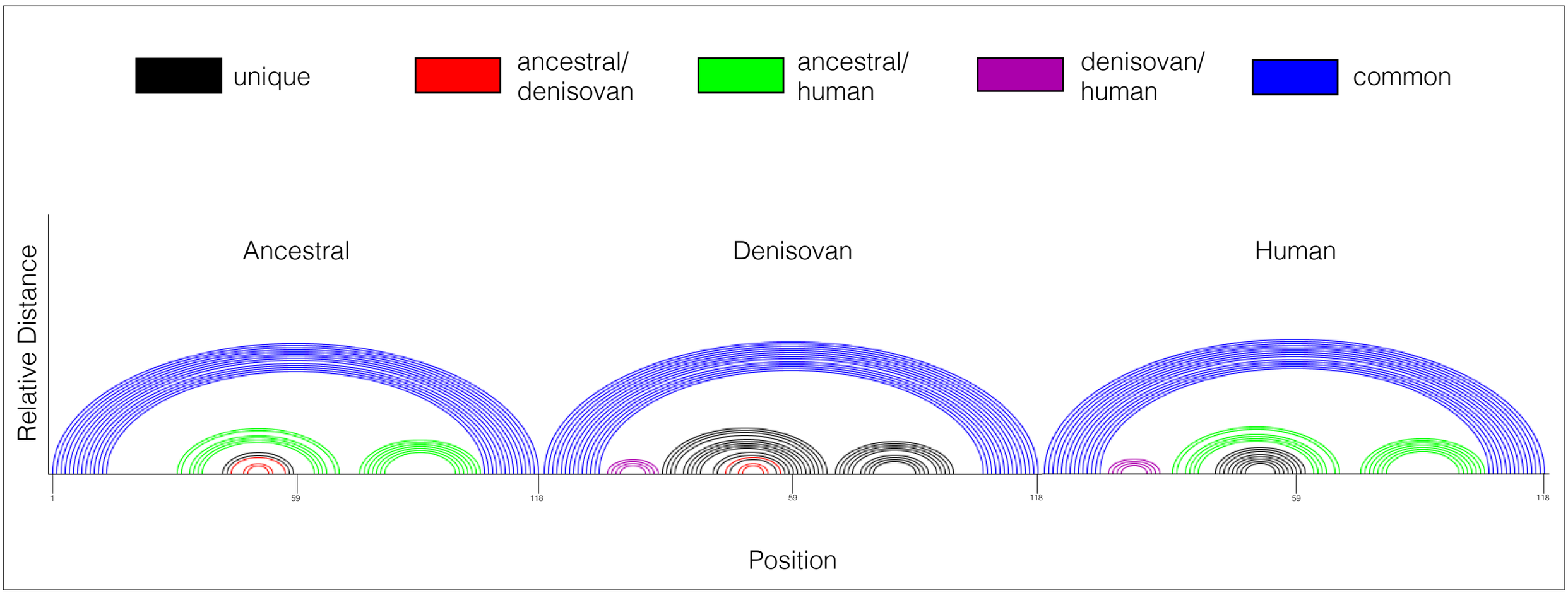
$$(| k_1 - k_2 | + 1) / L$$

The three graphs in the figure display half of an ellipse between two binding positions. Nothing is displayed if a position does not bind. The blue pairings are conserved among ancestra, denisovan, and human genes. The green pairings are conserved from ancestral to denisovan. The red pairings are common among ancestral and human, the fuscia pairings are conserved from denisovan to human, and finally, unique pairings are shown in black.

References

1. (2015). Design challenge. 5th Symposium on Biological Data Visualization. Retrieved from http://www.biovis.net/year/2015/design-contest
2. Arora, A., Panwar, B., Raghava, G. PS. (2014). Prediction and classification of ncRNAs using structural information. BMC Genomics, 2014 (15): 127. Retrieved from http://www.biomedcentral.com/1471-2164/15/127
3. Makunin, I. V., Mattick, J. S. (2006) Non-coding RNA. Human Molecular Genetics, 15 (suppl 1): R17 - R29. Retrieved from http://hmg.oxfordjournals.org/content/15/suppl_1/R17.short
4. The W.C. Ray Lab. (2015, January 22). Understanding RNA folding energy dot-plots. [Video file]. Retrieved from https://www.youtube.com/watch?v=v1UbIUZ8k9o
5. Zuker, M. (1995). Prediction of RNA secondary structure by energy minimization. Washington University Institute for Biomedical Computing. Retrieved from http://mfold.rna.albany.edu/doc/old-mfold-manual/index.php

Relative distance vs. MFE base pairing for ancestral, denisovan, and human HAR1 genes



The Circular Secondary Structure Uncertainty Plot (CS^2-UPlot) - Visualizing RNA Secondary Structure with Base Pair Binding Probabilities

Dan Tulpan

Abstract—The **C**ircular **S**econdary **S**tructure **U**ncertainty **Plot** (**CS^2-UPlot**) is an intuitive visual representation of an RNA secondary structure that includes the uncertainty of all possible base pairings. The **CS^2-UPlot** uses a chord diagram layout and is comprised of 3 concentric graphical layers representing the three main information components of an RNA secondary structure required by the BioVis 2015 Design challenge: (i) the RNA sequence (outer layer), (ii) uncertainty and free energy (*mfe*) scatter plots for each base (middle layer) and, (iii) uncertainty and minimum free energy (*mfe*) base pairings (inner layer). The **CS^2-UPlot** challenges the classical ways of representing RNA secondary structures and combines base pairings with dot-plot values in a single graphical representation capable of assisting biologists in quickly spotting similarities and differences among a large number of secondary structures and their corresponding RNA sequences. **Availability:** Figures can be downloaded from: `http://www.nrcbioinformatics.ca/cs2uplot/`

1 The Challenges

The BioVis 2015 Design Competition includes two challenges: uncertainty visualization and sequence evolution visualization. My solution is designed to mainly address the first challenge, while it can be part of a solution for the second challenge, too, as it will be described in the following sections.

First, I will extrapolate lists of requirements that capture the main desired features for solutions to each challenge. I will use these requirements to design a solution and to determine to what extent the solution addresses each challenge.

The required features for *Challenge 1* are: **(1.1)** display the RNA sequence, **(1.2)** visualization of base pairing probabilities represented in the top-right triangle of the table associated with the dot-plot, **(1.3)** visualization of base-pairings corresponding to the *mfe* structure represented in the bottom-left triangle of the table associated with the dot-plot (optional since the *mfe* secondary structure already presents this), **(1.4)** must be a static picture.

The required features for *Challenge 2* are loosely defined as follows: **(2.1)** supports comparison of predicted RNA structures, **(2.2)** ability to identify changes in the RNA sequence that influence its structural stability, **(2.3)** must be one or more static pictures.

2 The Solution

Challenge 1 mainly requires an intuitive mean to visualize sparse tabular information, with the constraint that the (x,y) coordinates for each uncertainty value in the table corresponds to a base pair in an RNA sequence, whose *mfe* secondary structure is known and typically represented as a graph. Based on these constraints, my choice is to open-up and map the typical *mfe* secondary structure representation of the RNA sequence on a circular plot much in the same fashion as Circos [1] represents linear genomes and their interactions, inter-dependencies and features. In the interest of time the Circos graphical library was employed to implement an automatic process that produces the plots presented in this manuscript. Alternatively, the *D3.js* JavaScript library can also be used for the same purpose. I also acknowledge that simplified versions of circular plots were used for RNA structure representations as early as 1978 [2], nevertheless they represented only one type of information such as *mfe* secondary structures.

Figures 1, 2, 3 and 4 provide a first glimpse of the more advanced visualization method proposed in this manuscript. For a lack of better words, I called this plot type a **C**ircular **S**econdary **S**tructure **U**ncertainty **Plot** (**CS^2-UPlot**).

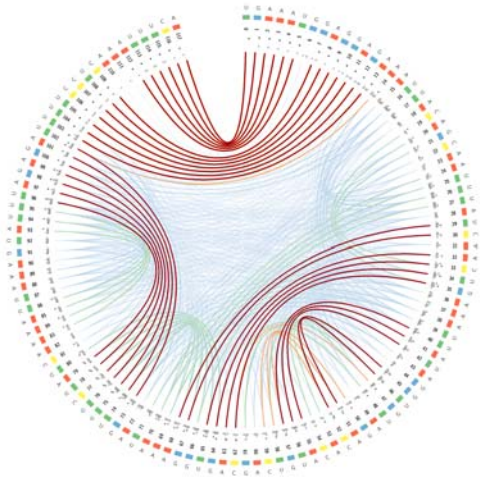


Fig. 1. The **CS^2-UPlot** of the HAR1 ncRNA in ancestral chimp.

A **CS^2-UPlot** consists of 3 concentric layers of information: (outer layer) the RNA sequence, (middle layer) the uncertainty and minimum free energy (*mfe*) mini scatter plots for each base and, (inner layer) the uncertainty and *mfe* base pairings.

The outer layer (RNA sequence) consists of 4 types of equally spaced blocks colored corresponding to their base content (A, C, G and U). Each base and its corresponding position in the sequence (counting starts at 0) is displayed around each block. This addresses requirement **1.1**.

The middle layer represents a mini scatter plot for each RNA base 4. The height and color intensity (gray scale) of each dot on the plot mark its pairing probability, while its x-location within a narrow segment that signifies the whole sequence suggests the relative location of the corresponding base pair on the RNA sequence. The density of the dots in a scatter plot is tightly related to the interaction capacity of each base in a given structural conformation. The higher the number of dots in the scatter lot, the more potential base pairings that base can form. This partially addresses requirement **1.2**, which will be fully addressed by the inner layer.

• *Dan Tulpan is with the National Research Council Canada. E-mail: dan.tulpan@nrc-cnrc.gc.ca.*

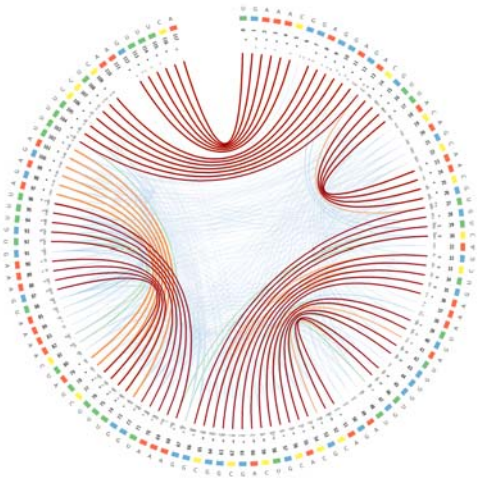


Fig. 2. The **CS^2-UPlot** of the HAR1 ncRNA in Denisovan.

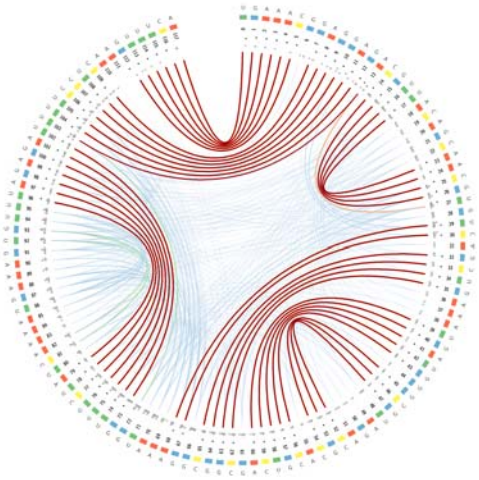


Fig. 3. The **CS^2-UPlot** of the HAR1 ncRNA in human.

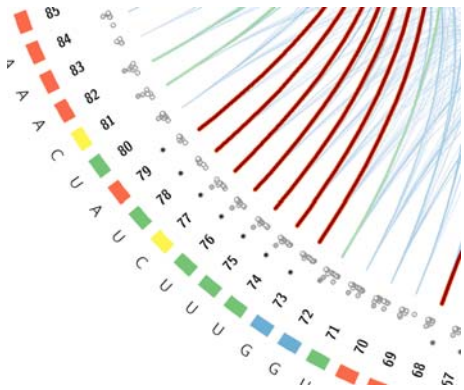


Fig. 4. A close-up view of the **CS^2-UPlot** of the HAR1 ncRNA in human.

The inner layer consists of a set of colored arches that connect base pairs on the RNA sequence, which signify hydrogen bonds depicted by short segments in a typical graph-based RNA secondary structure representation. The arch colors (currently using a 5-color palette assigned to uncertainty values within 0.2 unit intervals span the interval [0,1]) are assigned based on the corresponding uncertainty probabilities, ranging from blue (less stable) to dark orange (more stable) with green representing medium stable base pairs. The thickness of each arch is also proportional with the uncertainty probabilities using 5 incremental sizes ranging from 1 (less stable) to 9 (more stable). Complemented by scatter plots, this representation addresses requirement **1.2**. The dark red arches represent the *mfe* base pairing corresponding to the most stable interactions. Thus requirement **1.3** is also addressed.

The next section will explain how to interpret the information depicted in the **CS^2-UPlots** and will shed light on how the plots can be used to address the requirements for *Challenge 2*.

3 The Interpretation

The information represented in Figures 1, 2 and 3 can be interpreted based on predominance and localization of various visual cues related to color and position. For example, the **CS^2-UPlot** of the HAR1 ncRNA ancestral chimp sequence shows a fairly large number of low and medium base-pair probabilities (blue and green arches) localized in two areas of the RNA sequence between base pairs 19 and 35, and between base pairs 47 and 95. Subsequently, the plot also displays a fairly large subsequence with no base pairings between positions 14 and 28. By comparison, the corresponding subsequences in Denisovan and human HAR1 ncRNAs, contain a solid 4 base-pair stem formed due to 2 consecutive base pair mutations at positions 14-15 and 25-26 replacing the AA/UU base pairs in chimp with the more stable CG/GC in Denisovan and human.

A total of 17 single and consecutive base pair mutations (positions 5, 14-15, 25-26, 32, 40, 43, 53, 56, 63, 65, 72, 87, 93 and 112) can be identified between the ancestral chimp and Denisovan HAR1 ncRNAs, while only one mutation (U replaced by C) at position 46 occurred between the Denisovan and human HAR1 ncRNAs. The large number of mutations that distinguish the Denisovan from the chimp RNA sequences apparently caused not only the apparition of a new 4 base pair stem between positions 14-17 and 23-26 in Denisovan, but caused also the breakdown of an existing 7 base pair stem between positions 73-79 and 96-102 in ancestral chimp and the creation of 2 neighbouring 4 and 5 base pair stems in Denisovan bordered by positions 68-72/92-96 and 74-77/87-90. The orange arches in the Denisovan HAR1 **CS^2-UPlot** suggest the existence of a powerful energetic pressure to recreate the ancestral chimp stem between positions 73-79 and 96-102. This stem re-appears in the human HAR1 ncRNA secondary structure due to changes in stem structures between positions 27-46 and 51-66.

We can also notice in the **CS^2-UPlot** that the human secondary structure is by far the most stable out of all three HAR1 structures given the sparsity of stronger base pairing probabilities (less orange and green arches).

4 Future Improvements and Usability

The comparison of multiple RNA secondary structures using the **CS^2-UPlot** can be also achieved by combining on a single plot all concentrical layers representing the corresponding sequences and the uncertainty base pairings. The only modification that is required is to use different color palettes (one for each RNA sequence) for arch coloring, which in turn might limit the total number of sequences that can be represented and compared in this fashion to 6 or 7. The superposition of concentrical rings representing the base sequences will also allow a fast identification of base pair mutations as opposed to looking at separate plots.

References

[1] M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

[2] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978.

Visualizing RNA Secondary Structure Base Pair Probabilities

William K. Jannen and Daniel P. Aalberts

Abstract— RNAbow diagrams are a versatile tool for visualizing and comparing ensembles of RNA secondary structures. Previously, RNAbows have proved useful when investigating individual ensembles of folds, performing cluster analysis, and identifying conformational changes caused by single nucleotide polymorphisms [Aalberts and Jannen, RNA, 19, 475–478 (2013)]. This contest submission highlights their usefulness in (1) making visual comparisons of the Minimum Free Energy structure to the full structural ensemble and (2) understanding of the effects of mutations on RNA folding stability.

Index Terms—RNA secondary structure, Visualization, Partition function, Mutations

1 Introduction

As authors of visualization tools, we make editorial decisions that — both implicitly and explicitly — attach relative importance to features in the data. Our challenge is to create visualizations that accurately represent physical reality and provide insight.

An RNA secondary structure identifies which bases form pairs, and which unpaired bases form hairpin or internal or multi-branch loops. The Minimum Free Energy (MFE) state is the most probable secondary structure. The MFE or any other single state can be depicted effectively with airport diagrams, bracket notation, and rainbow diagrams. The strengths of these representations — cleanliness and clarity — make them both popular and dangerous. Their use suggests a one-to-one mapping between sequence and structure. However, thermodynamic equilibrium samples many structures.

The partition function is a weighted average of all of the structures at a specified temperature [4]. Encoding the partition function probabilities as a heat-map atop an airport diagram's structure is a recent improvement, but the heat map measures the certainty of the MFE structure rather than suggesting the reality of thermal fluctuations among multiple structures.

There are fewer methods to visualize ensembles of states. Dot plots have often been used to display the partition function's base pairing information. Dot plots compactly represent the probabilities P_{ij} of pairing base i with base j. For coexisting multiple structural classes, however, dot plots often become puzzling. It is difficult to pick out compatible structures within an ensemble, and comparing structures across ensembles requires the viewer to translate between matrices or to reflect across the main diagonal.

Our RNAbow diagrams [1] approach to visualizing RNA secondary structure combines the intuitive qualities of rainbow diagrams with the information density of dot plots to encode the entire partition function in an easily-digestible graph. Further, RNAbows use color, weight, and brightness, along with vertical juxtaposition, to ease the comparison of different ensembles or clusters.

2 RNAbow Diagrams

The rainbow (or arc) diagram graphically represents a single secondary structure state. In a rainbow diagram, bases in the primary sequence form nodes along the graph axis, and an edge between base i and base j represents a bond.

- *William K. Jannen is at Stony Brook University. E-mail: wjannen@cs.stonybrook.edu*
- *Daniel P. Aalberts is Prof. of Physics at Williams College. E-mail: aalberts@williams.edu*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

The RNAbow diagram [1] is a generalization which represents an ensemble of states, using edge thickness and darkness to represent pair probabilities.

Our RNAbows webserver (http://rnabows.com) permits users to select from ViennaRNA, UNAFold, or RNAstructure to compute the partiton function [2, 3, 5] and provides several tools:

AllPairs is a generalization of the rainbow diagram and represents base pair probabilities with line thickness and darkness. *AllPairs* is a drop-in replacement for the dot plot. Because our visual processing system naturally groups parallel lines, the *AllPairs* method makes the compatible stems of multiple structures easy to identify.

Difference RNAbows facilitate the comparison of the folds of different ensembles in just a glance by coloring the regions of difference. *Difference RNAbows* are useful when comparing macrostates (Figure 1) and analyzing the effects of mutations (Figure 3).

Cluster RNAbows allow users to split the *AllPairs* partition function to reveal two sub-partition functions (or "macrostates") describing local minima. Figure 1 is an example.

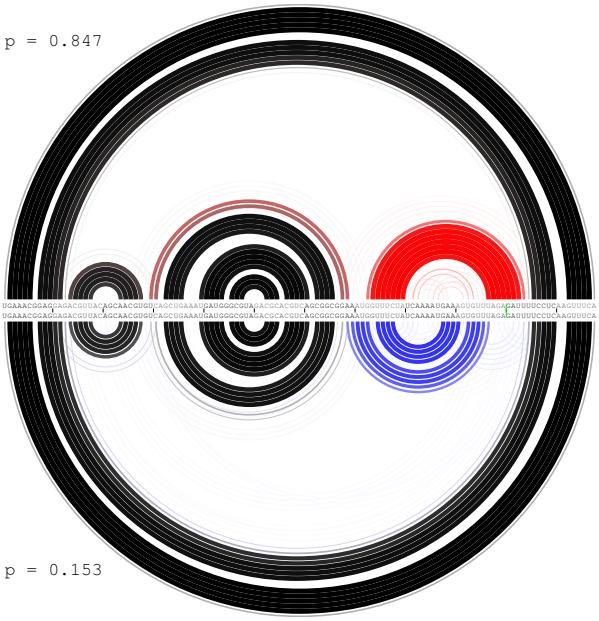


Fig. 1. This *Cluster* RNAbow shows two macrostates within the HAR1 ncRNA Human sequence, along with their relative probabilities. Base pairs unique to either cluster are colored proportional to their probability difference. Common pairs are in black.

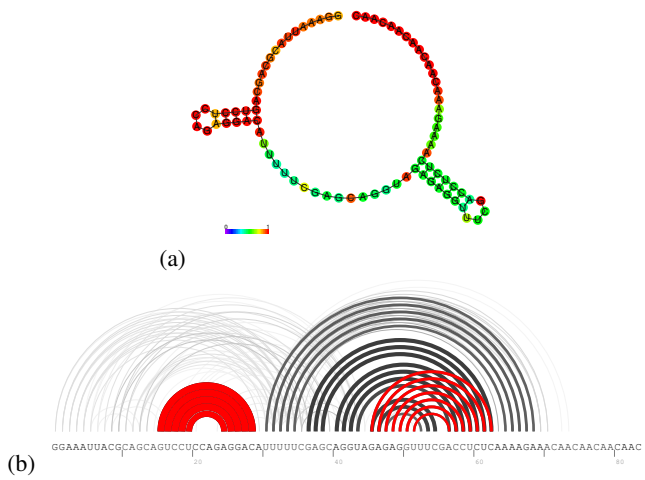


Fig. 2. (a) An "MFE structure drawing encoding base-pair probabilities" image from ViennaRNA [2]. It is clear that half of this structure is uncertain based on the partition function information, but the diagram does not identify why. (b) Our *AllPairsMFE RNAbow* diagram colors pairs from the MFE structure in red and depicts the other pairs in black. Line thickness and darkness are proportional to P_{ij}. The *AllPairsMFE RNAbow* diagram immediately reveals the second local minimum structure.

For this contest submission, we introduce an additional tool:

AllPairsMFE enhances the *AllPairs* representation by signifying the pairs of the MFE structure in red. See Figure 2(b). This tool is a replacement of the heat-map airport diagram, Figure 2(a). *AllPairsMFE* presents the MFE structure, but by also depicting competing structures, it more clearly identifies the sources of structural uncertainty.

3 Contest Challenge 1: Visualizing Uncertainty

Focus on MFE representations may mislead researchers into picturing RNA as a single structure, and not as a thermally fluctuating ensemble of structures. One first step towards visualizing the uncertainty of this approximation is to overlay pairs of the MFE structure with partition function information as a heat map. This approach is seen in Figure 2(a), a visualization from the ViennaRNA package [2]. This diagram makes clear that half of the predictions are fairly certain, but the other half are not. The cause of this uncertainty is not shown.

The *AllPairsMFE* RNAbow for the same sequence, shown in Figure 2(b), reveals the cause of this uncertainty to be a competing structure. The full thermal ensemble of folds is represented clearly, with the pairs of a particular structure (here the MFE, though it could be the consensus or any suboptimal structure) highlighted by color. We see the uncertainty of the pairs *within* the MFE structure, as well as the uncertainty from competing structures.

This example further highlights the dangers of single-state-centric representations. With MFOLD [6] we find $G_{\text{MFE}} = -18.8$ kcal/mol and a second local minimum $G_{MFE2} = -17.7$ kcal/mol. After including the entropic weight, $p_{\text{MFE}} = 0.41$ and $p_{MFE2} = 0.59$. In other words, the entropic weight of the second macrostate makes it more probable than the MFE's macrostate. *Cluster RNAbows* is our tool to resolve macrostates, Figure 1 shows the two Human HAR1 ncRNA macrostates and their probabilities.

4 Challenge 2: Visualizing Sequence Evolution

We investigated the HAR1 ncRNA sequences from Chimp and Human. The *Difference RNAbow* allows us to juxtapose the folds, and reports the folding free energy (here, as calculated by UNAFold [3]). We observe that mutations stabilize the Human sequence's folding free energy by $\Delta G = -11.6$ kcal/mol.

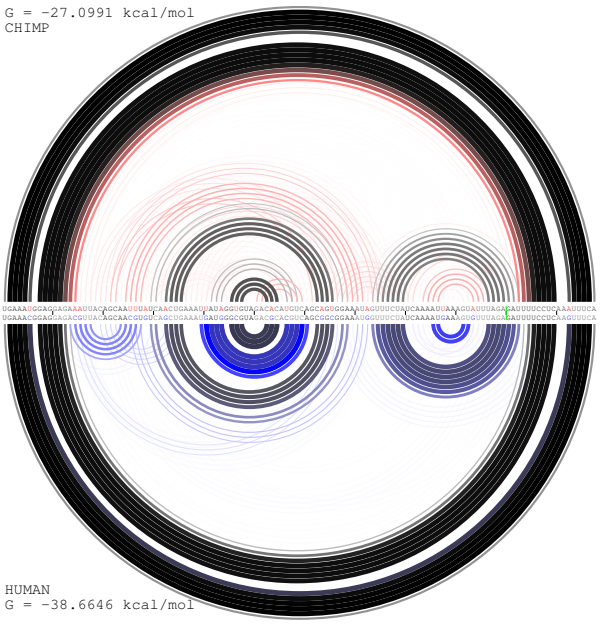


Fig. 3. A *Difference RNAbow* highlights the differences in the RNA secondary structures of the Chimp and Human HAR1 ncRNA sequences. Base pairs unique to either structure are colored proportional to the probability difference. Base substitutions are also color coded. The mutations substantially increase the RNA folding stability of the Human fold relative to the Chimp.

The *Difference RNAbows* tool not only allows for easy comparisons of two known sequences, but suggests where mutations would most influence the fold. Mutations that extend stems would stabilize the fold, while mutations that disrupt stems would destabilize the fold. For example, in Figure 3, the U41G mutation stabilizes the stem and permits the U40-A61 pair in the Human.

5 Conclusions

The *AllPairsMFE* RNAbow of Figure 2(b) is an intuitive visual representation of the uncertainty of RNA secondary structures — not just of pairs within single structures, but of entire structural classes within an ensemble. The *Cluster RNAbow* of Figure 1 isolates structural classes and shows the uncertainty within each.

Our *Difference RNAbow* diagram of Figure 3 facilitates comparisons of the structures of related sequences.

Acknowledgments

This work was supported by the National Institutes of Health [R15GM106372 to DPA]. Earlier development of RNAbows was supported by the National Institutes of Health [GM080690 to DPA] and the National Science Foundation [MCB-0641995 to DPA].

References

[1] D. P. Aalberts and W. K. Jannen. Visualizing RNA base-pairing probabilities with RNAbow diagrams. *RNA (New York, N.Y.)*, 19(4):475–8, 2013.

[2] R. Lorenz, S. H. Bernhart, C. H. Z. Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6.

[3] N. R. Markham and M. Zuker. *UNAFold - Software for nucleic acid folding and hybridization*, volume 453 of *Methods in Molecular Biology*, pages 3–31.

[4] J. S. McCaskill. The equilibrium partition-function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119.

[5] J. S. Reuter and D. H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11.

[6] M. Zuker. On finding all suboptimal foldings of an rna molecule. *Science*, 244(4900):48–52, 1989.

Visualizing Uncertainty of RNA Sequence Base Pairing Variants

Fleur Jeanquartier, Claire Jean-Quartier and Andreas Holzinger

Abstract—This work describes a design oriented approach to visualizing uncertainty of RNA secondary structure probabilities. We address the challenge of finding an intuitive visual representation of encoding uncertainty in RNA secondary structures. We highlight certain limitations and present three different but not exclusive approaches for tackling this challenge.

1 Introduction

In molecular biology researchers have to deal with a decreasing certainty when predicting secondary structures of RNA sequences. Practical testing is limited, computational methods fill the gap in the data with predicted and hence uncertain data. Computational biologists have developed methods to predict the secondary structures (2D folding views of RNA) from a primary sequence of RNA. The outputs of this calculation includes the minimum free energy structure (MFE), the thermodynamically favored and most likely structure, and equilibrium base pairing probabilities. These outputs are typically visualized as a "dot plot", where a box on a square grid of nxn (n is the sequence's length) encodes the base pair binding probability in its area on a logarithmic scale. In addition, the predicted MFE structure is often represented as a secondary structure graph.

2 Background

Dot plots (base pair probability matrices) are a common way for visualizing secondary structure calculations. The squares in the plot area represent a pair (x,y), while either color, transparency, blur effects or size of a dot is used to indicate the probability of a base pair [13]. For today, conservation consensus dot plots can even be interactively controlled to some extent: For example, Sorescu et al. [12] describes a mechanism to specify a threshold probability for dynamic visualization adaptation. However, dot plot representations for base pair probabilities are also said to be confusing when complexity rises, and therefore alternative representations exist too. Base pairings visualization can also be found as linear and circular representations. Alberts et al. [1] introduced so called "RNAbow" diagrams. Hofacker [6] described a software package for analyzing secondary structures and rendering structures as mountain plot and other representations.
When speaking of uncertainty, uncertain data sets may have diverse sources, including data acquisition (signal-to-noise ratio), data mapping (pre-processing and post-processing) and the visualization method itself. Uncertainty can be described as a composite of different concepts, such as errors, accuracy, and subjectivity [4]. Visualizing uncertainty is a difficult problem in all kinds of scientific domains too [5, 11, 2, 8]. Potter et al. [10] already identified uncertainty representations commonly used in visualization and presented a taxonomy of visualization approaches.
None of the mentioned research already dealt with visually encoding uncertainty of the complete set of folding possibilities into one single visualization.
Therefore, we submit this entry to the BioVis 2015 Design Contest [3], that addresses the challenge of visualizing uncertainty of RNA secondary structures. In the following, we describe our visual approaches to the challenge of visualizing uncertainty.

- *Fleur Jeanquartier, Claire Jean-Quartier and Andreas Holzinger are with the Research Unit HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz. E-mail:* {f.jeanquartier,c.jeanquartier,a.holzinger}@hci-kdd.org

3 Visual Approach to Challenge 1

We address the first contest's challenge, namely visualizing uncertainty. The problem is defined as follows:

3.1 Problem:

Design an intuitive visual representation of RNA secondary structure to encode the uncertainty within all the possible base pairing possibilities. The top-right triangle of a dot plot encodes base pairing probabilities and the bottom-left triangle represents the MFE structure. The RNA sequence of n nucleotides is shown on the edge of the nxn square grid. The MFE secondary structure is visualized as a graph, where the color of each nucleotides depicts the strength of base pairing. The challenge is to design a structural representation that is in line with the uncertainty.
To deal with this challenge, however, using the right visualization technique is a question of scaling: An unanswered question remains: What is the limit of possible base pairing probability matrices that can be visualized within one single visualization? Since the number ob potential secondary structures is exponential to the rna sequence's length n [9]. Therefore, we present the following three different approaches for (interactive) visual analysis of rna base pair configurations:

3.2 Approach 1:

One possible interactive visualization approach is sketched in Fig. 1:

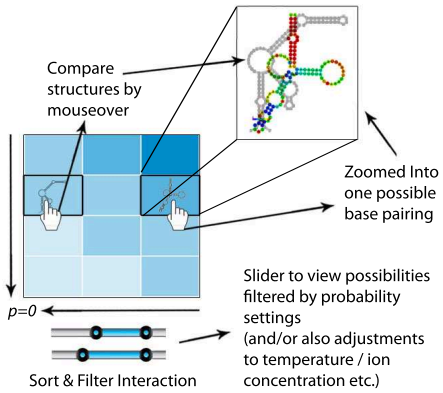


Fig. 1. Visualizing encoded uncertainty of RNA secondary structure possibilities as interactive heatmap including detail view

Holzhuter et al. [7] have shown that particularly heat maps can be dangerous as they can be over-plotted. It is possible, up to a certain amount, to visualize the ensemble organized in a heatmap. But, as common to information visualization, there will be the necessity to integrate interactive exploration features for zoom and filter. We also sketched such interaction integrations. The slider filter at the bottom supports viewing only those rectangles that are related to the most probable configurations but also allows for highlighting the unusual ones. Different perspectives support the interactive visual analysis approach. Additional interactions should be taken into account, like a

slider for filtering specific temperature areas and/or ion concentration settings and adding a switch for sorting not only by probability but also other data variables (i.e. number of base pairs, hairpins, free energy).

3.3 Approach 2:

To overcome some of the heatmap's limitations, another additional or alternative approach is visualizing the complete set of dot plot representations as interactive visual analysis approach making use of the "Rolodex"-art metaphor (also known from window manager in operating systems, apple's time machine or windows exposé), illustrated in Fig. 2. All possible structures are visualized as matrices one after another, while the most probable, the MFE, is the first one on top and behind lay the less probable ones. Interaction allows for toggling through all the possible structures seamlessly while clicking on upper right part of the dot plot all secondary structures are shown in a details view the following manner: All the possible configurations are shown at once, while the most probable is on top. Below all other configurations are shown but with increased transparency values. The most likely is therefore 100% opaque, while the less likely ones are more translucently renderer.
Additionally, Eterna's animation metaphor can be used: Single bases and base pairs within the details view can be animated insofar, as the base pairs movement in pixel per second is related to the structure's folding stability and probability.

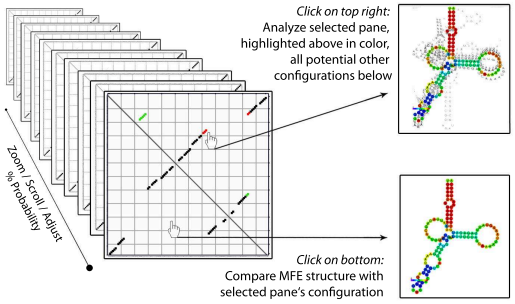


Fig. 2. Visualizing encoded uncertainty of RNA secondary structure base pairings by exploring complete set/ensemble at once

3.4 Approach 3:

Last but not least, another possible approach could be visualizing all possibilities not as box but as part of a network graph, sketched in Fig. 3. The graph is composed by the complete ensemble of structures as follows: Each node represents one possible folding structure, each edge stands for a user defined number x similar base pairs between two structures, while the whole graph integrates the complete "picture". Thereby, similar base pair areas can be marked with another color (compare sketched red area in Fig. 3)
The nodes' transparency (or color/contrast variance) depicts the probability of the particular structure. The node that stands for the MFE is highlighted (in darkest contrast or special color) as the root or center of the graph as the most probable base pairing combination. If the MFE is not the most probable configuration, the visualization can be adapted to distinguish between root, as most probable one, and MFE, as a node somewhere else within the graph highlighted by another color.
According to the dynamic programming algorithm for all subsequences (i, j) of a dot plot, the less probable folding possibilities can be traced back too. Less probable configurations are marked in a translucent manner: The more like configurations are represented by nodes with higher opacity while the more unlikely ones are rendered with less opacity.
Regarding the interaction: By adjusting x certain isles are highlighted, where the configurations represented by the nodes within an isle are more similar to each other. Additional network analysis approaches may further suite the rna analysis process.

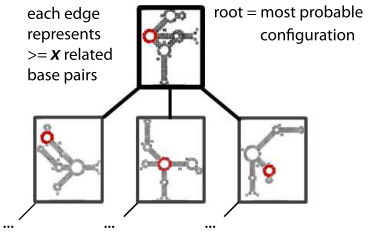


Fig. 3. Visualizing encoded uncertainty of RNA secondary structure by putting focus on the configurations' related base pairs as network graph

4 Material And Methods

Due to the fact, that the submission should be no more than 2 pages we include only a few figures into it. We also recommend watching a short animation, that depicts some details about the three different visualization approaches and the structural representation that is in line with the uncertainty: `http://youtu.be/PZp5GNpNZX4`.

5 Terms and Conditions

By submitting this entry, we give the BioVis 2015 organizers permission to publish it in conference-related materials. Any usage or reference to any submission will include full credit to its authors.

Acknowledgments

We gratefully acknowledge the dataset provided by Maria Beatriz Walter Costa, Henrike Indrischek, Katja Nowick and Christian Hner zu Siederdissen at The University of Leipzig for the purposes of the BioVis 2015 Contest.

References

[1] D. P. Aalberts and W. K. Jannen. Visualizing rna base-pairing probabilities with rnabow diagrams. *RNA*, 19(4):475–478, 2013.
[2] M. Albrecht, A. Kerren, K. Klein, O. Kohlbacher, P. Mutzel, W. Paul, F. Schreiber, and M. Wybrow. On open problems in biological network visualization. In *Graph Drawing*, pages 256–267. Springer, 2010.
[3] BioVis. Design contest, 2015.
[4] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *SimVis*, pages 143–156, 2006.
[5] C. D. Hansen, M. Chen, C. R. Johnson, A. E. Kaufman, and H. Hagen. *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*. Springer, 2014.
[6] I. L. Hofacker. Rna secondary structure analysis using the vienna rna package. *Current protocols in bioinformatics*, pages 12–2, 2009.
[7] C. Holzhüter, A. Lex, D. Schmalstieg, H.-J. Schulz, H. Schumann, and M. Streit. Visualizing uncertainty in biological expression data. In *IS&T/SPIE Electronic Imaging*, pages 82940O–82940O. International Society for Optics and Photonics, 2012.
[8] A. Holzinger, M. Schwarz, B. Ofner, F. Jeanquartier, A. Calero-Valdez, C. Roecker, and M. Ziefle. Towards interactive visualization of longitudinal data to support knowledge discovery on multi-touch tablet computers. In *Availability, Reliability, and Security in Information Systems*, pages 124–137. Springer, 2014.
[9] G. Pavesi, G. Mauri, M. Stefani, and G. Pesole. Rnaprofile: an algorithm for finding conserved secondary structure motifs in unaligned rna sequences. *Nucleic acids research*, 32(10):3258–3269, 2004.
[10] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer, 2012.
[11] J. Smith, D. Retchless, C. Kinkeldey, and A. Klippel. Beyond the surface: current issues and future directions in uncertainty visualization research. In *Proceedigs of the 26th International Cartographic Conference*, pages 1–10, 2013.
[12] D. A. Sorescu, M. Möhl, M. Mann, R. Backofen, and S. Will. Carnaalignment of rna structure ensembles. *Nucleic acids research*, page gks491, 2012.
[13] A. Wilm, K. Linnenbrink, and G. Steger. Construct: improved construction of rna consensus structures. *BMC bioinformatics*, 9(1):219, 2008.

RNA-SequenLens for Visualizing RNA Secondary Structures

Florence Ying Wang Arnaud Sallaberry Mathieu Roche

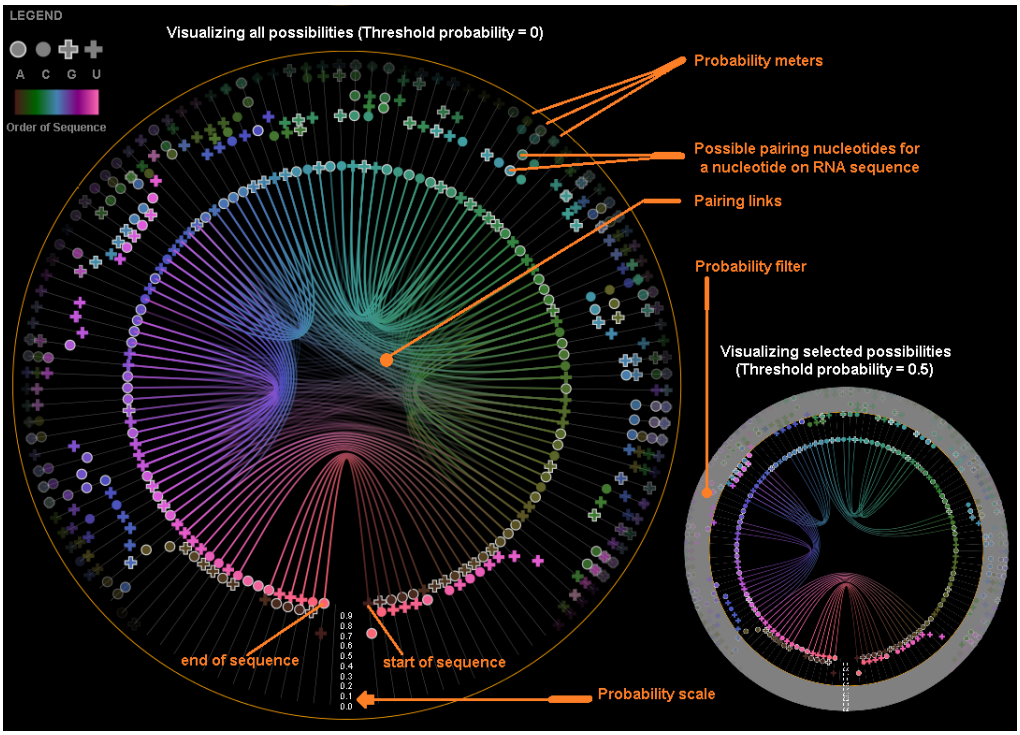


Fig. 1. Visualizing all the possible base pairing probabilities of ancestral(chimp) RNA with RNA-SequenLens. A threshold probability can be set interactively by the *probability filter* to determine the amount of pairing links shown in the center of SequenLens.

Abstract—In this paper, we present RNA-SequenLens to facilitate the visualization and comparison of RNA secondary structures. With RNA-SequenLens, all possible base pairings of a RNA sequence can be visualized at the desired probability threshold. Different RNA secondary structures can be easily compared. The interactive demo is available at `https://youtu.be/C6EDC8LZJXw`

1 Introduction

Traditionally, computational biologists use dot-plot to visualize the base pairing probabilities, where each grid on $N \times N$ grids encodes the base pair binding probability [3]. Furthermore, the predicted secondary structure is often represented as a graph or arc diagram [1]. Although these visualizations have been popular, they have limitations such as: (1) with dot-plot and arc diagram, possible pairing nucleotides are not easily perceived, therefore it is difficult to compare RNA structures; (2) rainbow colors are often adopted, which can not reveal the characteristics of data that has implied ordering [2]. In this paper, we introduce a visual design called "RNA-SequenLens" to address above problems and the visualization tasks raised in the re-design contest.

- *Florence Ying Wang, LIRMM & Universite de Montpellier, France. E-mail: ying.wang@lirmm.fr.*
- *Arnaud Sallaberry, LIRMM & Universite Paul Valery Montpellier, France. E-mail: arnaud.sallaberry@lirmm.fr.*
- *Mathieu Roche, TETIS & LIRMM& CIRAD, France. E-mail: mathieu.roche@cirad.fr*

2 Visualization Requirements for Challenge 1 and 2

The overall purpose of challenge 1 is to design an intuitive visual representation of RNA secondary structure to encode the uncertainty for all the possible base pairing possibilities. Detailed visualization requirements for challenge 1 are summarized as follows:
[T1-1] Visualizing a RNA sequence.
[T1-2] For each nucleotide (i.e. A, C, G, U) on a RNA sequence, visualizing its possible pairing nucleotides.
[T1-3] For a RNA sequence, visualizing all base pairing possibilities.
The overall purpose of challenge 2 is to design a visual representation that supports comparison of the predicted RNA structures. Detailed visualization requirements for challenge 2 are summarized as follows:
[T2-1] Comparing different RNA sequences.
[T2-2] Comparing different predicted RNA secondary structures.

3 Visual Mapping

The features of RNA-SequenLens are shown in Figure 1.
(1) For **[T1-1]**, a RNA sequence is displayed circularly in the inner most circle, then color is used to encode the order of nucleotides on the RNA sequence. As shown by the legend, RNA sequence start with brown and ends with pink (i.e. counterclockwise). For visualizing different nucleotides, we use "circle" to present A (with outline) and C (without outline), "cross" to represent G (with outline) and U (without outline). This is due to the fact that A is likely to bind with U and C is

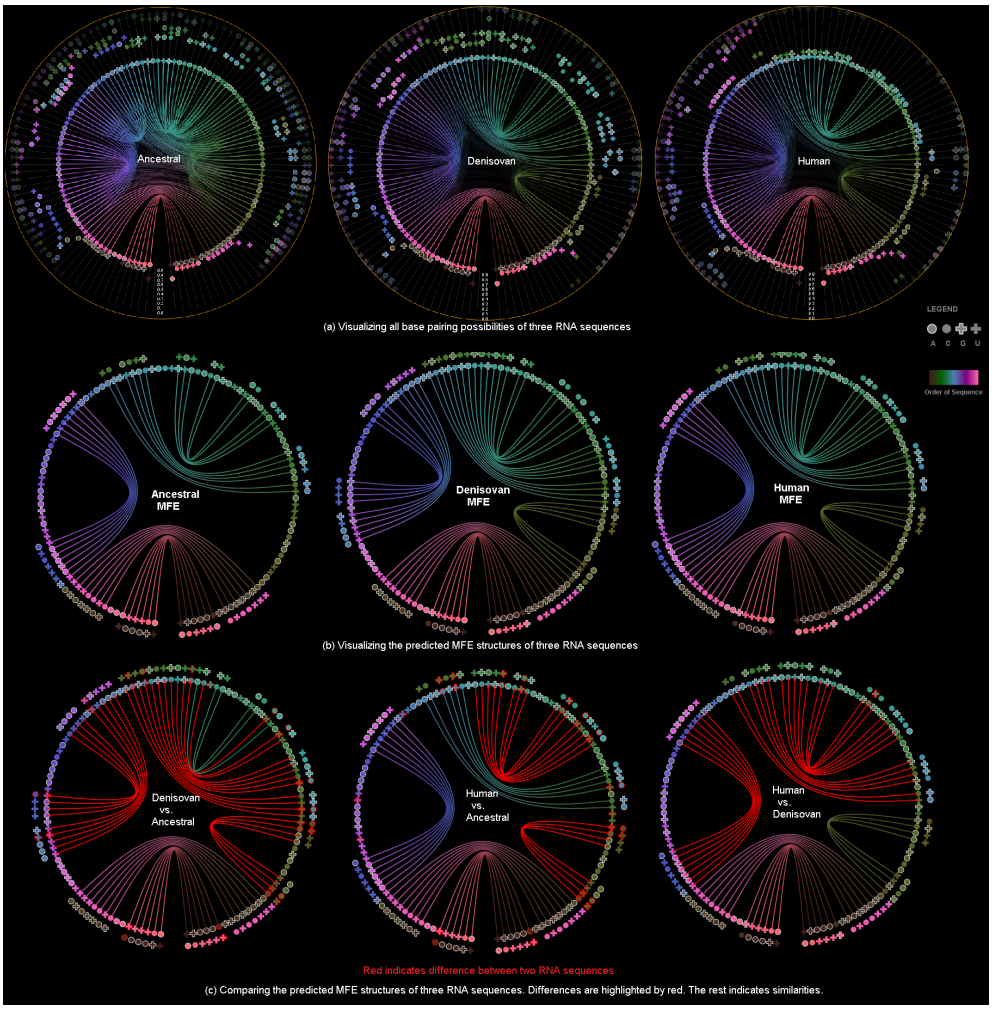


Fig. 2. Visualizing the changes of RNA primary and secondary structures of Ancestral(chimp), Denisovan and Human with RNA-SequenLens.

likely to bind with G. Hence, in SequenLens, circles often bind with crosses.

(2) For **[T1-2]**, we plot one *probability meter* for each nucleotide on a RNA sequence. On the *probability meter* of a nucleotide, all its possible pairing nucleotides are displayed sequentially based on probabilities. The closer the pairing nucleotides to the center, the higher the probability. *Paring links* are used to link possible binding nucleotide pairs. The color of a *Pairing link* is gradient color that gradually changes from one nucleotide color to its pair's color. The opacity of both *pairing links* and pairing nucleotides on *probability meters* indicate probabilities. The higher the opacity, the higher the probability.

(3) For **[T1-3]**, we introduce a *probability filter* (gray mask in Figure 1 right corner) to allow the interactive visualization of pairing possibilities at the desired threshold. By interactively changing the size of the *probability filter*, *paring links* of different probabilities will be displayed in the center of SequenLens.

(4) The visualization of predicted RNA secondary structures (e.g. MFE structure) can be considered as a special case of uncertainty visualization where the threshold probability is fixed (Figure 2(b)). Then, for **[T2-1][T2-2]**, we can highlight the differences between two RNA sequences and their predicted secondary structures (Figure 2(c)).

4 Case Studies

In this section, we use two case studies to further illustrate our visual design in solving problems raised in the re-design contest.

Case Study 1: Visualizing Uncertainty of Ancestral RNA In Figure 1, circles often pair with crosses, symbols with white outlines often pair with symbols without outlines, which validate the binding properties of A, C, G, U. Also, there is a strong likelihood of pairing between the start and end segments of the sequence. By changing the probability threshold to 0.5 (Figure 1 right corner), we can see that there are 3 major groups of base pairings indicated by 3 major *pairing link* colors: green, purple and pink. This finding also aligns with the 3 groups of *pairing links* in the MFE structure (1st image Figure 2 (b)).

Case Study 2: Visualizing Sequence Evolution In Figure 2(c), the first image shows that between the MFE structures of denisovan and ancestral, segments with the most variations in primary structure also have the most difference in secondary structure. The middle image shows that human and ancestral have less differences in the MFE structure than human and denisovan, whereas the last image indicates that human and denisovan have the least differences in terms of RNA primary structure (only 1 nucleotide is different).

5 Conclusion

In this paper, we have introduced RNA-SequenLens for visualizing RNA secondary structures. Our case studies show the effectiveness of our design in visualizing uncertainty and sequence evolution.

References

[1] D. P. Aalberts and W. K. Jannen. Visualizing rna base-pairing probabilities with rnabow diagrams. *RNA*, 19(4):475–478, 2013.

[2] D. Borland and R. M. Taylor II. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, 27(2):14–17, 2007.

[3] I. L. Hofacker. Vienna rna secondary structure server. *Nucleic acids research*, 31(13):3429–3431, 2003.

Visualizing RNA Secondary Structure Base Pair Binding Probabilities using Nested Concave Hulls

Joris Sansen and Romain Bourqui and Patricia Thebault and Julien Allali and David Auber

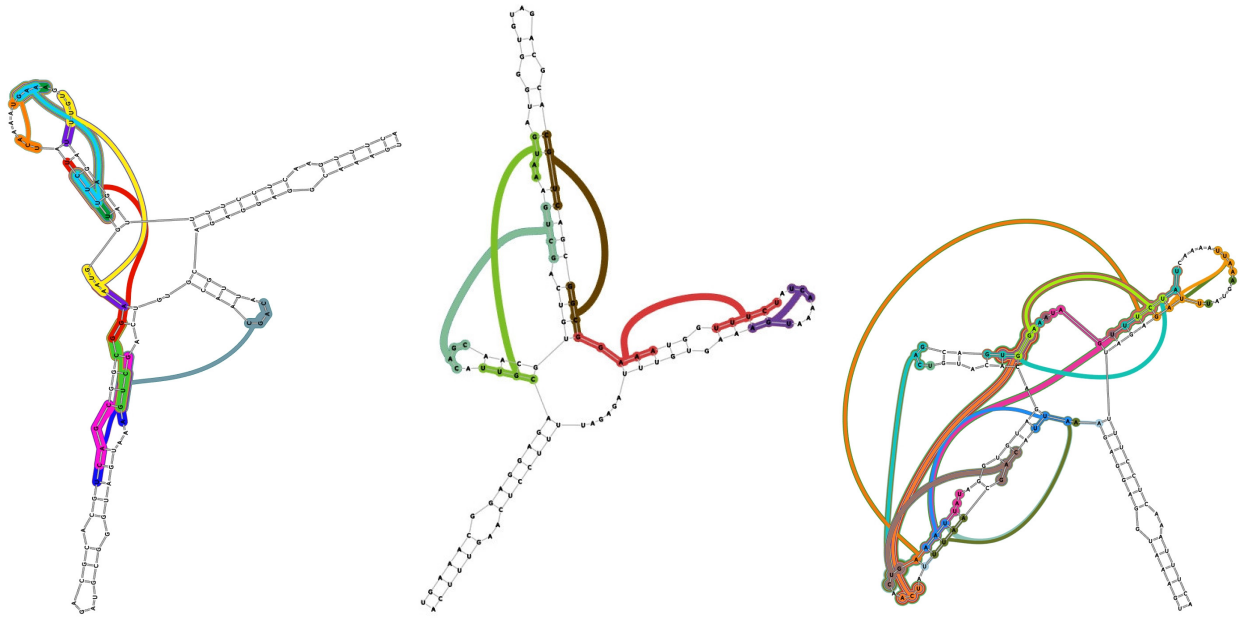


Fig. 1. MFE-structure of ncRNA whith potential pairs binding probabilities. From left to right: human, denisovan and ancestral chimp. The two generation parameters are the same for each sample: minimum number of base-pairs binding = 3 and number of levels =10 but the smallest probabilities have been filtered out for *Human* and *Ancestral chimp*.

Abstract—The challenge 1 of the BIOVIS 2015 design contest consists in designing an intuitive visual depiction of base pairs binding probabilities for secondary structure of ncRNA. Our representation depicts the potential nucleotide pairs binding using nested concave hulls over the computed MFE ncRNA secondary structure. Thus, it allows to identify regions with a high level of uncertainty in the MFE computation and the structures which seem to match to reality.

INTRODUCTION

Among all the discovered RNAs, a major part of them are not transformed into proteins, the non-conding RNAs (ncRNAs). Nevertheless, they are involved in many cellular processes and have a variety of catalytic properties. While primary structure of RNAs is a linear sequence of nucleotides, secondary structure refers to strong binding between pairs of nucleotides. This structure is often considered as characteristic of ncRNAs classes and their biological functions. Being able to predict a ncRNA secondary structure therefore helps to predict its functions. Actually, computational biologists have developed methods to compute the probabilities of base-pairing. These probabilities are then computed to constitute a Minimum Free Energy-structure (MFE) which represents the most likely base pairs binding. These two results are represented within a matrix called a dot-plot. In such matrix, half part represents all the potential base-pairs bindings while the other half contains the results that represent the MFE-structure. Indeed, dot-plots depict the MFE-structure and all potential bindings but it is almost impossible to understand the resulting RNA 2D structure and conclude about its biological functions. Another common technique lies in representing the MFE-structure as a graph. However, the MFE-structure is a prediction that might not depict the reality and that is why the strength of base pairs binding is depicted colors. Nevertheless, such representation does not depict the filtered out potential associations, *i.e.* the other potential pair bindings which might influence the resulting two-dimensional representation. The challenge 1 of 2015 BIOVIS Design Contest consists in designing a visual representation of the probabilities of base pairs binding. We choose to represent the rejected base pairs using nested concave hulls over the graph of the MFE-secondary structure of the ncRNA. In the following, we introduce our depiction, and detail its design. Then we describe the computational process and provide an analysis of the resulting picture. Finally we discuss the strengths and weaknesses of our technique and present conclusions.

- *Joris Sansen is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: jsansen@labri.fr.*
- *Romain Bourqui is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: bourqui@labri.fr.*
- *Patricia Thebault is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: thebault@labri.fr.*
- *Julien Allali is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: allali@labri.fr.*
- *David Auber is with Université de Bordeaux, LaBRI, UMR 5800, Talence, FRANCE. E-mail: auber@labri.fr.*

1 Design

Our method (see Fig. 1 and Fig. 2) depicts the major unselected potential bases pairs sequences over the RNA secondary structure. Using the bases pairs computed for the Minimum Free Energy structure, we depict the ncRNA secondary structure as a graph. Potential base pairs binding rejected for this structure are then emphasized with the addition of nested concave hulls which put forward the subsequence. The more a sub-sequence is wrapped into hulls, the more its probability is important. Paired sub-sequences are linked to improve the pairs tracking and the bundling algorithm described in [2] associated to bezier curve bending is used to improve the design and visibility of the base pairs binding. This makes possible to ease the identification of paired sub-sequences by reducing link crossings and clutter. Finally, we ease the identification of paired subsequences providing a different color to each hulls and corresponding links. Thus, while displaying the MFE ncRNA secondary structure, we made possible to depict the potential base pairs binding and their relative values. Furthermore, our depiction eases the identification of stems and loops that could be transformed if different base pairs binding were selected, or on the contrary those who would remain stable.

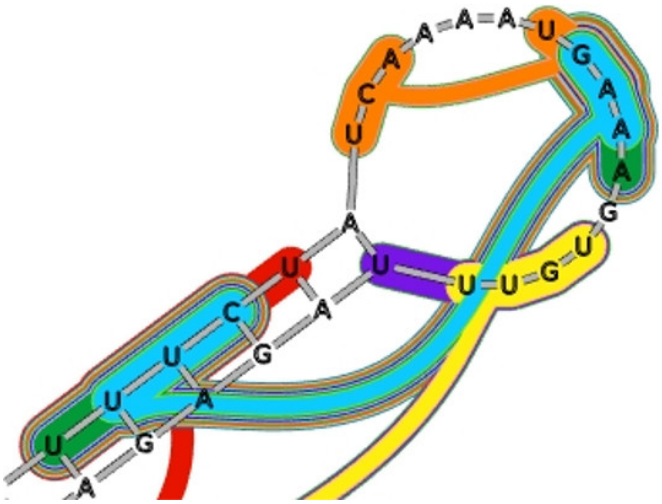


Fig. 2. Zoom on a loop of Human ncRNA secondary structure for design details.

2 Data processing

The first step to generate our depiction consists in generating the representation of the ncRNA secondary structure. Nucleotides are laid out using the primary sequence of nucleotides associated to the MFE-structure base pairs binding. This combination makes possible to draw the MFE-structure using a dedicated algorithm as the one described by Auber *et al.* in [1].

The second step is the search for the alternative pairs bindings. Our algorithm loop through the dot-plot and seek over the pairing probabilities for diagonal sets of minimum 3 matching bases missing in the computed MFE-structure. This algorithm needs two parameters: i) the minimum number of pairs to search in the set and, ii) the number of different levels of probability required. The second is used to define a threshold which determine the minimum base pair probability value authorized in the set. This threshold is simply defined as the range of probabilities ($max-min$) divided by the number of levels required. This process is repeated and the threshold progressively increased to pick smaller and smaller sets with a range of probability more and more restrained. Additionnally, we can tune the threshold in order to filter out of the selectable sets the ones with negligible matching probabilities.

The last step consists in the creation of nested concave hulls. The nucleotides of the selected sets are wrapped into hulls onto the MFE-structure and connected with a link to ease the identification of pairwise elements. The technique used to wrap the sets of nucleotides is the one described by Lambert *et al.* in [3]. This technique makes possible to order and nest hulls by size of sets. Thus, even sets with low pairing probabilities can be depicted and identified while putting forward the most likely sequences.

3 Results

To generate the Fig. 1, we used the previously described generating algorithm with a minimum number of base pairs of 3 and 10 different levels of probabilities. One can easily notice that some regions of the ncRNA present a few other potential base pairs binding indicating a high level of uncertainty, while other regions present just a few potential bindings. Indeed, the two upper stems seem quite uncertain since a few other bases pairing were possible. Moreover, three loops near these stems have an important amount of potential different bindings which could completely change the ncRNA secondary structure. For example, the left stem and its two inner loop could induce the appearance of a larger loop on its extremity by pairing the blue, pink or green sequences. On the contrary, the two bottom stems and loops does not have a lot of other potential nucleotides pairing possibilities. It seems to reveal that the prediction might be correct and correspond to reality.

The generation of this depiction was restrained by the tune of a few parameters which can improve or reduce the readability of the resulting visualization. Indeed, the result of the sequence detection process may vary depending on the two parameters of the generation algorithm: minimum number of base per sequence and value of the minimum probability threshold. Moreover, the bundling process induces variation in the readability of the structure and must be tuned too.

Conclusions

We have presented here an effective and intuitive depiction of a ncRNA secondary structure showing the most important potential base-pairs binding rejected during the MFE-structure computation. Our method puts forward the different possibilities while emphazing the most probable pair bindings and the influence they may have onto the predicted MFE-structure. The use of nested concave hulls produces a pleasant depiction of the potential base pairs binding which reveals the uncertainty in the predicted MFE-structure and the region of the structure that should remain stable with different base-pairs binding.

Acknowledgments

We gratefully acknowledge the dataset provided by Maria Beatriz Walter Costa, Henrike Indrischek, Katja Nowick and Christian Hner zu Siederdissen at The University of Leipzig for the purposes of the BioVis 2015 Contest.

References

[1] D. Auber, M. Delest, J.-P. Domenger, and S. Dulucq. Efficient drawing of rna secondary structure. *J. Graph Algorithms Appl.*, 10(2):329–351, 2006.

[2] A. Lambert, R. Bourqui, and D. Auber. Winding roads: Routing edges into bundles. *Computer Graphics Forum*, 29(3):853–862, 2010.

[3] A. Lambert, F. Queyroi, and R. Bourqui. Visualizing patterns in node-link diagrams. In *Information Visualisation (IV), 2012 16th International Conference on*, pages 48–53. IEEE, 2012.

Visualizing Ensembles of Predicted RNA Structures and Their Base Pairing Probabilities

Peter Kerpedjiev and Ivo Hofacker

Abstract—In this design contest submission we present an enhanced version of a traditional RNA dot plot containing a multitude of extra features and data, foremost among which is the inclusion of diagrams for the top Zuker sub-optimal RNA secondary structures. This new design facilitates and eases the interpretation of the dot plot by providing the viewer with an immediate representation of which structures the displayed base-pair probabilities belong to.

1 Introduction

The traditional RNA dot plot conveys the probability that a particular base-pair is present in the ensemble of predicted structures. This information is presented as a 2D scatter plot, where the size of the rectangular marks is proportional to the probability of a pairing between nucleotide *i* (on the x-axis) and nucleotide *j* (on the y-axis). The upper right triangle of the plot displays this information for the ensemble of predicted secondary structures whereas the bottom left displays only the pairs present in the minimum-free energy structure (MFE). The dot plot is useful in conveying to the viewer that some nucleotides may have a propensity to form differing base-pairs. At first glance, it shows whether there are stems which are consistent across the whole ensemble and which nucleotides they encompass.

Beyond this application, however, it becomes difficult (albeit far from impossible) to extract extra information. The key unanswered question, in our opinion, is which structures correspond to the indicated base-pairs? As previously mentioned, the pairs corresponding to the MFE structure are shown in the lower left hand corner. What does this structure look like, however? What about the other base-pairs in the upper right section? Which structures do those correspond to? How many different structures do they correspond to? Which can be found in the same sub-optimal structure?

With these questions in mind, we set about redesigning the dot plot to include actual secondary structure diagrams in the background. The result, shown in Figure 1, gives the viewer an answer to each of the questions posed above and more. It further provides a platform which can be extended to create an interactive tool to ease the exploration of the data presented in the visualization.

2 Design Considerations

Our design was created to answer some basic questions that researchers might ask about an ensemble of predicted RNA structures, as well as to provide some minor improvements to the way the traditional dot plot is laid out. In each section we describe what we did, why we did it, as well as how we feel it could be improved with an interactive version of our design.

2.1 What does the MFE structure look like?

Description: In the traditional use case, one receives a secondary structure diagram representing the minimum free energy structure in one file and the dot plot in another. We strive to unite these two representations by showing the MFE structure in the background of the dot plot. Such an approach is alluded to in a figure in [3], but we go one step further and arrange the MFE structure along with other sub-optimal structures and scale their size according to their expected population in the Boltzmann ensemble of predicted secondary structures.

Motivation: The give the viewer an immediate representation of the MFE secondary structure.

• *Peter Kerpedjiev (pkerp@tbi.univie.ac.at) and Ivo Hofacker (ivo@tbi.univie.ac.at) are both at the University of Vienna.*

2.2 Which other structures are predicted?

Description: RNA folding, being a kinetic process, leads to the presence of more than one particular structure in solution. We display a subset of these sub-optimal structures, along with the MFE structure, in the background of the dot plot. Based on the energy of each predicted structure, one can calculate its expected weight within the ensemble and use it to scale the size of its secondary structure using a squarified treemap layout [1]. Only structures which correspond to base-pairs with a probability above a certain threshold (see next section) are displayed.

Motivation: The MFE structure can quickly be compared to the other predicted structures in the ensemble in terms of not only structure, but also energy value.

Potential Improvement: Some structures can appear quite small. An interactive version of the plot can enlarge them when one hovers over a base pair belonging to that structure.

2.3 Which structures do the predicted base pairs correspond to?

Description: The upper right hand corner of the dot plot shows all of the potential predicted base pairs above a certain probability value (0.08 in our case, 0.00001 in the traditional dot plot). We chose a higher cut-off due to the simple fact that a lower cutoff would yield points so small as to be virtually indistinguishable without a magnifying glass. Each of the dots is colored to match the color of the best sub-optimal secondary structure containing that base pair. Recall that these structures are displayed in the background of the dot plot.

Motivation: This encoding helps to link the predicted base pairs with the structures they are expected to appear in.

Potential Improvement: Increasing the size on mouse-over, as suggested in the previous section, should help alleviate this issue. Clicking on a structure could also be employed to highlight/enlarge the base pairs belonging to it.

2.4 Which base pairs in a structure are displayed in the dot plot?

Description: The MFE and sub-optimal structures in the background are generated by finding the lowest energy structure given a base pair constraint. Within those structures, we highlight the pairs which, when constrained to being paired, lead to the prediction of that structure. These also correspond to the base-pairs displayed as dots on the dot plot.

Motivation: By highlighting the base pairs in the secondary structure, one can easily see not only how many, but which pairs in a sub-optimal structure are represented in the dot plot.

Potential Improvement: The identity of the base-pairs could be clarified by drawing lines between the secondary structure and the dots when users hover the mouse over the dots.

2.5 Minor improvements

Nucleotide Numbering: We added the positions of the nucleotides to the margins. To avoid clutter, we only add the numbers for nucleotides

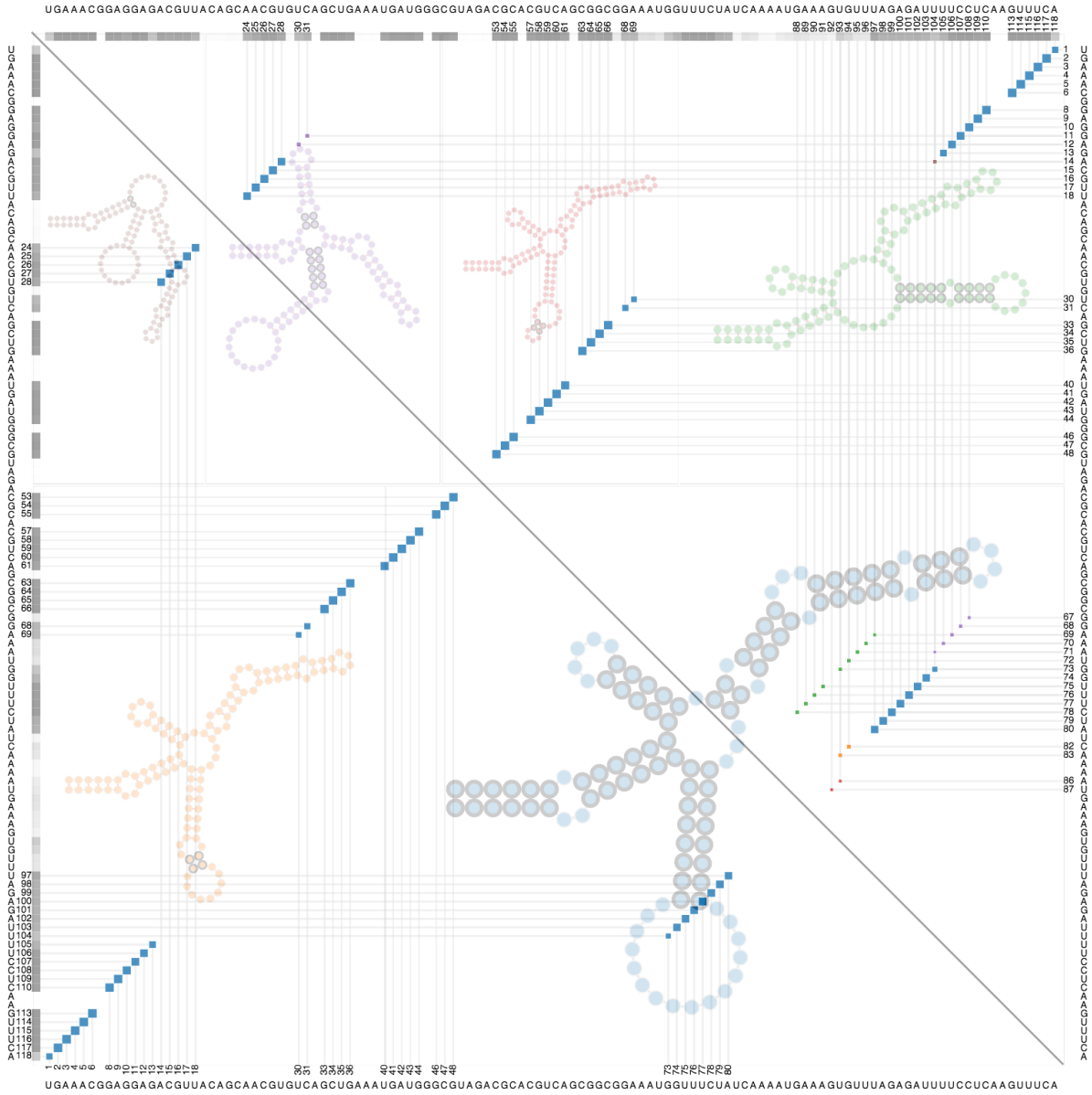


Fig. 1. Sample enhanced dot plot for the Human Highly Accelerated Region 1A provided in the contest data.

that have the potential to be in a base-pair (i.e. have a probability greater than the threshold).

Numbering Guide Lines: To guide the viewer in reading out the identity and position of each paired nucleotide, we have added faint lines from each dot on the dotplot to the numbers of the nucleotides in the margin

Total Pairing Probability: The summed pairing probability for each nucleotide is encoded as colored squares on the upper and left border of the plot. This provides an overview of which nucleotides are likely to be paired in the whole ensemble. It can be used as a comparison with data from probing experiments.

3 Generation

The data for the plot is generated by a python script which makes use of the python binding of the ViennaRNA package. The actual plot is rendered in the browser using the D3.js and fornac.js libraries. Such a format makes it easy to add interactivity to the current design.

4 Availability

The code for creating this visualization is available at:

`https://github.com/pkerpedjiev/dotstruct`

A higher resolution rendering of Figure 1 can be found at:

`http://www.tbi.univie.ac.at/~pkerp/dotplus/`

Acknowledgments

We wish to thank Ronny Lorenz for his help with the ViennaRNA package python interface and Stefan Hammer for his tireless work on creating the secondary structure visualization tool **forna** [2] which paved the way for this submission.

References

[1] M. Bruls, K. Huizing, and J. J. Van Wijk. *Squarified treemaps*. Springer, 2000.

[2] P. Kerpedjiev, S. Hammer, and I. L. Hofacker. forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *submitted*, 2015.

[3] R. B. Lyngsø, J. W. Anderson, E. Sizikova, A. Badugu, T. Hyland, and J. Hein. Frnakenstein: multiple target inverse RNA folding. *BMC bioinformatics*, 13(1):260, 2012.