TECH TRACK

Partitioning biological data with Transitivity Clustering

Tobias Wittkop[1], Sita Lange[2], Dorothea Emig[4], Sven Rahmann[3], Jan Baumbach[4,5]

Partitioning biomedical data objects into groups, such that the objects in each group share common traits, is a long-standing challenge in computational biology. Here we present an integrated data clustering framework based on weighted transitive graph projection: Transitivity Clustering. We illustrate a typical, biomedical clustering task that starts with a list of amino acid sequences, investigates similarity functions and parameter estimation problems, and finally deals with an integrated result interpretation; all of which can be done easily with Transitivity Clustering, but with no other clustering software. We will further demonstrate how to use Transitivity Clustering to identify protein complexes in protein-protein interaction data. With Transitivity Clustering, we provide the user with easy-to-use interfaces that significantly ease and improve each step of the typical data clustering workflow: (1) A web interface for a quick analysis of medium-sized data sets, (2) a powerful stand-alone Java implementation for large-scale data clustering, and (3) a collection of Cytoscape plugins that also provide further methods to answer typical follow-up questions.

Reference: Wittkop T, Emig D, Lange S, Morris JH, Boecker S, Rahmann S, Albrecht M, Stoye J, Baumbach J (2010) Partitioning biological data with Transitivity Clustering. Nature Methods. 2010 Jun;7(6):419-20.

URL

http://transclust.cebitec.uni-bielefeld.de

Speaker

Dr. Jan Baumbach[1,2,3,4,5]

Author affiliations

[1] Buck Institute for Age Research, USA.
[2] Albert-Ludwigs-University, Freiburg, Germany.
[3] Technische Universität Dortmund, Germany.
[4] Max Planck Institute for Informatics, Germany.
[5] International Computer Science Institute, University of California at Berkeley, USA

The Microsoft Biology Foundation

Michael Zyskowski, Research Program Manager[1]; Bob Davidson, Principal Software Architect[1]

The aim of the Microsoft Biology Foundation (MBF) project is to produce a well-architected and comprehensively-documented library of common functionality related to bioinformatics and genomics, with the intention of making it easier to write life science applications on the Windows platform. Using C# and the .NET 4.0 framework provides additional levels of flexibility for the developer – over 70 .NET programming languages are compatible, from Visual Basic and Python to C++ and F#. It also leverages the power of .NET – over 15,000 pre-written functions - and takes advantage of .NET Parallel Extensions, a new feature which can parallelize algorithms across all cores and processors of the local machine.

This demonstration will include a brief tour of the MBF library, including details of its free, open source, community-curated and community-owned philosophy and how scientists and developers can participate in future development. We will also demonstrate the flexibility and usability of the library through a range of applications, including a DNA sequence assembler using the Windows Presentation Foundation, an add-in for Microsoft Excel integrating bioinformatics functionality directly with the spreadsheet, access to webservices including demonstration of the Microsoft cloud computing solution Azure, and integration with HPC and scientific workflows.

URL

`http://mbf.codeplex.com`

Speaker

Simon Mercer, Ph.D. Director of Health and Wellbeing[1]

Author affiliations

[1] Microsoft Research

Data integration in proteomics through EnVsion and EnCore webservices

Pascal Kahlem[1], Henning Hermjakob[1]

This demo will introduce the EnCore infrastructure as collection of webservices to query proteomics data. We will describe how to use individual EnCore webservices and we will explain how to create worflow connecting different services. We will explain how the EnCore platform is technically structured and how it simplifies the way to use webservice using a common format named "EnXML". We will list and briefly describe webservices available in EnCore. Among these services we will explore how to query molecular interactions, protein identifications, biological pathways, protein sequence information, biological models, protein localization and ontology distribution. We will learn how to use EnVision, a web graphical user interface to query EnCore webservices and display elaborated information from its results. We will see examples to query EnCore using EnVision and we will describe in detail the results obtained by EnVsion.

URL

http://www.enfin.org

Speaker

Rafael Jimenez[1]

Author affiliations

[1] European Bioinformatics Institute

ELIXIR: A Sustainable European Infrastructure for Biological Information

Dr Andrew Lyall[1]

ELIXIR has recently completed a two year, European wide consultation involving academic and industrial users, data providers and international collaborators, including three stakeholders meetings, two surveys, and fourteen work packages.

ELIXIR will be a distributed infrastructure arranged as a hub and nodes, with the hub at the European Molecular Biology Laboratory’s European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK.

The ELIXIR Steering Committee has recently issued a Request for Suggestions for ELIXIR nodes, seeking input from organisations that are interested in hosting one of its ‘nodes’, in order to help shape ELIXIRs construction. Organisations wishing to contribute their input have been requested to consider how they might contribute: data resources; bio-computing capacity; infrastructure for data integration; and services for the research community, including training and standards development.

Responses from this round of stakeholder input will enable ELIXIR to: define the landscape of potential nodes; provide the coordination necessary to ensure interoperability; and identify any duplications and missing capabilities.

URL

www.elixir-europe.org

Author affiliations

[1] ELIXIR Project Manager, EMBL-EBI

NextGen Biology with TIBCO Spotfire

Prof. Peter v.d. Spek[2], *Dr. Andreas Kremer*[2]

Spotfire helps you to integrate and analyse Microarray and other –omics data and align it with your LIMS and screening data. The complexity of such an analysis is rising with every new data format you are adding. The analyst needs powerful and flexible software tools to get results out of the combined data.

The TIBCO Spotfire software platform fulfils all important business needs of data analysis in modern bioinformatics. Every bioinformatician knows that collections of .CEL and .csv files are not sufficient for analyses as well as simple relational data bases are not. Even Data Warehouses and associated reporting tools are not sufficient to turn data into knowledge. Spotfire DecisionSite is well known for high quality ad-hoc analysis, but also this is not sufficient anymore. We will show how the new TIBCO Spotfire brings all pieces together, reduces the complexity of the analysis and is compliant you're company IT infrastructure. We will demonstrate how Spotfire and the bioinformatics department of the Erasmus Medical Center in Rotterdam built an interactive human brain atlas. This project shows the combination of genomics, proteomics and cytogenetic data with medical imaging. The aim is to identify genes associated with neurological symptoms and diseases (see ECCB2010 poster, A.Kremer).

URL

http://spotfire.tibco.com

Speaker

Dr. Christof Gaenzler, Senior Solution Consultant [1]

Author Affiliations

[1] TIBCO Spotfire, Germany

[2] Erasmus Medical Center, Rotterdam

Study capturing: from research question to sample annotation

Machiel Jansen [2], Jeroen Wesbeek[1], Tjeerd Abma[3], Jahn-Takeshi Saito[4], Adem Bilican[4], Michael van Vliet[5], Prasad Gajula[6], Vincent Ludden[1] , Robert Horlings[1], Siemen Sikkema[7], Margriet Hendriks[3], Chris Evelo[4], Ben van Ommen[1], Jildau Bouwman[1]

The demonstration of the software tool will focus mainly on the part about capturing biological studies. During the demonstration, it will become clear how we achieve the following goals:

- Facilitate standardization of the description of biological research studies
- Leverage existing standards and ontologies and integrate them into one user interface
- Cover complex study designs, such as double-blind crossover designs
- Provide an overview of the different studies done in e.g. a consortium, to encourage cooperation and data exchange
- Track samples that are used for omics assays back to the source, the circumstances under which and the subject from which they were taken
- Provide the first cornerstone needed to do 'multi-omics' data analysis over multipe different studies and multiple omics platforms
- Focus on user experience, through extensive testing with biologists

The demonstration will have the following outline:

- Walk the attendants through the process of adding a study to the study capture tool
- Demonstration of the connection of samples in the study capture tool to omics data
- Give an example of a query involving both study design data and omics data

URL

http://www.dbnp.org

Speaker

Kees van Bochove [1,2]

Author affiliations

[1] TNO Quality of Life, Zeist
[2] NBIC BioAssist Engineering Team
[3] UMCU Metabolomics Centre, UMC Utrecht
[4] BiGCaT Department of Bioinformatics, Maastricht University
[5] Leiden/Amsterdam Centre for Drug Research
[6] Plant Research International, Wageningen University
[7] Biosystems Data Analysis Group, Swammerdam Institute for Life Sciences, University of Amsterdam

How robust are NGS whole-genome assemblies? A case study with plant genomes

Laxmi Parida[1]

A decade after the human genome was sequenced, a large section of the computational/bioinformatics community believes the general assembly problem to be “solved”, in spite of the recent changes in the underlying technologies. However, the “quality” of a draft genome of an organism continues to dog the community that studies the genomes closely with a fine-toothed comb for biological insights such as phenotypic associations with specific loci, strain-specific polymorphism detection and so on.

As genome assemblers come of age and their numbers grow, is it possible to consolidate multiple assemblies so that the quality of the resulting draft is of a higher quality than the individual constituents? We are developing a system to investigate these questions and their related issues. One such burning issue is to understand, model and detect erroneously assembled segments. Clearly the misassemblies require more attention than gaps in the assembly (the former is the usual problem of we-don’t-know-what-we-don’t-know) and is also perhaps more amenable to computational/algorithmic remedies. We are focusing on two directions to study these: one is on intrinsic single assembly statistics and the other on comparing assemblies.

Our work is motivated by the need for the generation of a high quality draft for Theobroma cacao, a plant with estimated 440 Mb genome size that produces cocoa beans, the basic ingredient in chocolate. The case study with the plant genomes originated within the USDA/Mars/IBM Consortium for sequencing the T. cacao genome. In this talk I will describe our assembly evaluation environment and present some preliminary results.

Speaker

Niina Haiminen[1]

Author affiliations

[1] IBM T. J. Watson Research Center, Computational Genomics, Yorktown Heights, NY, USA

Clinical genomic analysis at IBM: from HIV positive to Hypertension

Ehud Aharoni[1], Hani Neuvirth[1], Noam Slonim[1], EuResist GEIE partners[2], Hypergenes partners[3]

The domain of personalized health and genome-based therapy has flourished in recent years due to the significant reduction of information storage devices, the reduction in care delivery costs, and improved clinical outcomes. Building on IBM's leadership in areas like systems integrations, cloud computing, massive scale analytics and even emerging areas of science like nanomedicine, we focus on creating technologies and processes that will build an evidence-centric healthcare ecosystem.

IBM has defined four main areas of research: evidence generation, which uses scientific methods to turn raw health data into proof of effective treatment methods; the ability to deliver evidence in a context-dependant and personalized way at the point of care; improving service quality through simplifying the complex healthcare delivery process; and incentives and models to shift the healthcare industry to a system that rewards based on outcomes and healthier patients rather than only treatment and volume of care.

In recent years, IBM Research in Haifa, Israel has been involved in research related to evidence generation, with a focus on using the massive volumes of clinical and genomic data arriving from different hospitals to find new ways of optimizing patient treatment. Data mining and machine learning techniques were also applied to decipher the relationship between clinical status and genomic variations, with the goal of helping improve diagnostics and treatment.

From 2006 to 2008, IBM Research and the EuResist GEIE consortium developed a drug-interaction modeling tool that lets users predict the success rate of various drug combinations and their impact on virus evolution via an online portal. A set of prediction engines that leverage medical data (for example, viral gene sequences, patient histories) from seven sources are available in the portal and recommend which therapies are expected to be most efficient given viral, genomic, and other clinical and demographic measures. This engine predicts patient response to therapy with 78% accuracy, outperforming other common tools. Since then more data has been contributed and the algorithms were updated based on the new information. In 2010, IBM and the GEIE partners updated the recommendation to include new drugs.

IBM researchers are now working with the European HYPERGENES consortium to identify the genetic variations responsible for hypertension and associated organ damage. The team is working with clinical data accumulated from hypertensive and healthy people and genomic biomarkers provided by Illumina’s 1Million SNP chip. The aim is to create a comprehensive genetic-epidemiological model that takes into account how genomics and other factors help improve diagnostic accuracy and introduce new strategies for early detection, prevention, and therapy. This effort will help create more personalized treatment plans for individuals that suffer from hypertension.

URL

http://engine.euresist.org/ and *http://srv-rimon.haifa.il.ibm.com:8080/rimon_web/snpWeights.jsp*

Speaker

Dr. Michal Rosen-Zvi, manager[1]

Author affiliations

[1] machine learning and data mining group, IBM Haifa Research Lab

[2] EuResist GEIE

[3] Hypergenes

The Universal Protein Resource (UniProt)

The UniProt Consortium[1,2,3]

UniProt is the central resource for storing and interconnecting information from large and disparate sources, and the most comprehensive catalog of protein sequence and functional annotation. UniProt is built upon the extensive bioinformatics infrastructure and scientific expertise at European Bioinformatics Institute (EBI), Protein Information Resource (PIR) and Swiss Institute of Bioinformatics (SIB). It has four components optimized for different uses. The UniProt Knowledgebase (UniProtKB) is an expertly curated database, a central access point for integrated protein information with cross-references to multiple sources. The UniProt Archive (UniParc) is a comprehensive sequence repository, reflecting the history of all protein sequences. UniProt Reference Clusters (UniRef) merge closely related sequences based on sequence identity to speed up searches. The UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for the expanding area of metagenomic and environmental data. Other developments include the ID mapping service which allows users to map between UniProtKB and more than 30 other data sources; and UniSave, a comprehensive history service of UniProtKB entries.

The demonstration will cover:

- A brief description of the UniProt databases.
- Accessing UniProt using simple query syntax. The user will be presented with helpful suggestions and hints.
- Exploration of sequence similarity searches, alignments and ID mapping tools provided.
- Accessing UniProt data programmatically.

This demonstration will also encourage user interaction and feedback.

URL

http://www.uniprot.org/

Speaker

Maria J. Martin, PhD[1]

Author affiliations

[1] Team Leader, UniProt (Development), EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[2] Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St. NW, Suite 1200 Washington, DC 20007, USA

[3] Swiss Institute of Bioinformatics, Centre Medical Universitaire 1 rue Michel Servet 1211 Geneva 4, Switzerland.

The Protein Data Bank in Europe (PDBe) - Bringing Structure to Biology

Sameer Velankar[1], Gerard Kleywegt[1]

In 2011, the Protein Data Bank (PDB) celebrates its 40th anniversary. To date, the PDB has essentially been a historic archive capturing structures (and the underpinning experimental data) of biomacromolecules published in the primary literature. As the use of structural data by non-experts becomes commonplace, the demands on the archive (by both these users and funding agencies) and the way it is made accessible will inevitably change. It is therefore necessary to transform sites that serve the archive into resources that are directly relevant for scientists who work in biomedicine and related disciplines, while simultaneously taking care not to alienate the communities that produce the structures. EMBL-EBI's Protein Data Bank in Europe (PDBe), one of the founders of the Worldwide Protein Data Bank (wwPDB), is committed to becoming such a resource.

The first step towards this goal is the recent make-over of our webpages. During the ECCB session, we will demonstrate how to use the new PDBe pages to navigate to the five areas ("ALIVE") where we are traditionally strong: Advanced services, Ligands, Integration, Validation and Experimental data. We will also relate this to our core position within the European Bioinformatics Institute (EBI).

During the session we will demonstrate:

- New front pages and functionality
- Wizard for structure newbies
- PDBprints (visual overview of PDB entry)
- PDBeXplore (exploring biological data)
- Easy navigation to the five areas ("ALIVE") on which the PDBe focuses
- Advanced services
- SSM (Secondary Structure Matching)
- Pisa (Quaternary Structure)
- PDBeMotif (Motifs and Sites)
- Ligands
- PDBeChem (Ligand information)
- Integration with other databases and resources
- SIFTS project (in collaboration with UniProt)
- Validation (implementing new tools and resources)
- Experimental data
- The Electron Density Server (which will be ported from Uppsala to PDBe
- The Electron Microscopy Database (EMDB)
- Nuclear Magnetic Resonance (NMR) data

URL

www.pdbe.org

Speaker

Dr Wim Vranken[1]

Author affiliations

[1] Protein Data Bank in Europe (PDBe), EMBL-EBI

Super-scale Sequence Data Analysis with Hybrid Core Computing

It is clear that the latest generation of sequencing devices are a vast improvement over previous generations. However, the volume of data created by these new devices has significantly outpaced the ability to sufficiently analyze the data in a timely manner with traditional processors. As such, the only feasible methodology is the adoption of accelerator technologies. The current general purpose graphics processing methodologies provide interesting platforms for development, yet fail to deliver the sufficient order of speedup required to keep pace with the data explosion. Field programmable gate arrays [FPGAs] have typically performed quite well for bioinformatics applications. However, the difficulty in programming the soft cores and the difficulty in data management have historically failed to allow for widespread adoption.

Convey's Hybrid Core computing platform is an ideal platform for sequencing tasks in that it combines the traditional x86 economy of scale with an abstracted set of FPGAs for direct algorithmic synthesis coupled with a cache-coherent memory subsystem. The platform includes a significant set of compilers, tools and libraries such that the historical pain in programming FPGAs is fundamentally eliminated. In this way, Convey has the ability to implement very discretely managed algorithmic kernels [called *personalities*], which serve as logic and instruction set extensions to the x86 environment. In doing so, the Convey platform allows users and applications the ability to significantly parallelize and outperform traditional Von Neumann architectures.

Speaker

John Leidel[1]

Author affiliations

[1]Software Architect, Convey Computer Corp

CLC bio, A Comprehensive Platform for NGS Data Analysis

Not long ago high-throughput DNA sequencing (HTS) was reserved for a few dedicated genome centers that have dedicated and specialized bioinformatics staff and hardware for the analysis of sequencing data. But over the course of only a few years HTS has become broadly accessible to many non-specialized researchers and groups and the responsibility for analyzing very large HTS data sets may now rest upon researchers with a background in molecular biology or medicine and no formal computer training. For these groups the establishment of an appropriate software and hardware platform for bioinformatics analysis presents a very hard and often insurmountable challenge.

Most bioinformatics software for analysis of HTS data exist as stand-alone command line tools that are inaccessible to non-specialists. This creates the need for a dedicated bioinformatician in the data analysis workflow to perform simple and tedious tasks such as simply establishing the analysis software infrastructure, executing commands and parsing textual or binary output. This is unfortunate since it makes the access to bioinformatics personnel a serious and poorly scalable bottleneck in the data analysis work flow and directs valuable bioinformatics brain resources away from creative and proactive tasks and towards tedious and non value creating routine tasks. Furthermore, it dis-empowers the biomedical researchers that requested the data and formulated the biological hypotheses by making them unable to directly inspect, handle and analyze the data themselves.

Another serious bottleneck for data analysis is the access to appropriate hardware. The output of a single run of a sequencing instrument is now at around 50 giga bases and still increasing. This places very tough demands on the hardware used in the analysis regarding CPU power and accessible memory. Again, the recent broad dissemination of HTS data means that very large data sets are now accessible to research group that have insufficient financial or personnel resources to establish a dedicated and powerful computational infrastructure. In order to give these groups access to data analysis, software must be designed that can operate on and maximize the utility of standard, economic and omnipresent computer solutions. In organizations where powerful central computing facilities can be installed and dedicated to HTS data analysis these can only be accessed through technically challenging interfaces that are inaccessible to most employees and require specialist intervention and a drain on the bioinformatics personnel resources.

To resolve these issues, we have developed and here present the CLC bio integrated solution for analysis and data handling of high-throughput sequencing (HTS) data.

URL

www.clcbio.com

Speaker

Roald Forsberg, PhD[1]

Author affiliations

[1]Director of Scientific Software Solutions

Software for the data-driven researcher of the future

Alan Williams[2], Aleksandra Nenadic[2], Shoaib Sufi[2], Danius Michaelides[2], David Withers[2], Don Cruickshank[2], Franck Tanoh[2], Ian Dunlop[2], Jiten Bhagat[2], Katy Wolstencroft[2], Paolo Missier[2], Sergejs Aleksejevs[2], Stian Soiland-Reyes[2], Stuart Owen[2], Peter Li[2], Finn Bacell[2], Mannie Taggs[2], Rishi Ramgolam[2], Marco Roos[2], Eric Nzuobontane[2], Thomas Laurent[2], David De Roure[2], Robert Stevens[2], Steve Pettifer[2], Rodrigo Lopez[2],Carole Goble[2]

The myGrid project have developed a suite of open source software tools for the registration and discovery of Life Science Web Services (BioCatalogue), and the designing, execution, and sharing of scientific workflows (Taverna and myExperiment) for analytics and data processing.

The Taverna workflow workbench provides an environment in which scientists design and execute workflows, combining local and public distributed services and data resources into a single experimental protocol. The latest Taverna 2.2.0 release features: a command line tool for running workflows without the need to interact directly with a user-interface; access to the BioCatalogue of Life Science Web Services via a plug-in interface; and a Taverna Server for running workflows on a server as well as on the desktop.

myExperiment is a social networking site that provides a collaborative environment where scientists can safely publish their workflows, experiment plans, and standard operating procedures (SOPs). This allows researchers to share them with individuals, groups, and even discover those of others that can be subsequently re-used. myExperiment makes it easy for the next generation of scientists to contribute to a pool of scientific methods, build communities and form relationships – reducing time-to-experiment, sharing expertise and avoiding reinvention.

The BioCatalogue is a community driven registry of Life Science Web Services. It provides an open platform for Web Services registration, annotation and monitoring, with a comprehensive REST API. Moreover, BioCatalogue is a platform for Web Service providers to publish and advertise their services, providing a centralised and curated catalogue of Web Services, and to build a collaborative environment where the community can find, contact and meet the experts and maintainers of these services.

Taverna, BioCatalogue and myExperiment together create an integrated solution for bioinformatics, data analysis and analytics. We show how these tools together address the challenges of large scale data processing that comes with Next Generation Sequencing.

URL

http://www.mygrid.org.uk

Speaker

Dr. Paul Fisher[1]

Author affiliations

[1]University of Manchester

[2]University of Manchester, University of Southampton, University of Oxford, EMBL-EBI, University of Leiden

DNA Sequencing with Illumina Instruments and Chemistry: Current and Future Applications

Klaus Maisinger[1], Anthony Cox[1], Lisa Murray[1], Come Raczy[1]

The session will be separated into three parts, starting with the current range of Illumina sequencers and the latest performance figures achieved in the production sequencing facilities at Illumina Cambridge in Chesterford Research Park, UK.

The data will serve as a basis to describe the Illumina software pipeline from RTA real time analysis on the instrument PC to CASAVA for mapping and genome variation detection via grid computing.

This will be followed by a characterization of the current state of the art in high-throughput sequencing workflows. Starting from typical computing and storage volumes we will extrapolate application parameters for future research scenarios in de novo assembly, whole genome and whole exome re-sequencing as well as metagenomics. Finally, we will describe bioinformatics scientist positions currently open at our R&D site in Chesterford near Cambridge, UK.

Speaker

Dirk J. Evers, Director Computational Biology[1]

Author affiliations

[1]Illumina Inc.

Accurate Next Gen Sequencing Data Analysis on Cloud Computing

Miklós Csűrös[2], *Szilveszter Juhos*[1]

We present an internet based, automated, highly accurate genome variant analysis toolkit for next generation sequencing data called Omixon Variant Toolkit. This method is applicable for exome analysis or for analyzing highly variable genomes. Accurate, reliable and highly sensitive genome variant analysis based on next-generation sequencing short-read data poses a significant computational challenge. There is always a trade-off between computational speed and guaranteed high accuracy. (P. Flicek & E. Birney, S6| VOL.6 NO.11s| 2009| NATURE METHODS) A small increase in the desired accuracy can exponentially increase computational demand. Since many genomics applications do not require high precision, the most popular methods are fast but less accurate especially for identifying small insertions and deletions. These variants are important for variable genomes, in cancer research or biomarker discovery. We developed an accurate and computationally efficient method implemented in the Omixon Variant Toolkit which we offer as an internet-based service. In this presentation we demonstrate how to use our automated analysis tool for genome variant discovery. We also demonstrate the accuracy of the method through comparing results obtained with the most popular and the best open source tools available. We show that the Omixon Variant Toolkit is able to map short reads in the genome in areas of high density of insertions, deletions and SNP and longer (10-12 bp) deletions.

We carried out comparative analysis of 14 strains of a bacteria causing inflammatory disorder in humans including some pathogenic, non-pathogenic, drug-resistant, non-resistant strains. Our method was able to identify correctly variants and many of them were later confirmed by Sanger sequencing. We shall present additional example applications in human exome analysis.

Additional benefits of our internet-based method is that users do not need expensive computational facilities since the calculation is carried out on a cloud computing platform, where the user pays only for the particular calculation. It is significantly more cost efficient than owning your own infrastructure. In addition, our software is very easy to use and guarantees high accuracy without having to know large number of different parameter settings. Users can control the maximum number of mismatches and indels and can choose from three levels of sensitivity: fast, sensitive and ultra sensitive.

Accuracy is achieved by a spaced seeds mapping, followed by greedy extension and a precise statistical alignment based on a pair-hidden Markov model, combining DNA sequence evolution models and sequencing errors (from read quality files). The method was published in Csuros, Juhos, Berces "Fast Mapping and Precise Alignment of AB SOLiD Color Reads to Reference DNA" Springer Lecture Notes in Bioinformatics 6293:176-188, 2010.

URL

www.omixon.com

Speaker

Attila Bérces[1]

Author affiliations

[1]Omixon, Budapest, Hungary

[2]University of Montreal, Canada