Next Generation Sequencing Workshop

November 18-20, 2009. Rome, Italy

José R. Valverde

EMBnet/CNB, Centro Nacional de Biotecnología, CSIC. Madrid, Spain

Abstract

In November 2009, EMBRACE (FP6 NoE), CASPUR, EMBnet, the Italian Society of Bioinformatics and UPPMAX allied to organize a workshop on Next Generation Sequencing technologies and data analysis tools focusing on "Building Next Generation Sequencing platforms and pipeline solutions". The workshop took place in Rome and was addressed to bioinformaticians interested in discussing issues related to the management of platforms and analysis of next-generation sequencing (NGS) data. In addition to lectures from leading scientists in the field, the workshop also included a hands-on session, a "hack-a-thon" where participants could try by themselves some of the tools to solve an unpublished, real world problem in genome sequence assembly. The workshop web site is located at www.nextgensequencing.org.

Introduction

Next Generation Sequencing technologies have drawn the attention of the Molecular Biology scientific community by its unprecedented power and relatively low cost. It is now feasible to perform sequencing runs covering billions of base pairs in fragments from 35 to 400 bp that can be put to traditional (e.g. sequencing a genome with 50x coverage) and innovative (e.g. sequencing all transcripts in an organism or metabolome) uses.

The rapid advance of technology has raised concerns in many leading edge scientists, which see amazing new possibilities in this progress but find traditional tools designed to cope with conventional sequencing tasks wanting and newly developed tools too complex, difficult or specific to be conveniently applied to their interests. Out of this interest in current technological and scientific progress, arose a demand for bringing together the experts in the field to gather an overall view of the current status of the art and define future directions for dealing with the analysis problems.

In order to address this demand, the EMBRACE (EU FP6 NoE), UPPMAX and CASPUR with the support of the Italian Society of Bioinformatics and EMBnet organised a workshop in Rome, from the 18th to the 20th of November on Next Generation Sequencing technologies, addressed to bioinformaticians interested in learning more and discussing about these new techniques and the approaches to analyze the data they generate building appropriate infrastructures and data analysis pipelines.

Actual interest in the workshop was reflected in the fact that the original plans had to be altered in order to accommodate the large demand of scientists wanting to participate and that required reassignment and location of meeting facilities in order to be able to expand available space and increase the number of participants.

The workshop was organized with some major goals in mind: first of all, it should help participants get an overview of the applied scientific problems being addressed and the tools used to tackle them; secondly, we wanted to show as well the daunting magnitude of the problem posed by these new technologies and the complex platforms needed to be able to manage the large amounts of data generated; and finally, we wanted to provide both specialized and user views on the problem and provide a nurturing environment for discussions that fostered the exchange of experiences and feedback to developers shedding new light towards future directions to be pursued.

To this end, the organizers assembled an international team with varied expertise, interests and geographical provenance to constitute the organizing committee and ensure appropriate coverage of all the topics addressed (Erik Bongcam-Rudloff, Tiziana Castrignano, Eija Korpelainen, Inge Jonassen, Graziano Pesole, Nils-Einar Eriksson, Etienne deVilliers, Andreas Gisel, Laurent Falquet, Jose R Valverde and Gert Vriend). The committee in turn assembled a pan-

Figure 1. Group picture of workshop attendees.

el of speakers that we are glad to report successfully met in our view and, most important, in the eyes of all participants, the goals defined at the onset.

Workshop contents:

Welcome address, Erik Bongcam-Rudloff

The workshop opened with a welcome address of the chairman of the Organizing Committee, Dr. Erik Bongcam, who introduced rational for the meeting, the main lines to be developed, and thanked all the people and organizations that contributed to make it possible: EMBRACE (thanks to Gert Vriend), CASPUR (thanks to Tiziana Castrigliano), the Engineering Dept. of University of Rome, the Uppsala HTC facility UPPMAX (thanks to Ingela Nyström) and the EMBnet.

After this keynote address, the workshop moved on to address the issue of building sensible analysis platforms for NGS data analysis and where the main approaches used to address this problem where reviewed first hand:

Hardware and Storage, Tony Cox, Welcome Trust Sanger Institute, United Kingdom

In his talk, Tony Cox described the major challenges posed by NGS to Information Technologies and Data Management: data generated is huge, easily resulting in 500.000 images, with a weight of 8MB each, per run. These will need to be analyzed and converted to sequence data prior to any actual practical use. Once sequences are available, various uses will be possible depending on the input samples and the intended use of the information, resulting in turn in large analysis result datasets as well. From here, we have only reached the first step in the scientific inference process and further downstream analysis will come as scientists use these results to build upon new analysis.

All in all, this results in a logical division of the process into a pipeline, where output from one step feeds the next in the analysis process, all of them generating vast amounts of data that needs to be efficiently and safely managed.

Tony Cox described the approach taken by a major sequencing oriented institution such as the Sanger Center, the challenges they found, the solutions the came up with and the experiences they could drive out from the process, giving a series of rules of thumb useful for anybody planning to deploy an NGS data analysis infrastructure, probably the most relevant of which was the advice to plan for change: technology is evolving quickly and users are continuously coming up with new applications.

Data storage for HTS platforms. George Magklaras. Biotechnology Centre, Oslo, Norway

George Magklaras addressed the problem posed by High Throughput Sequencing (HTS) technologies from a similar standpoint, but with a different twist: the case of a centralized computing facility that has to cope with the needs of a

distributed network of users spread at a national level.

This presentation, like all the others that followed, confirmed the soundness of the pipeline approach to data analysis and management and reviewed the requirements and needs of a large community of high throughput sequencing users, presenting estimates well in line with the experience of Sanger Center and all the other speakers.

In his talk, George Magkalas concentrated on the various solutions available to deal with the management and storage of the data generated and pointed out the pros and cons of each solution, giving a set of recommendations to help define the hardware and software requirements of a sound NGS data analysis platform.

A multidisciplinary platform of computing resources, large scale storage and know-how in Uppsala, Jonas Hagberg, Uppsala University, Sweden.

The next presentation addressed the same problem, dealing with data analysis from scientists at a national level using a different approach: deployment of a coordinated, distributed national e-infrastructure over various centres.

Again, Jonas Hagberg confirmed the requirement estimates and experiences described on previous talks, agreeing in the basic pipeline approximation, but his talk provided an altogether different point of view to solving user needs, based on extended consultation with users and experts and a distributed solution to platform implementation, presenting rationale for the decisions they made and the solutions chosen to provide adequate support to users.

Sequence read archives at EBI, Guy Cochrane, EBI, United Kingdom

Guy Cochrane introduced the European Nucleotide Archive (ENA) being developed at EBI and approached the problem of data storage from a new perspective: that of pure data management instead of concentrating on solving the infrastructure problem (already addressed by G. Magklaras).

Guy Cochrane described the mechanisms and services being implemented in ENA, the kind of users it is addressed to and the uses currently envisioned for this archive as well as the services being implemented to deliver the functionality planned, addressing issues like data formats, standardization and coordination with other major players like NCBI. He also described the current approaches they are considering for data submission, retrieval and distribution, and gave us a glimpse to what they already have in the works for the future.

Bioinformatics for NGS Giorgio Valle, CRIBI, University of Padua, Italy

Giorgio Valle gave us yet another view on how to deal with the NGS data nightmare: that of an international consortium addressing major sequencing undertakings. His talk also acted as a knee play to the next topics in the workshop, moving attention towards the problem of data analysis itself.

The experiences derived from major sequencing enterprises, like the tomato genome was reviewed and put in historical context, showing the advantages that new NGS technologies can provide to classical sequencing approaches and the useful combination of both.

In addition, Giorgio covered the various approaches used to deal with a major undertaking and reviewed the software tools used and the development work they had to do to assemble, build scaffold paths and manage redundancy in the data to reconstruct the base genetic information needed to procced to their next logical step: gene prediction using sequence alignments and ab-initio methods, and subsequent whole transcriptome analysis, thus highlighting the dramatic nature of the problem: genomic sequencing is but the first step in a series of increasingly more complex and useful analysis where NGS techniques can also be of great help.

Mapping short reads as numbers, Alberto Policriti , Applied Genomics Institute Univeristy of Udine, Italy

The problem of sequence alignment, as demonstrated in the previous talk, is central to most of the analysis we want to perform on NGS data. In this talk, Albeto Policriti reviewed the basic theory related to the problem as a basis to describe new methods in development and use that aim at overcoming traditional shortcomings of classical algorithms. He gave a cursory review of many of the current tools of the trade. Then he moved on to present some of their most novel algorithms being developed to accurately match short tags and their applications.

Figure 2. Fontana di Trevi.

NGS alignment and applications, Zeming Ning, Sanger Center, United Kingdom

Zeming Ning pursued the methodological analysis of the tools used for NGS data analysis by presenting the new SSAHA2 algorithm. Again, he reviewed existing methods, their advantages and shortcomings and used this context to describe recent advances and applications as well as the typical analysis workflow.

Project HOPE: The last piece of the puzzle, Hanka Venselaar, University of Nijmegen, The Netherlands

With the methodological basis for sequence alignment covered, the workshop moved on to downstream data analysis. Hanka Venselaar described project HOPE, a new tool aimed to make easier the life of end users by allowing them to automate the comprehensive (to the limits of available knowledge) analysis of a protein, using both sequence comparisons, structural predictions and structural modeling and comparison to deliver ready to use, understandable knowledge.

Hanka's talk marked a new turning point in the workshop, shifting the topic from raw sequence comparison to downstream applications.

De novo assembly, Laurent Falquet, Vital_IT, SIB, Switzerland

In this talk, Laurent Falquet addressed the issue of de novo sequence assembly, reviewing the problems and challenges, the basic methodological and algorithmic approaches and the most popular tools in use, without forgetting to drive attention to the upcoming problem of ultra-high throughput sequencing of thousands or millions of genomes.

Given the core interest of this problem, a good understanding of the tools and how well they compare against each other, how and when they should be used and the limits of the state of the art is a most valuable knowledge to derive from experience. Laurent Falquet drove of their ongoing projects and on test experiments they carried out to distill this knowledge and present it in a useful way.

Vertebrate sequencing, Jim Stalker, Sanger Institute, United Kingdom

The current interest in NGS has driven many scientists to start a wealth of specific projects, as described by previous talks. But well beyond these "petty" problems, major sequencing institutes are already considering vast problems like the "1000 genomes" project that will collect 1000 human genomes from around the world... and more projects aimed at other organisms.

In his talk, Jim Stalker described how Sanger Center and EBI as coordinators manage the huge load of data, defining standards for storage and workflows that help involved scientists efficiently address these large projects that, although seemingly huge today, may become standard experiments sooner than expected. Their experience in dealing with these is most helpful to organize and plan work at any institution interested in NGS, and was summarized as a series of recommendations of general application.

Alternative splicing using NGS data, Graziano Pesole, University of Bari, Italy

Moving ahead with the problem of practical data analysis, one of the obvious steps after genome sequencing is transcriptional analysis, identification of expressed genes and characterization of their transcripts. Graziano Pesole proceeded to present the recent breakthroughs in determining transcriptional maps and identifying alternative splice sites, highlighting the relevance of what we know is a major issue that is expanding up to 10 times the human transcriptome.

Here again, deep sequencing has obvious advantages: it is now feasible to identify large numbers of transcripts by sequencing the cDNAs and mapping these back to the genome, hence gaining first hand experimental evidence of alternative gene expression. Graziano Pesole reviewed the various alternative approaches available to analyze the data, identify transcription start and termination and splice sites and reviewed the existing software that can be used to this end.

NGS transcriptomics, Marc Sultan, Max Planck Institute Berlin, Germany

Marc Sultan pursued the quest into transcriptomics started by Graziano Pesole in the next talk: with NGS not only can we identify transcripts, we can also quantify gene expression by measuring the presence of each transcript in the recovered sequences. Marc Sultan reviewed his experience and gave useful advice on how best to plan and pursue a gene expression experiment using NGS, reviewing the roadblocks, biases and artifacts we need to be aware of and the methods to evaluate statistical significance of resulting data.

Marc Sultan also provided an end user point of view and raised the issue of usability of existing tools for facilitating and accelerating their adoption and exploitation in the real world.

Chipster for NGS, Aleksi Kallio, CSC Finland

As an answer to the complains on usability, Aleksi Kallio described recent advances being added into CHIPSTER, a tool originally developed for microarray data analysis that is now moving into NGS with the integration of analysis and visualization extensions. To top it all, CHIPSTER, developed at the Finish EMBnet node, is an open source tool that can be freely installed and extended.

Sequence clustering (SeedMap and mBED), Des Higgins, University College, Dublin, Ireland

As other speakers had mentioned by now, the analysis of an individual species genome is but the first step in a long ladder to understand the tower of life. From here, an obvious jump is to compare sequences from many genomes or, given availability of NGS, of tens or thousands of sequences from different species.

The classic program, and probably the most widely used still nowadays for sequence alignment with phylogenetic intent is Clustal. In his talk, Des Higgins described recent methods being developed to deal with the problem of aligning and clustering such large numbers of sequences increasing efficiency and accuracy. After reviewing existing methods and recent developments, Des Higgins described the new methods being developed currently to deal with full genomes or large numbers of sequences.

Metagenomics, Daniel Huson, University of Tubingen, Germany

Once we have a method to cluster sequences we can also classify their respective organisms and move on to the thriving field of metagenomics and the problemof taxonomical and functional analysis. Daniel Huson introduced the problem of studying DNA of uncultured organisms (metagenomics) and the metagenome as the cumulative genetic information of a community of organisms.

Metagenomics can address many new problems but poses additional challenges as well. Thanks to NGS it is now possible to characterize the distribution of species in a community and to compare metagenomes to analyze populations and their evolution and changes in response to environment. And beyond: we can not only identify species, we can also identify the genes present in a community and understand its behavior as a coherent whole. Thanks to tools like MEGAN it is now becoming increasingly easy to perform these studies once all the preparatory comparison work has been carried out.

EGI LS-SSC, Tristan Glatard, CNRS Clermont-Ferrand, France

Despite its obvious interest, most of the preceding talks relied heavily on similar preparatory work requiring treatment and analysis of huge numbers of sequences, a task usually requiring unprecedented computing power rarely available in common scientific institutions. It is here where all the work that has been invested in Grid computing can play a major role.

Tristan Glatard introduced next EGI LS-SSC, part of a bit project proposal to EU, ROSCOE, which will try to provide support for Life Sciences in the Grid. He gave examples of what can be done and what is being developed on the Grid illustrating its significant potential to bring scientists the computing power they need.

Where from here? Open discussion.

The formal workshop talks closed with an open discussion where developers, users and in general all participants were able to exchange opinions, put forward projects and ideas and provide feedback and interactions that we hope will prove useful in the near future to orient work. Most

Figure 3. Hack-a-thon session: Laurent Falquet.

significantly, this session may become the seed for a new community with common interests that we hope will grow healthier and stronger in the future.

The hack-a-thon.

While the formal content of the workshop terminated with the presentations, the associated activities did not: for all participants interested, we also organized a practical hands on session where a sample problem in de novo sequencing was to be resolved using various different combinations of experimental datasets and methodological approaches.

The hack-a-thon, directed by Laurent Falquet, gave participants first hand experience on solving a real world problem using unpublished data, allowing them to compare their results with those of experts in the field and to demonstrate that each working group found another result dependent on which parameter settings were used - there are many of them - and to gather a direct feel of the challenges posed by NGS data analyis.

Figure 4. Participants at the hack-a-thon session.

Conclusion

The general opinion among all participants in the workshop was that it was a successful meeting, that covered in depth and breadth the main topics related with the status of the art in NGS data analysis, with an excellent balance in dealing with the many dimensions of the problem (hardware, data analysis, algorithms, tools, practical applications, real world problems and hands-on experience).

This has been a pioneering Workshop in this topic that has been able to achieve a broad coverage of the field with major scientists, allowing us to collect common points, advice and expertise. We intend to reflect all this knowledge in a white paper that will be available with all information we gathered during this workshop on the Next Generation Sequencing web site www.nextgenerationsequencing.org.

In addition, the organizers hope that it will have served as a meeting point for major workers and developers in the field, where they have been able to exchange opinions and experiences, find points in common and gather user feedback, helping them put their work in context and derive useful information to plan their work ahead. This process will be supported with a NGS forum soon available on our NGS web site.t

Finally, we all hope that this proves a useful seed for an emerging community that will grow from the original participants to a large group of professionals sharing a common interest and passion in science.

Acknowledgements

We want to thank first and foremost the institutions that have made this workshop possible through their support, specially, the European Commission through its funding for the EMBRACE Network of Excellence (the EMBRACE project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092), UPPMAX (the Uppsala Multidisciplinary Center for Advanced Computational Science), CASPUR (Consorzio interuniversitario per le applicazioni di supercalcolo per università e ricerca), the Italian Society of Bioinformatics, and EMBnet (the European Molecular Biology Network).

The organizers want to express their deep gratitude to all the invited speakers who graciously accepted to participate and contributed to make this a most interesting and scientifically sound event, and, of course, all the participants who, through their enthusiasm and interest made this a successful smash hit.

Finally, we want to apologize to all the scientists that could not be admitted to the workshop due to lack of resources and available space despite their earnest and keen expressions of interest. We are already planning new workshops to adapt to the increasing demand.