DOE Bioinformatics Panel Discussion

Dr. Matt Hahn
Chief Technology Officer, Accelrys
March 7, 2012
Introduction to Accelrys

An Open Scientific Informatics Platform

Partnering Example: Next Gen Sequencing

Areas of collaboration
Accelrys is a scientific informatics software company focused on serving organizations that exploit scientific innovation to more efficiently bring products to market.
What We Do

- We provide our 1,350 global customers with an enterprise scientific informatics platform that delivers:
  - Better manage scientific workflows, experiments and process optimization
  - Improved insight and management of complex scientific data across disciplines
  - Increased use of modeling and simulation to improve outcomes and reduce cost and cycle time
Many solutions require working across domain boundaries
Data Pipelining

A powerful, proven paradigm for data processing

Graphical pipelines guide the flow of data through a network of modular computational and visualization components

Access data from multiple, disparate sources

Process, manipulate, filter and analyze data

Output data to files, databases, and reports
Pipeline Pilot Enables...

- Integration of multiple disparate data types and sources
- Integration of disparate applications
- Capture and deployment of best practice processes
- Automated execution of routine processes
- Rapid creation of interactive reports and web applications
• Integration of multiple disparate data types and sources
• Integration of disparate applications
• Capture and deployment of best practice processes
• Automated execution of routine processes
• Rapid creation of interactive reports and web applications
Pipeline Pilot Enables...

- Integration of multiple disparate data types and sources
- Integration of disparate applications
- Capture and deployment of best practice processes
- Automated execution of routine processes
- Rapid creation of interactive reports and web applications
Pipeline Pilot Enables...

- Integration of multiple disparate data types and sources
- Integration of disparate applications
- Capture and deployment of best practice processes
- Automated execution of routine processes
- Rapid creation of interactive reports and web applications
Integration of multiple disparate data types and sources
Integration of disparate applications
Automated execution of routine processes
Capture and deployment of best practices
Rapid creation of web services, interactive reports, and web applications
Complexity of NGS Data Analysis
Single Greatest Impediment to Growth

GenomeWeb In Sequence Survey

Section 7: Looking

Figure 9. Improvements

Mardis Genome Medicine 2010, 2:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

Having recently attended the Personal Genomes meeting at Cold Spring Harbor Laboratories (I was an organizer this year), I was struck by the number of talks that described the use of whole-genome sequencing and analysis to reveal the genetic basis of disease in patients. These patients included a child with irritable bowel disease, a child with severe combined immunodeficiency, two siblings affected with Miller-Dieker syndrome, and several

BOX 3
Bioinformatics

Bioinformatics accounts for >40% of non-instrument NGS costs

Unpublished Survey From Major Instrument Vendor

NHGRI Perspective Feb 2011
Pipeline Pilot For NGS

NGS Component Collection

Biology
- Catalyst (Pharmacophore)
- CHARMm (Simulations)

Life Science M&S
- Catalyst (Pharmacophore)
- CHARMm (Simulations)

Chemistry
- Chemistry
- ADMET
- Cheminformatics

Materials M&S
- Materials
- Polymeric Properties
- Synthia

Pipeline Pilot For NGS
- NGS Component Collection
- Catalyst (Pharmacophore)
- CHARMm (Simulations)
- Materials
- Polymeric Properties
- Synthia
- Pipeline Pilot For NGS
NGS Collection Capabilities

**Quality Assessment and Processing**
- Assess run quality and coverage
- Filter data based on pairs proximity, quality and repeats
- Trim reads

**Assembly and Mapping**
- Velvet, MIRA3 for de novo assembly
- Mapping reads with BWA, Bowtie, mapreads
- Comprehensive mapping and assembly stats

**Variant Detection, Profiling, Comparisons**
- Identify SNPs, Indels, CNVs, Structural Variants
- Automated comparisons among samples and feature sets

**Transcript and Gene Regulation Experiments**
- Analyze RNA-Seq data to evaluate expression patterns and alternative transcript structures
- Analyze ChIP-Seq data to assess transcription factor binding patterns

**Powerful Repository Model**
- Uses industry standard formats (FASTA, BAM, GFF3/Tabix)
- Enables flexible storage and sharing options
- Powerful location-based query components
Partnership Example: Oxford Nanopore and Accelrys
The Next Generation Of Sequencing Innovation

OXFORD NANOPORE and PIPELINE PILOT

NGS in the palm of your hand...
“As data volumes grow and algorithms continue to rapidly evolve, the informaticians and scientists involved in NGS analysis are challenged to keep pace. The Accelrys Pipeline Pilot software is uniquely placed to deliver flexible workflow features for a broad range of users, from individual researchers to large scale installations.”

Clive Brown
Chief Technology Officer
Oxford Nanopore Technologies

PIPELINE PILOT and OXFORD NANOPOR
 Scaling NGS from the desktop to the largest installations.
GridLyn and Pipeline Pilot Enable Real-Time Analysis and Instrument Control

**Sequence Until:**

- Depth of coverage reached
- Desired level of accuracy
- Variants detected in genes of interest
- Retrospective experiments (clarify ambiguities)

- Most any computable result can be described within Pipeline Pilot and used to control the instrument
Consensus SNPs

SNP calls from the two algorithms.

Mouse over the Venn diagram to get the numbers of calls in each set.

The first table contains all common SNPs. (Dark brown = both algorithms have an allele score \( \geq 100 \).
The second table contains those SNPs just called by one algorithm. Both tables are currently sorted by

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>referenceBase</th>
<th>consensusBase</th>
<th>SNPQuality</th>
<th>depth</th>
<th>consensusQuality</th>
<th>mappingQuality</th>
<th>readBases</th>
<th>genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3712</td>
<td>3712</td>
<td>C</td>
<td>S</td>
<td>93</td>
<td>14</td>
<td>93</td>
<td>60</td>
<td>$650G0G0G_9$</td>
<td></td>
</tr>
<tr>
<td>4161</td>
<td>4161</td>
<td>A</td>
<td>M</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>40</td>
<td>$555C555C$</td>
<td></td>
</tr>
<tr>
<td>4849</td>
<td>4849</td>
<td>A</td>
<td>M</td>
<td>47</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>$555C555C$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SNPs called by both algorithms

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>referenceBase</th>
<th>consensusBase</th>
<th>SNPQuality</th>
<th>depth</th>
<th>consensusQuality</th>
<th>mappingQuality</th>
<th>readBases</th>
<th>genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3712</td>
<td>3712</td>
<td>C</td>
<td>S</td>
<td>93</td>
<td>14</td>
<td>93</td>
<td>60</td>
<td>$650G0G0G_9$</td>
<td></td>
</tr>
<tr>
<td>4161</td>
<td>4161</td>
<td>A</td>
<td>M</td>
<td>92</td>
<td>92</td>
<td>92</td>
<td>40</td>
<td>$555C555C$</td>
<td></td>
</tr>
<tr>
<td>4849</td>
<td>4849</td>
<td>A</td>
<td>M</td>
<td>47</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>$555C555C$</td>
<td></td>
</tr>
</tbody>
</table>

Easily Compare Results From Different Algorithms
Data Pipelining For NGS Analysis: Easily Deploy Validated Methods To End Users

Step 1: configure the protocol, choose parameters to expose

Step 2: save the protocol in the ‘Web Services’ Folder

Step 3: Now end users can run the protocol and view results from any browser
Final Thoughts

• Accelrys’ broad vision requires us to partner
• Open scientific platform facilitates integration with 3rd party solutions
• Many areas of interest and potential collaboration
  • Biology Mod/Sim (e.g. structure, binding, aggregation prediction)
  • Materias Mod/Sim (e.g. large scale QM, MesoScale modeling)
  • Image analysis (e.g. feature identification, classification)
  • “Scientific” big data (e.g. genomic, scientific documents)
• Demonstrated ability to partner
  • In-license 3rd technology
  • ISV partner program
  • Co-development efforts
  • Out-license and OEM our technology
  • Acquisition