Making the Web Work for Science: The Impact of e-Science and Cyberinfrastructure

CENDI/NFAIS/FLICC Workshop

December 8, 2008

Chris Greer
Director, US National Coordination Office
Networking and Information Technology Research and Development Program
Acronyms:

**NITRD**
Networking and Information Technology Research and Development Program

**NCO**
National Coordination Office
Science is global and thrives in a world that is not limited to 4-dimensions

Image: Andrew J. Hanson
www.cs.indiana.edu/~hanson/
Cyberinfrastructure

Computational capacity and capability

Information for innovation and discovery

Connectivity for access and interaction
Increasing Computational Capacity and Capability

ENIAC, dedicated in 1946, was one of the first fully-functional digital computers, using 17,000 vacuum tubes for up to 5,000 addition operations per second. Today’s petascale machines are designed to sustain more than one quadrillion (1,000,000,000,000,000) operations per second. A calculation these machines could complete in a week would take a machine operating at ENIAC speeds several billion years.
Simulation of 7.7 earthquake on lower San Andreas Fault

- Physics-based dynamic source model – simulation of mesh of 1.8 billion cubes with spatial resolution of 200 m
- Builds on 10 years of data and models from the Southern California Earthquake Center
- Simulated first 3 minutes of a magnitude 7.7 earthquake, 22,728 time steps of 0.011 second each
- Simulation generates 45+ TB data

Source: Fran Berman, Director, SDSC
Behind the Scenes – Enabling Infrastructure for TeraShake

- **Computers and Systems**
  - 80,000 hours on 240 processors of DataStar
  - 256 GB memory p690 used for testing, p655s used for production run, TG used for porting
  - 30 TB Global Parallel file GPFS
  - Run-time 100 MB/s data transfer from GPFS to SAM-QFS
  - 27,000 hours post-processing for high resolution rendering

- **People**
  - 20+ people involved in information technology support
  - 20+ people involved in geoscience modeling and simulation

- **Data Storage**
  - 47 TB archival tape storage on Sun StorEdge SAM-QFS
  - 47 TB backup on High Performance Storage system
  - SRB Collection with 1,000,000 files

Source: Fran Berman, Director, SDSC
A Critical Role of a Cellular Membrane Traffic Protein in Poliovirus RNA Replication

Goerge A. Beloq, Qin Fang, Kristina Nikolic, Catherine L. Jackson, Ellis Scharffl

Abstract
Regulation of viral RNA synthesis is accomplished by a complex interplay of cellular and viral factors. The presence of cellular inhibitors of RNA synthesis, such as inhibitors of RNA polymerase II, can lead to the inhibition of viral RNA synthesis. Therefore, it is critical to understand the mechanisms by which viral RNA synthesis is regulated and to identify potential targets for antiviral therapy. In this study, we analyzed the effects of cellular inhibitors of RNA synthesis on viral RNA synthesis and identified a novel cellular membrane traffic protein that plays a crucial role in viral RNA synthesis.

Introduction
Cellular inhibitors of RNA synthesis, such as inhibitors of RNA polymerase II, can lead to the inhibition of viral RNA synthesis. Therefore, it is critical to understand the mechanisms by which viral RNA synthesis is regulated and to identify potential targets for antiviral therapy. In this study, we analyzed the effects of cellular inhibitors of RNA synthesis on viral RNA synthesis and identified a novel cellular membrane traffic protein that plays a crucial role in viral RNA synthesis.

Figures

1. Model for the interaction of viral RNA with cellular inhibitors of RNA synthesis

2. Cellular inhibitors of RNA synthesis inhibit viral RNA synthesis

Table

<table>
<thead>
<tr>
<th>Protein</th>
<th>Number</th>
<th>Mass (Da)</th>
<th>Charge</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>1</td>
<td>15,440</td>
<td>60,993</td>
<td>40,699</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>15,297</td>
<td>61,381</td>
<td>39,289</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>15,706</td>
<td>60,349</td>
<td>38,282</td>
</tr>
<tr>
<td>O</td>
<td>4</td>
<td>16,675</td>
<td>59,665</td>
<td>38,561</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>16,879</td>
<td>62,723</td>
<td>39,019</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
<td>14,995</td>
<td>63,712</td>
<td>38,289</td>
</tr>
<tr>
<td>O</td>
<td>7</td>
<td>15,742</td>
<td>62,495</td>
<td>37,911</td>
</tr>
<tr>
<td>C</td>
<td>8</td>
<td>16,346</td>
<td>65,548</td>
<td>38,333</td>
</tr>
<tr>
<td>G</td>
<td>9</td>
<td>16,761</td>
<td>64,459</td>
<td>37,445</td>
</tr>
<tr>
<td>O</td>
<td>10</td>
<td>15,476</td>
<td>61,193</td>
<td>36,282</td>
</tr>
<tr>
<td>C</td>
<td>11</td>
<td>16,110</td>
<td>66,049</td>
<td>40,571</td>
</tr>
<tr>
<td>G</td>
<td>12</td>
<td>15,051</td>
<td>60,292</td>
<td>37,139</td>
</tr>
<tr>
<td>O</td>
<td>13</td>
<td>15,427</td>
<td>60,371</td>
<td>36,080</td>
</tr>
<tr>
<td>C</td>
<td>14</td>
<td>16,576</td>
<td>69,931</td>
<td>35,282</td>
</tr>
<tr>
<td>G</td>
<td>15</td>
<td>16,402</td>
<td>60,915</td>
<td>34,519</td>
</tr>
<tr>
<td>O</td>
<td>16</td>
<td>14,219</td>
<td>59,105</td>
<td>35,179</td>
</tr>
<tr>
<td>C</td>
<td>17</td>
<td>14,029</td>
<td>59,293</td>
<td>34,911</td>
</tr>
<tr>
<td>G</td>
<td>18</td>
<td>14,052</td>
<td>59,381</td>
<td>34,702</td>
</tr>
<tr>
<td>O</td>
<td>19</td>
<td>15,419</td>
<td>60,563</td>
<td>32,829</td>
</tr>
<tr>
<td>C</td>
<td>20</td>
<td>17,764</td>
<td>59,310</td>
<td>35,364</td>
</tr>
<tr>
<td>G</td>
<td>21</td>
<td>18,859</td>
<td>59,787</td>
<td>34,498</td>
</tr>
<tr>
<td>O</td>
<td>22</td>
<td>18,523</td>
<td>59,598</td>
<td>33,003</td>
</tr>
</tbody>
</table>

References


CRYSTAL STRUCTURE OF PUTATIVE NUDIX HYDROLASE FAMILY MEMBER FROM CHROMOBACTERIUM VIOLACEUM


NEW YORK SGX RESEARCH CENTER FOR STRUCTURAL GENOMICS (NYSGXRC)

27-OCT-08

Source: RCSB Protein Data Bank; www.rcsb.org
Redefining “Computer”

• “All the calculations that would ever be needed in this country could be done on three digital computers.”
  -- Douglas Hartree, Cambridge, 1954

• “There is no reason for an individual to have a computer in their home.”
  -- Ken Olsen, DEC, 1977

• “For the full year [2007], IDC said 269 million PCs were shipped worldwide”

• “In a sense, there are only five computers on earth.”

• “…some researchers at IBM believe that five computers may be four too many.”
Cyberinfrastructure

Information for innovation and discovery
“Sometime in the 2010s, if all goes well, the Large Synoptic Survey Telescope (LSST) will start to bring a vision of the heavens to Earth. Suspended between its vast mirrors will be a three billion-pixel sensor array, which on a clear winter night will produce 30 terabytes of data. In less than a week this remarkable telescope will map the whole night sky …. And then the next week it will do the same again … building up a database of billions of objects and millions of billions of bytes.”
Large Hadron Collider

Physicists will use the LHC to recreate the conditions just after the Big Bang, by colliding two beams [of hadrons] head-on at very high energy.
When LHC begins operations, it will produce roughly 15 Petabytes of data annually, which thousands of scientists around the world will access and analyse … The mission of the LHC Computing Project (LCG) is to build and maintain a data storage and analysis infrastructure for the entire high energy physics community that will use the LHC.

“In 2006, the amount of digital information created, captured, and replicated was $1,288 \times 10^{18}$ bits (or 161 exabytes) … This is about 3 million times the information in all the books ever written”
Information And Storage

Source: John Gantz, IDC Corporation
The Expanding Digital Universe

Transient information or unfilled demand for storage

Information

Available Storage

Petabytes Worldwide

2005 2006 2007 2008 2009 2010
Google's founding philosophy is that we don't know why this page is better than that one: If the statistics … say it is, that's good enough. No semantic or causal analysis is required. That's why Google can translate languages without actually "knowing" them (given equal corpus data, Google can translate Klingon into Farsi as easily as it can translate French into German).
Cyberinfrastructure

Connectivity for access and interaction
The Department of Defense’s ARPANET project, launched in 1966 to explore methods for “resource sharing among computers”, initially connected 4 nodes. Today’s Internet links more than 1.4 billion users over more than 200,000 networks worldwide; with 14 new users added every second.
“New types of scientific organizations [that] serve individuals, teams and organizations in ways that revolutionize ... 

... what they can do, how they do it, and who participates.”
Richard Jorgensen, Collaborative Director, Plant Sciences/BIO5, University of Arizona
Steve Goff, Director of Community Interactions, University of Arizona
Rick Blevins, Director of Cyberinfrastructure Development, University of Arizona
Martha Narro, Director of Education, Outreach and Training, University of Arizona
Greg Andrews, Team Leader, Computing Infrastructure Team, Computer Science, University of Arizona
Sudha Ram, Co-Team Leader, Integrated Solutions Team, Management Information Systems, University of Arizona
Lincoln Stein, Co-Team Leader, Integrated Solutions Team, Bioinformatics, Cold Spring Harbor Laboratory
Steve Rounsley, Team Leader, Synthesis Activities Team, BIO5, University of Arizona
Vicki Chandler, Team Leader, Education, Outreach and Training Team, Director of BIO5, University of Arizona
Susan Brown, Team Leader, Social Science Team, Management Information Systems, University of Arizona
Kobus Barnard, Computer Science, University of Arizona
Travis Huxman, Ecology and Evolutionary Biology, University of Arizona
Dan Stanzione, Director, Fulton High Performance Computing Center, Arizona State University
Ann Stapleton, Biology, University of North Carolina-Wilmington
Matt Vaughn, Bioinformatics, Cold Spring Harbor Laboratory
Doreen Ware, Bioinformatics, USDA/Cold Spring Harbor Laboratory
A Discovery Environment

Remote repositories
Local datasets
Software
Algorithm
Computational tools & resources
Visualization
Machine Level Interface (API)
Campus Clusters; TeraGrid
Web Services Interface

Community Collaboration Platforms
Distributed Content Creation (e.g. community annotation)
Community Interoperability frameworks
Web 2.0
INTEGRATION
Discovery Landscape

• Distributed
• Integrated
• Open source
• Evolvable
• Extensible
The Future: Next Generation Networking and Information Technology

- Emergent Cognition
- Beyond Virtual
- Pathways to Trust and Confidence
Emergent Cognition

Harness new levels of intelligence, intuition and perception that emerge as greater than the sum of the parts in the union of cyber, human, and social capabilities.

Example: Enable the combined use of logic, analysis, and modeling; expert intuition and perception; and market judgment and crowd wisdom in understanding the properties and behavior of a global economy.
Beyond Virtual

Extend the envelope of capabilities by making the virtual and the physical one.

Example: Cognitive computing – Provide the ad hoc capability to access the full sensing, information, and computational capabilities of the every day objects around you to respond to an emergency or to an ordinary need.
Pathways to Trust and Confidence

*Ensure the value to society of cyber systems is not limited by a lack of trust and confidence.*

**Example:** Provide capabilities for identity management, information security, privacy protection, and tiered access to enable an electronic health records system that can be confidently used for individual patient care, public health, and biomedical research.
eLibrary - Single point of access connecting resources worldwide:

• Digital and physical repositories

• Computational capabilities

• Digital environments for access and interaction by people, machines, and organizations

• Software tools and resources
Contact information:

greer@nitrd.gov