Apr 6th, 12:00 PM

Bioinformatics: Alive and Kicking

David L. Osterbur

Harvard Medical School

Follow this and additional works at: http://escholarship.umassmed.edu/escience_symposium

Part of the Bioinformatics Commons, and the Library and Information Science Commons

http://escholarship.umassmed.edu/escience_symposium/2009/program/5

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in University of Massachusetts and New England Area Librarian e-Science Symposium by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Bioinformatics: alive and kicking

David L. Osterbur
Too Central

• “Bioinformatics has become too central to biology to be left to specialist bioinformaticians. Biologists are all bioinformaticians now.”

Google Generation

NCBI/BLAST/blastp suite: BLASTP programs search protein databases using a protein query.

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Are you feeling lucky?

Or, upload file

Job Title

I'm Feeling Lucky

Search database nr using Blastp (protein-protein BLAST)

Algorithm parameters
History
Auto Designers
Auto Mechanics
Driver’s Ed
Keeping up to speed...

Not just how to drive but choosing the right tools.
...to reach your goal on time.
The Library’s Role

Why Libraries?

• We are a service organization.

• We are already good at organizing, distributing and teaching access to many different types of information.

• We are a shared organization, not “owned” by any one department or unit.

• “Librarians like to search…

  …everyone else likes to find”
Why Your Library?

(15) Translational Science

15-LM-101*  Presenting genome information in electronic health records. Develop approaches for presenting relevant genomic information in an understandable way, in the context of a patient's electronic health record. As genomic data becomes available for more individuals, these data must be integrated into electronic health records in ways that: help clinicians and patients understand the significance of the data; provide an avenue for alerting clinicians and patients when new knowledge from GWAS, etc. arises to the level of potential clinical impact; and enable linking to effective decision support. Contact: Dr. Jane Ye, 301-594-4882, yej@mail.nih.gov.

15-LM-102  Computational hypothesis generation for biology and medicine. Employing two or more sources, use advanced computational approaches to generate a new and meaningful hypothesis in biomedical science, capable of being tested by bench or clinical research. One source must be full-text published biomedical literature; the other source should be either (1) a database storing primary data from basic biomedical research or (2) data drawn from the electronic health records used for routine clinical care or from the data accumulated for a clinical research project. The user interface of an integrated hypothesis generation system should support easy use by the intended users (i.e., by biomedical researchers or clinicians). Mining techniques should involve minimal human intervention. Contact: Dr. Valerie Florance, 301-594-4882, florancev@mail.nih.gov.
Google Generation

NCBI/BLAST/blastp suite: BLASTP programs search protein databases using a protein query.

Enter Query Sequence

Are you feeling lucky?

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

I'm Feeling Lucky

Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters
BLAST Results

Gene ID: 28056911 | glutamic-oxaloacetic transaminase 1, soluble (aspartate aminotransferase 1) [Homo sapiens] (Over 10 PubMed links)

Score = 860 bits (2223), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 413/413 (100%), Positives = 413/413 (100%), Gaps = 0/413 (0%)

Query 1
MAPPSVFAPQAPFVTLKTDQFPDPKYNLVGAYRDCHFWLPFPVKVEQK 60
MAPPSVFAPQAPFVTLKTDQFPDPKYNLVGAYRDCHFWLPFPVKVEQK 60

Sbjct 1
IANDNSLHPEYLPILGAAFRSCASRIAGDDGSLPKKEKRGVGQSLGGETALRIGADFL 120
IANDNSLHPEYLPILGAAFRSCASRIAGDDGSLPKKEKRGVGQSLGGETALRIGADFL 120

Query 61
IANDNSLHPEYLPILGAAFRSCASRIAGDDGSLPKKEKRGVGQSLGGETALRIGADFL 120
IANDNSLHPEYLPILGAAFRSCASRIAGDDGSLPKKEKRGVGQSLGGETALRIGADFL 120

Sbjct 61
ARWNGTNNRFVSVMRENHSNVAASAGFKDIRYSXANEDAEKRGLDLQFLSNLEN 180
ARWNGTNNRFVSVMRENHSNVAASAGFKDIRYSXANEDAEKRGLDLQFLSNLEN 180

Query 121
ARWNGTNNRFVSVMRENHSNVAASAGFKDIRYSXANEDAEKRGLDLQFLSNLEN 180
ARWNGTNNRFVSVMRENHSNVAASAGFKDIRYSXANEDAEKRGLDLQFLSNLEN 180

Sbjct 121
ARWNGTNNRFVSVMRENHSNVAASAGFKDIRYSXANEDAEKRGLDLQFLSNLEN 180
ARWNGTNNRFVSVMRENHSNVAASAGFKDIRYSXANEDAEKRGLDLQFLSNLEN 180

Query 181
APRFEIVLCHACAHNPTGIDTPQEWQKASVMKHRFLFPFFDSAYQGFASGNEFV 240
APRFEIVLCHACAHNPTGIDTPQEWQKASVMKHRFLFPFFDSAYQGFASGNEFV 240

Sbjct 181
APRFEIVLCHACAHNPTGIDTPQEWQKASVMKHRFLFPFFDSAYQGFASGNEFV 240
APRFEIVLCHACAHNPTGIDTPQEWQKASVMKHRFLFPFFDSAYQGFASGNEFV 240

Query 241
IFYTVSKEGFEFCAQFSKFNQGYNVRGNTVQGEPESILQVLSQEMEIYRISWNSPP 300
IFYTVSKEGFEFCAQFSKFNQGYNVRGNTVQGEPESILQVLSQEMEIYRISWNSPP 300

Sbjct 241
IFYTVSKEGFEFCAQFSKFNQGYNVRGNTVQGEPESILQVLSQEMEIYRISWNSPP 300
IFYTVSKEGFEFCAQFSKFNQGYNVRGNTVQGEPESILQVLSQEMEIYRISWNSPP 300

Query 301
AGQRAIVTSLPENPEFWEGNVKTDARLIMTSERARLEALKTGETWTHIQTQIQM 360
AGQRAIVTSLPENPEFWEGNVKTDARLIMTSERARLEALKTGETWTHIQTQIQM 360

Sbjct 301
AGQRAIVTSLPENPEFWEGNVKTDARLIMTSERARLEALKTGETWTHIQTQIQM 360
AGQRAIVTSLPENPEFWEGNVKTDARLIMTSERARLEALKTGETWTHIQTQIQM 360

Query 361
FSFGTLPKQVYEVLVNEHYLLLPGRINSGLTTNLVDYATSHEATVITIQ 413
FSFGTLPKQVYEVLVNEHYLLLPGRINSGLTTNLVDYATSHEATVITIQ 413

Sbjct 361
FSFGTLPKQVYEVLVNEHYLLLPGRINSGLTTNLVDYATSHEATVITIQ 413
FSFGTLPKQVYEVLVNEHYLLLPGRINSGLTTNLVDYATSHEATVITIQ 413
NCBI has already done it…
### Pairwise Alignment Scores

<table>
<thead>
<tr>
<th>Species</th>
<th>Protein</th>
<th>DNA</th>
<th>d</th>
<th>dN/dS</th>
<th>dNS/dNC</th>
<th>Gene Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Homo sapiens</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GOT1</td>
</tr>
<tr>
<td>vs. Pan troglodytes</td>
<td>100.0</td>
<td>99.8</td>
<td>0.002</td>
<td>0.000</td>
<td>undef</td>
<td>GOT1</td>
</tr>
<tr>
<td>vs. Canis lupus familiaris</td>
<td>92.5</td>
<td>89.4</td>
<td>0.114</td>
<td>0.087</td>
<td>0.464</td>
<td>GOT1</td>
</tr>
<tr>
<td>vs. Bos taurus</td>
<td>91.5</td>
<td>89.5</td>
<td>0.113</td>
<td>0.100</td>
<td>0.490</td>
<td>Got1</td>
</tr>
<tr>
<td>vs. Mus musculus</td>
<td>91.0</td>
<td>86.9</td>
<td>0.144</td>
<td>0.069</td>
<td>0.541</td>
<td>Got1</td>
</tr>
<tr>
<td>vs. Rattus norvegicus</td>
<td>89.8</td>
<td>87.2</td>
<td>0.140</td>
<td>0.090</td>
<td>0.731</td>
<td>Got1</td>
</tr>
<tr>
<td>vs. Gallus gallus</td>
<td>80.4</td>
<td>76.1</td>
<td>0.287</td>
<td>0.076</td>
<td>0.538</td>
<td>Got1</td>
</tr>
<tr>
<td>vs. Dano renio</td>
<td>77.0</td>
<td>72.5</td>
<td>0.343</td>
<td>0.052</td>
<td>0.685</td>
<td>got1</td>
</tr>
<tr>
<td>vs. Drosophila melanogaster</td>
<td>56.9</td>
<td>59.1</td>
<td>0.591</td>
<td>0.182</td>
<td>0.912</td>
<td>Got1</td>
</tr>
<tr>
<td>vs. Anopheles gambiae</td>
<td>61.1</td>
<td>58.6</td>
<td>0.603</td>
<td>undef</td>
<td>0.867</td>
<td>AgaP_AGAP004142</td>
</tr>
<tr>
<td>vs. Caenorhabditis elegans</td>
<td>54.5</td>
<td>58.5</td>
<td>0.604</td>
<td>0.110</td>
<td>0.891</td>
<td>aminotransferase</td>
</tr>
<tr>
<td>vs. Schizosaccharomyces pombe</td>
<td>47.4</td>
<td>51.0</td>
<td>0.704</td>
<td>undef</td>
<td>0.828</td>
<td>SPAC10F6.13c</td>
</tr>
<tr>
<td>vs. Saccharomyces cerevisiae</td>
<td>48.1</td>
<td>51.6</td>
<td>0.777</td>
<td>undef</td>
<td>0.872</td>
<td>AAT2</td>
</tr>
<tr>
<td>vs. Klyuyveromyces lactis</td>
<td>50.0</td>
<td>50.8</td>
<td>0.801</td>
<td>undef</td>
<td>0.792</td>
<td>KLLA0F17754g</td>
</tr>
<tr>
<td>vs. Eremothecium gossypii</td>
<td>48.1</td>
<td>51.9</td>
<td>0.770</td>
<td>0.157</td>
<td>0.818</td>
<td>AGS_AFR211C</td>
</tr>
<tr>
<td>vs. Magnaporthe grisea</td>
<td>53.4</td>
<td>56.1</td>
<td>0.660</td>
<td>undef</td>
<td>0.731</td>
<td>MGG_04156</td>
</tr>
<tr>
<td>vs. Neurospora crassa</td>
<td>53.1</td>
<td>54.6</td>
<td>0.698</td>
<td>undef</td>
<td>0.795</td>
<td>NCU07941.1</td>
</tr>
<tr>
<td>vs. Arabidopsis thaliana</td>
<td>50.6</td>
<td>55.6</td>
<td>0.672</td>
<td>undef</td>
<td>0.764</td>
<td>ASP3</td>
</tr>
<tr>
<td>vs. Oryza sativa</td>
<td>51.4</td>
<td>55.9</td>
<td>0.664</td>
<td>undef</td>
<td>0.814</td>
<td>Os01g0760600</td>
</tr>
<tr>
<td><strong>Pan troglodytes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GOT1</td>
</tr>
<tr>
<td>vs. Homo sapiens</td>
<td>100.0</td>
<td>99.8</td>
<td>0.002</td>
<td>0.000</td>
<td>undef</td>
<td>GOT1</td>
</tr>
<tr>
<td>vs. Canis lupus familiaris</td>
<td>92.5</td>
<td>89.6</td>
<td>0.112</td>
<td>0.089</td>
<td>0.464</td>
<td>GOT1</td>
</tr>
<tr>
<td>vs. Bos taurus</td>
<td>91.5</td>
<td>89.7</td>
<td>0.111</td>
<td>0.103</td>
<td>0.490</td>
<td>GOT1</td>
</tr>
</tbody>
</table>

*Note: Blast scores are left blank where applicable.*
Why not Libraries?

Librarian Action Figure with Movable Arms
Criteria for Success

• Bioinformaticist interested in service
• Money to support software licensing
• No micromanaging
Bioinformatics Support at Countway
Countway Bioinformatics Education Program

- R/Bioconductor
- GeneGO - Metacore
- Biobase – ExPlain, TransFac and HGMD
- Ingenuity Pathway Analysis
- SNP Data
- Sequence Alignment – BLAST & Clustal
- Genome Browsing
- Beginning Unix
- ENSEMBL
- Matlab
- Rosetta Resolver
Community

• BITS – Bioinformatics Tutorial Series – In collaboration with Courtney Crummett at MIT.
Harvard’s Favorites

• Survey to find out applications that are used by various labs around campus
  – This will inform us of what software we need to support if we are not already
  – It will help researchers to see what their colleagues are using
  – A way for others to see what Harvard is doing
Impact
For the Library

• First time ever library invited to participate in both the graduate and medical curriculums.

• Collaborative opportunities
  – MIT
  – MLA – other libraries
  – Across Harvard
Impact
For HMS and the CTSA

• More productive researchers
• Better educated students and postdocs
• … (and faculty)

“The greatest obstacle to discovery is not ignorance - it is the illusion of knowledge.“ Daniel J. Boorstin (1914–2004) Historian and Librarian of Congress
When you think of libraries

Don’t think of Marian the Librarian
Think of Conan the Librarian
Bioinformatics: alive and kicking... in the library.