Algorithms, data structures and web computing for data mining in biomedicine

Jonas S Almeida
Dept Bioinformatics and Comp. Biol.
Univ Texas MDAnderson Cancer Center

OECD Workshop on Knowledge Markets in the Life Sciences
16-17 October 2008


→ not manageable
→ not affordable
→ not traceable
→ not evolvable
→ not feasible
This outcome was anticipated right at the onset of the Web [recall Tim Berners-Lee “weaving the web”]

Desired key features of a web-based data management system:

1. Syntactic interoperability
   Ability to get the data once told where it is.

2. Semantic interoperability
   Ability to use the data for a different purpose than the one that dictated its generation.
The path backwards.

**Model ID**

**Variable Selection**

**Discovery**

Models, transfer functions

[ $y = f(x)$ ]

Boosting, evolutionary algorithms, exhaustive search

[ $x \leftarrow X$ ]

Self-described structures, Ontologies, RDF, Description Logic, S3DB.

[ $x \leftarrow [X,Z]$ ]

Models ------------------------ Tools ----------------------------- Software Environment
Lesson learned: predictive independent variables are a needle in the haystack.

Model ID


Lesson learned: critical co-variables are often found in other haystacks.
Lesson learned: more than domain specific models or tools, integrative research requires a Knowledge Engineering environment.

The critical characteristic of that environment is semantic interoperability for both data and tools. Lack of syntactic interoperability is inexcusable.
A brief history of data
<table>
<thead>
<tr>
<th>Accession</th>
<th>Version</th>
<th>GI</th>
<th>Source</th>
<th>Organism</th>
<th>Reference Authors</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CX769772</td>
<td>1</td>
<td>58180125</td>
<td>Litopenaeus vannamei (Pacific White Shrimp)</td>
<td>Eukaryota; Metazoa; Arthropoda; Crustacea; Malacostraca; Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea; Penaeoidea, Litopenaeidae</td>
<td>Bartlett, J.E., Cuchbertson, B.J., Shepard, Z.F., Chapman, R.W., Gross, P.S., Robalino, J., Almeida, J.S., Chen, Y., McKillen, D.J., Wu, S. and Warr, G.W.</td>
<td></td>
</tr>
</tbody>
</table>

**COMMENT**

Litopenaeus vannamei EST Library (L99-05) at marinegenomics.org


Contact: J.E. Bartlett  
Department of Biochemistry and Molecular Biology  
Medical University of South Carolina  
Box 855550, 175 Hospitality Court, Charleston, SC 29425, USA  
Tel: 843 953 8500  
Fax: 843 792 4850  
Email: grossp@musc.edu.
LOCUS    CX769772
DEFINITION est_l_vannamei5798 LvG-gll27t32d Litopenaeus vannamei cDNA, mRNA sequence.
ACCESSION CX769772
VERSION   CX769772.1 GI:58180125
KEYWORDS  EST.
SOURCE    Litopenaeus vannamei 
           (Pacific white shrimp)
ORGANISM  Litopenaeus vannamei
           Eukaryota; Metazoa; Arthropoda; Crustacea; Malacostraca;
           Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea;
           Penaeidae; Litopenaeus.
REFERENCE 1 (bases 1 to 446)
AUTHORS   Bartlett, T.C., Cuthbertson, B.J., Shepard, E.F., Chapman, R.W., Gross
           , P.S., Robalino, J., Almeida, J.S., Chen, Y., McKillen, D.J., Wu, S.
           and Warr, G.W.
TITLE     Litopenaeus vannamei EST Library (L99-05) at marinegenomics.org
COMMENT   Contact: Gross P.S.
           Department of Biochemistry and Molecular Biology
           Medical University of South Carolina
           Box 250509, 173 Ashley Avenue, Charleston, SC 29425, USA
           Tel: 843 792 8503
           Fax: 843 792 4850
           Email: grossp@musc.edu.
FEATURES
          Location/Qualifiers
source     1..446
           /organism="Litopenaeus vannamei"
           /mol_type="mRNA"
           /db_xref="taxon:6689"
NCBI Sequence Viewer v2.0 - Mozilla Firefox

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Nucleotide

Limits Preview/Index History Details

Display GenBank Send all to file

Range: from end

GenBank

Locus: 1: CX769

Accession: 58180125

Version: LvG-gil127t32d

Keywords: Litopenaeus vannamei cDNA, mRNA

Organism: 

Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea; Penaeidae; Litopenaeus.

Reference:


Title: Litopenaeus vannamei EST Library (L99-05) at marinegenomics.org


Comment: Contact: Gross PS

Department of Biochemistry and Molecular Biology
Medical University of South Carolina
Box 250509, 173 Ashley Avenue, Charleston, SC 29425, USA
Tel: 843 792 8503
Fax: 843 792 4850
Email: grossp@musc.edu.

Features:

Location/Qualifiers

source 1..446
/organism="Litopenaeus vannamei"
/mol_type="mRNA"
/db_xref="taxon:6689"
vannamei579...[gi:58180125]

446 bp    mRNA    linear    EST 25-JAN-2005

5798 Lvg-gill27t32d Litopenaeus vannamei cDNA, mRNA

LOCUS DEFINITION

vannamei (Pacific white shrimp)

Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoida;
Penaeidae; Litopenaeus.

ACCESSION

gi:58180125

REFERENCES

1 (bases 1 to 446)

Bartlett, T.C., Cuthbertson, B.J., Shepard, E.F., Chapman, R.W., Gross,
P.S., Robalino, J., Almeida, J.S., Chen, Y., McKillen, D.J., Wu, S. and
Warr, G.W.

TITLE

Litopenaeus vannamei EST Library (L99-05) at marinegenomics.org

JOURNAL


COMMENT

Contact: Gross PS
Department of Biochemistry and Molecular Biology
Medical University of South Carolina
Box 250509, 173 Ashley Avenue, Charleston, SC 29425, USA
Tel: 843 792 8503
Fax: 843 792 4850
Email: grossp@musc.edu.

FEATURES

Location/Qualifiers

1..446

/organism="Litopenaeus vannamei"
/mol_type="mRNA"
/db_xref="taxon:6689"
S3DB:
Project: Shultz
Rules:
<V2><Person><has><Name>
<V3><Dog><has><Name>
<V4><Person><has><Dog>
<V5><Person><has><Age>
Statements
<S12><P1><R6><V2>"Charlie Brown"
<S13><P1><R6><V4><R7>
<S14><P1><R6><V5>"56 years old"
<S15><P1><R7><V3>"Snoopy"
Resources
<R6> “This is Charile Brown”
<R7> “This is Snoopy, Charlie’s Dog”

N3:
<P#1><s3:project>"Shultz".
<RC#8><s3:resource><P#1>,<s3:name>"Person".
<RC#9><s3:resource><P#1>,<s3:name>"Dog".
<V#2><s3:rule><P#1><s3:subject><RC#8>,<s3:verb>"has",<s3:subject>"Name".
<V#3><s3:rule><P#1><s3:subject><RC#9>,<s3:verb>"has",<s3:subject>"Name".
<V#4><s3:rule><P#1><s3:subject><RC#8>,<s3:verb>"has",<s3:subject><RC#9>
<V#5><s3:rule><P#1><s3:subject><RC#8>,<s3:verb>"has",<s3:subject>"Age".
<R#6><s3:rsrClInstance><RC#8>,<s3:notes>“This is Charlie Brown”.
<R#7><s3:rsrClInstance><RC#9>,<s3:notes>“This is Snoopy, Charlie’s Dog”.
<S#12><V#2>[<R#6>,"Charlie Brown"].
<S#13><V#3>[<R#7>,"Snoopy"].
<S#14><V#4>[<R#6>,<#R7>].
<S#15><V#2>[<R#6>,"56 years old"].
A brief history of data
Unique Identifiers of entities:
- Durl rdf:type s3db:Deployment
- Pid rdf:type s3db:Project
- Cid rdf:type s3db:Collection
- Rid rdf:type s3db:Rule
- Sid rdf:type s3db:Statement
- Iid rdf:type s3db:Item
- Uid rdf:type s3db:User
- Gid rdf:type s3db:Group

Annotation of s3db entities:
- {Doublin Core:}
  - dc:created_by Uid
  - dc:created_on date
  - dc:service {term of cv}
  - etc ...

Needed only if sharing with Project that is hosted by a distinct S3DB Deployment.
A brief history of integrative architectures
S3DB

WebS3DB
Generic Web-based GUI for S3DB

Specialized Applications (stand alone)

Client machine (in the lab)

Web server at IBL

S3DB

(I/O for machines)

ibl.mdanderson.org
Unique Identifiers of entities:
- Durl rdf:type s3db:Deployment
- Pid rdf:type s3db:Project
- Cid rdf:type s3db:Collection
- Rid rdf:type s3db:Rule
- Sid rdf:type s3db:Statement
- Iid rdf:type s3db:Item
- Uid rdf:type s3db:User
- Gid rdf:type s3db:Group

Annotation of s3db entities:
- {Dublin Core:}
  - dc:created_by Uid
  - dc:created_on date
  - dc:service {term of cv}
  - etc...

Annotation of s3db entities (continued):
- Needed only if sharing with Project that is hosted by a distinct S3DB Deployment.
Snapshots of interfaces using S3DB's API (Application Programming Interface). These applications exemplify why the semantic web designs can be particularly effective at enabling generic tools to assist users in exploring data documenting very specific and very complex relationships. Snapshot A was taken from S3DB's web interface, which is included in the downloadable package. This interface was developed to assist in managing the database model and, therefore, is centered on the visualization and manipulation of the domain of discourse, its Collections of Items and Rules defining the documentation of their relations. The application depicted on snapshots B-D describe a document management tool S3DBdoc, freely available as a Bioinformatics Station module (see Figure 6). The navigation is performed starting from the Project (C), then to the Collection (B) and finally to the editing of the Statements about an Item (D). The snapshot B illustrates an intermediate step in the navigation where the list of Items (in this case samples assayed by tissue arrays, for which there is clinical information about the donor) is being trimmed according to the properties of a distant entity, Age at Diagnosis, which is a property of the Clinical Information Collection associated with the sample that originated the array results. This interaction would have been difficult and computationally intensive to manage using a relational architecture. The RDF formatted query result produced by the API was also visualized using a commercial tool, Sentient Knowledge Explorer (IO-Informatics Inc), shown in snapshot E, and by Welkin, F, developed by the digital interoperability SIMILE project at the Massachusetts Institute of Technology. See text for discussion of graphic representations by these tools. To protect patient confidentiality some values in snapshots B and D are scrambled and numeric sample and patient identifiers elsewhere are altered.

A Semantic Web Management Model for Integrative Biomedical Informatics

Helena F. Deus¹,², Romesh Stanislaus¹, Diogo F. Veiga¹, Carmen Behrens³, Ignacio I. Wistuba³,⁴ John D. Minna⁵, Harold R. Garner⁵,⁶,⁷,⁸,⁹, Stephen G. Swisher¹⁰, Jack A. Roth¹⁰, Arlene M. Correa¹⁰, Bradley Broom¹, Kevin Coombes¹, Allen Chang¹, Lynn H. Vogel¹,¹¹, Jonas S. Almeida¹*¹*

¹ Department of Bioinformatics and Computational Biology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, United States of America, ² Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Lisboa, Portugal, ³ Department of Thoracic/Head and Neck Medical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, United States of America, ⁴ Department of Pathology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, United States of America, ⁵ Hamon Center for Therapeutic Oncology Research, Simmons Cancer Center, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America, ⁶ Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America, ⁷ Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America, ⁸ Center for Biomedical Inventions, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America, ⁹ Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America, ¹⁰ Department of Thoracic and Cardiovascular Surgery, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, United States of America, ¹¹ Department of Biomedical Informatics, Columbia University, New York, New York, United States of America

Abstract

Background: Data, data everywhere. The diversity and magnitude of the data generated in the Life Sciences defies automated articulation among complementary efforts. The additional need in this field for managing property and access permissions compounds the difficulty very significantly. This is particularly the case when the integration involves multiple domains and disciplines, even more so when it includes clinical and high throughput molecular data.

Methodology/Principal Findings: The emergence of Semantic Web technologies brings the promise of meaningful relationships to datasets, and of extending the relationships to include those derived from multiple sources and collections. We present the initial findings of a project that uses Semantic Web technologies to manage and structure, in a cross-institutional context, the datasets and data sources used in human cancer research. In particular, we present the use of a semantic web framework to manage a database of cancer samples, including data about gene expression, clinical and pathology features, annotations, and relationships to other databases. This presentation highlights our experiences in the design, construction, and use of a semantic web framework for managing these datasets, and identifies important challenges to be addressed in the future.

Conclusion: The framework described here shows promise in improving the management of complex data collections and in enhancing our understanding of the relationships between data sources.
# Entity
1 Spa typing
2 Doubling time
3 monthly fee
4 RAPD
5 collection site
6 patient admittance data
7 patient (or subject) demographic data
8 MLST
9 Clal mecA::Tn554
10 PFGE
11 disk inhibition
12 project, station
13 leukocidins
14 hemolysins
15 other
16 Ribotyping
17 Phagetyping
18 Smahybridization bands
19 Antibiotic
20 abbreviation
21 class
22 full name
23 subject type
24 disk inhibition
25 collection date
26 project, period
27 setting, hospital/DCC/heard, service/room, ICU
28 MIC
29 isolates from same subject
30 beta lactamase
31 susceptibility
32 Agr
33 PCR genes amplification
34 country, state/province/count, city
35 name
36 3-4 letter code
37 alternative name
38 MIC
Ontology-centric web client

S3DB is equipped with REST application programming interface (API), that is, client applications can be easily weaved by composing URL calls with variable values.

Document-centric clients

...and client side applications can be easily developed, relying only on the REST protocol to interoperate with the S3DB DBMS service.

S3DB is being used for a variety of molecular epidemiology domains, for example, for Cancer Research:

- **Seeding:** The first stage of usage of the semantic database is characterized by a focus on the domain of discourse. In this seeding stage many Rules are inserted without validation by submission of actual data (Statements).

- **Calibration:** once the submission of data triples (Statements) intensifies, the seed data model is reconsidered and is significantly edited. This second stage is characterized by heavy activity both regarding expanding or updating the domain of discourse and also regarding submission of data. We found this to be the right time to engage the user community with training programs.

- **Maturation:** The end of the data acquisition program that motivated the creation of the database is sometimes associated with a decrease in the insertion of new data (Statements) and a near stop in the editing of the domain of discourse (Rules). This period of maturation therefore produces a stable data service that remains useful and is accessed regularly. We found this period to be ideal for harvesting: exporting the database schema for analysis of the knowledge domain, including the designing of intuitive Graphic User Interfaces.

- **Growth:** This third pattern of usage is much longer than the previous two and corresponds to a relative light activity editing the domain of discourse while, on the contrary, an intensification of the database access by the target community of users. This is distinct from the preceding Calibration state where data submission is frequently aided or even mediated by the database developers.
S3DB

TCGA portal

Other sources and usages

CaBIG interoperable initiatives

CaBIG interoperable initiatives

Other sources and usages
a) Manual data input and retrieval

b) Automatic data submission by BiS applications at high throughput screening facilities.

c) Daemon application using S3DB as a web-service. These are typically BiS modules, open source bioinformatics applications or R scripts.

d) Public data and web services, for example, at NCBI, Cancer Genome Atlas, etc.

for the same functionality as web-applications see prototype at docs.s3db.org
Ontology-driven web-service oriented architecture
Composite web-based applications
Desired key features of a web-based information management system:

1. Syntactic interoperability
   Ability to get the data once told where it is.
   RESTful WOA

2. Semantic interoperability
   Ability to use the data for a different purpose than the one that dictated its generation.
   SPARQL endpoints (reified to native API exposed through REST)
   Separation of domain of discourse from its instantiation
   Permission migration built-in core data model