Atlas

a data warehouse for integrative bioinformatics

Shorab P Shah, Yong Huang, Tao Xu, Macaire MS Yuen, John Ling, BF Francis Ouellette
Atlas

- Motivation by Example
- Goal of Atlas
- Atlas Architecture
  - Atlas Ontologies
- Pros and Cons of Atlas
• Motivation by Example
• Goal of Atlas
• Atlas Architecture
  • Atlas Ontologies
• Pros and Cons of Atlas
Motivation by Example

Definitions:

DNA (Deoxyribonucleic acid): In 1957, Watson, Crick and Wilkins, in their Nobel prize winning paper, described the process how DNA is “read” to produce proteins.

Genes:
The “read” regions of DNA are commonly referred to as Genes.
Case Scenario

• What if a Biologist approaches you with DNA region in Humans, and they want you to describe it as best possible, what do you do?
Case Scenario

- GenBank: Find all known genes
- Taxonomy: Find all related species
- GenBank: Find all known genes in similar Species
- RefSeq: Find predicted Genes
- Gene Ontology: Find all functionally related genes.

...
Case Scenario

- What happens if we have 100 regions?
- Or how about 1000 regions?
Atlas

- Motivation by Example
- Goal of Atlas
- Atlas Architecture
  - Atlas Ontologies
- Pros and Cons of Atlas
Goal of Atlas

• Integration of biological databases, with ability to perform complex queries
• Efficient storage and handling of data
Atlas

- Motivation by Example
- Goal of Atlas
- Atlas Architecture
  - Atlas Ontologies
- Pros and Cons of Atlas
15 data sources.

Loaders periodically (user specified) updates data to local data warehouse.
Architecture: Databases

- 4 Types of Databases:
  - Ontology
  - Sequence
  - Molecular Interactions
  - Gene Related
• Ontologies describe the relationships (mappings) between two databases.

• Describes the legal operations available when querying more than one database.
Atlas Ontologies

• Possible Relationships:
  • “is_a”
  • “part_of”
  • “inverse_of”
  • “is_synonym_of”
  • “refers_to_PubMed”
  • “feature-includes-qualifier”
  • “gene-contains-promoter”
Atlas Ontologies

- 2 Types of Ontologies used by Atlas:
  - External: Ontologies downloaded externally.
  - Internal: Lists ontologies described by Atlas.
Atlas Ontologies

- Ontologies are described using:
  - Ontology: holds descriptions in plain english
  - Ontology_Type: holds the source and description of the source type
  - Ontology_Ontology: hold binary relationships
  - (fig 2: SQL schema)
Discussion

1) Atlas uses a combined approach to data integration and data warehousing. The data is queried, and can be queried together, but it is maintained separately. What are the advantages and disadvantages of this? What other applications could benefit from this blending of data integration and data warehousing?

2) What other domains could use ontologies? Suggest specific applications. How is it going to help? How would relationships in that ontology be defined (such as “is-a” or “part-of”)?
• 3 querying levels:
  • directly from SQL
  • through API languages
  • or from application toolbox (pre-defined queries) from the command-line
Atlas

- Motivation by Example
- Goal of Atlas
- Atlas Architecture
  - Atlas Ontologies
- Pros and Cons of Atlas
Positives

- Shallow learning curve (for toolbox usage)
- Small and portable to promote efficiency
- Solid ontology (relationships are well defined)
Negatives

- Steep learning curve (for extended usage)
- Small, not easily extended to other data sources
- No automatic expansion to alternative data sources
The authors give as a challenge that conflicts occur if there are different representations of the same semantic entity. They provide the following solution “store the information from all sources as is, and also annotate that information with the source from which it came, so as to not have any information loss”.

a) Is this solution only relevant/feasible to bioinformatics? Why or why not?

b) Would it work for the other applications you thought of earlier? Why or why not?

c) Are these the challenges that you would expect, or would you expect other challenges?