Limitations of BLAST

- Can only search for a single query (e.g. find all genes similar to TTGGACAGGATCGA)
- What about more complex queries?
- “Find all genes in the human genome that are expressed in the liver and have a TTGGACAGGATCGA (allowing 1 or 2 mismatches) followed by GCCGCG within 40 symbols in a 4000 symbol stretch upstream from the gene”

Evaluating complex queries

- Idea: write a script that make multiple calls to a BLAST database
- An example query plan:
  1. Search for all instances of the two patterns in the human genome
  2. Combine results to find all pairs within 40 symbols of each other
  3. Consult a gene database to see if pair is upstream of any known gene
  4. Consult database to check if gene expressed in liver

What about using BLOBs?

- Idea: store sequences in column in an object-relational database
- Problems:
  - No opportunities for optimization
  - May need to execute queries by dumping data outside database

Some possible improvements

- Several extensions proposed for relational databases to support sequences
- PiQL (“Pickle”)- Tata, Patel, Friedman, Swaroop, ICDE 2006
- MoBIoS - Miranker, Xu, Mao, SSDBM 2003

PiQL (Tata et al. 2006)

- “Protein Query Language”
- Extension of SQL to support complex sequence similarity
- New operators:
  - Match –find approximate matches to query string
  - Match augmentation – union of matches if operands are within specified distance

PiQL example

- “Find all matches that are of the form ‘VLLSTTSA’ followed by ‘REVWAYLL’ with a gap of 0-10 symbols between them”

- SELECT AUGMENT(M1.match, M2.match, 0, 10)
  FROM MATCH (prots.p, “VLLSTTSA”, MM(PAM30)) M1, MATCH (prots.p, “REVWAYLL”, MM(PAM30)) M2

- Each component found using MATCH operator, then combined using AUGMENT operator

MoBIoS (Miranker et al. 2003)

- Storage manager based on metric-space indexing
- Idea: clustering a dataset without mapping data to coordinates
- Supports distance metrics and indexes to identify similar sequences

MoBIoS SQL (M-SQL)

- Contains built in abstract data types for DNA, RNA, and protein sequences
- Includes subsequence operators, local alignment
- Concept of similarity between instances of data types (metric distances)

M-SQL Example

SELECT Prot.accesion_id, Prot.sequence

FROM protein_sequences Prot,
digested_sequences DS, mass_spectra MS

WHERE

MS.enzyme = DS.enzyme = E and

Cosine_Distance(S, MS.spectrum, range1) and DS.accession_id = MS.accession_id = Prot.accesion_id and DS.ms_peak = P and MPAM250(PS, DS.sequence, range2)

- Experimental inputs: E – enzyme, S – spectrum, P- parent peak, PS- sequence for parent peak

Web Services Revisited

- http://xml.nig.ac.jp/wsdl/index.jsp
- http://www.ncbi.nlm.nih.gov/BLAST
- http://www.ebi.ac.uk/clustalw/

Kepler Scientific Workflow Management System

Overview

- Introduction to Scientific Workflows
- Overview of Kepler
- Kepler Demo
- Scientific workflow details

What is a workflow?

- General definition: series of tasks performed to produce a final outcome
- Scientific workflow – “data analysis pipeline”
  - Automate tedious jobs that scientists traditionally performed by hand for each dataset
  - Process large volumes of data faster than scientists could do by hand

Background: Business Workflows

- Example: planning a trip
- Need to perform a series of tasks: book a flight, reserve a hotel room, arrange for a rental car, etc.
- Each task may depend on outcome of previous task
  - Days you reserve the hotel depend on days of the flight
  - If hotel has shuttle service, may not need to rent a car

What about scientific workflows?

- Perform a set of transformations/ operations on a scientific dataset
- Examples
  - Generating images from raw data
  - Identifying areas of interest in a large dataset
  - Classifying set of objects
  - Querying a web service for more information on a set of objects
  - Many others…

More on Scientific Workflows

- Formal models of the flow of data among processing components
- May be simple and linear or more complex
- Can process many data types:
  - Archived data
  - Streaming sensor data
  - Images (e.g., medical or satellite)
  - Simulation output
  - Observational data

Challenges

- Questions:
  - What are some challenges for scientists implementing scientific workflows?
  - What are some challenges to executing these workflows?
  - What are limitations of writing a program?

Challenges

- Mastering a programming language
- Visualizing workflow
- Sharing/exchanging workflow
- Formatting issues
- Locating datasets, services, or functions

Kepler Scientific Workflow Management System

- Graphical interface for developing and executing scientific workflows
- Scientists can create workflows by dragging and dropping
- Automates low-level data processing tasks
- Provides access to data repositories, compute resources, workflow libraries

Benefits of Scientific Workflows

- Documentation of aspects of analysis
- Visual communication of analytical steps
- Ease of testing/debugging
- Reproducibility
- Reuse of part or all of workflow in a different project

Additional Benefits

- Integration of multiple computing environments
- Automated access to distributed resources via web services and Grid technologies
- System functionality to assist with integration of heterogeneous components

Some examples