

Innovative Databases for Bioinformatics

Greg Butler

Department of Computer Science

Centre for Structural and Functional Genomics

Concordia University, Montreal

gregb@cs.concordia.ca

<http://www.cs.concordia.ca/~faculty/gregb>

Outline

What is Bioinformatics?

Some Pragmatic Bioinformatics Databases

ACEDB

Heterogeneous Databases

- Entrez, SRS, BioKleisli, OPM, TAMBIS

Pathway/Network Databases

- EcoCyc, MPW, KEGG, CSNDB

Conclusion

What is Bioinformatics?

Bioinformatics is

- (formally) IT for biological sciences
 - (de facto) IT + CS for genome projects
- cf *computational biology* which implies numerical modeling and simulation

Kinds of computing tasks

- Lab notebooks: word processing, document management
- Project management: workflow, documents, databases
- Scientific analysis: databases, data mining, modeling

Explain major kinds of experiments

- **Sequencing**
extract dna, chop up dna, sequence fragments, assemble fragments
- **Gene Expression**
locate genes, similarity search for tentative function
series of mutagenesis experiments
microarray
- **Drug Discovery**
select genes responsible for disease, configure as “target” for HTS
high-throughput screening of chemical library against target
improve chemical hits (for potency, selectivity, efficacy in vivo)
pre-clinical tests for safety and delivery means of “leads”

Lab notebooks

- usually paper documents, manually kept, then entered into word processor for document managing
- record of each experiment: date, equipment, materials, conditions (temperature etc), who, where, ...
- *vital* for FDA, and to debug lab protocols, and to justify research publications

Possibilities:

- “palm pilot” handheld document entry
- barcode readers for equipment, reagents, cell cultures, etc
- digital photography of results
- infrared upload for document management

Project management

- track all experiments and documents related to all projects
- large-scale, long-term
- usually document management, with separate project management tools
- provides “go-stop” decision support

Possibilities to integrate:

- workflow
- document management, databases
- project management
- decision support systems

eg *PharMatrix* from *Base4 Bioinformatics*

Scientific analysis

- very varied
- some high-volume analysis and screening is part of project management

Past and present vognes:

- drug design using 3D models of molecules
- data collections, curating, and accessing
- web intranets, hyperlink databases and documents (Entrez)
- data mining
- machine learning (neural nets, hidden Markov models)
- microarray image processing and analysis

Where to Find More on Bioinformatics

Layperson information

- Larry Gonick & Mark Wheelis, “*Cartoon Guide to Genetics*”
- Bioinformatics primer (www.bis.med.jhmi.edu/Dan/DOE/intro.html)

Text and reference books

- Gusfield, “Algorithms on Strings, Trees, and Sequences”, CUP, 1997.
- Durbin, Eddy, Krogh, Mitchison, “Biological Sequence Analysis”, CUP 1998.
- Baxevanis, Ouellette, “Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins”, Wiley, 1998.

Organisations

- NCBI (www.ncbi.nlm.nih.gov)
- EBI (www.ebi.ac.uk)
- GenomeNet Japan (www.genome.ad.jp)

Overview of Databases

Dbcat online catalogue (www.infobiogen.fr/services/dbcat)

- lists over 400 bioinformatics databases
- about 5% relational, 1% object

Kinds

- Literature citations, annotations, classifications
- Sequences for RNA, DNA, proteins
- Structure: 3D models, Xray/NMR of crystals, images
- Maps of chromosome layouts
- Chemical properties, reactions and their properties

Every January issue of *Nucleic Acids Research* about 100 articles (www.oup.co.uk/nar/) summarizes state-of-the-art

Pragmatics for Biosciences

General lack of IT expertise and support, lack of money for software, unknown data fields at start of research

Flat files: flexible, portable, cheap

- Genbank www.ncbi.nlm.nih.gov
- Protein DB (now Sybase) www.rcsb.org

Relational: proven technology, not cheap

- Genome Sequence DB (Sybase) www.ncgr.org/gsdb/

Object: natural fit, but many problems

- LabBase (ObjectStore)

Wide Variety of Data

- sequences, genes, regulatory regions
- enzymatic behaviour
- protein-protein interaction
- molecular structure
- microarray data
- metabolic pathways
- gene control networks
- kinematic models of reaction rates

What is required of databases?

Modeling notation(s)

- tailored to datatype
- tailored to scientists

Intuitive ways to query the data

- diagrams, forms, point-and-click, text

Support for efficient answering of queries

- query optimization, indexes, compact physical storage

Support for multiple models

- co-existence of models
- integration
- resolution of uncertainty and incompleteness

ACEDDB: A *Caenorhabditis elegans* Database

1989- Richard Durbin (Sanger) & Jean Thierry-Mieg (Montpellier)

<http://probe.nalusda.gov:8000/acedocs/>

Single-user, object DB for genome projects

- C, X11, highly graphical, web/java interfaces too
- persistence, caches, session control, own query language
- dynamic schema (ie extend fields, incomplete objects OK)
- class = hierarchical structure, object = tree
- can annotate any node in the tree
- “map” objects represented as tables (for efficiency)

Not true multi-user (no transactions, recovery, concurrency)

Free, 50+ projects, about 12% of bioinformatics DBs

ACEDB: Data Model

Models are nested relational
(in *models.wrm* file)

fields are set-valued by default

? indicates object with oid

?Author	Address	E_mail	Text
Paper		?Paper	
?Paper	Reference	Title	UNIQUE ?Text
		Journal	UNIQUE ?Journal XREF Paper
		Year	UNIQUE Int
		Volume	UNIQUE Text Text
		Page	UNIQUE Text UNIQUE Text
Author		?Author	XREF Paper

Subclasses defined as filters on
classes (in *subclasses.wrm*)

Class Prolific_Author
Visible
Is_a_subclass_of Author
Filter "COUNT Paper > 6"

“Magic tags” control display
layout

?Map	Display	Non_graphic	
		Title	UNIQUE ?Text
		Flipped	
		Centre	UNIQUE Float UNIQUE Float
		Extent	UNIQUE Float UNIQUE Float
		Default_view	UNIQUE ?View
		View	?View
Inherits	From_map	UNIQUE	?Map
		Author	Text
Main_Marker	Main_Locus	?Locus	
Contains	Locus	?Locus	

ACEEDB: Data Model

Data is in *.ace* files

Author gb

Paper gb86

Paper gb97a

Paper gb97b

Author pdjames

Paper gb86

Paper gb86

Title “On a clear day”

Author gb

Author pdjames

Year 1986

Page 16 18

Journal “Music of Our Times”

ACEEDB: Querying

Three ways to access ACEEDB databases:

- `browse`
- `current pipeline-and-filter query language`
- `AQL`, `OQL`-like query language, new in 1999

ACCEDB: Query Language

Queries process sets, and filter these through a pipeline

List all papers with “music” in their title
find Paper Title = *music*

Commands include

List all authors who published in 1997 JACM

find Paper Year = 1997 **and** Journal = JACM; **follow** Author

- **find** *class pattern*
- **follow** *tag*
- **grep** *pattern*
- *tag*
- logical ops
- relational ops
- **count**

List all coauthors of authors whose name begins with John

find Author John*; **follow** Paper; **follow** Author

ACEEDB: AQL

AQL will appear in
ACEEDB 5.0 in 1999,
but no optimizer
Based on OQL

List all papers with “music” in their title
select p
from p in class Paper
where p->Title = *music*

Commands include

- **select .. from .. where**
- path access
- logical ops
- relational ops
- **count**

List all authors who published in 1997 JACM
select p->Author
from p in class Paper
where p->Year = 1997 and p->Journal = “JACM”

List all coauthors of authors whose name begins with John
select a->Paper->Author
from a in class Author
where a = John*

Heterogeneous Databases

None are true databases, more data warehouses

- Entrez
- OPM
- SRS
- BioKleisli
- TAMBIS

Entrez

1994- NCBI (<http://www.ncbi.nlm.nih.gov/Entrez/>)

Web interface to linked databases

- Genbank, seqs of proteins and dna, with neighbourhood of similar seqs
- PDB, protein structure, and 3D structures
- Medline, literature citations and abstracts
- Genome maps for completed genome projects

Pre-computed links

Web forms for queries: ids, patterns, similarity + logical ops

Very, very heavy use

OPM: Object Protocol Model

1993-1997 Victor Markowitz, Amy Chen (Lawrence Berkeley Nat. Lab)

1997 - commercially at Data Logic division, Gene Logic, Inc
(<http://gizmo.lbl.gov/opm.html>)

Object Protocol Model includes “protocol” classes

Maps OPM schema to RDBMS (Sybase, Oracle7)

Web schema browser for query formulation

Translates OPM queries to simple SQL + extra computation

Web presentation of results

1996- supports heterogeneous databases

query must specify source database explicitly

OPM Protocol Classes

PROTOCOL CLASS Construct

ID: construct_id

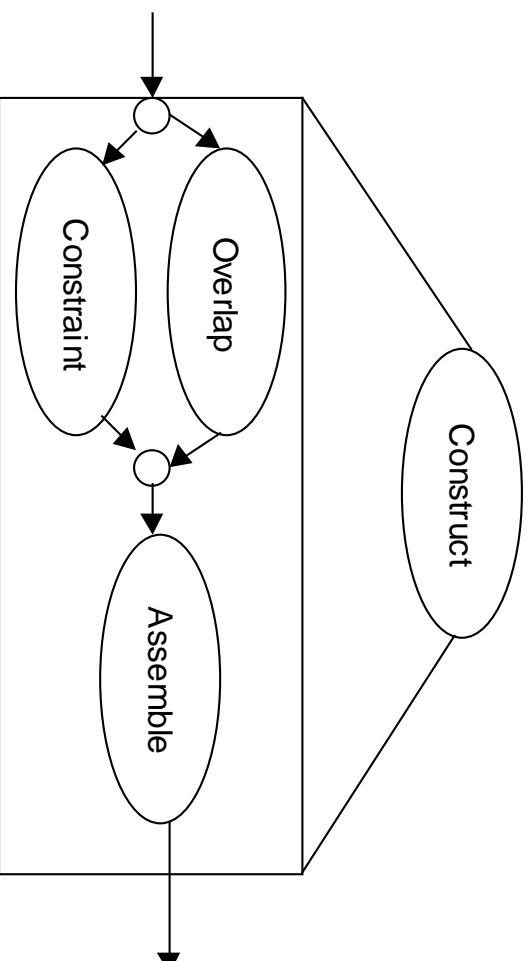
EXPANSION: (Overlap or Constraint), Assemble

SUBPROTOCOLS DELETE NULLIFIES

ATTRIBUTE construct_id: INTEGER not null

ATTRIBUTE fragments: set-of Fragment not null input

ATTRIBUTE contig_map: Contig_Map not null output



SRS: Sequence Retrieval System

1993-1998 Thure Eitzold (European Bioinformatics Institute)

1998- commercially at Lion Biosciences (Heidelberg, Cambridge)
(<http://srs.ebi.ac.uk/>)

Links flat-file databases

Queries must explicitly identify database, but can combine sources

Cross-reference links are pre-computed, and indexed

BioKleisli

1994 - Peter Buneman, Susan Davidson, Christian Overton (U. Penn),

Limsoon Wong (NUS) (<http://sdmc.krdl.org.sg/kleisli/MoreInfo.html>)

Collection Prog Lang (CPL) underlies Kleisli DB

BioKleisli integrates flat-file (and some RDBMS)

No common schema

Query must explicitly identify source DB

Programmer must explicitly code transformations from
source DB to answer queries

TAMIBIS

1997 - Computer Science/Biological Sciences, University of Manchester
(<http://img.cs.man.ac.uk/tambis/>)

BioConcept Knowledge base is common ontology/schema in
GRAIL description logic

Web browser support for query formation

Models describe source capabilities

Uses mediator/wrapper architecture

BioKleisli provides back-end wrappers

Pathway/Network Databases

- EcoCyc
- MPW
- KEGG
- CSNDB

EcoCyc

1993-98 Peter Karp and Suzanne Paley (SRI)

1998- commercially by Pangea

(<http://ecocyc.PangeaSystems.com/ecocyc/>)

KBS of genes, compounds, reactions in E. Coli

Suzanne Paley read literature and extracted knowledge of E. Coli

Built on Othello KBS for GFP (generic frame protocol)

Moving to OKBC (Open Knowledge Base Connectivity)

Automatic graph layout for results presentation

MPW : Metabolic Pathways Database

199? - Evgeni Selkov (Novosibirsk)

(<http://beauty.isdn.mcs.anl.gov/MPW/>)

KEGG: Kyoto Encyclopedia of Genes and Genomes

1997- Minoru Kanehisa (Kyoto) (<http://www.genome.ad.jp/kegg/>)

Pathways from Boehringer collection as manual GIF images

Hotspots on GIF link to compounds, reactions, genes

Data warehouse of binary relations processed by C++ code

Query answering using subsumption/relaxation for hierarchy

DBGET/LinkDB integrated DB retrieval of existing DBs
populates the warehouse

CSNDB: Cell Signaling Networks DB

1997- Takako Igarishi, Tsuguchika Kamimuma (NIHS)

(<http://geo.nih.go.jp/csndb.html>)

Database of genes, compounds, reactions using ACEDB

Extract relations from DB create warehouse

Rules about pathways act on relations in warehouse using

CLIPS inference engine

Resulting pathways automatically drawn (Letovsky's code)

CSNB Data in ACCEDB Format

- Standard reactions

Signal_Reaction: "EGF receptor -> Grb2"
From_Molecule "EGF receptor"
To_Molecule "Grb2"
Tissue "liver"
Effect "activation"
Interaction "SH2+phosphorylated Tyr" ...
Activity "growth-hormone-induced ..."
Reference "[Yamauchi_1997]"

- Polymerization reactions

Signal_Reaction: "-> AMPA receptor + GRIP"
Component "AMPA receptor"
"GRIP"
Tissue "brain"
Effect "association"
Interaction ...

- Disassociation type reactions

Signal_Reaction: "Ah receptor + HSP90 ->"
Component "Ah receptor"
"HSP90"
Effect "disassociation"
Interaction...

- Metabolic reactions

Signal_Reaction: "phospholipase C-beta + PIP2 -> IP3"
From_Molecule "PIP2"
To_Molecule "IP3"
Effect "metabolism"
Enzyme "phospholipase C-beta"
Reference "[Dove_1997]"
"[Alberts_1994]"

Conceptual Models for Biosciences

- N.W. Paton et al, “Conceptual modelling of genomic information”, *Bioinformatics* 16 (2000) 548-557
- P.D. Karp, “An ontology for biological function based on molecular interactions”, *Bioinformatics* 16 (2000) 269-285
- V.L. Junker, R. Apweiler, A. Bairoch, “Representation of functional information in the SWISS-PROT Data Bank”, *Bioinformatics* 15 (1999) 1066-1077
- G.D. Bader & C.W.V. Hogue, “BIND - a data specification for storing and describing biomolecular interactions, molecular complexes and pathways”, *Bioinformatics* 16 (2000) 465-477

Conclusion

- Lots of pragmatics decisions, lots of useful data collections
- Theme seems to be ODBMS model that maps to some implementation (flat-file, relational, object, object-relational) that might well change
- Web!

References

- Dimitrij Frishman, Klaus Heumann, Arthur Lesk, Hans-Werner Mewes, “Comprehensive, comprehensible, distributed and intelligent databases: current status”, *Bioinformatics* 14, 7 (1998) 551-561.
- Stan Letovsky (ed.), “Bioinformatics: Databases and Systems”, Kluwer Academic Press, 1999.