

Bioinformatics Challenges in a Fungal Genomics Project

Greg Butler

Department of Computer Science

Concordia University, Montreal

www.cs.concordia.ca/~faculty/gregb

gregb@cs.concordia.ca

Aims of Talk

- introduce myself and my interests
- introduce bioinformatics
- illustrate the bioinformatics needs of a “typical” genomics project
- highlight some research challenges in bioinformatics

Outline

- Brief Bio of Greg Butler
- Bioinformatics Activities at Concordia
- Bioinformatics in Overview
- Fungal Genomics Project
- Bioinformatics Challenges
 - Easy
 - Moderate
 - Difficult
 - Holy Grails
- Conclusion

Abstract

At Concordia University we have just began a project in fungal genomics after more than two years of preparation. The general aim of the project is to identify novel enzymes in 14 species of fungi that have applications in industrial processes, particularly in the processing of wood fibre into paper. In many ways, this is a typical genomics project, covering sequencing, sequence analysis, gene expression analysis using cDNA microarrays, and follow-up work in the wet lab to verify conjectures generated "in silico". Where the project differs from many others is the breadth of species, chosen for the environments in which they flourish, and for the focus on a particular set of enzyme families related to the industrial processes of concern. Both of these factors offer us distinct advantages in terms of bioinformatics data analysis.

Short Bio

Gregory Butler is Professor of Computer Science at Concordia University, Montreal, Canada. His main research activities are methodologies for framework evolution, the development of a framework for databases and knowledge-bases, and applications to bioinformatics. Dr Butler is the author of over 50 technical papers. He has consulted on object-oriented design, object-oriented technology, database technology, and large-scale software architecture.

Dr Butler is a founder of the Centre for Structural and Functional Genomics in Montreal where he directs the development of the bioinformatics platform for a large-scale fungal genomics project.

Dr Butler obtained his PhD from the University of Sydney in 1980 for work on computational group theory. He worked in computer algebra for 20 years developing algorithms, constructing software systems, designing languages, and investigating the integration of databases and knowledge-bases with computer algebra systems. He is a major contributor to the Cayley/Magma system for computational group theory, modern algebra, and discrete mathematics.

Dr Butler was on the faculty of the Department of Computer Science at the University of Sydney from 1981 to 1990. He has held visiting positions in Delaware, Bayreuth, Karlsruhe, and Heidelberg.

Brief Biography of Greg Butler

Professor of Computer Science, at Concordia since 1992

1974–1991 Computer algebra research at Sydney
Many cutting-edge algorithms.

1986– Scientific knowledge based systems

- Cayley, v4 (Magma) language and system design
- First mathematical database — TwoGroups
- First mathematical KBS — SmallSimpleGroups

Lesson: must build own infrastructure!

1990– Object-oriented technology

- Reuse : what is effective
- Frameworks
- Generative techniques
- Agents

1995– Bioinformatics

- Tuning BLAST
- Secondary structure prediction
- 3D structure prediction (John Gunn's TRIP)
- Biochemical pathways, database technology (EML)

1997 began Know-It-All framework for DBs

- ... lots here ...

2000 – Fungal Genomics

- sequencing project for *A. niger*
- microarray and gene expression project for *A. niger*
- high-throughput fungal genomics project

Bioinformatics at Concordia

Laboratory for Bioinformatics Technology

Greg Butler — software, databases, bioinformatics

Gosta Grahne — databases, data mining

Shiri — databases, data integration

Clement Lam — large-scale computing, algorithms

Joey Paquet — internet technology, languages

Volker Haarslev — description logic, ontology

Ahmed Seffah — UI design, usability lab

S.P. Mudur — visualization

Centre for Structural and Functional Genomics

Biology: Tsang, Storms, Gulick, Varin, ...

Biochemistry: Joyce, Turnbull, Powlowski, ...

Computer Science: Butler, Lam, ...

Centre for Research in Molecular Modeling

Chemistry and Biochemistry: Peslherbe, English, ...

Bioinformatics Overview

Modeling

— physical, statistical, mathematical

Algorithms

— strings, trees, graphs, optimization

— machine learning, computational geometry

Data management, and visualization

— variety, volume, data integration, data mining

Workflow and computational grids

— high-throughput analysis pipelines, ad hoc analysis

Intelligent agents

— ontology, semantic web, knowledge-based systems

— user modeling

... and general systems building issues

— usability, UI design, rapid development and evolution

Overview of Bioinformatics Architecture

Infrastructure Platforms

Data management

Computational task management

Web access and visualization

Intelligent computing

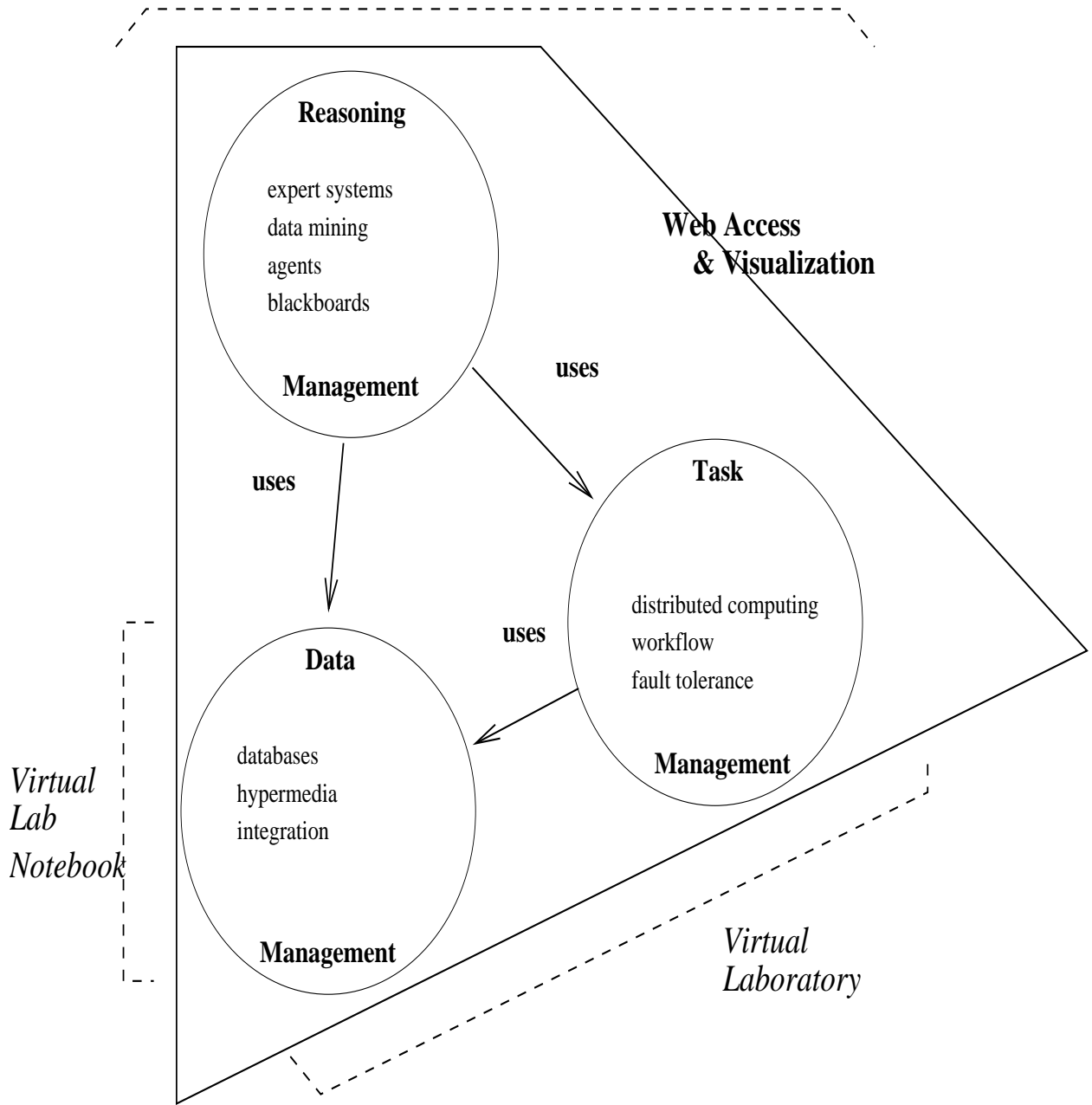
Sample Applications

Virtual Lab Notebook

Virtual Laboratory

Virtual Research Assistant

Virtual Research Assistant



Major Bioinformatics Concerns

Major Activities

Project and Laboratory Management

Automated Sequence Analysis

Gathering a Dataset

Data Mining

Brainstorming

Issues

Growth in Size of Datasets

Variety of Data Kinds

New Instruments, Techniques, Data Kinds, Analysis

Autonomous Databases

Autonomous Application Servers

Concerns

Hide the technology, Do the science

Confidence in results/interpretation

Flexibility, staying ahead of the curve

Bang for the buck

computational resources can be huge

Fungal Genomics Project

Genomic approach to identify fungal enzymes for industrial processes

Adrian Tsang (Biology) — **Principal Investigator**

Aim: Sequence 14 species of fungus, study expression of 70,000 genes, find and characterize about 300 enzymes useful for

- pulp and paper industry
- synthesis of fine chemicals
- destruction of pollutants

Investigators

Greg Butler (Computer Science)

Paul Joyce (Biochemistry)

Peter Lau (BRI, NRC)

Michael Paice (Paprican)

Justin Powlowski (Biochemistry)

Ian Reid (Paprican)

Reg Storms (Biology)

Michel Sylvestre (INRS-Institut Armand Frappier)

Luc Varin (Biology)

Richard Villemur (INRS-Institut Armand Frappier)

Collaborators

Salvadurai Dayanandan (Biology)

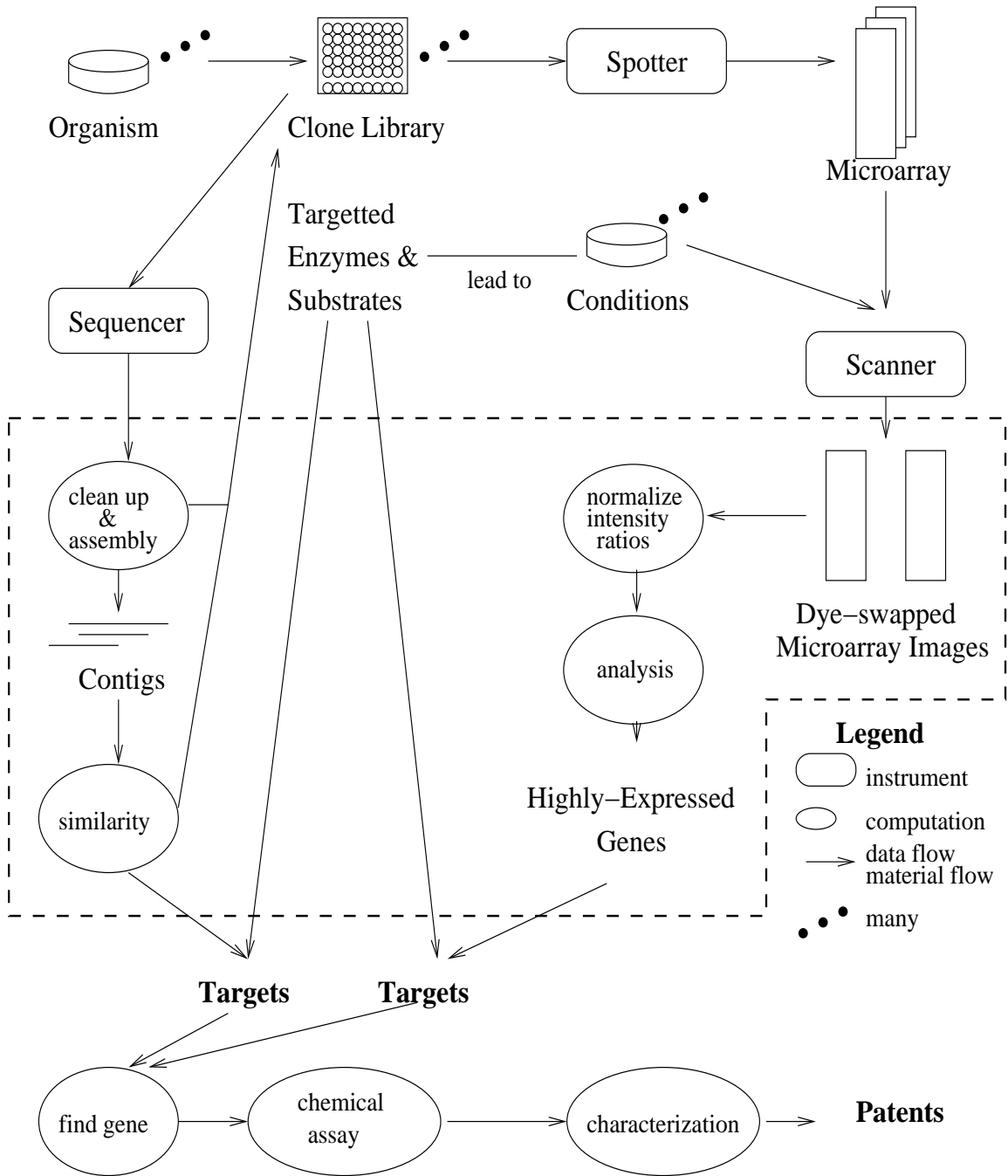
Gosta Grahne (Computer Science)

Clement Lam (Computer Science)

Ahmed Seffah (Computer Science)

Francois Sharek (INRS-Institut Armand Frappier)

Bioinformatics Processing



Bioinformatics Processing

14 species of fungus

18,000 cDNA clones per species (= 252,000 total)

70,000+ new genes

2,000 targets (approx. 130 per species)

80% identified from similarity info

— the “easy” ones with known relatives
low potential payback

20% identified from gene expression

— the “difficult” ones with new activity
high potential payback

300 secreted enzymes characterized

30-45 intracellular enzymes characterized

14 Species of Fungi

1. available from American Type Culture Collection
2. we know how to cultivate them (easily)
3. they seem to have interesting activity
4. make a diverse selection

... I cannot tell you about all of them ...

Bioinformatics Infrastructure and Tasks

Build up knowledge bases

- each known enzyme in the interesting categories
 - sequences
 - structure, domains, ...
 - function, activity, conditions
 - literature
- yeast (*Saccharomyces cerevisiae*) model organism
- each fungus that other people are working with
 - *Aspergillus niger*
 - *Neurospora crassa*
 - *Dictyostelium discoideum* (slime mould)
 - ...
- each targetted species

Construct analysis pipelines

- Sequence quality control, assembly, library normalization
- Sequence analysis
- Microarray data analysis
- Analysis for chemical assay and enzyme characterization

Must learn how to identify targetted “categories” of enzymes, preferably from sequence characteristics

Bioinformatics Challenges for Fungal Genomics Project

Easy — at least *technically* easy, if not *organisationally*

- ensure high quality data, record metadata
- data analysis for pipelines 1, 2, 3
Software techniques and tools exist.
Need to be selected and integrated.

Moderate

- data analysis for pipeline 4
New to “genomics” projects.
Should be like pipeline 3.
- LIMS (laboratory information management system)
There is need for an open-source LIMS.
- managing communication within the project
We are working with BioCore team (Illinois).

Difficult

- how to identify targetted “categories” of enzymes
Want to relate sequence alignments, and phylogenetic distances, to enzymatic activity data.
Have 30,000 examples.

Difficult — but not really required for project

- extend pipeline 3 (microarray data analysis) to relate gene expression data to metabolic pathways and regulatory networks
- apply comparative genomics within the pipelines since we will have the data on 14+ species

Bioinformatics Holy Grail I

Ab Initio 3D Structure Prediction

Blackboard architecture, as multi-agent system,
to combine agents with knowledge of

1. *Sequence*
2. *Multiple sequence alignment*
3. *Secondary structure prediction*
using amino acid properties too
4. *3D structure determination*
in many layers: backbone, sidechains, ...
5. *Docking*

Bioinformatics Holy Grail II

Metabolism, Regulation, Control

Infer mechanisms of cellular processes and control
implies modeling at many levels

- gene
- function
- interaction
- pathway
- regulatory network
- reaction kinetics
- simulation

Bioinformatics Holy Grail III

Systems Biology

Systems Biology takes a holistic view of an organism

Experimentally, it monitors cell processes in vivo

Demands on **bioinformatics** are not yet clear

— probably large-scale Holy Grail II

— maybe more emphasis on real-time analysis

Conclusion and Future Directions

Bioinformatics for the fungal genomics project is do-able.

- it will generate a lot of interesting data
- useful for deeper analysis and scientific investigation

Some challenges within the project

- better multiple sequence alignments (MSA)
- better capture biologically relevant structure
- easier to relate MSA to enzymatic activity (?)
- relate sequence in a family to its enzymatic activity

Future Possibilities

- FungalWeb, a prototype semantic web
- algorithm libraries — reusable, OO, and generic
- database technology
- usability and visualization issues

Thank You!

Questions?