

Challenges for Bioinformatics Systems

Greg Butler

Department of Computer Science

Centre for Structural and Functional Genomics

Concordia University, Montreal

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

gregb@cs.concordia.ca

Abstract

Bioinformatics is a very diverse field, and as an applied field it is judged by how effectively it delivers tools to scientists working in genomics, proteomics, and other post-genomics areas of research. This demands advances in software systems, database systems, and intelligent systems in order to provide a foundation for rapid development of bioinformatics systems.

Our work at Concordia is addressing a range of issues across several key areas: generic C++ algorithm libraries for workstations and clusters; a database framework to provide technology for a variety of data models and allow intuitive data access for scientists; workflow and computational grids; ontologies, agents, and the semantic web; and usability of tools, web sites, and visualizations for bioinformatics.

Our active involvement in a large-scale fungal genomics project helps keep our focus firmly on the needs of scientists.

The seminar will provide a tour of our projects and highlight research issues. This is joint work with many colleagues at Concordia and within Quebec.

Outline

Introduction to Genomics and Bioinformatics

Overview of Fungal Genomics Project

Bioinformatics Systems — Issues

Bioinformatics Systems — Projects

Conclusion

Introduction to Genomics and Bioinformatics

Biology

Genomics

Bioinformatics

“Big Science” Genomics

External Context of the Fungal Genomics Project

Long-term Challenges in Bioinformatics

Biology

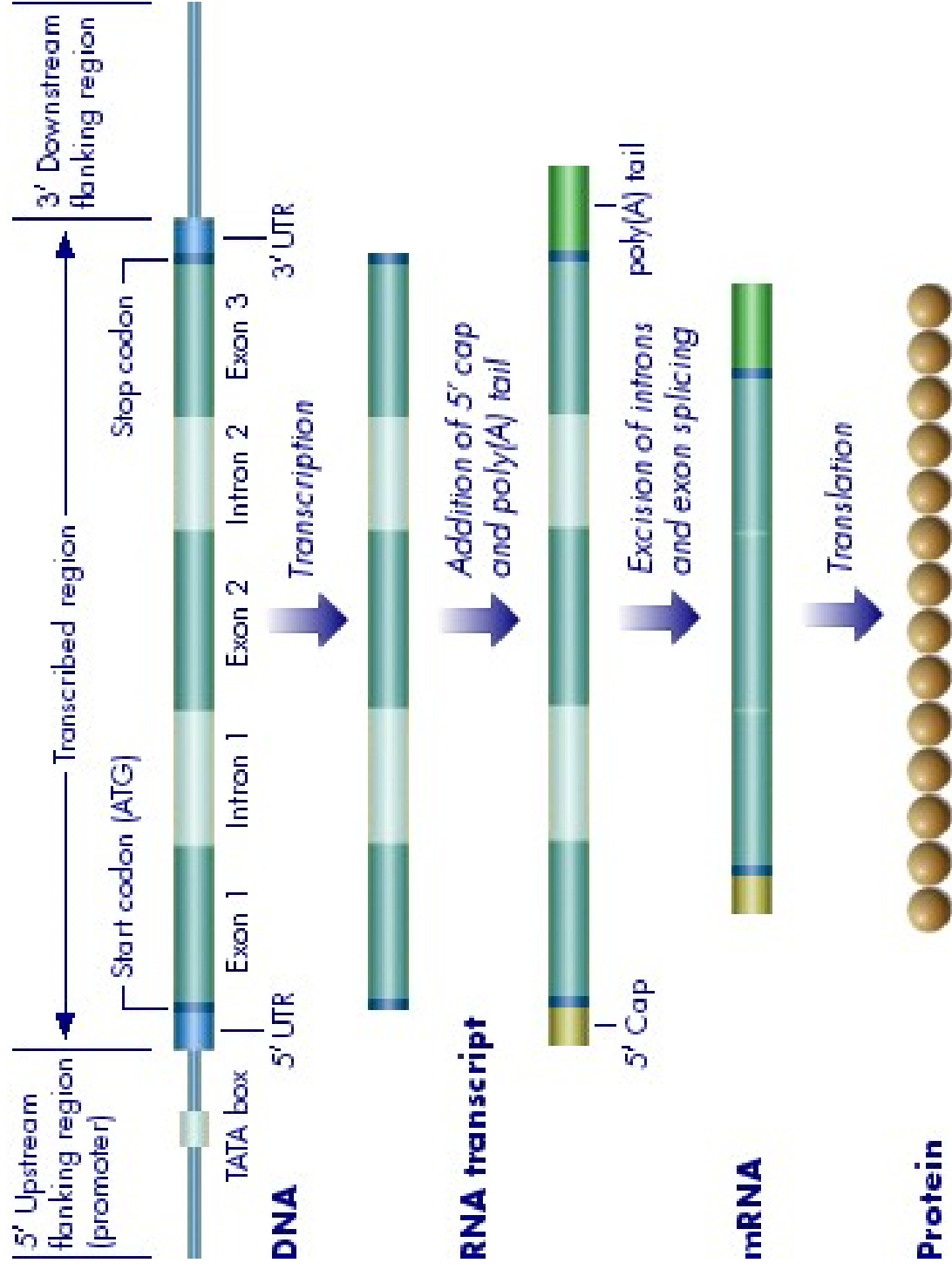
Long history of “the scientific method”

- make observations, collect data
- organize data
- create hypothesis
- design experiment to validate hypothesis
- do experiment, analyse data, interpret results

eg, data and organization of protein’s chemical and structural properties

Linnaeus, Mendel, Darwin ... role of evolution

Biology — Transcription and Translation



Biology — Transcription Regulation

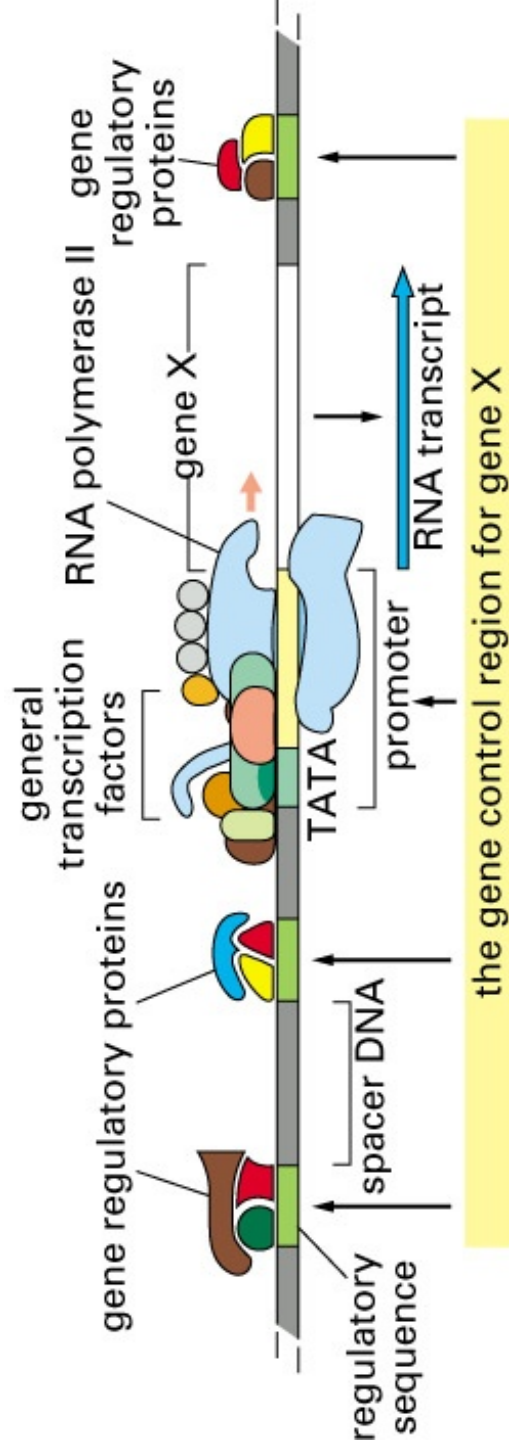
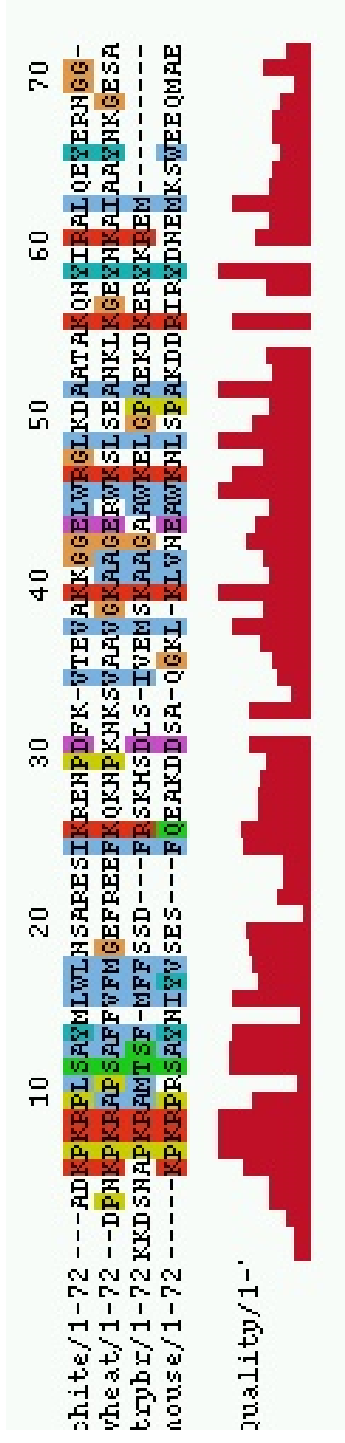


Figure 7-41. Molecular Biology of the Cell, 4th Edition.

Biology implies ...

... use Multiple Sequence Alignment (MSA)

MSA Problem: Given a set of protein sequences, and an *objective function*, determine the *optimal* alignment of the sequences.



Why?

Amino acid sequence

determines protein structure

determines enzyme function

Genomics

high-throughput, collecting or using all genes

Genomics project: determine full chromosome sequence, predict genes

EST Project: mine mRNA, assemble into "unigenes"

Microarray (Gene Expression) Project: genome-on-a-chip, study gene activity

Expression Project: splice gene into host, produce enzyme

Enzyme Assay and Characterization Project: chemistry and biochemistry

Bioinformatics

To store, organize & analyze biological data

Models — mathematical, statistical, physical

Algorithms — string, tree, graph, patterns

Database technology

Workflow and computational grids

Intelligent reasoning — expert systems, agents

Web access, visualization, usability

Rapid system development and evolution

Big Science Genomics

... to balance risk and reward

Short-Term Aims

- support massive data collection

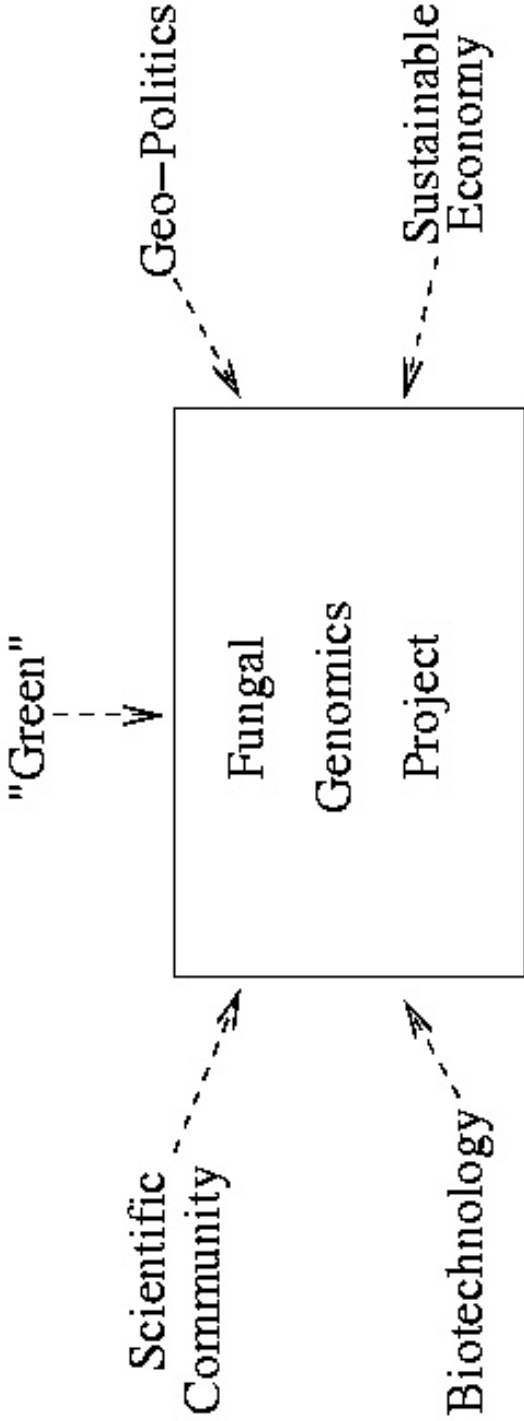
Medium-Term Aims

- exploit standard analysis of data, and
- validation in wet lab

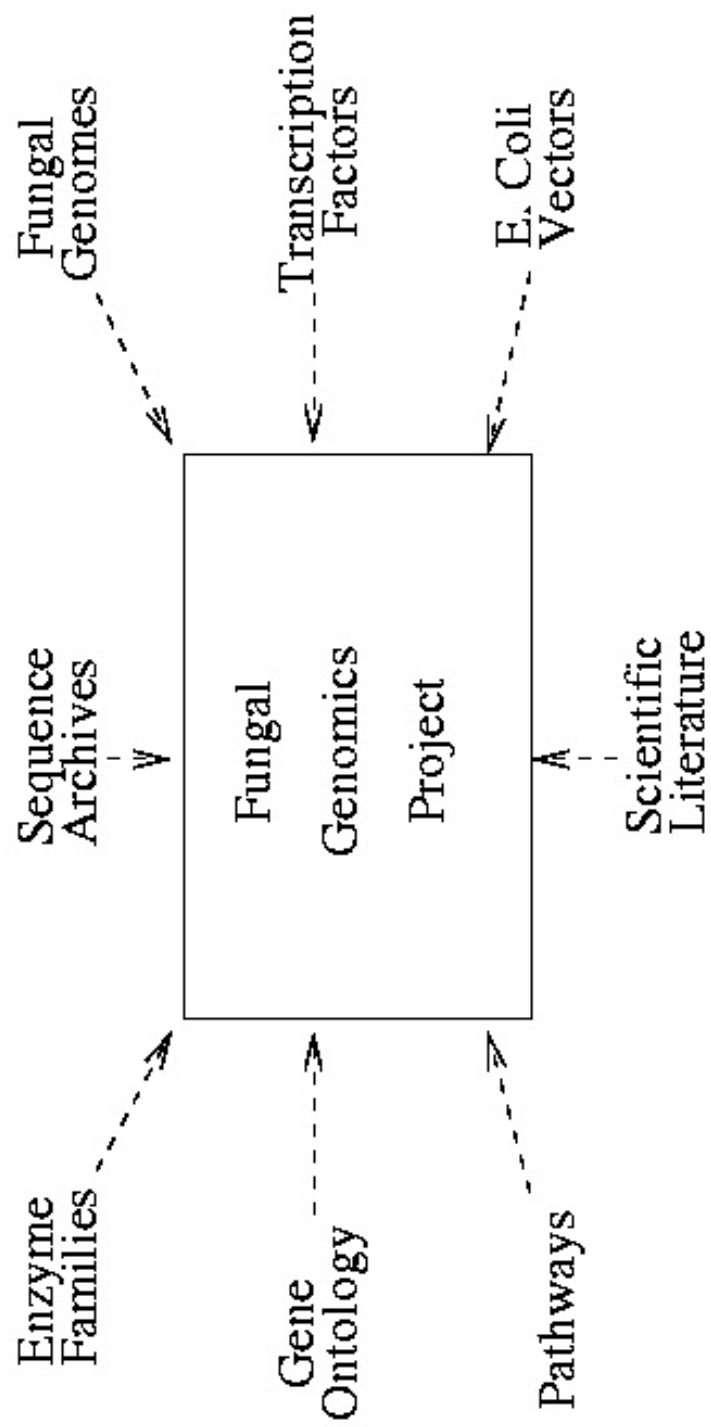
Long-Term Aims

- create knowledge
- create new analysis techniques
- be more efficient in wet lab validation
often new biotechnology

External Context of Fungal Genomics Project



External Context — Public Data



Gene Ontology — Entry

lyase activity

Accession: GO:0016829

Synonyms: None.

Definition:

Lyases are enzymes cleaving C-C, C-O, C-N and other bonds by other means than by hydrolysis or oxidation, or conversely adding a group to a double bond. They differ from other enzymes in that two substrates are involved in one reaction direction, but only one in the other direction. When acting on the single substrate, a molecule is eliminated and this generates either a new double bond or a new ring.

Term Lineage

[GO:0003673](#) : [Gene Ontology \(79026\)](#)

[GO:0003674](#) : [molecular function \(64313\)](#)

[GO:0003824](#) : [enzyme activity \(21805\)](#)

[GO:0016829](#) : [lyase activity \(1071\)](#)

External References

[EC \(1\)](#)

[PROSITE \(5\)](#)

[Pfam \(6\)](#)

[SP_KW \(1\)](#)

[ProDom \(2\)](#)

[InterPro \(11\)](#)

GO Term:

Gene Symbol:

[GO:0016829](#) : [lyase activity](#)

Datasource:

Evidence:

Full name:

[PCL1_HUMAN](#) ^{CO_01} [SPT1](#) [NAS](#) Prenylcysteine oxidase precursor

[PCL1_HUMAN](#) ^{CO_01} [SPT1](#) [NAS](#) Prenylcysteine oxidase precursor

[HPCL_MOUSE](#) ^{CO_01} [SPT1](#) [IDA](#) 2-hydroxyphytytanoyl-CoA lyase

[H0806H05.2](#) [Gmeme](#) [ISS - IPR001926](#) Cysteine synthase

[2410077I05Rik](#) [MGI](#) [ISS](#) RIKEN cDNA 2410077I05 gene

[4933425L11Rik](#) [MGI](#) [ISS](#) RIKEN cDNA 4933425L11 gene

[9030221M09Rik](#) [MGI](#) [ISS](#) RIKEN cDNA 9030221M09 gene

[Aco2](#) [MGI](#) [ISS](#) acoutrate 2, mitochondrial

[Acy3](#) [MGI](#) [ISS](#) adenylate cyclase 3

InterPro

InterPro Na⁺/K⁺ ATPase, beta subunit

[?](#) = help

| | |
|---|---|
| IPR000402 Na_K_ATPase_beta | Matches: 70 proteins View matches: [DiverView] sorted by name Detailed view Table view |
| Name ? | Na ⁺ /K ⁺ ATPase, beta subunit |
| Signatures ? | PF00287 :Na_K-ATPase (69 proteins) P500390 :ATPASE_NA_K_BETA_1 (53 proteins) P500391 :ATPASE_NA_K_BETA_2 (48 proteins) IIGR01107 :Na_K_ATPase_beta (52 proteins) |
| Type ? | Family |
| Dates ? | 1999-10-08 17:07:25.0 (created) 2001-03-12 16:43:42.0 (modified) |
| Process ? | potassium ion transport (GO:0006813) sodium ion transport (GO:0006814) |
| Function ? | sodium:potassium-exchanging ATPase activity (GO:0005391) |
| Component ? | membrane (GO:0016020) |
| Abstract ? | <p>The sodium pump (Na⁺,K⁺ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane. Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35 residues, followed by a transmembrane region, and a large extracellular domain that contains three disulfide bonds and glycosylation sites. This structure is schematically represented in the figure below.</p> <pre> +---+ +---+ +---+ xx -cys- TW -----Extracellular----- </pre> <p>'c': conserved cysteine involved in a disulfide bond.</p> |
| Database links ? | Blocks IPB000402 PROSITE doc PDOC00328 |

InterProScan

| InterProScan | Pic |
|--|--|
| <input type="checkbox"/> InterProScan:P33133 | <p> IPR005837 Family </p> <p> PR00951 — FLGBIOSNELIP TIIR01103 — fliP </p> <p> IPR005838 Family </p> <p> PD002586 — TypeIII_P PF00813 — FlpP PR01302 — TYPE3IMPPTOT PS01060 — FLIP_1 PS01061 — FLIP_2 </p> <p> Flagellar transport protein FliP Type III secretion system inner membrane P protein </p> |

Long-term Challenges in Bioinformatics

Enzyme Family Data — be able to predict kinetics

Scientific Literature — be able to mine knowledge

Gene Expression Data — be able to predict function
— be able to predict regulation

... Systems Biology

Fungal Genomics Project

Overview of Fungal Genomics Project

Bioinformatics for the Project

- Materials Tracking
- Data Collection
- Quality Control of Data and Process
- Analysis of Sequences — Automated Annotation
- Analysis of Target Sequences — Manual Curation
- Analysis of Microarray Data — Basic Expression

Bioinformatics — Future Needs

- Analysis of Microarray Data — Profiling
- Analysis of Microarray Data — Integrating other Evidence
- Enzyme Families — Phylogeny, Multiple Alignment, Classifiers
- Enzyme Families — Predicting Kinetics

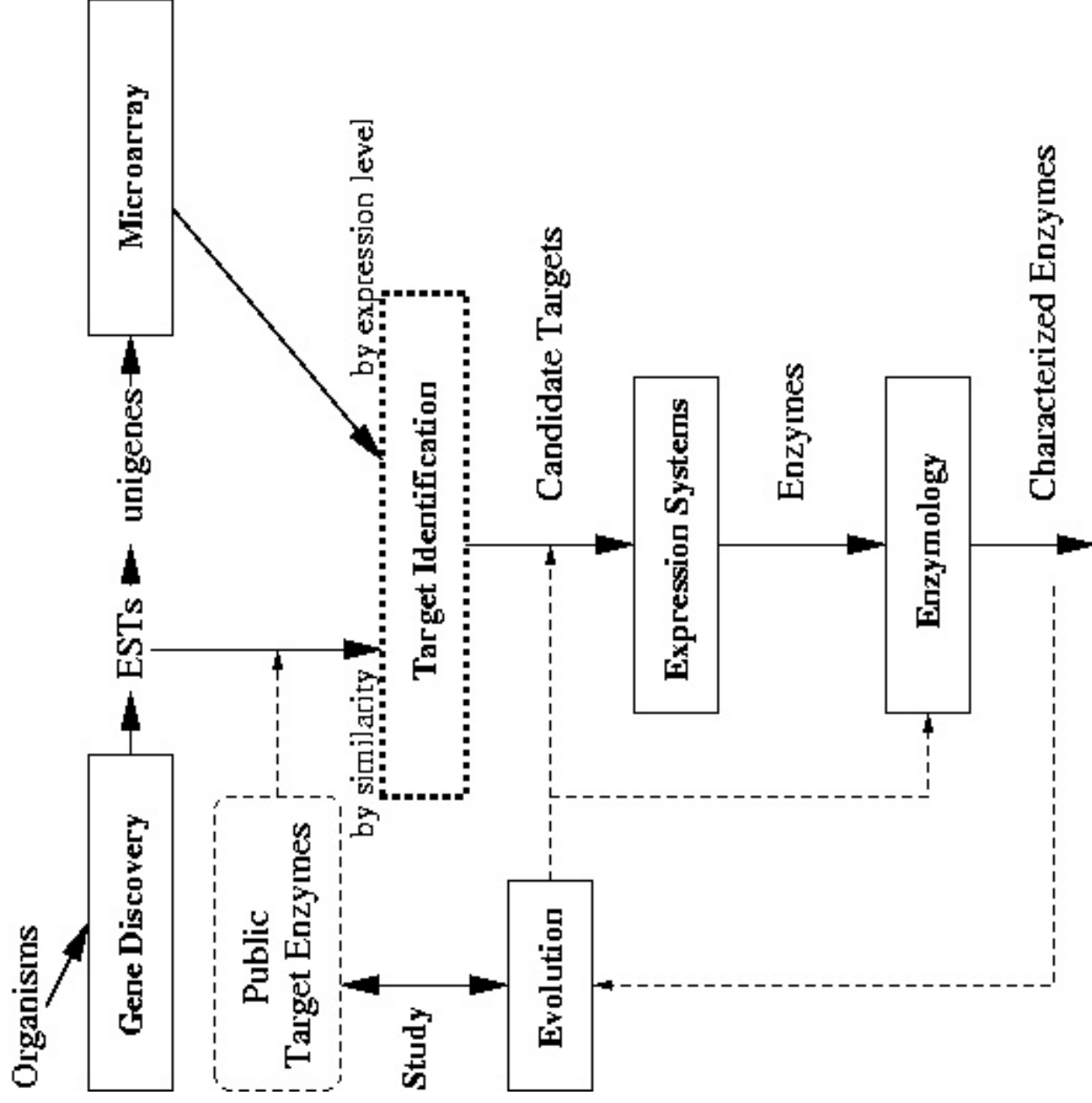
Bioinformatics Platform

The Fungal Genomics Project

Identify useful enzymes for industrial applications secreted by 14 species of fungi by constructing cDNA libraries with 5000 genes per species and detecting genes similar to known enzymes and by constructing cDNA microarray for each species and detecting genes expressed under related conditions

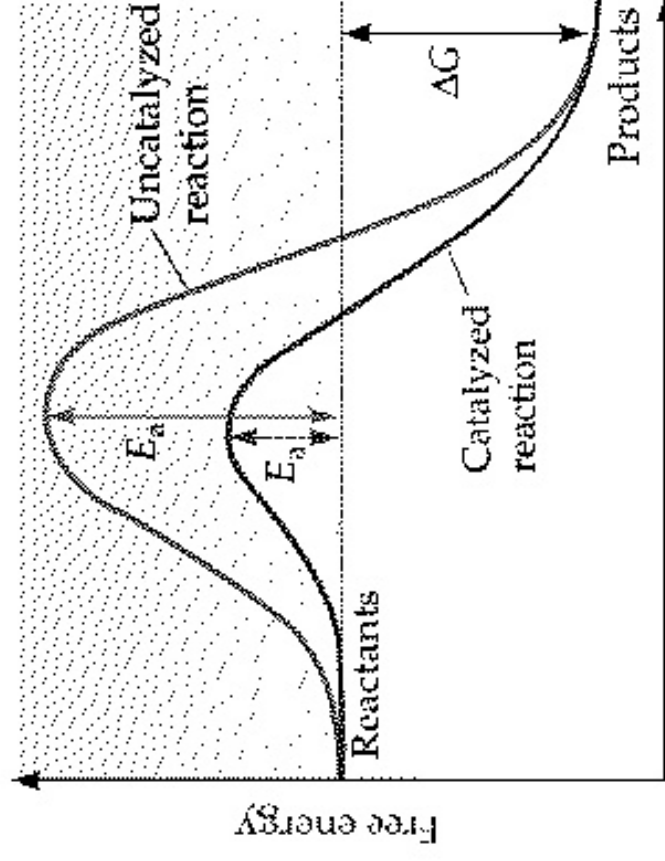
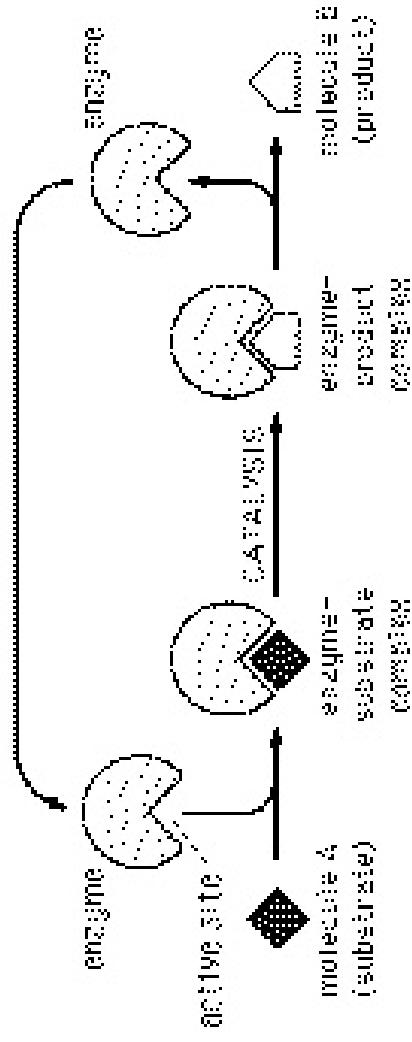
Characterize selected enzymes

The Fungal Genomics Project



What is an Enzyme?

Enzyme is a *protein* that *catalyses* a reaction.



What is an Enzyme?

Enzymes are very *specific*.

Enzymes are very *efficient catalysts*.

Some Rate Enhancements Produced by Enzymes

| | |
|---------------------------------------|-----------|
| Cyclophilin | 10^5 |
| Carbonic anhydrase | 10^7 |
| Triose phosphate isomerase | 10^9 |
| Carboxypeptidase A | 10^{11} |
| Phosphoglucomutase | 10^{12} |
| Succinyl-CoA transferase | 10^{13} |
| Urease | 10^{14} |
| Orotidine monophosphate decarboxylase | 10^{17} |

Bioinformatics — Overview

Aims

ensure high-quality experimental data is collected
know quality, confidence, provenance of data
automate data analysis & report generation

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

Bioinformatics — Housekeeping Tasks

Materials Tracking

- identify (bar code labels) and record materials
- know location of all relevant physical materials
- know mappings caused by physical transfer of materials

Data Collection

- keep all data secure and accessible
- know quality and provenance of all data
- support analysis and interpretation of data

Quality Control

- know the quality of data
- provide reports to monitor quality of lab processes
- assist in diagnosis of problems with lab processes

Bioinformatics — Analysis Tasks

Sequence Analysis

determine high-quality sequence segment
... base call quality, remove contaminants, trim
assemble ESTs into unigenes

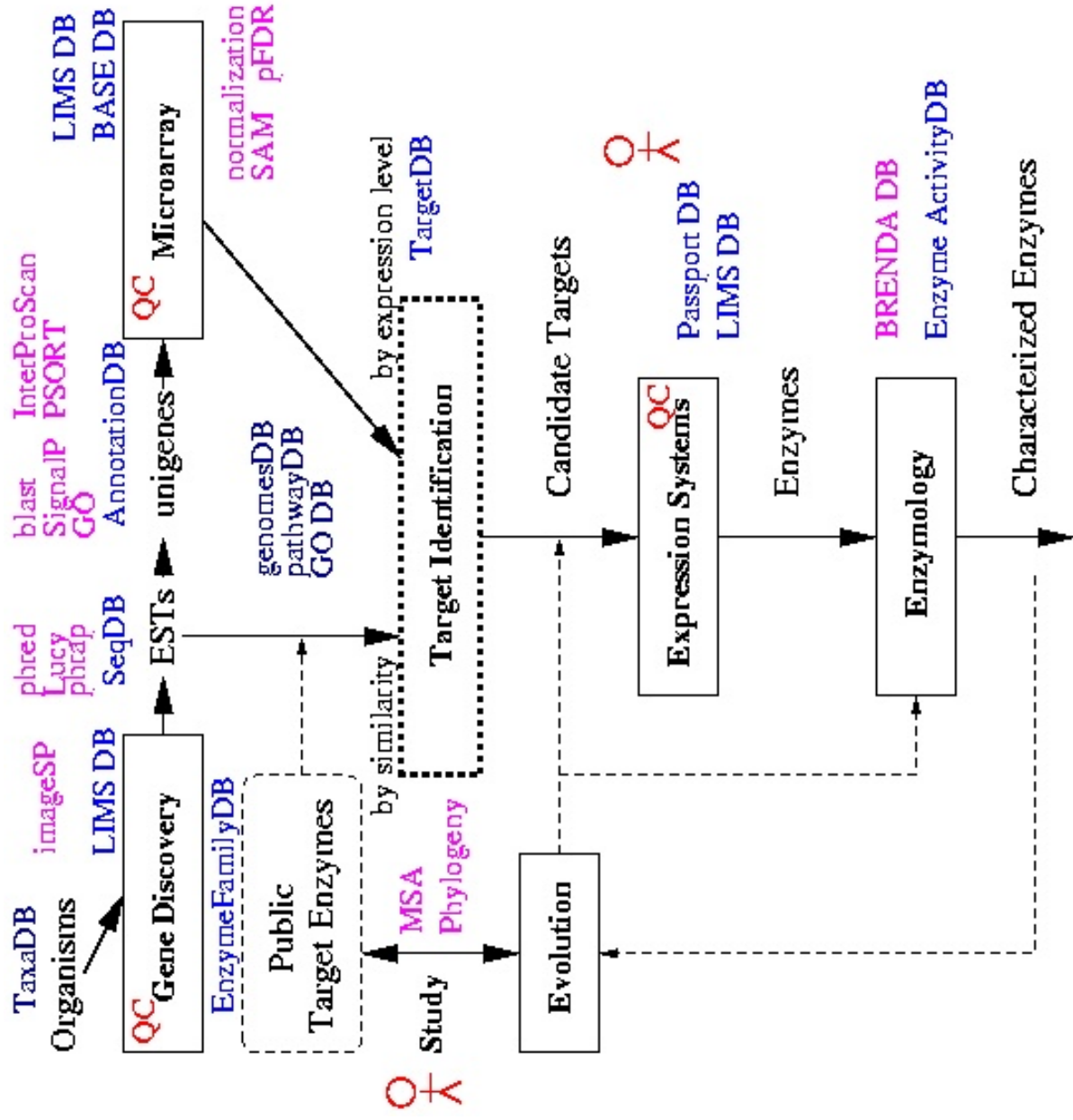
Sequence Annotation

similarity against known nucleic and protein sequences
... especially against targeted enzyme families
search for protein motifs and domains
is it secreted enzyme? other localization info?
classify to Gene Ontology category

Microarray Data Analysis

normalize: QC determines bad spots and dynamic range
set threshold for significant expression levels
determine highly expressed genes

Bioinformatics Platform



Bioinformatics Systems — Issues

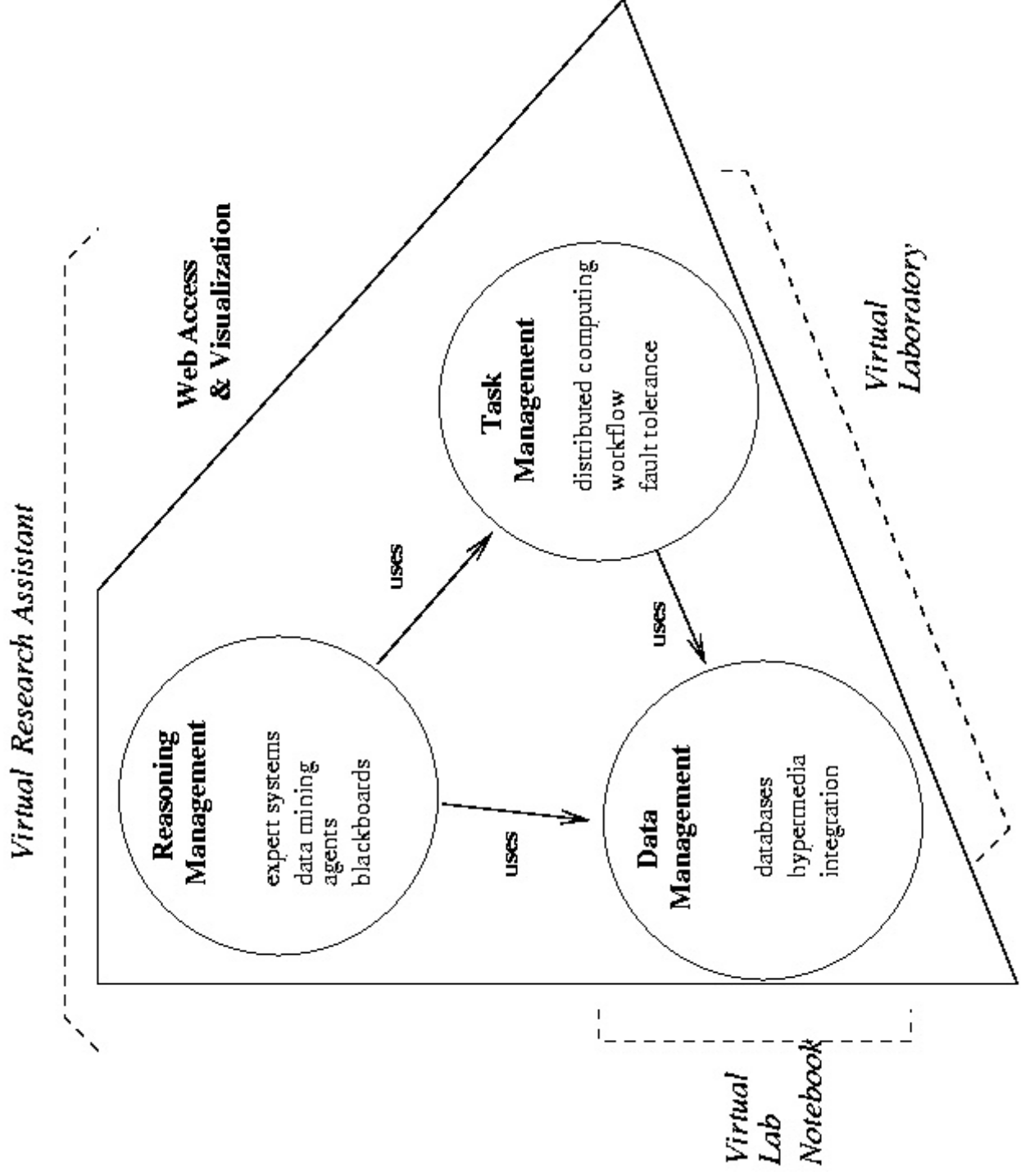
The Abstractions

Data Management Layer

Task Management Layer

Reasoning Layer

Abstractions of Bioinformatics Platform



Data Management Layer

Store large data sets

Store many different kinds of data

Access them efficiently

Integrate and cross-reference them

Respect autonomy

Browsing, querying more important than transactions

Data integrity and security

“Turing” Equivalence for Databases

Everything can be stored in a relational database

... or flat file

... or XML document

Data Management Layer

Issue: is *effective* use of data

A modeling notation
tailored to datatype and
tailored to scientists

Intuitive ways to query the data

Support for efficient answering of queries
query optimization
indexes
compact physical storage

Support for multiple models/datatypes in one database
co-existence of models
integration
resolution of uncertainty and incompleteness

Task Management Layer — Workflow and Grids

Requirements:

- Define a dataset
- Define a computation on a dataset
- Describe combinations of computations
- Describe available computational resources
- Monitor status of computations
- Examine intermediate results
- Notification (of completion, poor quality data, ...)
- Re-do subsets of computations

Examples

- MobiDick (Chris Hogue, Toronto)
- BioOpera (G. Alonso, ETH)
- Nimrod Grid Resource Broker (D. Abramson, Monash)

Reasoning Layer

These are *tasks* where the goal of the task is not so clearly known.

Need *knowledge*, i.e. “semantic web”

Characteristics

Fuzzy description of tasks

Prioritized results (rather than single result)

Mediation to merge disparate results/priorities

Modeling of user’s interests, preferences, ...

Feedback from user & reinforcement learning

⇒ “better” user models

Issues

- data quality, quantifying uncertainty, incompleteness
- algorithms need confidence values
and alternative results (by priority/confidence)
- databases that deal with uncertainty, incompleteness

Summary of Layers

| Objects | XML Level | Behavioral Requirements | Uses of Lower Layers |
|---------|-----------|-------------------------|---|
| agents | semantics | fuzzy reqs | computation analysis persistent state |
| tasks | schema | precise reqs | persistent state |
| data | document | no behaviour | |

Bioinformatics Systems — Projects

Overview of Projects

Framework for Database Technology

Intuitive Data Access

FungalWeb

Bioinformatics Systems — Projects

Bioinformatics platform for fungal genomics

FungalWeb, a prototype semantic web for genomics

A framework for database technology for genomics

An open-source generic C++ library for bioinformatics

High-quality multiple sequence alignments

Improving usability of bioinformatics tools and systems

Framework for Database Technology

Develop reusable software technology for rapid customization of advanced databases

Make available cutting-edge query languages, optimization, indexing, and storage techniques for advanced data types: relations, objects, spatial, images, time-series, GIS

Address issues of ease of access, data integration, and large-scale data storage

Intuitive Data Access

not SQL !

Querying:

WISH — form-based queries, schema browser
diagrammatic queries — GraphLog and Hy+

Object Comprehension Language

ontology-based queries — TAMBIS

example-based queries — cheminformatics

Browsing

linear — eg, genomes

trees — eg, ontologies

networks — eg, pathways

spatial — eg, molecules

temporal — eg, time-series gene expression

FungalWeb, a Prototype Semantic Web

Research with agents, ontologies, data integration, probabilistic relational models (PRM) reasoning, reinforcement learning

Practical application of cutting-edge IT and CS to analysis of gene expression data from fungal genomics project using PRMs to integrate other data sources

Future work incorporating text mining of scientific literature

PI Volker Haarslev

Concordia: Butler, Shiri, Kosseim, Bergler, Tsang, Powlowski

McGill: Doina Precup, Mike Hallett

Collaboration with Keith Decker (Delaware) - Decaf and BioMAS systems

Funding: Quebec Bioinformatics Network

FungalWeb, a Prototype Semantic Web

Agents programmed by plans

Ontologies capture knowledge

Tasks to validate effective use of knowledge

- finding sources
- querying sources
- integrating data from several sources
- reasoning with data
- learning parameters of PRM from data
- learning structure of PRM from data availability
- improving through reinforcement learning
- using service ontology to create plan for agent

Mining of scientific literature

- using formal ontologies
- to help construct ontologies

Conclusion

Bioinformatics is a rich, challenging area for interdisciplinary research in

- information technology
- artificial intelligence
- algorithms

Exciting times ahead !!!

Thank you!

Questions?