

# Intelligent Use of Ontologies: Towards a Semantic Web

**(Planning for FungalWeb: A prototype semantic web)**

Greg Butler

Department of Computer Science

Concordia University, Montreal

[www.cs.concordia.ca/~faculty/gregb](http://www.cs.concordia.ca/~faculty/gregb)

[gregb@cs.concordia.ca](mailto:gregb@cs.concordia.ca)

with Volker Haarslev, Justin Powlowski, Adrian Tsang

## **Abstract**

The vision of the semantic web is to extend the world wide web from a collection of data and documents, that are often hard to find and use, into a collection of knowledge that is very convenient to use. One pillar of the semantic web is to associate an ontology with each web site: the ontology describes the information that is available on the site, and it describes the information in a way that is precise and formal enough that the description can be manipulated by computer software.

In this paper, we explore the range of tasks in genomics and the kinds of intelligent reasoning required for those tasks, in order to gain a better understanding of the knowledge that must be captured in formally described ontologies.

This is a planning document for the FungalWeb project. FungalWeb will be a prototype semantic web for fungal genomics within the Montreal region that uses DAML+OIL description logic to represent ontologies, and the RACER reasoner to perform T-Box and A-Box reasoning. FungalWeb will complement a large-scale fungal genomics project, develop a variety of ontologies for genomic knowledge, and use the knowledge to determine the role of a gene from a variety of experimental evidence, including microarray expression data.

## **Aim of Talk**

- Learn about the range of ontologies in genomics
- Inform you about the tasks we wish to perform
- Brainstorm

## **Outline**

- Fungal Genomics Project: Variety of Data
- Semantic Web: Agents, Ontologies, Reasoning
- Formal Ontologies, Description Logics, RACER
- Intelligent Tasks

## 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

### The Numbers

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

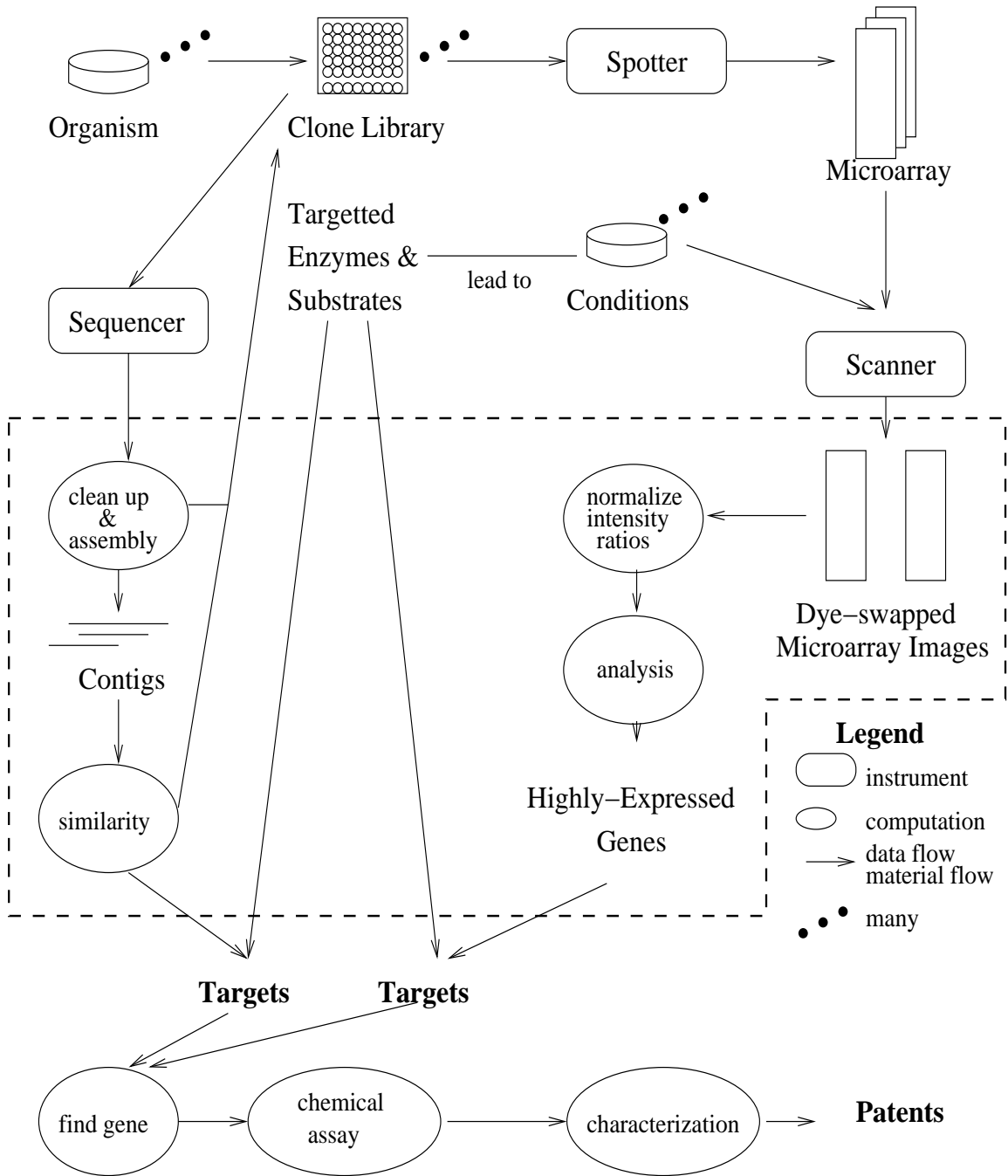
20% identified from gene expression

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

300 secreted enzymes characterized

30-45 intracellular enzymes characterized

# Bioinformatics Processing



## **Data Variety**

**cDNAs** from 14 fungal species

**genomic DNA** for some fungal species

automated analysis: similarity, phylogeny, motifs, ...  
GO categories

genomic DNA to cDNA matching; genomic DNA  
regulatory sites; gene clusters ...

**Enzyme Families and Pathways** of interest

sequences (EST, genomic), structure, domains, ...;  
function

enzymatic activity, conditions, kinetic data

multiple sequence alignments, phylogeny,  
family "structure"

Yeast model organism: annotation, mutants, protein-  
protein interaction data

**Gene expression data:** microarrays  
(about 200 conditions per species)

**Assay and characterization data** for selected targets  
(and others)

## Our Plan: A Prototype Semantic Web

**Ontologies** for each data source

DAML+OIL formal models

with T-Box and A-Box reasoning using Racer

build on Manchester work with Gene Ontology and  
TAMBIS

**Agents** for sequence analysis

build on Decaf

**Agents** for analysis of microarray data

*probabilistic relational models* (PRM) allow

- the use of other biological information along with the microarray data;
- the ability to identify clusters supported by a subset of the conditions;
- the inference of pathways; and
- the inference of regulatory subnetworks.

PRMs need to *learn* their model from available data

*reinforcement learning* possible after our wet lab work

## Opportunities for Agent Technology

**Information Extraction** — DB and scientific literature

### **Low Level Signal Processing**

*base calling*

determining the location of *vector contamination*

### **Patterns, Features and Structure**

Classic applications of *machine learning*;

- gene finding on genomic DNA
- cDNA sequences < – > genomic sequences
- low complexity regions
- motifs, domains, and families of proteins
- cell location of a protein
- secondary structure prediction

### **Annotation Support and High-Level Analysis**

*sequence analysis* in support of annotators

*sequence analysis* for assignment of function

*analysis of microarray data*

*biosynthesis pathways* from gene clusters

### **Wizards**

Guide parameter selection: eg, microarray scanning

Guide task: eg *primer design*

### **Reinforcement Learning**

## **Common Uses for Agent Technology**

**Data Transparency**

**Data Integration**

**Distributed Databases**

as well as ontology, optimization requires *cost information* from each data source

**Knowledge Discovery**

**Thank You!**

**Questions?**