

The ACGT Bioinformatics and Computational Biology Unit & The NBN Pretoria Node



History of the Bioinformatics and Computational Biology Unit (BCBU)

- Initiated from the Department of Biochemistry
- A Unit in the Faculty of Natural and Agricultural Sciences
- One of the ACGT's facilities
- The Main Pretoria Node of the **National Bioinformatics Network**, including the CSIR, UL and ARC



What does the BCBU do?

- Post-graduate training
- Under-graduate training
- Short courses
- Research
- Services & Support



Who are we?

- Senior Bioinformaticist
 - 'Fourie Joubert
- Node Manager
 - 'Oleg Reva
- System Administrator
 - 'Andrew Bwalya
- Part-time Statistician
 - 'Loveness Dzikiti



Current students

- 1 Post-doc
- 6 PhD
- 10 MSc
- 6 BSc Hons



Academic courses

- BSc IT Bioinformatics (2007)
- BSc Hons Bioinformatics
- MSc Bioinformatics
- PhD Bioinformatics



Short Courses

- Introduction to Bioinformatics on a regular basis
- Genome annotation
- Molecular Dynamics
- FAB: Phylogenetics

- Homology Modeling
- Bioconductor



Services & Support

- Major databases in SRS
- Specialist databases on request
- Nucleotide databases require high bandwidth



BLAST

- Parallel BLAST on cluster set up with a limited number of databases
- Will be expanded near future



EMBOSS

- European Molecular Biology Open Source Software
- General analysis tools
- wEMBOSS and command-line



BASE

- BioArray Software Environment
- Microarray data archiving and basic data analysis



General Tools

- Alignment tools, phylogeny tools, presentation tools, homology searching, statistics, microarrays, population genetics, RNA modeling, gene prediction, genome annotation, protein structure modeling and many others





Linux cluster



Post-grad lab



Seminar Room



Training Lab



Infrastructure

- 4 x processor Sun V880 server for general analysis and web portal apps



- 16 x processor SGI server for structural modeling and general analysis



- 64 x CPU Linux cluster for high-throughput analysis



<http://deephought.bi.up.ac.za>



▫ 1-terabyte storage array



▣ Backup robot



▫ 24 x PC training lab



High-throughput analysis on the cluster

- Currently running
 - ' BLAST
 - ' FASTA
 - ' MEME
 - ' HMMER
 - ' Molecular Docking
 - ' Molecular Dynamics
 - ' Gaussian
 - ' Phylogenetics



Current research

- Annotation of the *E. ruminantium* genome
- Structural biology in malaria
- A high-throughput pipeline for structural annotation of genomes
- Comparative genomics in malaria
- Automated modification of lead ligands
- Microarray downstream data analysis
- A Functional-Genomics Information Management System



Annotation of the *E. ruminantium* genome

- Sequenced by Prof Allsopp's group at the OVI / Veterinary Tropical Medicine
- Assembled in Staden Package
- Annotation done together with Sanger Centre
- Published in PNAS in January 2005



Structural Biology in malaria

- Investigate enzymes from the folate and polyamine pathways as potential drug targets
- Prepare homology models of enzymes, or use available X-ray structures
- Elucidate mechanisms of resistance, and develop new potential lead compounds



A high-throughput pipeline for structural annotation of genomes

- User provides all protein sequences
- Sequences are analyzed for structural features in various programs
- Promising candidates are predicted using threading / homology modeling
- Data is visualized
- Candidates for structural studies may be identified



Comparative Genomics in Malaria

- Web-based system for comparing proteins between malaria species and human
- Filtering based on homology, function, interactions, ligands, text mining, etc.



Automated modification of lead ligands

- Existing ligands are submitted
- Mutable positions are indicated
- All possible combinations and permutations of functional groups are added to positions
- A small library of compounds with suitable properties is generated
- This may be used for docking, screening, etc.



Microarray downstream data analysis

- A system has been developed to annotate malaria microarray gene clusters data in terms of:
 - ' GO terms
 - ' Metabolic pathways
 - ' Chromosomal location
 - ' Transcription regulation
 - ' Malaria-specific features
- Being expanded to plants



A Functional-Genomics Information Management System

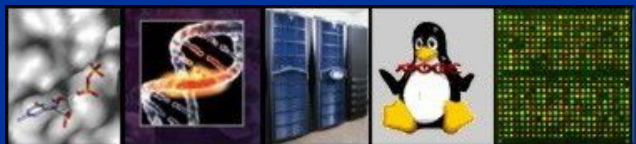
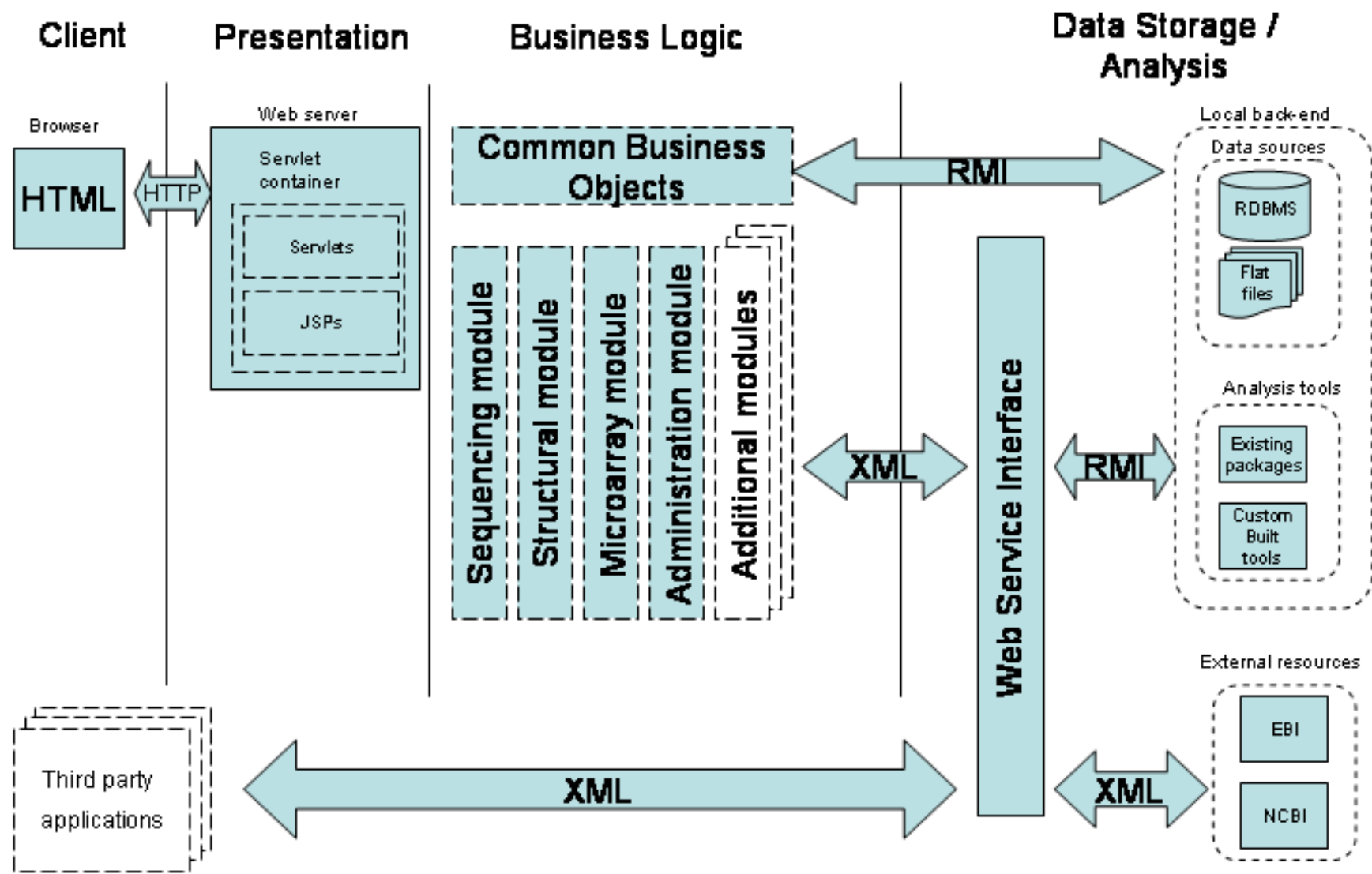
- Requests by bench scientists
 - System for managing their own data
 - System for comparing their data to public data by means of various tools
 - System for linking and integrating data across different experimental types



Some Criteria

- Centralized management
- Remote access
- A user-friendly interface
- Easy extensibility on functionality and sub-functionality levels





Modules Currently Under Development

▣ Sequence module

- ' Import trace data from local trace server and perform pre-processing
- ' Facilitate standard types of sequence analysis
- ' Enable personal annotation and versioning of results
- ' Enable automated re-analysis and notification
- ' Provide novel and effective visualization of sequence features



▣ Structural Module

- ' Provide basic and advanced structure visualization
- ' Provide mapping of sequence features to structure
- ' Provide interface to homology modeling for predicting mutation effects
- ' Provide interface to molecular dynamics
- ' Provide interface to docking tools



▣ Genotyping Module

- ' Integrated interface will be developed to:
 - ▣ Construct a database of marker alleles, allele sizes and allele fingerprints
- ' Calculation and recalculation of allele frequencies for fingerprinting projects will be enabled
- ' Interface layers will be created to export data to a series of commonly-used mapping, phylogenetics and fingerprinting analysis packages.
- ' The project will cater for data from AFLP, SSR, SNP and other projects, and will provide facilities for the management of group-based projects, storage of experimental methods and annotation of results.
- ' The project will further focus on the development of an allele fingerprint matching tool for matching unknown subjects to known individuals in a database, as well as paternity, maternity, and sibling matching.



▣ Microarray Module

- ▣ Complete interface for storage of microarray protocols and results in MIAME-compliant MAGE-ML format
- ▣ Provide analysis capability through a R / Bioconductor interface



▣ Literature Mining Module

- ' Storage / retrieval of publications
- ' Support linking to related data-types
- ' Develop API for literature mining applications □ □



▣ Comparative Genomics Module

