

A Life Science Grid System Case Study

18th May 2006

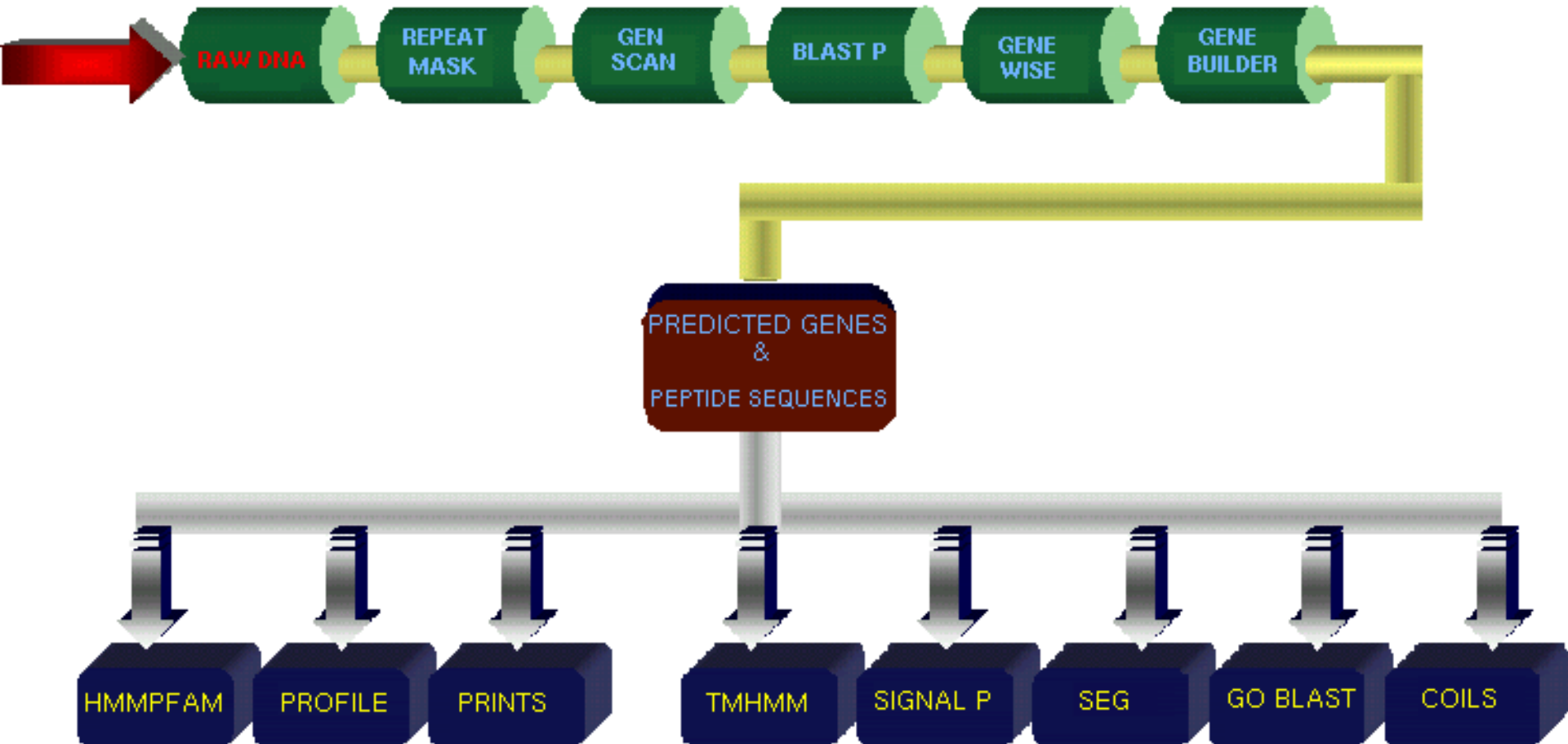
LAI Loong-Fong

Fugu Genome Project



The Fugu Genome Project is an international program aimed at determining the complete DNA sequence of the genome of the Japanese pufferfish, *Fugu rubripes*. Despite the obvious differences between fish and humans, it is expected that comparisons of the human genome with that of Fugu will shed light on the common genetic systems shared by these two species, and help us understand the information encoded in the human genome.

The Pipeline Process



The Pipeline Process

REPEAT MASK	screens DNA sequences in fasta format against a library of repetitive elements and returns a masked query sequence ready for database searches as well as a table annotating the masked regions
GENSCAN	allows prediction of complete gene structures in genomic sequences, including exons, introns, promoters and poly-adenylation signals
BLASTP	compares an amino acid query sequence against a protein sequence database
GENE WISE	compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.
GENE BUILDER	an integrated computing system for protein-coding gene prediction
HMMPFAM	search for protein families against a HMM database
COILS	compares a sequence to a database of known parallel two-stranded coiled-coils and derives a similarity score.
SEG	replaces low complexity regions in protein sequences with X characters
SIGNAL P	predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms.
TMHMM	predicts transmembrane helices in protein sequences using HMM
PRINTS	compares against a compendium of protein fingerprints
PROFILE	scan sequence from PROSITE entries
GO BLAST	Gene Ontology Mapping

System of the Past

The Hardware

- Compaq Alpha High Throughput Compute Farm
- Compaq Storageworks SAN

The Software

- System Software - Tru64 UNIX, Trucluster
- DRM Software - Platform LSF
- Database Software - MySQL Database

System of the Present

The Hardware

- HP Oteron High Throughput Compute Farm
- EMC & NetApp High Performance NAS
- HP SAN

The Software

- System Software - Linux
- DRM Software - Platform LSF
- Database Software - MySQL Database

System of the Future



The Hardware

- Multiple Heterogeneous High Throughput Compute Farms
- Distributed Grid Storage

The Software

- DRM Software - Platform LSF & Others
- Database Software - Distributed Grid Database
- Grid Middleware Software
- Enhanced Pipeline Software