Ensembl, a database for comparitive genomics.

http://www.ensembl.org

William Rossie

Highlights

- Introduction
- Data handling
- New features
- Enhancements
- Future directions

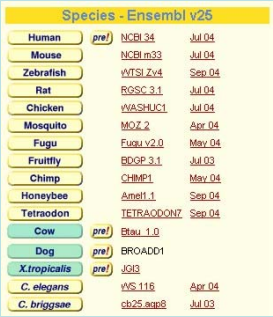
Introduction

- Used by comparative genomic scientists.
- Framework to genomic sequence analyses.
- Intergrates manual annotated genes from external sources.
- Leading source of genome annotation.
- Open source project of genome annotation, for
  - Sequence analysis
  - Data storage
  - Visualisation

Data handling:

1.Searching and viewing genomes

- Compares data from:
  - Vertebrates
  - Invertebrates
  - Worms

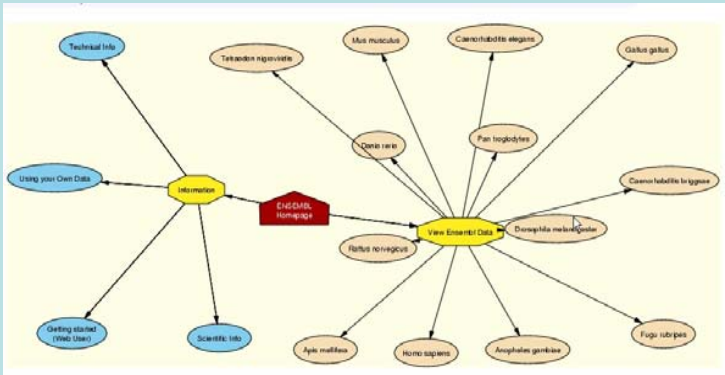


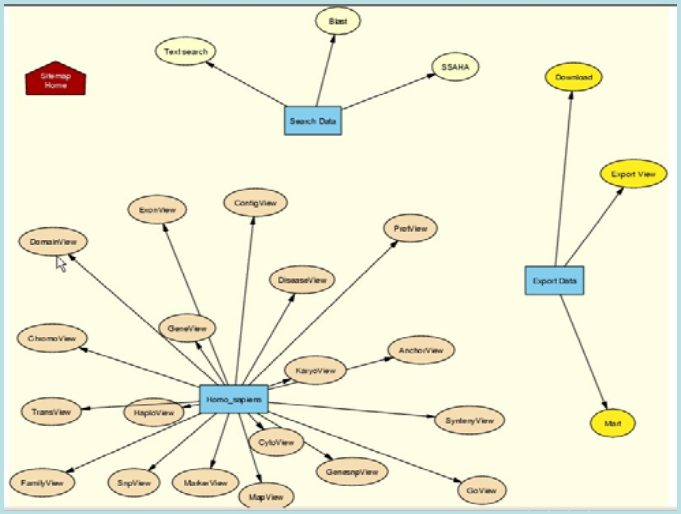
Species - Ensembl v25

Human	pre!	NCBI 34	Jul 04
Mouse		NCBI m33	Jul 04
Zebrafish		WTSI Zv4	Sep 04
Rat		RGSC 3.1	Jul 04
Chicken		WASHUC1	Jul 04
Mosquito		MOZ 2	Apr 04
Fugu		Fugu v2.0	May 04
Fruitfly		BDGP 3.1	Jul 03
Chimp		CHIMP1	May 04
Honeybee		Amel1.1	Sep 04
Tetraodon		TETRAODON7	Sep 04
Cow	pre!	Btau_1.0	
Dog	pre!	BROADD1	
X.tropicalis	pre!	JGI3	
C. elegans		WS 116	Apr 04
C. briggsae		cb25.agp8	Jul 03

Data handling:

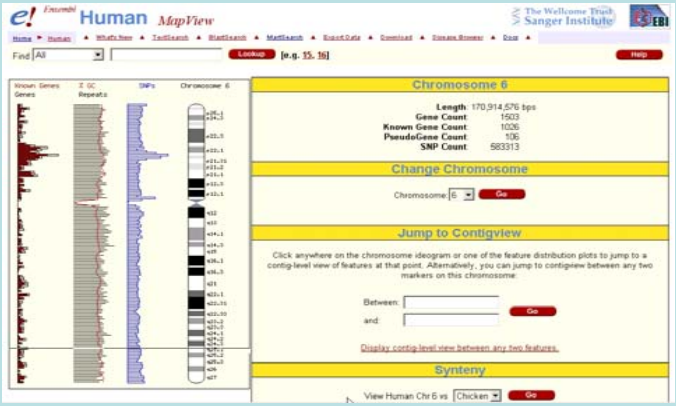
1.Searching and viewing genomes





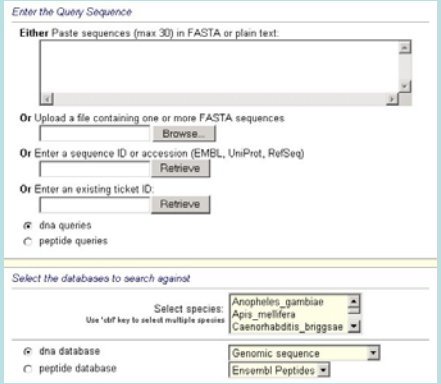
Data handling:

1.Viewing; Human mapview



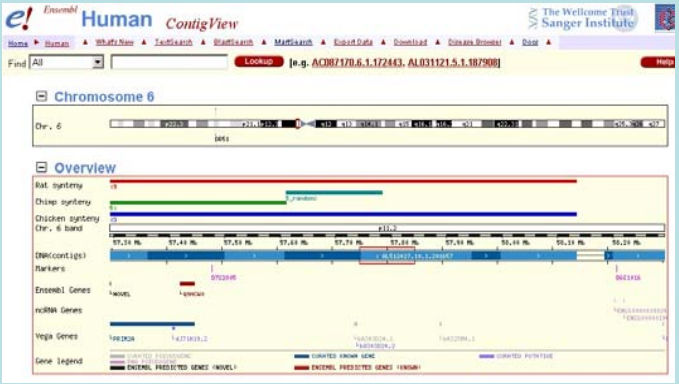
Data handling:

1.Viewing; BLAST view



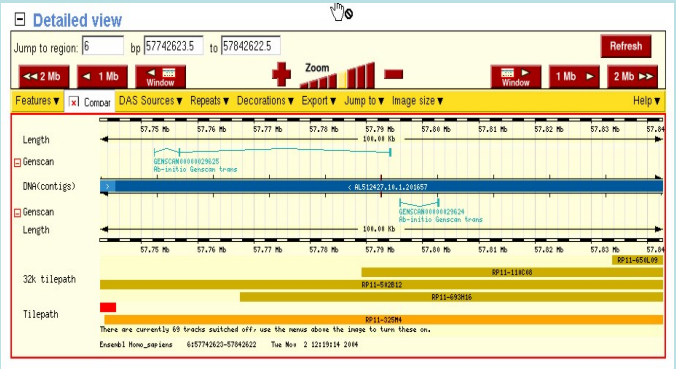
Data handling:

1.Viewing; Contig view (overview)



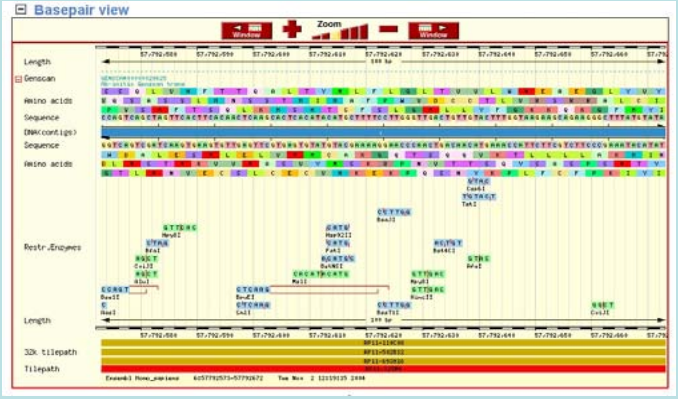
Data handling:

1.Viewing; Contig view (detailed)



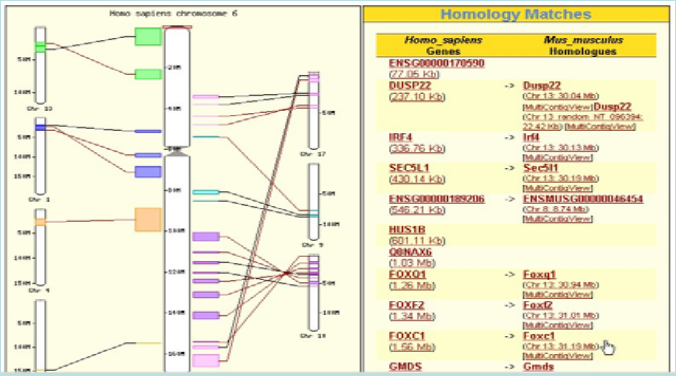
Data handling:

1.Viewing; Contig view (basepair)



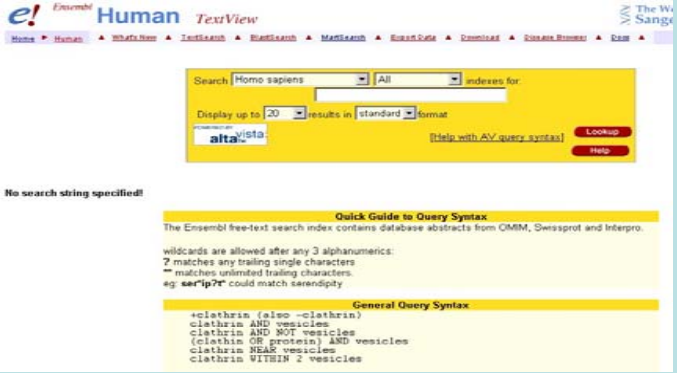
Data handling:

1.Viewing; Synteny view



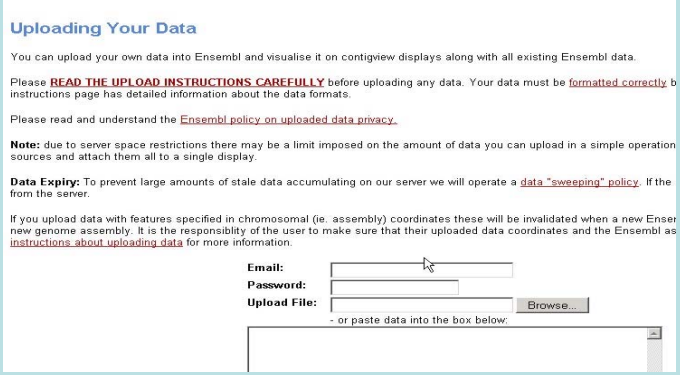
Data handling:

1.Viewing; Text view



Data handling:

2. Data uploading



Data handling:

3. Data downloading

Features

- 1. Comparative genome analysis.
- 2. Apollo annotator viewer/editor.
- 3. EnsemblMart (datamining).
- 4. Otter (extended schema for gene curation).

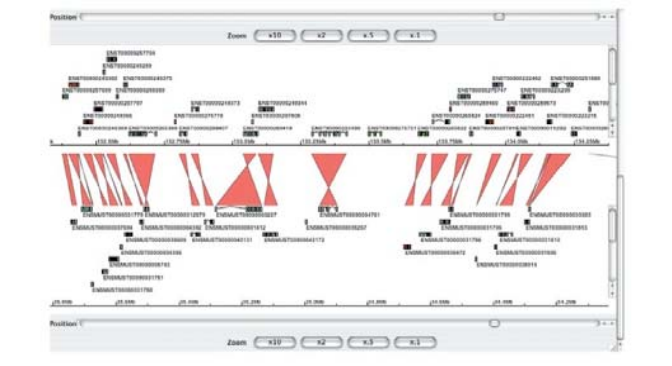
Features: 1. Comparative genome analysis.

- Fugu project and Ensembl.
- Linking and defining synteny regions.
- Constructing a catalogue of orthologous genes.
- Types of comparative information:
  a. Fine grained DNA-DNA analysis.
  b. Orthologous protein information.
  c. Large scale synteny data.

Features: 2. Apollo.

- Viewer/editor
- Java based browser.
- Open source.
- Support DAS (distributed annotator system)
- Annotator client for Otter.
- Viewing of DNA-DNA alignment, contigview and protein comparison.

Features: 2. Apollo; Synteny view



Features: 3.EnsemblMart

- Datamining for genomes.
- MySQL database, query optimised.
- View via martview.

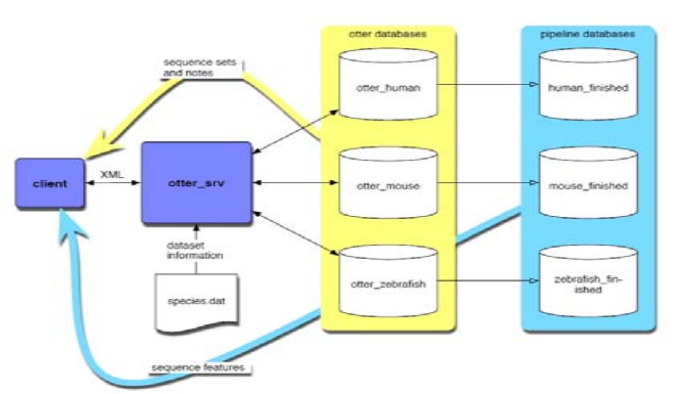
Features: 4. Otter

- Ensembl database with extended schema for gene curation.
- Associated client/server system.
- Communicate with clients via XML format.

```
<otter>
  <sequence_set>
    <sequence_fragment>
      <accession>
    <locus>
      <transcript>
        <exon>
    <feature_set>

<stable_id>
<author>
<start> <end> <strand>
```

Features: 4. Otter; overview



Enhancements

- Ensembl genome annotation.
- Website.
- Software.
- Data analysis pipeline.

Enhancements

1. Ensembl genome annotation.

- Ensembl EST genebuilder: gene prediction.
  - Combination of exonarate, BLAST and EST2Genome.
  - Transcript processed by Genomewise (finding the largest ORF across each one).
- Predicting pseudo genes automatically.
- DNA synteny generated between each group automatically.

Enhancements

2. Website.

- New tracks: Eponine track (shows transcription start site prediction).
- New interface:
  - Martview (data mining interface).
  - Goview (ontology of gene function process and location terms).
  - Haploview (interface to haplotype data contigview).
- DAS integration server:
  - DAS tracks on contigview include;
    - NCBI transciption models.
    - NCBI genomescan prediction.
    - Assembly transcript models.
    - Ensembl mapped RefSeqs.

Enhancements

2. Website.

- Basepair panel to contigview:
  - Showing nucleotide
  - Six frame amino acid translation
  - Restriction enzym site features.
- Pre-processing of SNP data.
- Contigview: include labelled sysntenic blocks.
- Dotterview: web interface to DOTTER program (shows dotplot of DNA similarity).

Enhancements

3. Software.

- Reused to build Contigview like webviews of a virtual database composed entirely of different DAS sources.
- Ensembl pipeline used to support gene curation by Wormbase or Havana (stores also gene annotation from Otter).
- Power the Vega website (curated annotation of vertebrate genome collection from a number of annotated groups into a singel database).

Enhancements

3. Software.

- Hold consistenty between biological objects and aware adaptor objects consistent style of function names.
- Perl code base, adaption of parallel java code base, with a common design between the two language bindings.
  - Perl for the biological part.
  - Java for the stable ID transfer and as a backend data adaptor for Apollo.

Enhancements

4. Data analysis pipeline.

- Improvement of processing of ESTs and cDNA as part of the EST gene analysis.
- Adjustment of maximum intron size between vertebrates and invertebrates.
- Improved handling from complex data condition
- Use of different scheduler systems (like PBS GridEngine and LSF).
- Compact storage of gapped alignments.

Future directions

- Providing genome information.
- Providing baseline annotation for a number of genomes.
- Data expanding of new genomes.
- Technology improvement (complete comparitive information between multiple vertebrates).
- Professional uses.
- Technical documentation manuals.
- Intergration with rich ontology based systems (like GenomeKnowledgeBase).

Conclusion

- Great database for comparitive biologists.
- Good links with other databases.
- Up to date.