Tools for Comparative Genomics and Sequence Analysis

Michael I. Jensen-Seaman

Human & Molecular Genetics Center
Medical College of Wisconsin
mseaman@mcw.edu
Why be “Comparative”? 

- Animal models of human disease
- Genomic sequence is only available for a few species
- Not everyone happens to be studying your favorite animal
- More power in inferring function
Genomic Sequence Projects

• “Finished” or nearly so
  – Human
  – Mouse
  – Rat
  – *Drosophila*, mosquito
  – Yeast
  – *C. elegans*
  – *E. coli* and lots of other bacteria
  – *Ciona intestinalis*
  – Pufferfish

• Near future (?)
  – Sea urchin
  – Dog
  – Cow
  – Pig
  – Horse
  – more

• In progress
  – Chimp
  – Honeybee
  – Monkey
  – Chicken
  – Zebrafish
  – Several fungi
Comparative Mammalian Genomics

Sequenced genomes

On the way
Rat Genome Sequence

- Began in Spring, 2001
- First assembly in November, 2002
  - Current assembly “v3.1” in June, 2003
- Approach
  - BAC-based sequencing
  - Whole genome shotgun sequencing
  - BAC-end sequencing
  - Fingerprint mapping
- Status of v3.1:
  - 419 ultrabactigs (large sequence contigs)
  - Average size of ultrabactigs is 6.5Mb
Maps and their problems

- All maps have mistakes
- Genome sequence assemblies have mistakes
- Best to rely on several ‘independent’ maps
Genomics, more than just sequencing

- EST sequencing projects
- Full-length cDNA sequencing projects
- Mapping
  - Linkage, RH, FISH, Physical maps, Haplotypes
- Expression data
- Proteomics
- SNP database
- Literature on function and disease

In other words ... even if your model organism is sequenced, there may still be much more data for other species.
Sources of Genome Sequence Data

- UCSC (a.k.a. Golden Path)
  - www.genome.ucsc.edu
- NCBI
- Ensembl (e!)
  - www.ensembl.org
- Celera
  - www.celera.com
Sources of Other Genome Data

- UCSC (a.k.a. Golden Path)
  - [http://genome.ucsc.edu](http://genome.ucsc.edu)
- NCBI
- Ensembl (e!)
  - [www.ensembl.org](http://www.ensembl.org)
- Celera
  - [www.celera.com](http://www.celera.com)
- Rat Genome Database
  - [www.rgd.brc.mcw.edu](http://www.rgd.brc.mcw.edu)
- RatMap
  - [http://ratmap.gen.gu.se/](http://ratmap.gen.gu.se/)
- Mouse Genome Database
  - [www.jax.org](http://www.jax.org)
- Mouse Knockout Database
  - [www.bmn.com](http://www.bmn.com)
Definitions

- **EST**
  - Expressed tag sequence: a sequenced portion of a cDNA clone

- **UniGene**
  - A theoretical cluster of ESTs that represents a unique gene

- **RefSeq**
  - NCBI’s manually curated known genes

- **Homolog**
  - DNA or protein sequences that share similarity and ancestry

- **Ortholog**
  - Homologous genes in different species descending from the same gene in the nearest common ancestor

- **Paralog**
  - Genes within the same species descending from an ancestral gene by duplication and divergence

- **HomoloGene**
  - NCBI’s automated assignment of homologs by UniGene

- **Genetic map**
  - Markers are ordered based upon genetic recombination

- **RH mapping**
  - Map markers based upon presence/absence in radiation hybrid cell line

- **SNP**
  - Single nucleotide polymorphism

- **SSLP, STRP, microsatellite**
  - Polymorphic marker based on short tandem repeat

- **Knockout**
  - Animal with a targeted gene disrupted

- **CNS**
  - Conserved noncoding sequence, presumably functionally important
Concepts

- Comparative genomics
- Genetic markers
- Repetitive elements
- Conserved noncoding sequences
Rationale for Comparative Genomics

- Evolutionary biology is the basis of all comparative biology
- Over evolutionary time, species become different to a greater or lesser extent
- Need to keep in mind evolutionary history and relationships among species
Evolutionary misconceptions

• Humans did not evolve from monkeys
  – Rather, we both evolved from a common ancestor

• No such thing as “higher” or “lower”
  – But there is such thing as “less complex” and “more complex”

• All species are equally “evolved”
  – Change between species could have happened along the way to either species
Comparative Genomics of mammals

Functional Genomics

Therapeutic Screening

Congenic Stains

Consomic Panel

Physiology

QTL Mapping

Comparative Sequencing and Mapping

Association Studies

Clinical Research

Medical Genomics

Linkage Studies

ENU Mutagenesis

Physiology

Knock-outs

Functional Genomics
Chromosome conservation in mammals

Gene conservation in mammals

- Same genes in all species
  - Well, more or less…
  - Clusters of genes within a family
  - Occasional loss of genes
  - Occasional gain of genes/exons
- Similar number of exons
- Similar size of exons and introns
Concepts

- Comparative genomics
- Genetic markers
- Repetitive elements
- Conserved noncoding sequences
Genetic Markers

- Any locus that is polymorphic
- SSLPs/STRPs/microsatellites
  - e.g., …AGCATGCTACACACACACACACACACACACACACACACACATGTCGTATGCGATGGCA...
- SNPs
  - e.g., …CCGTTAGGCTAAGC(C/T)CGTTATATTTGGCAGATAGCTCCAAGGCGATAGCTA...

- ~10,000 genetic markers available in rat
- But, hundreds of thousands more could be developed
  - 15,723,564 bp of rat assembly v3.1 involved in (CA)_n repeat, where n ≥ 20
Concepts

- Comparative genomics
- Genetic markers
- Repetitive elements
- Conserved noncoding sequences
Repetitive elements

• Tandem repeats
  – e.g.,
    CGTAGCTGGGATAGATAGATAGATAGATAGATAGATAGATACATCGT
    AGCTAGACGATCAGCTACGATCGATCGATCGTA

• Interspersed repeats
  – LINEs
  – SINEs
  – LTRs
  – etc.
Concepts

• Comparative genomics
• Genetic markers
• Repetitive elements
• Conserved noncoding sequences
Conserved Noncoding Sequences (CNS)

- Genes are generally conserved between species.
- Noncoding (introns, etc.) are not very well conserved between species.
- Conserved regions are conserved because they are functional.
- Therefore, conserved noncoding sequences probably represent functional elements.
Genome Sequence
Identifying CNSs

• Large scale cross-species alignments
• Scan for highly similar regions
• Two main software packages:
  – PipMaker (http://bio.cse.psu.edu/)
  – VISTA (http://www-gsd.lbl.gov/vista)
• Both provide static Human-Mouse-Rat maps
• Both allow for user to provide own data
PipMaker
VISTA (Godzilla)

http://pipeline.lbl.gov/
VISTA

putative coding region