

A Lite Introduction to (Bioinformatics and) Comparative Genomics

Based on the *Genomics in Biomedical Research* course at the Berkeley PGA
<http://pga.lbl.gov/>



Chris Mueller

November 18, 2004

Biology

- Evolution
 - Species change over time by the process of natural selection
- Molecular Biology Central Dogma
 - DNA is transcribed to RNA which is translated to proteins
 - Proteins are the machinery of life
 - DNA is the agent of evolution
- Key idea: Protein and RNA structure determines function

Genome Stats

organism	estimated size	estimated gene number	average gene density	chromosome number
<i>Homo sapiens</i> (human)	3000 million bases	~30,000	1 gene per 100,000 bases	46
<i>Mus musculus</i> (mouse)	3000 million bases	~30,000	1 gene per 100,000 bases	40
<i>Drosophila melanogaster</i> (fruit fly)	180 million bases	13,600	1 gene per 9,000 bases	8
<i>Arabidopsis thaliana</i> (plant)	125 million bases	25,500	1 gene per 4000 bases	5
<i>Caenorhabditis elegans</i> (roundworm)	97 million bases	19,100	1 gene per 5000 bases	6
<i>Saccharomyces cerevisiae</i> (yeast)	12 million bases	6300	1 gene per 2000 bases	16
<i>Escherichia coli</i> (bacteria)	4.7 million bases	3200	1 gene per 1400 bases	1
<i>H. influenzae</i> (bacteria)	1.8 million bases	1700	1 gene per 1000 bases	1

* from http://www.ornl.gov/sci/techresources/Human_Genome/faq/compngen.shtml

Comparative Genomics

- Analyze and compare genomes from different species
- Goals
 - Understand how species evolved
 - Determine function of genes, regulatory networks, and other non-coding areas of genomes

Tools

- Public Databases
 - NCBI: clearing house for all data related to genomes
 - Genomes, Genes, Proteins, SNPs, ESTs, Taxonomy, etc
 - TIGR: hand curated database
- Analysis Software
 - Database “query” (find similar sequences), alignment algorithms, family id (clustering), gene prediction, repeat finding, experimental design, etc
 - Expect for query routines, these are generally not accessible to biologists. Instead, results are made available via databases and browsers
- Browsers
 - Genome: Ensembl, MapViewer
 - Comparative Genomics: VISTA, UCSC
 - Can query on location, gene name, everyone plays together!

Browser Links

- UCSC Genome Browser
 - <http://www.genome.ucsc.edu/>
- VISTA
 - <http://gsd.lbl.gov/VISTA/index.shtml>
- Map Viewer
 - <http://www.ncbi.nlm.nih.gov/mapview/static/MVstart.html>
- Ensembl
 - <http://www.ensembl.org/>

(try using each one to find your favorite gene)

Queries and Alignments

- Find matches between genomes
- “Queries” find local alignments for a gene or other short sequence
- Global alignments attempt to optimally align complete sequences
 - “Indels” are insertions/deletions that help construct alignments:

```
AGGATGAGCCAGATAGGA---ACCGATTACCGGATAGC
|||||  |||||  |||||
AGGATGA-CCAGATAGGAGTGACCGATTACCGGATAGC
```

Large Genome Alignments

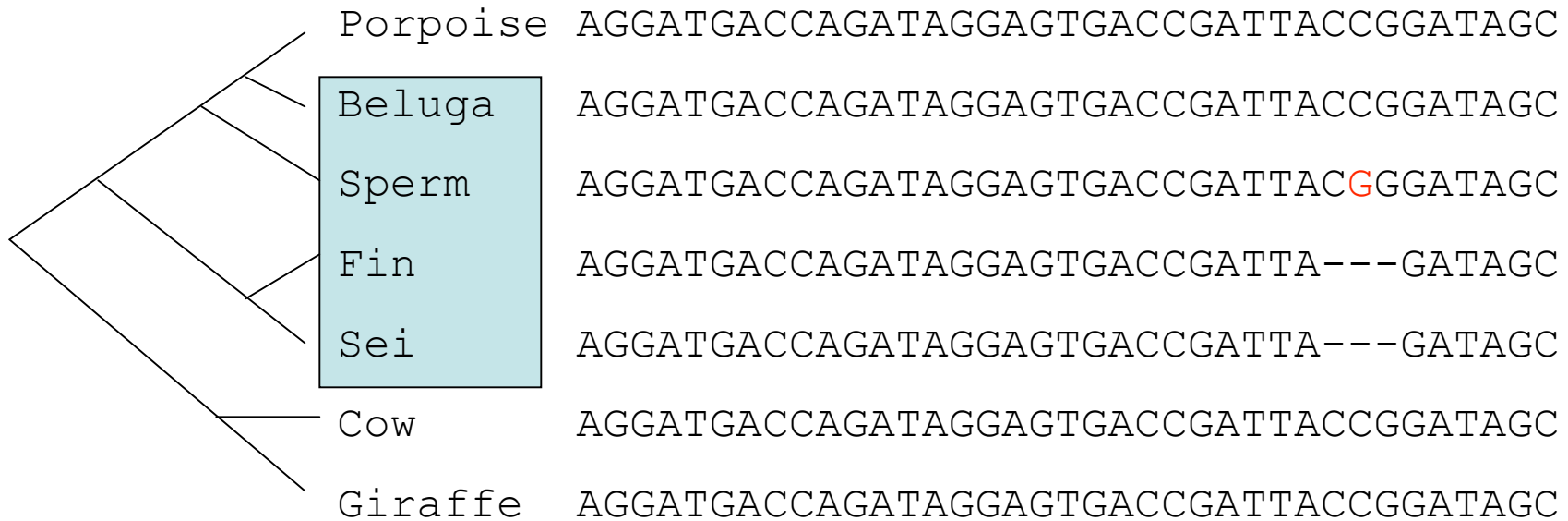
- LAGAN
- MLAGAN
- Shuffle LAGAN

Application: Phylogenetic Analysis

- Determine the evolutionary tree for sequences, species, genomes, etc
- Theory: natural selection, genetic drift
- Traditionally done with morphology
- Techniques
 - Model substitution rates
 - Statistical models based on empirically derived scores
 - Works well for proteins, but is difficult for DNA
 - Phylogenetic reconstruction
 - Distance metrics*
*No evolutionary justification!
 - Parsimony (fewest # of subs wins)*
 - Maximim likelihood

Example

What is the evolutionary tree for whales?



Application: Phenotyping Using SNPs

- SNP: Single Nucleotide Polymorphism - change in one base between two instances of the same gene
- Used as genetic flags to identify traits, esp. for genetic diseases
- CG goal: Identify as many SNPs as possible
- Challenges
 - Data: need sequenced genomes from many humans along with information about the donors
 - Need tools for mining the data to identify phenotypes
- dbSNP is an uncurated repository of SNPs (many are misreported)
- (this was the one talk from industry)

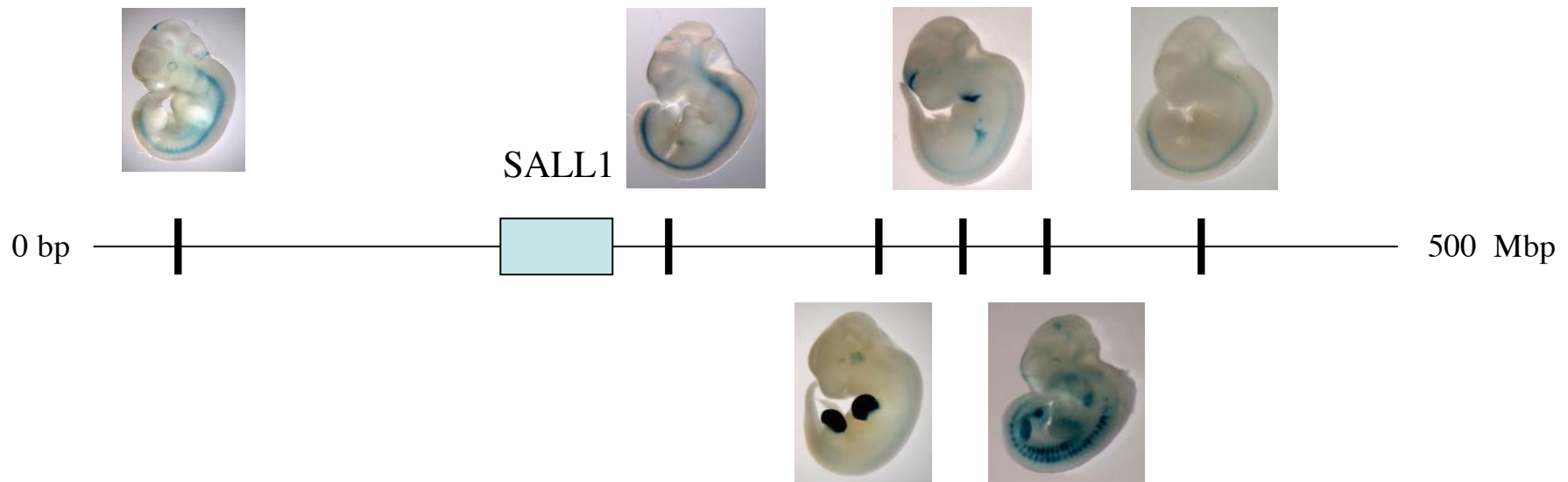
Application: Fishing the Genome

- Look for highly conserved regions across multiple genomes and study these first
- Only 1-2% of the genome is coding, need a way to narrow the search
- Driving Principle: regions are conserved for a reason!

(VISTA Plot of SALL1 Human-
Mouse-Chicken-Fugu)

Chromosome 16 Enhancer Browser

Find conserved regions between genes in human
fugu (pufferfish) alignments and systematically
study them



DOE Joint Genome Institute

(or, this stuff is cool, sign me up!)

- “Industrialized” genomics
 - High throughput genomic sequencing
 - Technology development
 - Computational Genomics
 - Functional Genomics
- Model: Partner with researchers to on sequencing and technology projects
- All data freely available
 - <http://genome.jgi-psf.org/>
- <http://www.jgi.doe.gov>

CS Challenges

- “Engineering”
 - Scalability! (nothing really scales well right now)
 - Stability! (Interactive apps crash way too often)
 - **Timeliness of data**
 - **Biologists don’t use Unix!** (and the Web is not the answer)
 - Better/faster algorithms
 - **Interoperability among tools and better analysis tools**
 - It’s hard for biologists to use their own data with existing tools
- “Basic”
 - Automated curation, error checking
 - Computational models that biologists can trust
 - Structure/Function algorithms (this really is the grail)
- Education! (both ways)