DNA sequence assembly

The human genome project, etc.

DNA sequencing

• Determining order of nucleotides in DNA sequence
  — process by which gene sequences determined in laboratory
  — started with individual genes; has grown to entire genomes
• Goal: knowing sequence & function of every gene leads to defining all functions of cells/organisms in terms of genes & their interaction with environment

Human genome project: case in point

• Raw data
  — 23 chromosomes
  — about 25,000 genes in 3.2 billion base pairs

What we know ...

• The following facts (and more) can be found at: http://www.ornl.gov/sci/techresources/Human_Genome/project/journal/legends.shtml
  — The human genome sequence is almost exactly the same (99.9%) in all people.
  — About 2% of the genome encodes instructions for the synthesis of proteins.
  — Repeat sequences that do not code for proteins make up at least 50% of the human genome.

What we know ...

• The human genome’s gene-dense "urban centers" are predominantly composed of the DNA building blocks G and C.
• Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between.
• Stretches of up to 30,000 C and G bases repeating over and over often occur adjacent to gene-rich areas, forming a barrier between the genes and the "junk DNA." These CpG islands are believed to help regulate gene activity.
• Chromosome 1 (the largest human chromosome) has the most genes (3,168), and Y chromosome has the fewest (344).

What we don’t know ...

• Functions are unknown for more than 50% of discovered genes.
• Excerpts from the Future Research Checklist at the previously-cited link:
  — Exact gene number, exact locations, and functions
  — Gene regulation
  — DNA sequence organization
  — Noncoding DNA types, amount, distribution, information content, and functions
  — Coordination of gene expression, protein synthesis, and post-translational events
  — Predicted vs experimentally determined gene function
  — Correlation of SNPs (single-base DNA variations among individuals) with health and disease
  — Disease-susceptibility prediction based on gene sequence variation
  — Genes involved in complex traits and multigene diseases
Anticipated Benefits of Genome Research

Molecular Medicine
Microbial Genomics
DNA Identification (Forensics)
Agriculture, Livestock Breeding, and Bioprocessing

Gene Testing, Pharmacogenomics, Gene Therapy

Anticipated Benefits:
- improved diagnosis of disease
- earlier detection of genetic predispositions to disease
- rational drug design
- gene therapy and control systems for drugs
- personalized, custom drugs

Genomes to Life: A DOE Systems Biology Program
Exploring Microbial Genomes for Energy and the Environment

Goals
- identify the protein machines that carry out critical life functions
- characterize the gene regulatory networks that control these machines
- characterize the functional repertoire of complex microbial communities in their natural environments
- develop the computational capabilities to integrate and understand these data and begin to model complex biological systems

How it’s done
- Automated sequences, technology improvement have sped up rate of sequencing
- Basic method (dideoxy sequencing) relatively unchanged since 1980s
  - each sequencing run provides no more than 1000 nucleotides of information
  - each segment must be sequenced multiple times for accuracy
- Assembly: process of putting together all these sequences

Assembly methods: two approaches
- The “official” group, Francis Collins and IHGP, preferred a hierarchical technique, beginning with long fragments cloned into BACs (as portrayed in the video)
  - Random fragments of BACs cloned into plasmid vectors, sequenced, assembled into contigs
  - Since relationship of BAC fragments known, fairly easy to assemble BACs into genome sequence
  - Accurate, not computationally taxing – but slow

GTL Applications in Energy Security and Global Climate Change

Within a Decade
- 2030
- Save billions of dollars in fossil fuel costs
- Clean up hazardous waste sites
- Increase sustainable bioenergy production
- Develop biological systems for enhanced carbon capture

Long Term
- 2050
- Contributed U.S. economy highly competitive
- Biotechnology industry in place
- Balance global atmospheric temperatures
- Reduce greenhouse gas emissions
- Increase energy efficiency in transportation
- Promote sustainable development

2040

2050
IHGP’s method

Assembly methods: two approaches

- J. Craig Venter & Celera used whole-genome shotgun sequencing:
  - No BAC cloning: broke up entire genome into small fragments sequenced from plasmid vectors
  - Had to assemble complete genome without advance relationship info (from BACs)
  - Advantage: speed
  - Computational difficulties:
    - errors in sequences, so overlaps don’t line up perfectly
    - large number of overlaps in complex sequences makes determining fragment overlaps difficult

Celera’s approach

Sequence Assembly

- Computational advances made whole-genome shotgun sequencing practical – now preferred approach
- Computational problem; strings of nucleotides with
  - no way of knowing location of fragments with respect to original sequence
  - orientation unknown (could have come from either strand)
- Assembly program has to determine how fragments fit together by finding overlaps

Sequence assembly

- Still an open problem – not completely solved
- Most programs
  - determine amount of overlap between fragments
  - merge fragments with greatest degree of overlap
  - not accurate for all assembly problems!
- 2003 version of human genome had error rate of 1/10,000 bp, down from 1/1000 in 2000 draft

Electropherogram

- Also known as DNA trace
- Graph produced in automated sequencing: interpretation of color & intensity of fluorescence from each fragment
- Base-calling program (part of sequencing software) assigns most likely base to each peak of electropherogram
Software for analyzing base calls

- Chromas Lite: described in text
- FinchTV: available from Geospiza: http://www.geospiza.com/Products/finchtv.shtml
  – recommended by Greg

Installing FinchTV

- Download the software: http://www.geospiza.com/ftvdlinfo.html
- Click link for your operating system:
- Run Setup

Using FinchTV

- Download the sample sequences from the textbook website (files are named SAMPSQ1.AB1 and SAMPSQ2.AB1)
- Open first sequence in FinchTV
- You should see something like this:

Using FinchTV

Middle section contains long stretch of strong signals – base calls are obvious from data. In this illustration, the horizontal scale is compressed to get more of the sequence on screen at once.

Using FinchTV

You can edit individual base calls by clicking on the base or on the sequence signal, then typing a replacement or by inserting or deleting at that point.
Using FinchTV
Ambiguous base calls are marked with an N (undetermined). You can find these using the search box marked “Find Sequence.”

Using FinchTV
As you move horizontally through the sequence, you will see that the signals start to become less strong.

Using FinchTV
... to the point that ambiguity is all that is left; at this point, just delete the rest of the sequence.

Using FinchTV
Select the target section, then choose Edit / Delete. Keep in mind that a single sequencing run rarely produces more than 1000 readable base calls.

Using FinchTV
• You can save your corrected file to FASTA format by using the Export tool on the File menu.

Using FinchTV
• In the Save Sequence dialog box that opens as a result, be sure to put quotes around the file name and add the .TXT extension to make the file readable by text editors.
Dealing with file incompatibility issues

- Because FinchTV was probably developed on a platform you are not using (most likely UNIX, although it might have been a Mac), your text file might look a little strange when you bring it up in NotePad:
  - You can clean up the little rectangles as shown on the next slide

Dealing with file incompatibility issues

- Start by copying one of the little rascals (highlight and hit Ctrl-C)
  - With your cursor at the start of the sequence, open the Edit menu and select Replace
  - In the resulting dialog box, paste your rectangle (Ctrl-V) in the “Find What” box, leave the “Replace” box blank, and click “Replace All”

Sequence Assembly

- No perfect program exists
  - But several good ones are freely available
  - A program called CAP is available both as standalone freeware (to download and install) and as a web-based service, at the following URL:
    http://host9.bioinfo3.ifom-ieve-campus.it/cap/

Using CAP

- As is often the case, you can either paste in the sequences or provide a file name (with several sequences in one file)

CAP results

Assembly results:

- Number of sequence pairs = 1 number of sequence intersections = 1

- CAP results:

  - Number of sequence pairs = 1
  - Number of sequence intersections = 1