Assembling genomes from shotgun sequencing

Kevin Karplus

karplus@soe.ucsc.edu

Biomolecular Engineering Department

Graduate Director, Biomolecular Engineering and Bioinformatics

University of California, Santa Cruz

Outline of Talk

- What is Bioengineering? Biomolecular Engineering? Bioinformatics?
- What is a genome?
- What sequencing technologies are currently used?
- The assembly problem
- Algorithms for assembly

What is Bioengineering?

Three concentrations:

- Biomolecular
  - Drug design
  - Biomolecular sensors
  - Nanotechnology
  - Bioinformatics
- Rehabilitation
- Bioelectronics

What is Bioengineering?

Three concentrations:

- Biomolecular
- Rehabilitation
  - Systems to held individuals with special needs
  - Cell-phone-based systems to reach large numbers of people.
  - Novel hardware to assist the blind
  - Robotics for rehabilitation and surgery applications.
- Bioelectronics

What is Bioengineering?

Three concentrations:

- Biomolecular
- Rehabilitation
- Bioelectronics
  - Implantable devices
  - Interfacing between organisms and electronics
  - Artificial retina project

What to take early

- Mathematics
- Chemistry and then biology
- Introductory bioengineering courses:
  - BME80G, Bioethics (F)
  - BME5, Intro to Biotechnology (W, S)
  - CMPE80A: Universal Access: Disability, Technology, and Society (W, S)
- Declare your major immediately!!
  You can always change to another one latter.
  Bioengineering is one of the most course-intensive majors on campus Many courses have prerequisites.
  It's important to get advising office and faculty advise early.

What is Biomolecular Engineering?

Engineering **with**, **of**, or **for** biomolecules. For example,

with: using proteins (or DNA, RNA, . . .) as sensors or for self-assembly.

of: protein engineering—designing or artificially evolving proteins to have particular functions

for: designing high-throughput experimental methods to find out what molecules are present, how they are structured, and how they interact.

What is Bioinformatics?

Bioinformatics: using computers and statistics to make sense out of the mountains of data produced by high-throughput experiments.

- Genomics: finding important sequences in the genome and annotating them.
- Phylogenetics: “tree of life”.
- Systems biology: piecing together various control networks.
- DNA microarrays: what genes are turned on under what conditions.
- Proteomics: what proteins are present in a mixture.
- Protein structure prediction.

What is a genome?

- Complete sequence of all DNA in a cell (exceptions for plasmids, viruses, organelles).
- Varies from cell to cell, so we usually approximate to get a “typical” genome.
- Usually want an *annotated genome* which has genes and other features labeled and indexed.

Current sequencing technologies

- Sequencing by size sorting
- Sequencing by ligation
- Sequencing by replication
- Single-molecule sequencing

Sequencing by size sorting

- Need need pure sample: many copies of one DNA molecule.
- Generate “prefixes” of DNA, with known last base.
  - Maxam-Gilbert sequencing (obsolete): cuts DNA at specifc base.
  - Sanger sequencing: copies DNA stopping at specific base.
  - Hood variant: copies DNA using flourescent label for last base.
- Measure lengths of prefixes by electophoresis.
- About $1.50/read, 800 bases/read
- Error rate about 0.05% (1 in 2000)

Sequencing by ligation

- Only 1 platform (SOLiD)
- Shreds DNA, then does emulsion PCR to get beads with pure DNA fragments.
- Ligates small stretch of DNA to template.
- Unusual “color-space” reads. Color encodes 2 bases, but only 4 colors:
  - 0 (blue): AA, GG, CC, TT
  - 1 (green): AC, GT, CA, TG
  - 2 (yellow): AG, GA, CT, TC
  - 3 (red): AT, GC, CG, TA
- Takes a week to process a sample
- Get about 200–300 million 50-base reads.
- Error rate about 1.6%

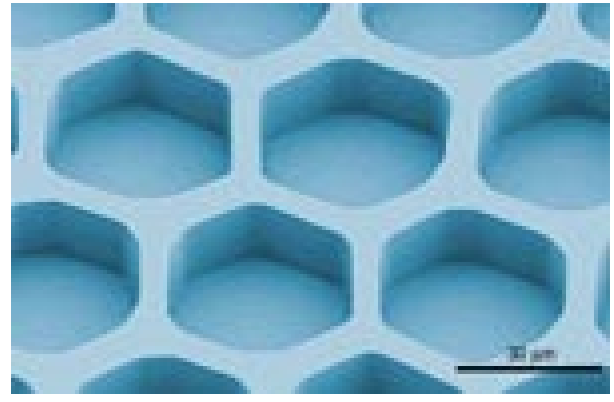
Sequencing by replication

- Bases added one at a time, with detector to tell whether a base is added (or which base is added).
- Pyrosequencing (454)
- Illumina/Solexa (Genome Analyzer)
- Ion Torrent

Pyrosequencing (454 machine)

- After shearing and size-selecting DNA, attach to beads.
- Do emulsion-PCR to get a polony on each bead.
- Put beads into one-bead wells in picotiter plate.

Nature Biotechnology 21, 1425–1427 (2003) doi:10.1038/nbt1203-1425



- Do polymerization with one base type at a time.
- Use light emission to determine how many copies of base are added to end of chains.

Pyrosequencing (454 machine)

- 1,000,000 reads, 400–500 bases/read
- about $3k for a run
- Error rate about 0.9%
- When several bases in a row are identical, determining exactly how many bases of that type were present can be difficult. (homopolymer errors)

Illumina/Solexa

- Polonies grown as spots on a slide rather than separate beads.
- One base at a time reading, all 4 bases read at once (different color fluorophors).
- ≈ 120 million 75-long reads.
- Error rate about 1.5%

Ion Torrent

- not on market yet
- small, cheap machine (expected to be about $50,000)
- Electronic readout, no flourescent molecules, no optics
- medium throughput, fast, low cost per run
- same homopolymer problems as 454 technology

Single-molecule sequencing

- Several new technologies that don't require amplifying DNA:
  - Pacific Bioscience (SMRT)
  - Helicos Bioscience (Helicos)
  - nanopores
- All have super high error rates (10–20%).
- Same molecule must be read repeatedly to get useful data.
- Pac Bio claims very long reads, but has to circularize molecule and use long read to re-read same molecule many times, so effective read length is moderate.

Characterisitics of data

platform	reads/run	read length	error rate
Sanger	1–384	500–1000	very low
454	1,000,000	300–500	low
Illumina	100,000,000–200,000,000	35–100	high
SOLiD	300,000,000	50	high

Different data representations

- base space
- flow space (454, Ion Torrent)
- color space
- Each sequencer and each program uses different data formats and different quality information.

The assembly problem

- Jigsaw puzzle with millions of pieces that overlap.
- Need much more DNA sequence than target genome (generally 15–100X)
- Want to end up with single sequence for each chromosome

Problems

- Sequence data is noisy.
- Repeats can have identical sequences in different parts of genome.
- DNA sample may have variations within sample.
- Data is huge (larger than computer memory).

Algorithms for assembly

- Overlap-consensus graph (needs long reads)
- de Bruijn graph (has trouble with high error rates and long reads)

Overlap consensus

- Each node is a single read. Edges represent overlaps between the end of one read and the beginning of another.
- Clusters of connected nodes can be used to build consensus contigs.
- Overlap must be large enough to be unique location in genome, or chimeric contigs can get built.
- Finding overlaps is expensive part.
- Clusters have to be broken where continuation of contig is ambiguous, so repeats tend to be represented by single consensus contig.
- Best method for 454 and Sanger data.

de Bruijn graph

- Each node is a k-mer. Edges connect window $[i, i+k)$ to window $[i+1, i+k+1)$ of read, and have counts of occurrence.
- Each read becomes a path in the graph.
- Contigs build from strongly supported paths.
- Errors create “bubbles” and “dead-ends” that need to be merged into main paths.
- No need to find overlaps, but graphs get huge.

Web sites

These slides: `http://users.soe.ucsc.edu/~karplus/papers/assembling-genomes-jul-2010.pdf`

Banana Slug Genomics wiki: `http://banana-slug.soe.ucsc.edu/`

UCSC bioinformatics (research and degree programs) info:

`http://www.bme.ucsc.edu/`