Molecular Biology Tools

CS 234

Roadmap

• Restriction Enzyme Digests
• Gel Electrophoresis
• Blotting, Hybridization, Microarrays and RNA-seq
• Cloning
• Polymerase Chain Reaction
• Mass spectrometry
• DNA sequencing (Sanger and next-gen)
• Genome Engineering (CRISPRS/CAS9)
Restriction Enzyme Digests

- In the 70’s H. Smith et al. discovered that bacteria produce enzymes that cut double-stranded DNA
- These enzymes are called restriction enzymes
- The very first restriction enzyme was EcoR1 that cleaves DNA molecules between G and A whenever it encounters a sequence 5’-GAATTC-3’

Restriction Enzyme Digests

- 5’-GAATTC-3’
  3’-CTTAAG-5’

  Digestion with EcoR1

- 5’-G
  3’-CTTAAG-5’ +

  Sticky ends

  5’-CTTAA + 3’-AATTC-3’

  G-5’
Restriction Enzyme Digests

- Restriction enzymes that do not give rise to *sticky ends* create *blunt ends*
- If one assumes uniform distribution, we can expect that EcoR1 cleaves its *restriction site* once every $4^6 = 4,096$ base pairs
- There exists hundreds of restriction enzymes
- Using a restriction enzyme to digest a DNA molecule gives some insights on the sequence (*restriction mapping*)

Gel Electrophoresis

- Complete digestion of a genome with a restriction enzyme can yield hundreds of thousands of DNA fragments
- A typical problem is to separate them based on their length
- In gel electrophoresis, DNA (or RNA or protein) fragments are loaded into wells at one end of a porous gel-like matrix (agarose)
Gel Electrophoresis

- An electric field is applied to the gel
- Molecules migrate toward the other end of the gel
- The distance traveled is inversely proportional to the molecules’ size (i.e., short/small molecules travel further)

DNA fingerprinting
Blotting and Hybridization

- A typical problem is to determine whether a DNA molecule contains a specific sequence $x$ (string matching)
- In blotting, DNA is transferred from the gel to a filter (nylon membrane or nitrocellulose paper) and permanently attached to it
- A solution containing radioactively labelled probes (i.e., the sequence $x$) is then washed over the filter
- If hybridization occurs, then there is a match
Gene expression analysis

- Gene expression does depend on “space location” and “time location”
  - Cells from different tissues produce different proteins
  - Certain genes are expressed only during development or in response to changes to environment, while others are always active (housekeeping genes)
  - …

Comparative hybridization

- Comparative hybridization can reveal genes which are preferentially expressed
  - in specific tissues
  - during specific phases of cell cycle (e.g., mitosis, sporulation, death)
  - during specific changes in the environment (e.g., cold/heat shock, nutrient availability, …)
  - in the context of heterogeneous diseases (e.g., certain types of cancer, diabetes, …)
DNA microarrays

- Monitor the activity of several thousand genes simultaneously
- They exploit the property of DNA to hybridize
- DNA “chips” with a number of probes in the order of 100,000 are common nowadays
- Tiling/genotyping microarrays: probes tile all the genome of an organism to detect genomic variations, i.e., SNPs, deletions/insertions

DNA microarrays

- They “measure” the amount of mRNA in the cell
- One cannot measure directly the mRNA because it is quickly degraded by RNA-digesting enzymes
- Use reverse transcription to get cDNA out of the mRNA: the assumption is that amount of cDNA will be proportional to the mRNA
RNA-seq

- An alternative to microarray is to use sequencing directly on cDNA
- Sequenced reads (corresponding to fragments of transcripts) have to be mapped first to the reference genome, allowing for splicing
- More accurate, more expensive (?) than microarrays
- Possible to reconstruct isoforms

Cloning

- Cloning involves the insertion of specific DNA fragments into chromosome-like carriers called vectors that allow their replication
- All copies of the fragment are identical and they can be purified or stored in a library
- Vectors types: plasmids, cosmids, BACs, YACs
Cloning

<table>
<thead>
<tr>
<th>VECTOR</th>
<th>Size of insert (bps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasmid</td>
<td>2,000-10,000</td>
</tr>
<tr>
<td>Cosmid</td>
<td>40,000</td>
</tr>
<tr>
<td>BAC (Bacterial Artificial Chromosome)</td>
<td>70,000-300,000</td>
</tr>
<tr>
<td>YAC (Yeast Artificial Chromosome)</td>
<td>&gt; 300,000</td>
</tr>
</tbody>
</table>

Slide by Serafim Batzoglou, Stanford U.
Genomic library

- Genomic library
  - Digest a genome with a restriction enzyme
  - Clone all the fragments
- In order to make sure that each fragment is present one need to store multiple copies of same region in the genome
- For example, 4x-5x genome equivalent has a 95% chance to contain at least one copy of each fragment

Polymerase Chain Reaction (PCR)

- Developed by K. Mullis in the mid 80’s
- Used to “amplify” a region of DNA
- Requires that we know a few bases of the sequence of the region to be “amplified” (primer)
- PCR uses DNA polymerase to make $2^n$ copies of a specific region in $n$ rounds
Polymerase Chain Reaction (PCR)

Mass spectrometry

- Proteins are digested into fragments \((\text{protease})\) which cuts the peptide bond between a lysin and an arginine
- Fragmented proteins are charged, then the instruments measures the mass (“weight”) of charged fragment
- The output is a “fingerprint” of peaks corresponding to the molecular weights of each fragment
- Fingerprint are then matched to theoretical fingerprints to derive the corresponding primary sequence
Mass spectrometry

DNA sequencing

- **Objective**: obtain the sequence of A/C/G/T for a fragment of DNA
- Sanger invented the chain-termination method in the 70’s
- Standard DNA sequencing strategies steps
  - Generate of all suffixes of the DNA fragment
  - Label each suffix with one of four tags depending on the leading nucleotide of each fragment
  - Separate the suffixes by gel electrophoresis
A “Sanger read”

A Sanger read: 500-1000 nucleotides

A C G A A T C A G …. A
16 18 21 23 25 15 28 30 32 21

Quality scores: -10*\log_{10}\text{Prob(\text{Error})}
20 corresponds to \text{p}=1/100
30 corresponds to \text{p}=1/1000

**Paired-end sequencing:**
Both leftmost & rightmost ends are sequenced. The distance between the left and the right read is distributed as a \text{N}(\mu,\sigma)

*Slide by Serafim Batzoglou, Stanford U.*
A “Sanger read”

Genome sequencing (BAC-by-BAC)

Slide by Serafim Batzoglou, Stanford U.
Genome sequencing (BAC-by-BAC)

- Assemble
- Fragment, select for size, sequence
- Assemble
- Fragment

Genome sequencing (shotgun)

- Cut many times at random (shotgun), select for size
- Sequencing: get one or two reads from each fragment (paired-end sequencing)

~500 bp ~500 bp

Slide by Serafim Batzoglou, Stanford U.
Sequence with $x$-fold redundancy or sequencing depth/coverage (e.g., 7X) According to Lander-Waterman [1988] the fraction of genome not covered by reads is $e^{-x}$ (e.g., if $x=7$, about 0.1% will be missing)

Goal: Overlap reads and extend to reconstruct the original genomic region

Computational Challenges

- Sequencing errors: $\approx 1\%$ of bases are wrong, chimeric reads, which can induce false overlaps (worse with short reads)
- SNPs, Variants, Repeats (create over-compression)
- All-against-all computation is $O(n^2)$ where $n = \#$ reads
The quest for the $1,000 genome

- **2001**: $2,700M to sequence the human genome
- **2002**: $50M to sequence the mouse genome
- **2004**: $25M to sequence a mammalian genome
- **2008**: $1M to sequence Watson
- **2014**: $1,000 to sequence the human genome (Illumina HiSeq X)

- Archon X-Prize $10M: 100 genomes in 10 days for less than $10,000 per genome

### 2nd gen sequencing technologies

<table>
<thead>
<tr>
<th>Sequencing platform</th>
<th>Sequencing biochemistry</th>
<th>Read length (bp)</th>
<th>Usable throughput (Gbp/day)</th>
<th>Read base error rate (%)</th>
<th>Dominant error type</th>
<th>Rate of inconsistencies with genotyping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche 454 Titanium</td>
<td>Emulsion PCR and polymerase pyrosequencing</td>
<td>4000&lt;sup&gt;1&lt;/sup&gt;</td>
<td>0.8–1.2</td>
<td>0.5–1.5</td>
<td>Insertion &amp; deletion</td>
<td>&gt;0.5% (ref. 7)&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>Illumina Genome Analyzer II</td>
<td>Cluster PCR and reversible terminators polymerase sequencing</td>
<td>75–125</td>
<td>1.5–2.5</td>
<td>0.2–2</td>
<td>Substitution</td>
<td>&lt;0.1% (refs. 8 and 9)&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>AB SOLID 3.0</td>
<td>Emulsion PCR and sequencing by ligation</td>
<td>50</td>
<td>0.8–3.0</td>
<td>&lt;0.1 (2 bp coding)&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Substitution</td>
<td>NA</td>
</tr>
<tr>
<td>Helicos SMS Heliscope</td>
<td>Amplification-free and single-molecule sequencing</td>
<td>32</td>
<td>2.5</td>
<td>4–5</td>
<td>Insertion &amp; deletion</td>
<td>0.2% (ref. 1)&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Illumina HiSeq 2500 v4 chemistry: 1 Tb of sequence (4B 2 x 125 base reads) from two flow cells in about 6 days with error rates < 1%

[https://www.youtube.com/watch?v=77r5p8IBwJk](https://www.youtube.com/watch?v=77r5p8IBwJk)
[https://www.youtube.com/watch?v=womKfikW1xM](https://www.youtube.com/watch?v=womKfikW1xM) (3:10)
3rd gen sequencing technologies

• **Pacific Biosciences** uses a *zero mode waveguide* which has a cylindrical hole etched on a silicon chip; the size of the hole is so small that the ZMW can carry only one DNA molecule; a silicon chip consists of thousands of such holes, each carrying a different DNA molecule; a DNA polymerase is affixed at the bottom of each ZMW; when a DNA molecule passes through the polymerase, DNA synthesis takes place one nucleotide at a time; each bases floating around carry fluorescent dyes of different colors, and so the incorporation of each nucleotide into the chain results in emission of light pulse of the corresponding color; those colored pulses are detected by an optical sensor linked to the ZMW.
  
  • [https://www.youtube.com/watch?v=NHCJ8PtYCFc](https://www.youtube.com/watch?v=NHCJ8PtYCFc)

3rd gen sequencing technologies

• **Oxford Nanopore** uses a protein nanopore attached to a synthetic membrane; a potential is applied across the membrane resulting in a current flowing through the aperture of the nanopore; as DNA flows through the nanopore different nucleotide cause characteristic disruption in the current which allows the base to be identified

• Both PacBio and MinION generate very long reads (10K-100K) but the error rate is very high (Pacbio is 15% errors: 11% insertions, 4% deletions and 1% mismatches), however there is no deterioration of accuracy with longer and longer reads (errors are uniform)

• Throughput is not as high as Illumina instruments

Throughput/read length

Chromosome Conformation Capture

Lieberman-Aiden et. al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science (2009)
Genome (chromosome) 3D structure

- Conformation of chromosomes in nuclei is critical to many cellular processes such as gene regulation, DNA replication, maintenance of genome stability.
Chromosome Conformation Capture

The Observed Contact Frequency:

\[ f_{i,j} \propto d_{i,j}^\gamma, \quad \gamma \approx -1.2 \]

Contact Frequency Matrices

Figure 2 | Topologically associating domains. a) Hi-C profiles reveal that the mammalian genome is organized into topologically associating domains (TADs): regions that show high levels of interaction within the region and little or no interaction with neighboring regions. The heat map represents normalized Hi-C interaction frequencies. b) Schematic of putative TAD structures. The central regions of TADs show high levels of chromatin interaction and coincide with the presence of tissue-specific genes and their associated enhancers, the interactions of which with their cognate promoters are facilitated by the presence of cohesin and CTCF-binding factor (CTCF). The border regions between TADs are enriched for housekeeping genes, which are often clustered together and generally lack the widely dispersed distal enhancers that are found around tissue-specific genes. The border regions show high levels of CTCF and cohesin binding, although only CTCF seems to prevent interactions between TADs. Figure, part a, is reprinted from REF. 38, Nature Publishing Group.

CRISPR-Cas9

- A CRISPR (Clustered, Regularly Interspersed Short Palindromic Repeats) sequence is a near-palindrome string of non-coding DNA (~35bp), followed by a spacer (~35bp), followed by the same near-palindrome, followed by a different spacer, followed by the same near-palindrome, followed by yet a different spacer, etc.
- 50% of bacteria, 90% of archaea have CRISPRs
- The spacers are chunks of genomes from an invading virus/phages/plasmids (a kind of stored memory of the “invaders”)
- Microbes can add new chunks when they encounter new viruses/phages/plasmids
- Microbes use the spacers as search strings to detect “invaders”; when they do they deploy an enzyme — e.g., Cas9 — that matches the CRISPR spacer sequences to sequences in an invading virus and then cuts the virus at a targeted site, which destroys the virus

Dekker et al., Nature Rev Genetics, 2013
Genome “editing” tools

- **Objective:** introduce a new piece a DNA into a genome (*in vivo*)
- Past: Zinc finger nuclease, Transcription activator-like effector nucleases (TALENs)
- Since 2013, CRISPR (clustered regularly interspaced short palindromic repeats)/Cas9 system has been used for genome editing
- CRISPR/Cas9 more efficient that Zinc fingers and TALENs
- [https://www.youtube.com/watch?v=2pp17E4E-O8](https://www.youtube.com/watch?v=2pp17E4E-O8)
Genome “editing” tools

- Science: “Breakthroughs of the Year” in 2013, cover in Nature
- In the last 2.5 years, more than 300 articles have been published on CRISPR-Cas9 applications
- This technique is revolutionizing biology: wide range of applications in medicine (genetic and infectious diseases), agriculture, biotechnology, etc.
- Critical ethical issues

Sander, Joung, Nature Biotech, 2014
Some of the fundamental questions

• Find the entire sequence of DNA [genome]
• Find all the transcribed genes [transcriptome]
• Find all the proteins and characterize their function [proteome]
• Find all the transcription factors (TF) and the full regulation network [interactome]
• Find all the inheritable non-DNA-encoded factors [epigenome]
• …

Some computational questions

• Predict secondary and/or tertiary structure of proteins
• Predict docking between proteins, DNA, and/or small molecules
• Find protein-coding genes in a genome (UTR, transcription start site, start codon, intron/exon splice junction)
• Discover transcription factor binding sites (promoters and enhancers)
• Discover other functional elements/features: ncRNA “genes”, CpG islands, bendability, methylation, nucleosome positioning, etc.
• Construct genome-wide physical map from fingerprinting data
• Analysis of gene expression (microarray and RNA-seq)
• Analysis of mass-spec data (protein “sequencing”)
• De novo genome assembly from (short/long) reads
• Mapping reads to a reference genome (resequencing)
• Detection of SNPs and other structural variations
• 3D genome reconstruction from contact maps