CMSC423: Bioinformatic Algorithms, Databases and Tools

Dr. Todd Treangen
Lecture #2
9/4/2012

Molecular Biology primer
• Review website
  • Recommend books
  • Supplemental reading material

• Have you tried your glue accounts?
  • Part of HW1 will require you to use glue system
    – Let’s try it out

• Issues/concerns/questions about class and policies?
  • Attendance policy important
Admin continued...

• TA
  • Milad Gholami
    • Mondays & Wednesdays 9-11am
  • Everyone know where the TA office is located?

• Grades will be available through the grade server grades.cs.umd.edu

• Frontiers in Genomics symposium (free!)
  • Excellent speakers
  • October 18th, from 8:45am to 3pm
  • RSVP to igs-event@som.umaryland.edu
Bioinformatics

- Biological Application
- Software and computational tools
- Algorithms and Data Structures
DNA

[Diagram of DNA structure with base pairs labeled: Adenine (A) with Thymine (T), Guanine (G) with Cytosine (C)]
Documents of evolutionary history

The four nucleotides in DNA contain the bases: guanine (G), cytosine (C), adenine (A), and uracil (U), read as thymine (T) in DNA.

The genome of an organism contains its hereditary information and is encoded in its DNA.

In 1965, Zuckerandl and Pauling described an organisms DNA as “Documents of Evolutionary History”

See reading material!
However, there are some exceptions
Genetic Exchange in bacteria

There are three main ways which a bacteria can gain new genetic material (DNA) from donor bacterium:

1. Conjugation: Transfer from one bacteria to another via a pilus. (Lederberg and Tatum, 1946)

2. Transduction: Transfer of DNA is mediated by a bacteriophage. (Zinder and Linderberg, 1952)

3. Transformation: DNA is taken up from the environment (Griffith 1928)

REVISED “TREE” OF LIFE retains a treelike structure at the top of the eukaryotic domain and acknowledges that eukaryotes obtained mitochondria and chloroplasts from bacteria. But it also includes an extensive network of untreelike links between branches. Those links have been inserted somewhat randomly to symbolize the rampant lateral gene transfer of single or multiple genes that has always occurred between unicellular organisms. This “tree” also lacks a single cell at the root; the three major domains of life probably arose from a population of primitive cells that differed in their genes.
Extended view of ourselves as a lifeform

- We are composite of species: a ‘supra-organism’
- Our microbial census exceeds the total number of our own human cells by ~10 fold
- Our largest collection of microbes resides in the intestine (~10-100 trillion organisms)
- The aggregate genomes of these gut species (microbiome) may contain >100 fold more genes than our ‘own’ genome
- The microbiome is an integral part of our genetic landscape (‘human metagenome’) and of our genetic evolution
Central dogma

Replication
DNA duplicates

Transcription
RNA synthesis

Translation
Protein synthesis

DNA
RNA
nuclear envelope
nucleus
cytoplasm

Protein

The Central Dogma of Molecular Biology

AGGTACGCGTACCTGACAGG

DNA Replication

The Cartoon Guide to Genetics
Larry Gonick & Mark Wheelis, 1983
Genes, transcription, translation

• DNA – RNA - Thymine replaced by Uracil (T-U)
• The transcribed segments are called genes

ACCGUACC[**AUG**UUA]...AUAGGCU**UGA**GCA

• AUG – start codon (also amino-acid Methionine)
• UAA, UAG, UGA – stop codons
• Genes are read in sets of 3 nucleotides during translation – \(4^3 = 64\) possible combinations
• Each combination codes for one of 20 amino-acids – the building blocks for proteins
# Amino-acid translation table

<table>
<thead>
<tr>
<th>First letter</th>
<th>Second letter</th>
<th>Third letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>U</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
<td>C</td>
</tr>
<tr>
<td>U</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>U</td>
<td>G</td>
<td>U</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino-acid</th>
<th>Phe</th>
<th>Ser</th>
<th>Cys</th>
<th>Stop</th>
<th>Trp</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UUC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UUA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UUG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino-acid</th>
<th>Pro</th>
<th>His</th>
<th>Arg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino-acid</th>
<th>Ile</th>
<th>Thr</th>
<th>Asn</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino-acid</th>
<th>Met</th>
<th>Lys</th>
<th>Ser</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino-acid</th>
<th>Ala</th>
<th>Asp</th>
<th>Gly</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino-acid</th>
<th>Glu</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAU</td>
<td></td>
</tr>
<tr>
<td>GAC</td>
<td></td>
</tr>
<tr>
<td>GAA</td>
<td></td>
</tr>
<tr>
<td>GAG</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amino-acid</th>
<th>Gly</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGU</td>
<td></td>
</tr>
<tr>
<td>GGC</td>
<td></td>
</tr>
<tr>
<td>GGA</td>
<td></td>
</tr>
<tr>
<td>GGG</td>
<td></td>
</tr>
</tbody>
</table>
DNA in the computer

• FASTA/multi-FASTA file format

>gi|110227054|gb|AE004091.2| Pseudomonas aeruginosa PAO1, complete genome
TTTAAAGAGACGGCGATTTCTAGTGAAAATCGAACCAGGGCAGGTCAATTTCCTCCAACCAGCGATGACGTAATAGATAGATA
CAAGGAAGTCATTCTTTTTTTTTTTTTTTTAAAGGATAGAACAACGGTTAATGCTCTTGGGGACGACGGCTTTCTTCT
GTGCATTAACTCGATGAGGCCAGCAATTTCGTGTCTCCGGACGGCAAAAGGGTTTGTAGAAGAACCACCGGTGT
CGAGGGCTGTGTTTCCCTCCTGAGCGAAGGCGCTGGGATGAACGAGATGGTTATCCACAGCGGTTTTTTCTTCA
CGCAGCGGATGTACCCCTTTAAAGGACAGGTTATCCACAAAGTCAGGACCGTCCCTCG

• Parsers easy to write, also available in a variety of software libraries
Genes/proteins in the computer

- >gi|15596155|ref|NP_249649.1| basic amino acid,
- MKVMKWSAIALAVSAGSTQFAVADAFVSDQAEAKGFIEDSSLDDLRLRNNYFNRRDGKSGSAGDRVDFWTQGFL
- TTYESGFTQGTGFGVGVDAGYLGLKLGDGTSDKTGNTGNLPVMNDBGKPRDDYRSRAGGAVKVRISKHTMLKWGE
- MQPTAPVFAAGGSRLFPQTATGFQLQSEFEGLDLEAGHFTEGKEPTTVKSRGELYATYAGETAKSADFI
- GGRYAIDNLSASYGAILEDIYRQYYLNSNYTIPLASDQLGDFNIIYRTNDEGKAKAGDISNTTWSLA
- AAYTLDASHTFTLAYQKVHGDQPFDDYIGFGRNDSAGGDSIFLANSVQYSDFNGPGEKSWQARYDLNLASY
- GVPGLTFMVRYINGKIDGTMKSMDDNVGYKYNYGEGDGHETNLREAKYVQSECAPKDLSFRIRQAWHRA
- NADQGEGDQNEFRLIVDYPLSIL

• Same FASTA/multi-FASTA but with bigger alphabet
Genes/proteins in the computer

- gene complement(1043983..1045314)
  - /gene="oprD"
  - /locus_tag="PA0958"
- CDS complement(1043983..1045314)
  - /gene="oprD"
  - /locus_tag="PA0958"
  - /note="Product name confidence: Class 1 (Function experimentally demonstrated in P. aeruginosa)"
  - /codon_start=1
  - /transl_table=11
  - /product="Basic amino acid, basic peptide and imipenem outer membrane porin OprD precursor"
  - /protein_id="AAG04347.1"
  - /db_xref="GI:9946864"

- GenBank file format
Translation – complications

5’ UTR  Exon  Intron  Exon  Intron  Exon  3’ UTR

pre-mRNA

mRNA
Alternative splicing examples

(a) Alternative selection of promoters (e.g., myosin primary transcript)

(b) Alternative selection of cleavage/polyadenylation sites (e.g., tropomyosin transcript)

(c) Intron retaining mode (e.g., transposase primary transcript)

(d) Exon cassette mode (e.g., troponin primary transcript)
RECAP

• DNA is a string formed with letters A, C, T, G (called nucleotides or bases)
• DNA is double-stranded – allows replication: transfer of genetic “code” from parents to offspring
• DNA is naturally oriented from 5’ to 3’ and the two strands are anti-parallel
• If you know the sequence of one strand, you can obtain the sequence of the other by reverse-complementation

5’ AGACCTAGTGCACGGCTACTACC 3’

5’ CCATCATCGGCACGTGATCCAGA 3’  Reverse

5’ GGTAGTAGCCGTCGTGACTAGGTCT 3’  Complement
RECAP

• Central Dogma of molecular biology:
  – DNA – RNA (transcription)
  – RNA – Protein (translation)

• The transcribed segments of DNA are called “genes”

• Translation occurs in sets of 3 nucleotides – codons

• Each codon encodes one of 20 amino-acids and 3 stop-codons

• In eukaryotes the genes may be split into multiple exons, separated by introns: DNA segments that will not get translated

• The protein is translated from an RNA representing the concatenation of the exons of the gene
The “new” biology

• DNA is not the only heritable information
  – Epigenetic information: RNA molecules, DNA methylation patterns (affects coiling on DNA on histones)

• Complex regulation patterns
  – Genes turn on other genes
  – Genes inhibit other genes
  – RNA interference – small RNA molecules can destroy specific transcripts (down-regulate production)
Playing with DNA

- Biologists can:
  - Cut the DNA – restriction enzymes (often palindromes) (Nobel prize – Arber, Nathans, Smith)
    
    5′GAATTC  
    3′CTTAAG
    5′---G          AATTC---3′
    3′---CTTAA      G---5′
  
  - Attach “things” to DNA (either single or double-strand)
    
    TAGGCACGTTGCAACTACGGC
    TGCAACGCT

- “Amplify” DNA – Polymerase Chain Reaction (Nobel prize – Mullis)
Polymerase chain reaction (PCR)

1. Denature

2. Anneal (attach primer)

3. Extend

4. Repeat
Taq polimerase
DNA sequencing

• Most techniques “trick” the polymerase into revealing the sequence
• The traditional method – Sanger sequencing – based on “terminator” bases – prevent the polymerase from extending the DNA
• Sanger sequencing is essentially PCR + terminator bases
• Other methods “spy” on the polymerase as it incorporates nucleotides
Milestones in Molecular Biology

Nucleotide sequence of bacteriophage \( \Phi X174 \) DNA

http://schatzlab.cshl.edu/teaching
Milestones in Molecular Biology

1995
Fleischmann et al.
1st Free Living Organism
TIGR Assembler: 1.8Mbp

2000
Myers et al.
1st Large WGS Assembly.
Celera Assembler: 116 Mbp

2001
Venter et al. / IHGSC
Human Genome
Celera Assembler: 2.9 Gbp

ABI 3700: 500 bp reads x 768 samples / day = 384,000 bp / day.
"The machine was so revolutionary that it could decode in a single day the same amount of genetic material that most DNA labs could produce in a year."
J. Craig Venter

http://schatzlab.cshl.edu/teaching
The future of sequencing

• Single molecule sequencing - current technology requires many copies of DNA being sequenced - requires DNA amplification
• Massively-parallel sequencing - 100k sequencing reactions occurring at the same time

Sequencing by synthesis

Micro-fluidics

The future of sequencing

Massively parallel sequencing

- each spot is a molecule or amplified from one molecule
- image processing used to track molecules during sequencing by synthesis
- often micro-fluidics/lab-on-a-chip used

http://arep.med.harvard.edu/

• More on this in two weeks!
## The evolution of DNA sequencing

<table>
<thead>
<tr>
<th>Since</th>
<th>Technology</th>
<th>Read length</th>
<th>Throughput/run</th>
<th>Throughput/hour</th>
<th>cost/run</th>
</tr>
</thead>
<tbody>
<tr>
<td>1977-</td>
<td>Sanger sequencing</td>
<td>&gt; 1000bp</td>
<td>4hr 400-500 kbp</td>
<td>100 kbp</td>
<td>$200</td>
</tr>
<tr>
<td>2005-</td>
<td>454 pyrosequencing</td>
<td>250-400bp</td>
<td>4hr 100-500 Mbp</td>
<td>25-100 Mbp</td>
<td>$13,000</td>
</tr>
<tr>
<td>2006-</td>
<td>Illumina/Solexa</td>
<td>50-100bp</td>
<td>3 days 2-3 Gbp</td>
<td>25-40 Mbp</td>
<td>$3,000</td>
</tr>
<tr>
<td>2007-</td>
<td>ABI SOLiD</td>
<td>35-50bp</td>
<td>3 days 6-20 Gbp</td>
<td>75-250 Mbp</td>
<td>est. $3-5,000</td>
</tr>
<tr>
<td>2010-</td>
<td>Illumina HiSeq 2000</td>
<td>2X100 bp</td>
<td>11 days 600 Gbp</td>
<td>2200 Mbp</td>
<td>~$30,000</td>
</tr>
<tr>
<td>2010-</td>
<td>Pacific Biosciences single molecule</td>
<td>1-10 kbp</td>
<td>1 day 2.4 Gbp</td>
<td>100 Mbp</td>
<td>$1,000</td>
</tr>
<tr>
<td>2012-</td>
<td>Oxford Nanopore</td>
<td>80-100 kbp</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
DNA sequenced from bone is 30,000 to 50,000 years old!
Documents of evolutionary history

NEW MEMBER OF THE HUMAN FAMILY

DNA LINK

80,000 YEARS AGO

NEW DNA LINK

40,000 YEARS AGO

Denisovans

Neanderthals
30,000 YEARS AGO

Denisovan tooth found in a Siberian cave

Homo floresiensis ("hobbits")
10,000 YEARS AGO

Source: Nature

African
French
Han Chinese
Melanesian

PRESENT DAY

HOMO SAPIENS
Recap

- Central dogma of biology: DNA -> RNA -> Proteins
  - DNA encodes genes, most of which encode for proteins (via the genetic code)
  - Proteins perform much of the work of the cell.
  - RNA acts as an intermediate step (it also has other functions as well)
- Huge amount of data now available, need algorithms to make sense of it.
Notes & next lecture

- Homework #1 is posted
- Slides will be posted shortly
- Reading material is posted

- Next lecture:
  - Bioinformatics programming, DBs
  - Go over project specification
  - Might start string comparison