Introduction to Bioinformatics

3. DNA editing and contig assembly

Benjamin F. Matthews
United States Department of Agriculture
Soybean Genomics and Improvement Laboratory
Beltsville, MD 20708
matthewb@ba.ars.usda.gov

What we will cover today

- DNA editing
  - Phred
- Sequence assembly (Contig building)
  - Phrap
  - Consed
  - CAP3
  - DNA Star - commercial software
  - http://www.phrap.org/
What we will cover today

- DNA Sequencing software
- DNA sequence assembly
- Similarity searching with a DNA sequence
- BLAST

You cloned a cDNA

- Isolated mRNA
- Reverse transcribed
- Placed into vector
- Transformed and grew bacteria
- Harvested plasmid
- Sequenced insert
DNA sequence analysis

- Is full-length cDNA cloned?
- What are its properties
- What is function of encoded protein?
- Are there family members?
- Is it cloned from other organisms?
DNA sequence analysis

- DNA sequencing software
  - Phred
    - Reads DNA sequencer trace files, calls bases, assigns quality values to each called base
      (for Phred, Phrap, Consed)
  - Phrap
    - A program for assembling shotgun DNA sequence data into contig sequence; provides consensus quality estimates
  - Consed/Autofinish
    - A tool for viewing, editing, and finishing sequence assemblies
  - CAP3 Assembly program
    - Sequence assembly
    - [http://genome.cs.mtu.edu/](http://genome.cs.mtu.edu/)
DNA editing and sequence assembly - building contiguous sequences (contigs)

- Phil Green
- Genome Sciences, University of Washington
- [http://www.phrap.org](http://www.phrap.org)
- Provides software and documentation
Phred

- Software reads sequencing trace files
- Calls bases
- Assigns a quality value to each called base
  - Correct and incorrect base calls
  - Quality values allow sequence trimming
- Works with Amersham Biosciences,
  Applied Biosystems, Beckman, LI-COR
  Life Sciences instruments

Which base reads are reliable

- GATCATGAGCTC
• Vector sequences must be trimmed from both ends
• Poor quality bases must be edited
• PolyA tail indicates 3’ end
• PolyT tail indicates 3’ end reverse sequence

5’ end
TITATCATGGCTGCCCCTAGGGCGAT
GAATGATCGTATGCCAGCTAAAAAAA
AAAATCCGCCG

3’ end

From 5’ end:
ATG = methionine - possible start site
TGA=STOP site
AAAAA…= possible polyA tail
Remaining 3’ sequence may be cloning vector sequence
**Phrap**

- Assembling shotgun DNA sequence data
- Improves assembly accuracy in presence of repeats
- Provides extensive assembly information to assist in trouble-shooting assembly problems
- Handles large data sets

**Consed**

- Automatically chooses finishing reads
- Speeds up finishing
- Integrated with Phrap
- DNA editing more efficient
Sequence from your cDNA clone

```
TACAGCGCGTCGCCCTCCGGCGTCGCGCTTCTCGATTCCAAGGGGAATGTTTTTAAAGGCTCTTACATTGAGTCCGCTGCTTATAACCCCAGCTTGGGACCGCTTCA
GGCCGCCATCGTCGCCTTCATCGCCGGCGGCGGTGGGGATTATGAAGAGATTGTTGCGGCGGTGTTGGTGGAGAAGGAAGGGGCGGTCATCAAACAGGATCACAC
TGCAAGGTTGCTGCTCCATTCCATAGCGCCACGCTGCCACTTCAACAATTTTCTTGCTTCTCAATCTC
```
DNA sequence assembly

- PHRAP
- ConSED
- CAP3
- DNASTAR by Lasergene
  - Commercial - http://www.dnastar.com/
- Sequencher - commercial automated sequencers
  - http://www.genecodes.com/
  - Sequencher protocol
    - http://bip.weizmann.ac.il/sequencher/sequencher.html

**cystein protease**

```
GGAGCTCCACCCGCGGTGGCGGCCGTTCTAGAAGCTAGTGGATCCCCCGGGCTGCAGGAATTCGGCACCAGAACAGTGGGAG
GGAGATCCAAAAAGAGAGAGTGGAAAAGATGGCGCGGTTGATCACAGTGGTGTGGTGCGCGGTGGCGGTGCTATTATGC
CCCGCGGCCGGTGCGTCGTGGGTGGAGGAGGCGGCGAACCCGATACGAATGGTGTCTGGCGTGGAGGCGGAGGTGGTTGG
GTGATCGGGGAGTGCCGGCGTGCGTTGAAGTTTGCTAGGTTCGTGAGCAGGTTCGGGAAGAGTTACCAAAGCGAGGAA
GATGAAGAGAGGTACGAGATATTCTCGCAGAATCTCAGGTTCATCCGCTCCCACAACAAGAAGCGTTTGCCCTATACTCT
CTCTGTTAATTCATTTGGACCTTGCTGGAGCACTGGGAGGTTCAAAAGACACACACACTGGAGCTGCCCCCCATGCTGCC
A
CTCTTTACACACCCATTGATGACCTTGAGATAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
```

**cystein protease**

```
GGAGCTCCACCCGCGGTGGCGGCCGTTCTAGAAGCTAGTGGATCCCCCGGGCTGCAGGAATTCGGCACCAGAACAGTGGGAG
GGAGATCCAAAAAGAGAGAGTGGAAAAGATGGCGCGGTTGATCACAGTGGTGTGGTGCGCGGTGGCGGTGCTATTATGC
CCCGCGGCCGGTGCGTCGTGGGTGGAGGAGGCGGCGAACCCGATACGAATGGTGTCTGGCGTGGAGGCGGAGGTGGTTGG
GTGATCGGGGAGTGCCGGCGTGCGTTGAAGTTTGCTAGGTTCGTGAGCAGGTTCGGGAAGAGTTACCAAAGCGAGGAA
GATGAAGAGAGGTACGAGATATTCTCGCAGAATCTCAGGTTCATCCGCTCCCACAACAAGAAGCGTTTGCCCTATACTCT
CTCTGTTAATTCATTTGGACCTTGCTGGAGCACTGGGAGGTTCAAAAGACACACACACTGGAGCTGCCCCCCATGCTGCC
A
CTCTTTACACACCCATTGATGACCTTGAGATAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
```
Sequence Alignments

Why do DNA sequence alignments?

• If your sequence is not full length, then add other expressed sequence tags (ESTs) to build full-length clone
• Can identify mismatches for single nucleotide polymorphism (SNP) discovery
• Provide a measure of relatedness between nucleotide sequences
• Usually protein alignments with other proteins are used to determining relatedness that allows the drawing of biological inferences regarding
  – Structural relationships
  – Functional relationships
  – Evolutionary relationships
Similarity

- A quantitative measure
- Based on an observable
- Usually expressed as percent identity
- Quantifies changes that occur as two sequences diverge
  - Substitutions
  - Insertions
  - Deletions
  - Identifies residues crucial for maintaining a protein’s structure or function

Similarity

- High degrees of similarity *might* imply
  - A common evolutionary history
  - A possible commonality in biological function
Homology

- Implies an evolutionary relationship
- Hay apply to the relationship
  - Between genes separated by the event of speciation (orthology), ie. orthologous genes
  - Between genes separated by the event of genetic duplication (paralogy), ie. paralogous genes

- Orthologs
  - Sequences are direct descendants of a sequence in a common ancestor
  - Most likely have similar domain structure, three dimensional structure, and biological function

- Paralogs
  - Related through a gene duplication event
  - Provides insight into evolution, ie. adapting a pre-existing gene product for a new function
Global Sequence Alignments

- Sequence comparison along the entire length of two sequences being aligned
- Best for highly similar sequences of similar length
- As the degree of sequence similarity declines, global alignment methods tend to miss relationships

Local Sequence Alignments

- Sequence comparison intended to find the most similar regions in two sequences being aligned
- Regions outside the area of local alignment are excluded
- More than one local alignment could be generated
- Best for sequences that share some similarity or for sequences of different lengths
Scoring Matrices

- Empirical weighting scheme to represent biology
- DNA only has A,T,G,C
- Protein has amino acids; relatedness among amino acids; function; charges; side groups

Matrix Structure: Nucleotides

- Simple match/mismatch scoring scheme
- Assumes each nucleotide occurs 25% of the time
Sequence alignment – Building a contiguous sequence “contig”

Building a contig

• ESTs must be from the same gene, not a paralog (gene duplication event)
• ESTs must be of high quality sequence
• After a contig is constructed, the sequence should be confirmed by cloning and sequencing
EST 1: gagcctatgccggtcagattacgggcttacaggattcatggaccaagtttcacgtc

EST 2: ggacccaaagtttcagttcaaatattgtgttgaccatagaaaaaaaa

EST 3: acgggcttacaggattcatggaccaagtttcacgtc

Consensus sequence:

gagcctatgcgtccgagattacgggcttacaggattcatggaccaagtttcacggatagaaaaaaaaa

In this example EST 3 forms a bridge to connect EST 1 and EST 2.
DNA STAR EXAMPLE

Making a contig from EST sequences

This is your sequence from a clone

1 aactattagg ctctcgtcct ccggtgccc ctggccaaag acctcgctga aacaacaag ggtgctcgcg
 61 cgggtgcca ggtgcctcgt tggccaaag acctcgctga aacaacaag ggtgctcgcg
 121 tgcgtgctg tlgctttcag atcaccgcag tcacattccg cggcccaact gacacccatc
 181 ttgataact tgtgggtcaa gccttgtttg gagatggtgc agccgctgtc attgttggtgat
 241 cagacccctt accagttgaa aagcctttgt ttcagcttat ctggactgcc caaacaatcc
 301 ttccagacag tgaagggcct attgttttcg ccctccgctg aagctttttc gattgagctgc
 361 tcctcaagga tcgtcttgtg ctttcgtcct ccggtgccc ctggccaaag acctcgctga aacaacaag ggtgctcgcg
 421 tccaacctt gggaaatcct gattacaatt ctacattctg gattgcacac cct
Step 1: Go to BLAST search

Step 2: Paste sequence
Step 3: Set constraints and options
Step 4: BLAST!
Collect sequences into a series of files so they can be aligned

File 1.
1 aactattagg cccttcgtcc tccgtcaagc gttacatgat gtaccaacaa ggctgctttg
61 ccggtggcac ggtgcttcgt ttggccaaag acctcgctga aaacaacaag ggtgctgctg
121 tgtgtgctg tgtgtcagat atacccgcca ataccccgaa ctagcctcaac atacccctggt
181 gtaccaacaa ggctgctttg ccggtggcac ggtgcttcgt ttggccaaag acctcgctga
241 aaacaacaag ggtgctcgcg tgcttgccgt ttgttctgag atcaccgcag tcacattccg
301 cggcccaact gacacccatc ttgatagcct tgtgggtcaa gccttgtttg gagatggtgc
361 agccgctgctc attgttggat cagacccctt accagttgaa aagcctttgt ttcagcttat
421 ctggactgcc caaacaatcc ttccagacag tgaaggggct attgatggcc accttcgcga
481 agttggactc actttccatc tcctcaagga tgttcctgga ctcatctcta agaatattga
541 gaaggccttg gttgaagcct tcccaccctt gggaatctcc gattacaatt ctatcttctg
601 g

File 2.
1 ggcaatcaag gaatggggtc aacccaagtc caagattacc catctcatct tttgcaccac
61 tagtggtgtc gacatgcctgt gtagttgtga tcagctcaact aaactattag gctcctggcc
121 ctccgctaag cgtcacatgt ttcaccaaca aacgtgcttt ggctgcctgg cggtgcttcg
181 ctccgctaag cgttcctccg cagctcacta aactattag gccttcgtcc ctggccaaag
241 accttcggtg tttgccaaag acctcgctga aaacaacaag ggtgctcgcg
ttgcttgccgt ttgttctgag atcaccgcag tcacattccg
cggcccaact gacacccatc ttgatagcct tgtgggtcaa gccttgtttg gagatggtgc
361 agccgctgctc attgttggat cagacccctt accagttgaa aagcctttgt ttcagcttat
421 ctggactgcc caaacaatcc ttccagacag tgaaggggct attgatggcc accttcgcga
481 agttggactc actttccatc tcctcaagga tgttcctgga ctcatctcta agaatattga
541 gaaggccttg gttgaagcct tcccaccctt gggaatctcc gattacaatt ctatcttctg
601 g

File 3.
1 tgtgaaagta ccaaaagtgg gaaaaggggc tgtcaactaag gcaatcaagg aatggggtca
61 acaccaagtc aagatattcc atctcatcttt tgcaccacta agtgggtctg acatgccttg
121 tctgattatt cagctcacta aactatagg cccttcgtcc tccgctcaacgcttcatgat
181 gccaacaac cgggtgcttg cccttggccac ggtgcttcgt ttgcacaag acctgcctgt
241 aaacaacaag ggtgctgctg tgtgttctggat atacccgcca atacccgctgaa
cggcccaact gacacccatt gtagttgtca gtagtttttg gagatggtgc
361 aacgtgcttt agtctgagac agatgggcttaa aatcgatgtt ttcaccaacg
301 ctgtagctgc cagatgggctt gtagttgtga tcagctcaact aaactattag
gccctccaaa ggtgctcgcg tttgccaaag acctcgctga
421 ctgcttcgtg tttgccaaagacctcgctga aagtcgctccg cccttcgagc ttcaccaacg
481 agatggtgc accttcgctc ccaccaagtt ggttcctgcg ctacatctta agaatattga
541 gaaggccttg gttgaagcct tcccaccctt gggaatctcc gattacaatt ctatcttctg
601 g
DNA STAR

- **Edit sequence**
  - Allows you to import and edit DNA and protein sequences
- **Megalign**
  - Allows you to align DNA and protein sequences
CAP EST Assembler
Contig Assembly Program

- [http://bio.ifom-firc.it/ASSEMBLY/assemble.html](http://bio.ifom-firc.it/ASSEMBLY/assemble.html)
- Can use up to 30,000 EST sequences
- Fragment maximum is 30,000 bp
- Sequences must be in FASTA format
Some DNA and protein alignment software requires a specific format:

- FASTA format
  - Header HAS TO start with ‘>’
  - A description should follow
  - For DNA only five letters A,C,T,G,N allowed
  - No numbers

```plaintext
>soybean1
ATTCCTTAGGATC…
>soybean 2
TCCGTCAGGTGT…
>soybean3
GGCTATGGCCTAAT…
```
What we learned today

• DNA editing
• Phred
• Phrap
• Consed
• DNA Sequencing software
• DNA sequence assembly
• Similarity searching with a DNA sequence
• BLAST