Introduction to Microbial Genomics

Alignments & sequence information

Dave Ussery
Comparative Microbial Genomics Workshop
BIOTEC building
Pathumthani, Thailand
Module 1, second talk
2 June, 2008
The complete sequence of the genome of ..... has been determined by the whole genome shotgun method .... The entire length of the genome was assembling the sequences .................... The entire length of the genome was 1,669,695 bp. The authenticity of the entire genome sequence was supported by restriction analysis of long PCR products, which were directly amplified from the genomic DNA. As the potential protein-coding regions, a total of 2,694 open reading frames (ORFs) were assigned. DNA Res., 6:83-101, (1999).

The complete sequence of the genome of ..... has been determined by assembling the sequences .................... The entire length of the genome was restriction analysis of long PCR products, which were directly amplified from the genomic DNA. As the potential protein-coding regions, a total of 2,061 open reading frames (ORFs) were assigned, DNA Res., 5:55-76, (1998).
Figure 2.2. Alignment of two DNA sequences at the top does not display similarity. When the complementary strand of the subject is used (the second alignment) the similarity is apparent. At the bottom both strands of the subject are given.
Comparative Microbial Genomics group  
Center for Biological Sequence analysis  
Department of Systems Biology, Technical University of Denmark  

Aligning two sequences........

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>C</th>
<th>U</th>
<th>G</th>
<th>C</th>
<th>U</th>
<th>U</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1.0</td>
<td>0.7</td>
<td>0.7</td>
<td>1.0</td>
<td>0.3</td>
<td>1.3</td>
<td>0.3</td>
<td>1.0</td>
<td>0.3</td>
<td>2.0</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>1.0</td>
<td>0.7</td>
<td>0.0</td>
<td>1.7</td>
<td>3.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>


also see http://en.wikipedia.org/wiki/Smith-Waterman_algorithm
Figure 2.6. Multiple sequence alignment of 20 DNA sequences for IHF binding sites. Nucleotides in the region that was experimentally proven to contain the binding site are color-coded. Nucleotides outside the binding site are not defined (N). Below is a consensus sequence given for this alignment.
Aligning multiple sequences........

How can one visualize an alignment of >100 sequences?
### Comparative Microbial Genomics group

#### Center for Biological Sequence analysis

#### Department of Systems Biology, Technical University of Denmark

---

**B** Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Reference</th>
<th>Description</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ref</td>
<td>ZP_01810188.1</td>
<td>putative oxidoreductase subunit [Campylobacter jejuni]</td>
<td>154</td>
</tr>
<tr>
<td>ref</td>
<td>ZP_01070613.1</td>
<td>oxidoreductase, putative [Campylobacter jejuni]</td>
<td>154</td>
</tr>
<tr>
<td>ref</td>
<td>NP_281605.1</td>
<td>putative oxidoreductase subunit [Campylobacter jejuni]</td>
<td>154</td>
</tr>
<tr>
<td>ref</td>
<td>YP_178483.1</td>
<td>oxidoreductase, putative [Campylobacter jejuni]</td>
<td>153</td>
</tr>
<tr>
<td>ref</td>
<td>YP_001000126.1</td>
<td>oxidoreductase, putative [Campylobacter jejuni]</td>
<td>152</td>
</tr>
<tr>
<td>gb</td>
<td>AAX99091.1</td>
<td>unknown [Campylobacter jejuni subsp. jejuni 81-17]</td>
<td>133</td>
</tr>
<tr>
<td>gb</td>
<td>AAF25214.1</td>
<td>chimeric flagellin A/B [Campylobacter jejuni]</td>
<td>125</td>
</tr>
<tr>
<td>gb</td>
<td>AAC25637.1</td>
<td>flagellin A [Campylobacter jejuni]</td>
<td>125</td>
</tr>
<tr>
<td>gb</td>
<td>AAF25214.1</td>
<td>chimeric flagellin A/B [Campylobacter jejuni]</td>
<td>125</td>
</tr>
<tr>
<td>gb</td>
<td>AAC25644.1</td>
<td>flagellin A [Campylobacter jejuni]</td>
<td>123</td>
</tr>
<tr>
<td>gb</td>
<td>AAC25648.1</td>
<td>flagellin A [Campylobacter jejuni]</td>
<td>123</td>
</tr>
<tr>
<td>ref</td>
<td>NP_281605.1</td>
<td>putative oxidoreductase subunit [Campylobacter jejuni jejuni NCTC 11168]</td>
<td>154</td>
</tr>
<tr>
<td>ref</td>
<td>ZP_01067167.1</td>
<td>oxidoreductase, putative [Campylobacter jejuni subsp. jejuni CF93-6]</td>
<td>154</td>
</tr>
<tr>
<td>ref</td>
<td>ZP_01072110.1</td>
<td>dehydrogenase subunit [Campylobacter jejuni subsp. jejuni HB93-13]</td>
<td>125</td>
</tr>
<tr>
<td>ref</td>
<td>ZP_01100548.1</td>
<td>oxidoreductase, putative [Campylobacter jejuni subsp. jejuni 84-25]</td>
<td>125</td>
</tr>
<tr>
<td>gb</td>
<td>EAQ57550.1</td>
<td>oxidoreductase, putative [Campylobacter jejuni subsp. jejuni CF93-6]</td>
<td>125</td>
</tr>
<tr>
<td>gb</td>
<td>EAQ60080.1</td>
<td>dehydrogenase subunit [Campylobacter jejuni subsp. jejuni HB93-13]</td>
<td>125</td>
</tr>
<tr>
<td>gb</td>
<td>EAQ94348.1</td>
<td>oxidoreductase, putative [Campylobacter jejuni subsp. jejuni 84-25]</td>
<td>125</td>
</tr>
<tr>
<td>emb</td>
<td>CAL34565.1</td>
<td>putative GMC oxidoreductase subunit [Campylobacter jejuni jejuni NCTC 11168]</td>
<td>125</td>
</tr>
</tbody>
</table>

---

**C**

Length = 573 bits

<table>
<thead>
<tr>
<th>Score</th>
<th>Expect</th>
<th>Identities</th>
<th>Positives</th>
<th>Gaps</th>
<th>Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>154</td>
<td>1e-36</td>
<td>71/72 (98%)</td>
<td>72/72 (100%)</td>
<td>0/72</td>
<td>+2</td>
</tr>
</tbody>
</table>

**Query**: QLVMVSNIQEQYDPKGTGLGRNYCAQNNMGTTAQFEQFNTFMGSGALGTSDDDFNGDNFDHSEKEKLHG 400

**Sbjct**: KLVMVSNIQEQYDPKGTGLGRNYCAQNNMGTTAQFEQFNTFMGSGALGTSDDDFNGDNFDHSEKEKLHG
coliphage phiX174
5,386 bp

G Content
0.00
0.40

A Content
0.00
0.40

T Content
0.00
0.40

C Content
0.00
0.40

Annotations:
- CDS +

AT Skew
-0.20
0.20

GC Skew
-0.20
0.20

Percent AT
0.40
0.60

Resolution: 3

BASE ATLAS
Comparative Microbial Genomics group

Center for Biological Sequence Analysis

Department of Systems Biology, Technical University of Denmark

BASE ATLAS
Intrinsic Curvature
dev 0.09
avg 0.51

Stacking Energy
dev -10.31
avg -5.40

Position Preference
dev 0.03
avg 0.25

Annotations:
- CDS +
- Perfect Palindromes
  fix 0.00
  avg 0.30
- Local Inverted Repeats
  fix 3.00
  avg 8.00
- AT Skew
  fix -0.30
  avg 0.30
- Percent AT
  fix 0.35
  avg 0.63

Resolution: 3

GENOME ATLAS
The Race for the $1000 Genome

Fast, cheap genetic analyses will soon become a reality, and the consequences—good and bad—will affect everybody

MARCO ISLAND, FLORIDA—Computers aren’t the only things getting better and cheaper every time you turn around. Genome-sequencing prices are in free fall, too. The initial draft of the first human genome sequence, finished just 5 years ago, cost an estimated $300 million. (The final draft and all the technology that made it possible came in near $3 billion.) Last month, genome scientists completed a draft of the genome sequence of the second nonhuman primate—the rhesus macaque—for $22 million. And by the end of the year, at least one company expects to turn out a full mammalian genome sequence for about $100,000, a 3000-fold cost reduction in just 6 years.

It’s not likely to stop there. Researchers are closing in on a new generation of technology that they hope will slash the cost of a genome sequence to $1000. “Advances in this field are happening fast,” says Kevin McKernan, co-chief scientist at Agencourt Bioscience in Beverly, Massachusetts. “And they are coming more quickly than I think anyone was anticipating.” Jeffrey Schloss, who heads the sequencing-technologies grant program at the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland, agrees. “People are roundly encouraged and nervous,” Schloss says—encouraged because their own technologies are working, and nervous because their competitors are too.

A host of these novel sequencing technologies were on display last month at a meeting here. Although no one at the meeting claimed to have cracked the $1000 genome sequence yet, researchers are getting more confident that it’s a real possibility. “From what I’ve listened to the last few days, there is no physical principle that says we shouldn’t be able to do a $1000 genome,” says Harvard University sequencing pioneer George Church.

Even today, the declining cost of genome sequencing is triggering a flowering of basic research, looking at broad-ranging topics such as how the activation of genes is regulated and understanding genetic links to cancer. And as prices continue to drop, sequencing will revolutionize both the way biologists hunt for disease genes and the way medical professionals diagnose and treat diseases. In fact, some researchers say cheap sequencing technology could finally usher in personalized medicine in a major way. “The promise of cheap sequencing is in the understanding of disease and biology, such as cancer, where the genome changes over time,” says Dennis Gilbert, chief scientist of Applied Biosystems, the leading gene-sequencing-technology company based in Foster City, California. “It will enable different kinds of science to be done.” Of course, as with other forms of advanced technology, some people are concerned about the outcome.

* Advances in Genome Biology and Technology Conference, Marco Island, Florida, 8–11 February 2006.

**Free fall.** As with computer technology, the plunging cost of DNA sequencing has opened new applications in science and medicine.
1. “First Human Genome”
   $3,000,000,000 + 15 years

2. Celera genome (a.k.a. J. Craig Venter)
   $100,000,000 + 0.75 years (9 months)

3. Jim Watson’s genome
   $900,000 + 0.17 years (2 months)

4. Jens Jensen’s genome
   $1,000 + 0.0002 years (0.1 day)
A decade of sequencing prokaryotic genomes.

36 hours, from purified DNA to GenBank file, ready to submit...

- Library preparation
- Sequencer run
- Assembly
- Gene finding
- Gene annotation
- Genome analysis, visualization
- Sequencer set-up
- Emulsion set-up
- Emulsion breaking
- Bead enrichment
- Purified genomic DNA
- Genome Atlas
- NCBI
- Submitting Sequence Data to GenBank
Rapid-sequencing technology has also revealed unimaginable degrees of diversity. Paradoxically, the more diversity we find among microbes, the less simple it becomes to assign microorganisms to species......

The shock of finding seemingly endless variation in microbial genomes has thus quietly revolutionized thinking about evolution and the species concept, as well as strengthened realization of the importance of the ecological services supplied by the microbial universe.

– CAROLINE ASH, JOHN FOLEY, ELIZABETH PENNISI
Comparative Microbial Genomics group

Center for Biological Sequence analysis
Department of Systems Biology, Technical University of Denmark

rRNA tree

Univ

eral phylogenetic tree showing the relationships among Bacteria (e.g., most bacteria and blue-green algae), Archaea (e.g., methanogens and halophiles) and Eucarya (e.g., protists, plants, animals, and fungi).
Distribution of genomic properties in 690 complete prokarytic genomes