Pre-Genomics          Genomics

DNA                   Genome sequencing
Southern blot         Chip hybridization
PCR-based markers     SNP detection

RNA                   Luminex beads (millions!)
Northern blot         Oligo or cDNA chips (10,000s)

Protein               2-D gel -> mass spectrometry
c Western blot        tandem mass spec

1 probe at a time
reuse membrane ten times

What is bioinformatics and who uses it?

We all use bioinformatics (using our brains):
- is that a band or a smudge?
- is this a homozygote or heterozygote?
- is this a real difference?
- is A more related to B or C?

Bioinformatics to
- automate
- standardize
- attach measure of statistical significance
- process large numbers!!!

Key is to keep biology in BIOinformatics

The future of bioinformatics ...

- more computer power
- larger databases
- more links between databases
- more bioinformatic tools

Many of these are really INFORMATICS issues!

Biologists role is to
- understand what tools are available and how to use them
- which analysis to perform for a given problem

Why sequence rice?

Plant genome sizes in megabases:

- wheat 16,000
- barley 5,000
- sugarcane 3,000
- maize 3,000
- cotton 2,200
- soybean 1,100
- tomato 800
- sorghum 760
- rice 420
- Arabidopsis 125

human genome ~ 3,000
fungi ~15-50
bacteria ~1.5-5
**CO-LINEARITY IN CEREALS**

Adapted from Devos and Gale, PMB 1997

- liguleless
- shattering
- waxy
- dwarf

- rice
- maize
- wheat

Rice Genome Sequencing Framework
Clemson University Genomics Institute

1. Genomic BAC Libraries
   - Single-copy vector capable of holding very large insert (100-200 kb)
2. BAC end Sequences
   - Short sequence (ca 500 bases) from BAC vector into cloned DNA
3. Physical Map
   - Restriction digest of BAC DNA
   - Assembly with FPC program

**Framework Objectives:**

- Construct two large-insert BAC libraries
  - 25x genome coverage
- End-sequence clones (110,000)
- Construct a physical map:
  - fingerprint 64,638 BAC clones

**Manual Bandcalling using IMAGE (Sanger Center)**

- 1,536 clones per day

**The Sulston Score**

\[
\sum_{n=im}^{nL} \frac{nL!}{(nL-n)!} \cdot (1-p)^n \cdot p^{(nL-n)}
\]

where:

\[
p = \left(1 - \frac{2 \times \text{tolerance}}{\text{gel length}}\right)^{nh}
\]

im = number of matching bands
nL & nh = number of bands in each of the 2 clones

**DISTRIBUTION OF BAND SIZES - NIPPONBARE HindIII/LIBRARY**

TOTAL NUMBER OF BANDS: 938,318
ASSEMBLY OF A 1.1cM REGION FROM RICE-10

BAC Filter Hybridization

384 x 8 = 3,072 clones
x 6 = 18,432 clones

Physical mapping of rice chromosome 10s: current status

“Shotgun” Approach to Genome Sequencing

Syngenta rice shotgun sequencing project:

- 7.1 Million sequencing runs (Myriad)
- 119,954 BAC end sequences (CUGI)
- 60,000 BAC Fingerprints (CUGI)
- 60,000 cDNA Sequences
**Gene “Modeling”**

Homology to known genes

Prediction based on nucleotide distribution (HMM = Hidden Markov Models)

**Gene Function Prediction**

Homology to known genes

Domains (SCOP, Pfam, InterPro)

---

**Arabidopsis genome in color:**

Red (high density) - deep blue (low density). Gene density (‘Genes’) ranged from 38 per 100 kb to 1 gene per 100 kb.

ESTs ranged from more than 200 per 100 kb to 1 per 100 kb. TE densities ranged from 33 per 100 kb to 1 per 100 kb.

Mitochondrial and chloroplast insertions were assigned black and green tick marks, respectively.

Transfer RNAs and small nucleolar RNAs were assigned black and red tick marks, respectively.

Nature, 408, 796-815, 2000

---

**Functional analysis of Arabidopsis genes**

Nature, 408, 796-815, 2000

---

**Syngenta shotgun sequencing of rice, ctd:**

Random Fragment Sequence Assembly:

- 77,561 Sequence Contigs (5.6 Million Runs)
- 391.4 Mb Covered
- > 6X Coverage
- 500,000 Reads Repetitive DNA
  - 38.3 Mb – Removed
  - 1.3% Mitochondrial - Removed
  - 4.5% Chloroplast - Removed

**Syngenta Rice Project:**

Shotgun sequence rice genome (Myriad)

Assemble into ca 40,000 sequence contigs (Myriad)

Link sequence contigs to fingerprint contigs using BES

(and, by extension, to the genetic map)

First nearly complete physical map of rice genome

⇒ gene prediction and analysis
⇒ map knockout mutants
⇒ cluster ESTs
⇒ phylogenomics

---

**Genome Duplications in Arabidopsis**

http://maps.gsc.doe.gov/pv/tv_frame.html

---

**created by Dirk Haase**
ESTs – Advantages and Disadvantages

Advantages:
1. Cost & speed: sequence expressed portion (2%?) of genome only
2. expression information built into the sequencing project

Disadvantages:
1. no information on regulatory regions and introns
2. no information on genetic location
3. oversampling of highly expressed genes, rare transcripts elude detection
4. one gene or recent duplication?

How many bacterial EST projects?

Clustering ESTs using the rice genomic sequence

1) Which ESTs go together to make one gene?

2) Utilizing the genome sequence to anchor all ESTs
   a. Identify single optimal genome location using BLAST
   b. Assemble all clustered ESTs (PHRAP)
   c. Graph onto genome (est_genome)

Computer-generated gene predictions are often incorrect -

supplement with
- full-length gene isolation and sequencing
- mRNA hybridizations

~832,000 25-mer oligonucleotides/chip

Probe Validation for Bacterial Barcoding

in plant ……

Insertional Mutagenesis in Arabidopsis thaliana

1) Generate 100,000 knockout mutants (Agrobacterium)
2) Collect leaf tissue
3) Sequence borders following TAIL-PCR
4) Collect seeds
5) Make both seeds and sequence information available to collaborators

Many UC Berkeley, Scripps Research Institute and Salk Institute professors, post-docs, graduate students, and technicians, and numerous Syngenta scientists generously assisted with leaf sample and seed collection. This work could not have been accomplished without their help.

Phylogenomics

1) determine presence/absence of each Arabidopsis gene in all completely sequenced organisms

2) visualize by graphing genes vs genomes

\{ Synechocystis and plants only – likely chloroplast function

Applications:

- Enrichment for herbicide targets: single-copy Arabidopsis genes represent 8.6% of the total but make up 25% of confirmed lethal mutations
- Single-copy genes conserved in plants are priority genetic markers
- Elucidation of pathways
Genomics enables large-scale evolutionary studies:

- ortholog
- paralog

DNA Barcoding (Plastids)

DNA BARCODES

A unique identifier for a given species or variety based on organism’s genomic DNA

Traditional methods of identification

- morphology, physiology, DNA sequence analysis, antibodies, biochemical tests, fatty acid analysis, rDNA sequence
- requires adequate material & TIME!

Barcoding

Use bioinformatics to identify suitable target sequences
- a) gene that occurs in all species
- b) enough variation to differentiates subpopulations, pathogenic strains or cultivars

Applications in:
- biosecurity, environmental and ecological studies, invasive species detection

Why a nucleic acid-based barcoding system?

- Universal – nucleic acid present in all living organisms
- Coding capacity – 4^16: 10 divergent nucleotides can theoretically differentiate 1,048,576 organisms
- Proven Technology – widely utilized high-throughput sequencing and hybridization-based

Universal Plastid Primers

- 37 plastid genomes
- conserved primer pair amplifies short region

Plastid Barcode?
- universal amplification
- sufficient length to be unique, ideally at species level

Oligonucleotide Conservation Among 37 Plastid Genomes

- a) All 3,007 24-mers of the sugarcane chloroplast genome that are conserved in photosynthetic flowering plants.
- b) Sequence diversity within the URA region, a potential barcode for algae.
- c) Sequence diversity of the trnD-trnL intergenic spacer, a potential barcode region for closely related organisms.
- d) The plastid barcode proposed by Kress et al. for land plants

The degree of conservation for each 24-mer: 36, 27-23, and ≤21 nt

conserved in the database organism with respect to the sugarcane plastid sequence
A plastidial barcode for eukaryotic algae

universal primers

- species-level discrimination in red algal order Nemaliaceae

A collaboration with Dr. Alison Sherwood
University of Hawaii, Botany

Genome Sequence Project Resources

- informatics tools as varied as the problems
- understanding cellular processes
- Perl, Java, C++