Of Sea Urchins, Birds and Men
Darwin’s Finches

and Coco
"The Sequence of the Human Genome" Science, 2001

"Whole-genome shotgun assembly and comparison of human genome assemblies" Proc. Nat. Acad. Sci. USA, 2004
Algorithmic Functions of Computational Biology – Course 1
Professor Istrail

The Father of All Dot Plots

The Human Genome
Dot Plots 101

- a,b,c,d stand for letters
  A,B,C,D for words

- Where letters match, put a dot

- Where words match, put a line (words can be rc-ed)
Dot Plots 101

- When words line up
- Reversed
- Misplaced
- Something gained (relative to horizontal)
- Something lost (relative to horizontal)
Some large reversals in GP
NCBI has more of the centromere than anyone else (or is that N’s?)
Many reversals in GP, a piece of the end is re-ordered to the middle, celera assemblies boringly good.
Again everyone misses the first 10MB (or are those N’s) of NCBI31
Mixed character of the problem:

- continuous mathematics
- discrete mathematics

GENOMIC REGULATORY SYSTEMS
Emergency knife (S) is always handy in case opossum or the woodpecker gets sick and can't work.
A Tale of Two Networks

Sea Urchin

Drosophila
One gene, 30 years of study, 300 docs and postdocs

A Proposal for Nobel Prize

“Programs built into the DNA of every animal.”
Eric H. Davidson

Genomic Regulatory Systems
The Dogma

Figure 9.2 Schematic model for transcriptional activation. The TATA box-binding protein, which bends the DNA upon binding to the TATA box, binds to RNA polymerase and a number of associated proteins to form the preinitiation complex. This complex interacts with different specific transcription factors that bind to promoter proximal elements and enhancer elements.
Genomic Regulatory Regions

(A) CyIIIa complete system
-2.3 kb

(B) Endo16 complete embryonic system

mammals

(C) mouse pit-1 RA-response/autoregulation element

(D) mouse IL-2 inducible element

(E) rat myosin light chain 1/3 elements

(F) human β-globin locus

-10.7 kb

-0.3 kb

-0.6 kb

-61.0 kb

-53.0 kb
TF Binding Site Complexity
Genome Complexity

1 Billion DNA bases
20,000 Genes
cis-Regulatory Modules Complexity

A. GENE

Cis-Regulatory Modules

1 2 3 a b 4 c 5

1 kb

B. INPUTS

Time 2 Signal 2 Lineage 2

Time 5 Signal 5 Lineage 5

400 bp Module 2

400 bp Module 5

C. GENOMIC SITES

1 kb

D. ALTERNATIVE LOOPING

1 kb

200,000 cis-Modules
The DNA program that regulates the expression of *endo16* in sea urchin
THE FIRST NETWORK
The View from the Genome
The View from the Nucleus

View from the nucleus: Endomesoderm nuclei to hatching blastula stage; the Wnt8/Tcf signalling loop and its genes.

Apr. 19th, 2002

Notes:

1. β-catenin/Tcf input now produced by a zygotic signaling loop driven by Wnt8 expression in endomesoderm cells.

2. β-catenin/Tcf input required for expression of many regulatory genes that become active in the veg2 endomesodermal territory during early-mid blastula stage.
Building Protein-DNA Assemblies

- Inter-cismodule linkage
- Insulation
- Communication
- DNA
- Cismodule
- Cooperativity
- Linear-amp
- Gates
- Potentiality
The Building Blocks

- Free Energy
- Protein
- Protein-DNA Binding (free energy)
- DNA

Free energy is the “GLUE”
Information Processing
- **Boolean Circuit**
- **Synchronous** input and output
- **Completely** defined gates
- **Boolean Circuit**
- **Synchronous** input and output
- **Completely** defined gates

- **Booleaninear Circuit**
- **Asynchronous** input and output
- **Incompletely** defined gates
IF (\(x_1 = 1\) AND \(x_2 = 1\)) THEN

```plaintext
GTAGGATTAAG
......
CATCCTAATTC
......
GTATCTAGAAG
......
```
Web page: http://www.its.caltech.edu/~chyuh/cathy-mirsky-info.html

Caltech, Davidson Lab
October 2004