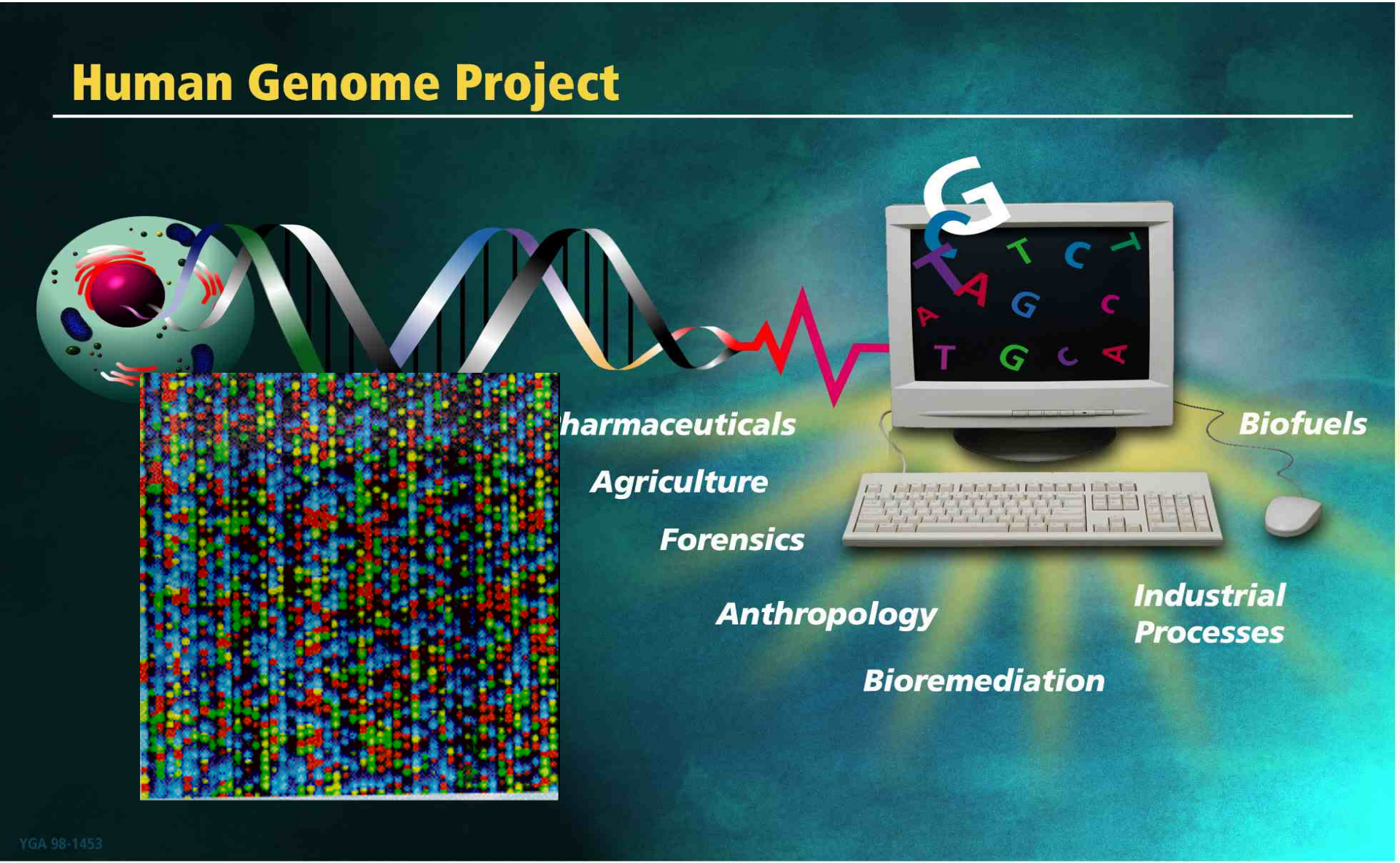
How Computing Science Saved The Human Genome Project

david.wishart@ualberta.ca

3-41 Athabasca Hall

Sept. 9, 2013

The Human Genome Project

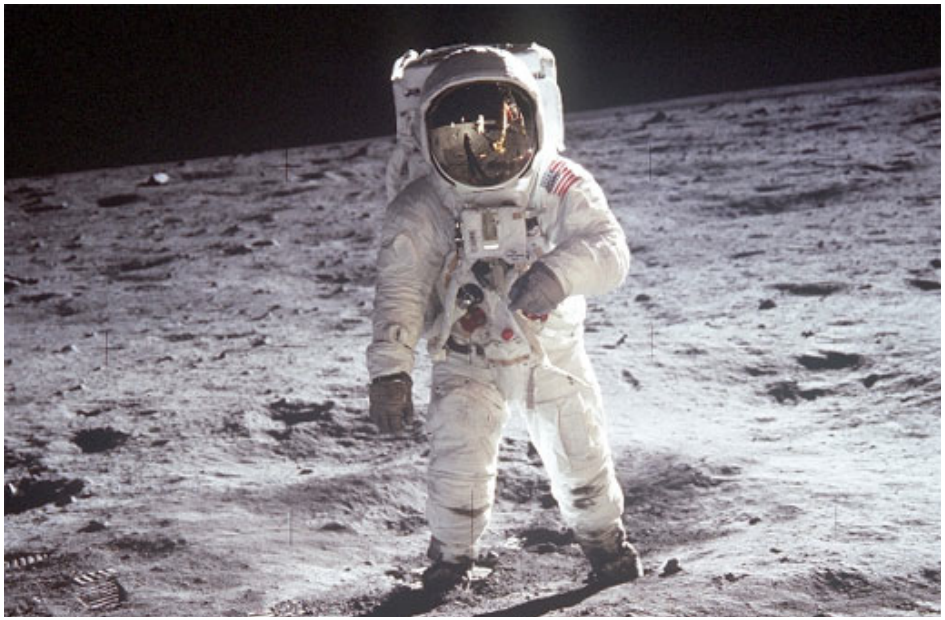


HGP Announcement
June 26, 2000***

J. Craig Venter
Francis Collins

Bill Clinton
Tony Blair

Most Significant Scientific Accomplishments of the Last 50 Years



June 26, 2000

July 20, 1969

The Human Genome Project

- First efforts started in October 1990
- Two competing efforts (private vs. public) Celera vs. NIH
- First Draft completed on June 26, 2000
- "Finished" on May 18, 2006 ($3.8 billion)
- Used hundreds of machines and 1000s of scientists to sequence a total of 3,283,984,159 bases on 24 chromosomes

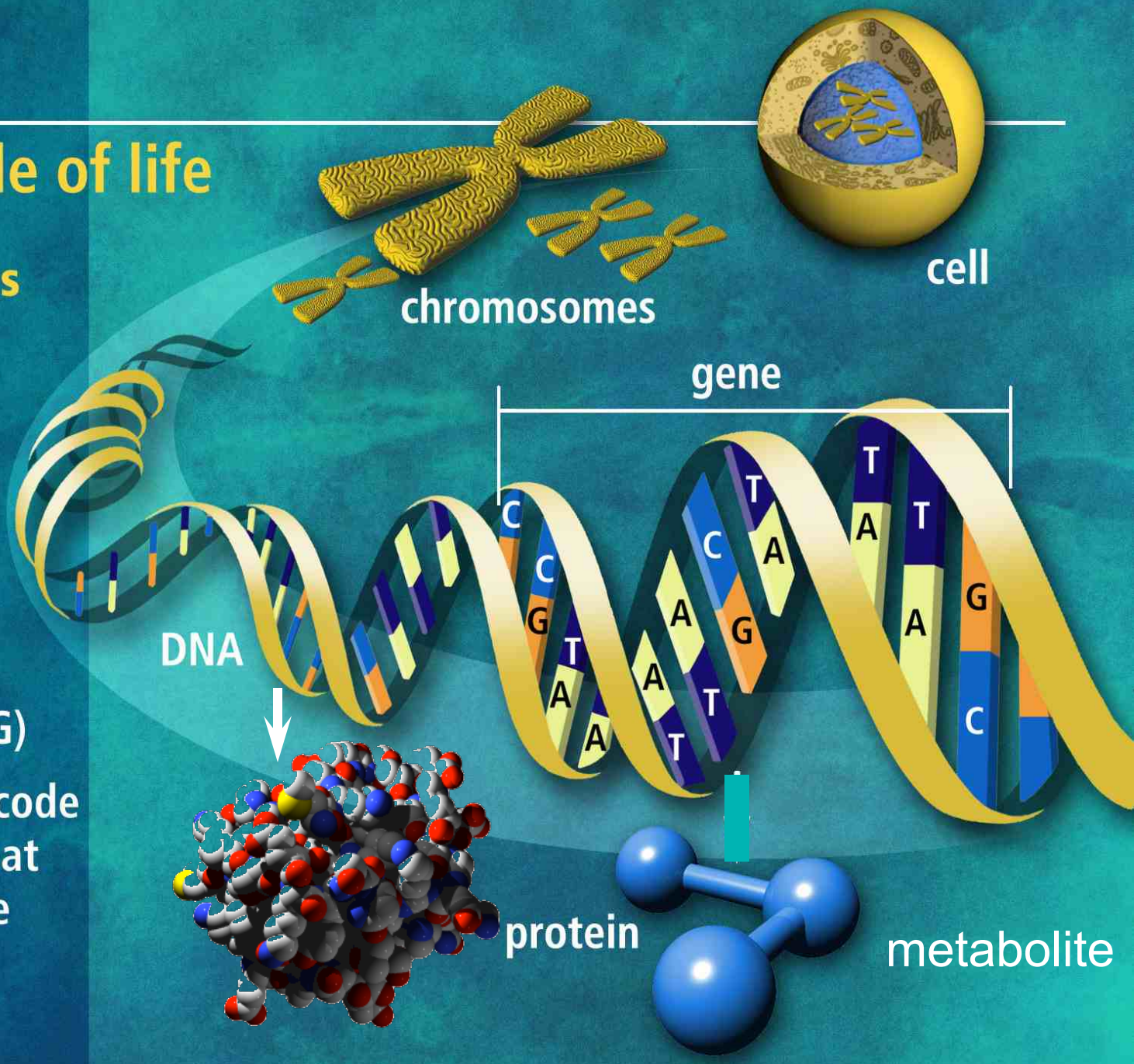
DNA

the molecule of life

Trillions of cells

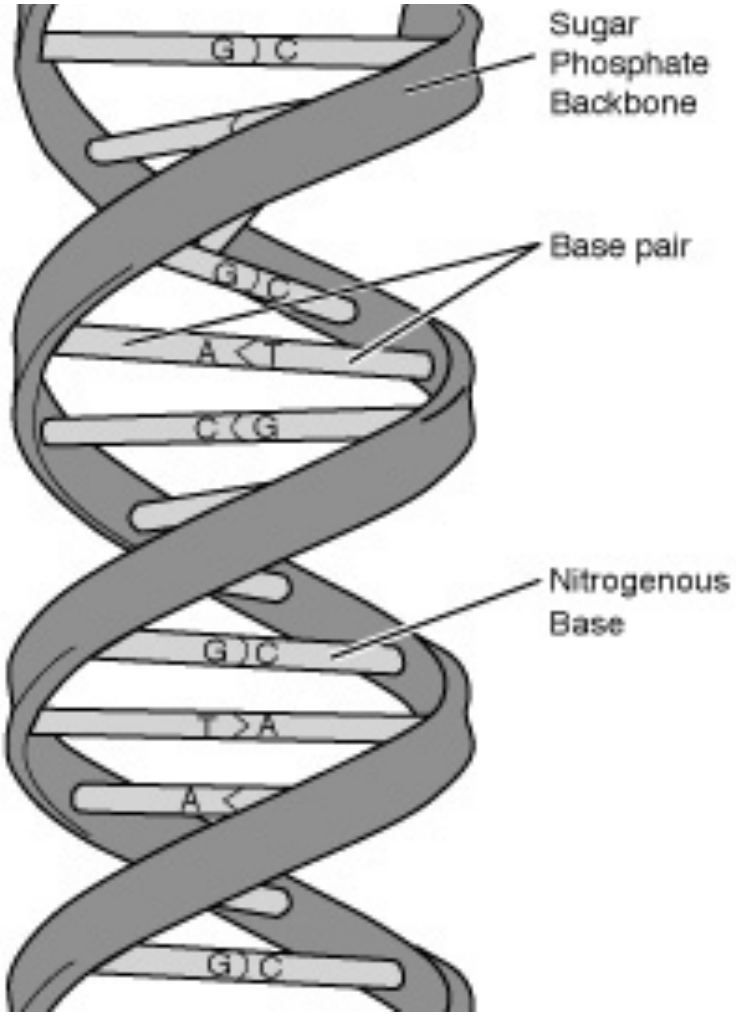
Each cell:

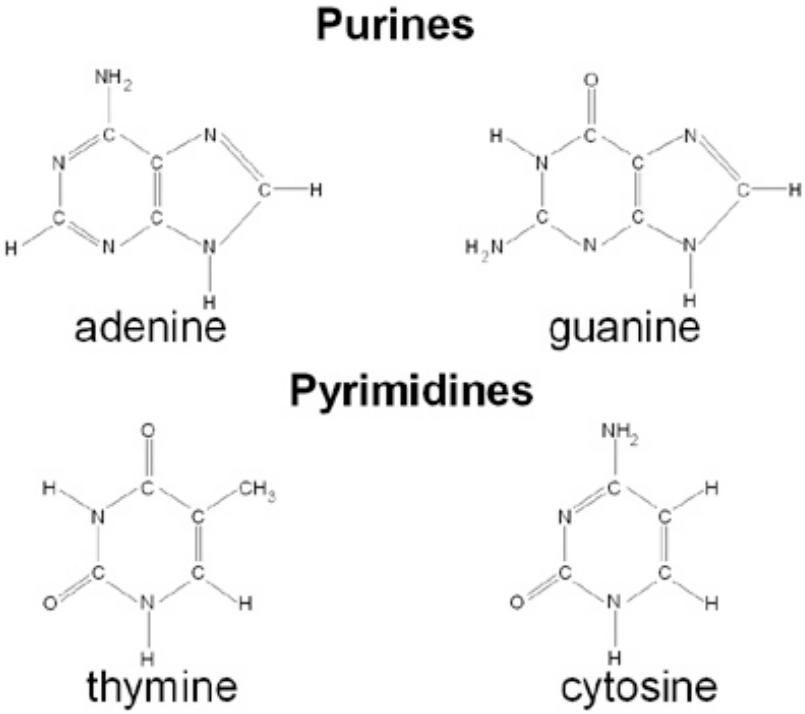
- 46 human chromosomes
- 2 m of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- 21,000 genes code for proteins that perform all life functions



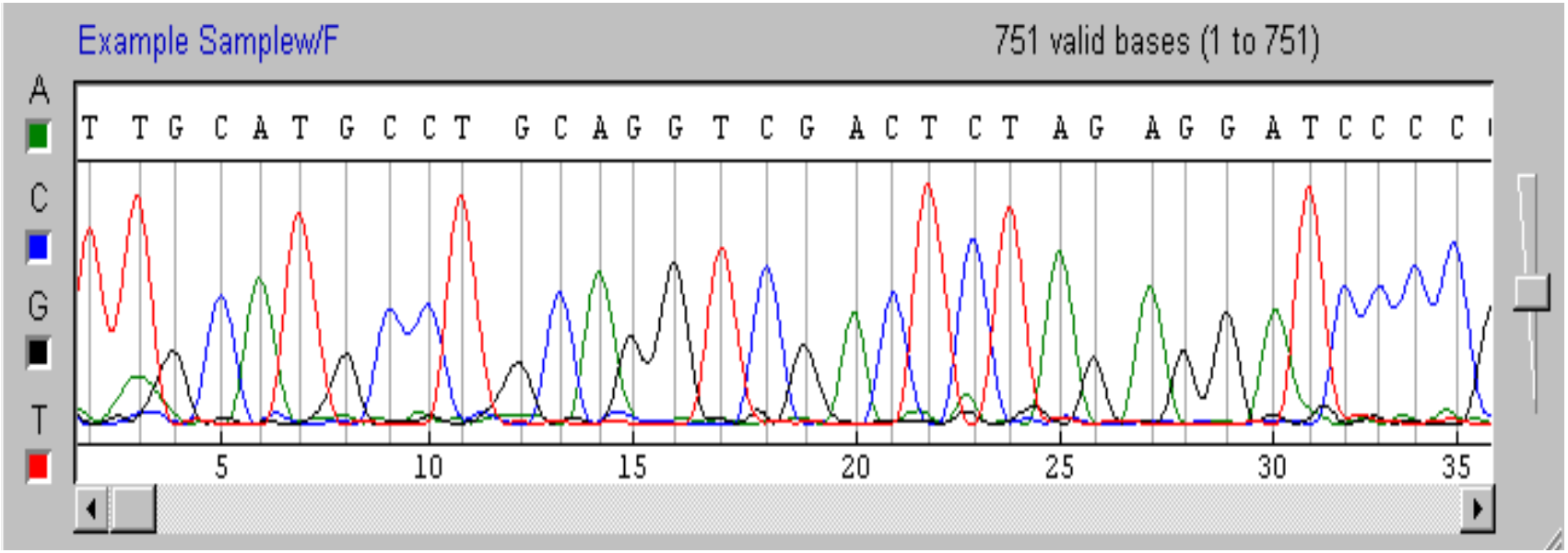
Y-GA 98-090R

DNA Structure & Bases

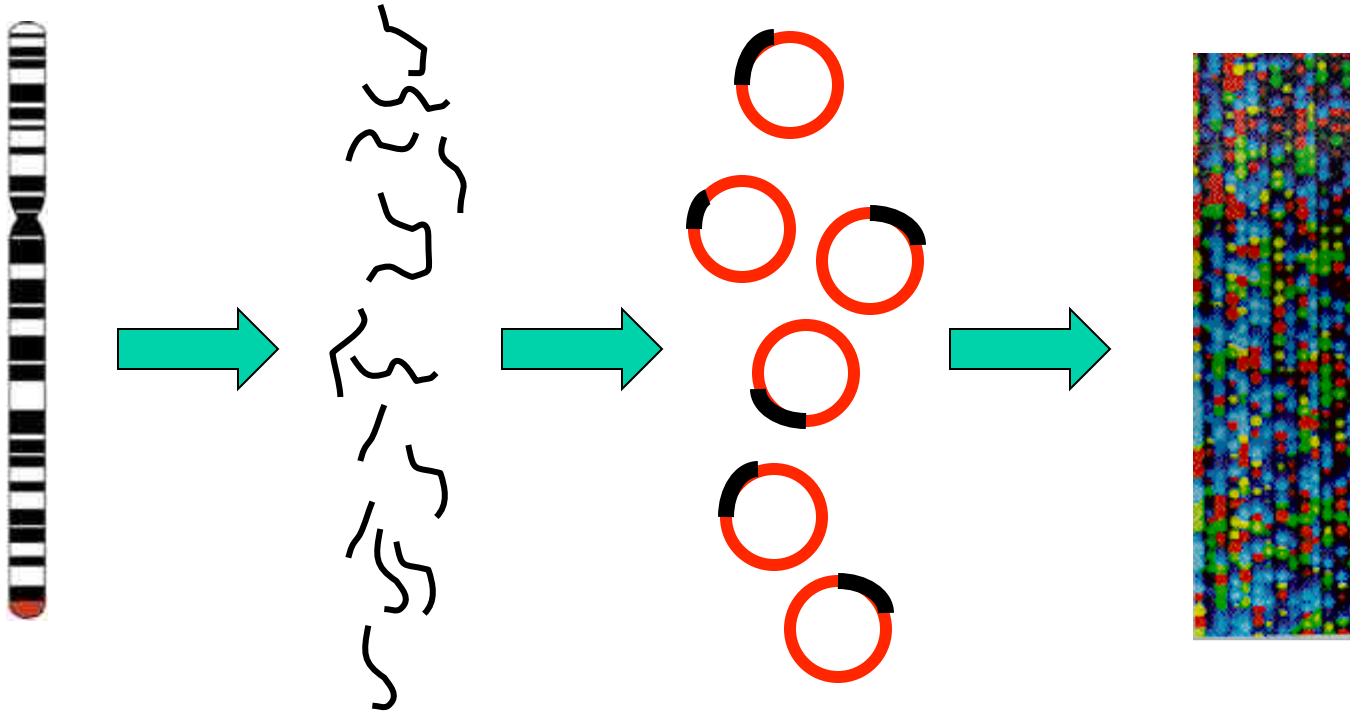




DNA Sequencing – The Key to the Human Genome Project



Shotgun Sequencing



Isolate Chromosome

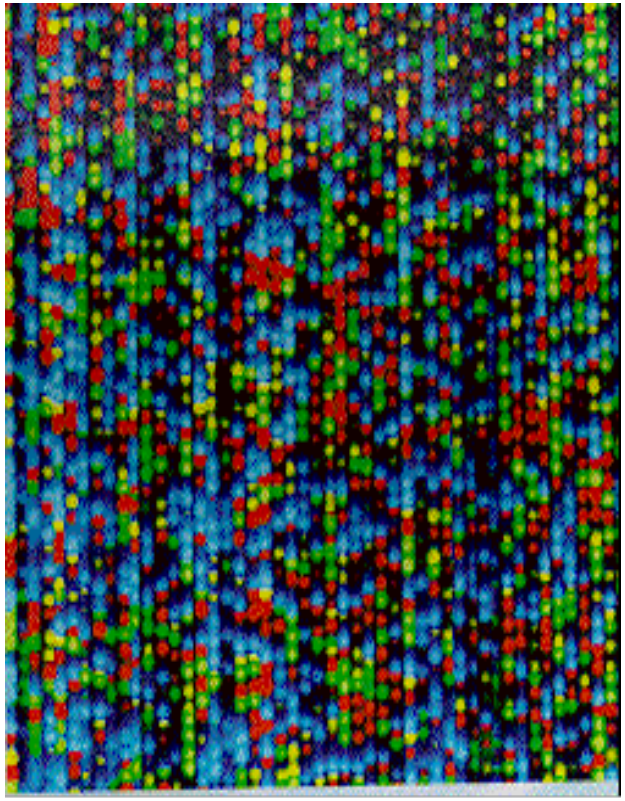
ShearDNA into Fragments

Clone into Seq. Vectors

Sequence

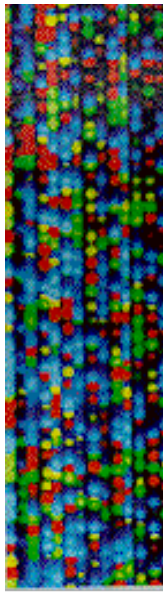
Multiplexed CE with Fluorescent detection

ABI 3700



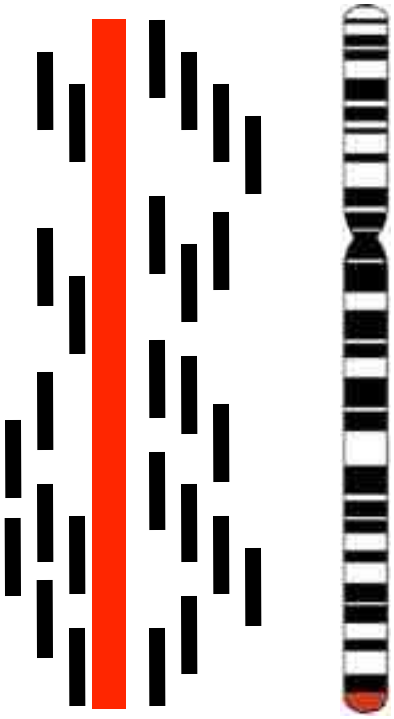
96x700 bases

Shotgun Sequencing



Sequence Chromatogram

Send to Computer



Assembled Sequence

HGP - Challenges

- Reading the DNA sequencer chromatograms (base calling)
- Putting millions of short “reads” together to assemble the genome (assembly)
- Identifying the genes from the DNA sequence (gene finding)
- Figuring out what each gene does

HGP - Challenges

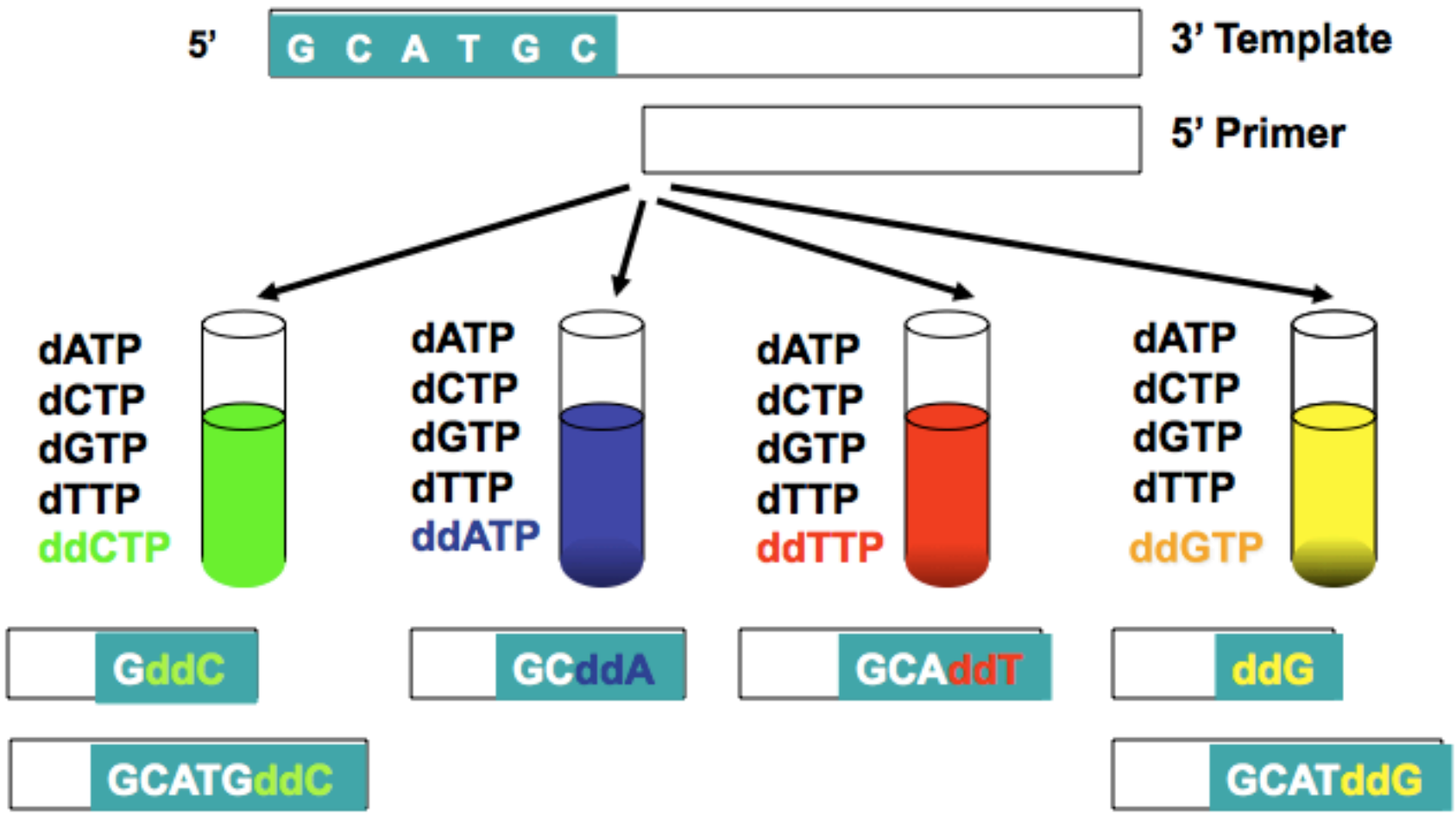
- Reading the DNA sequencer chromatograms 35 billion base calls
- Putting millions of short "reads" together to assemble the genome piecing 35 million reads together
- Identifying the genes from the DNA sequence Finding 1% signal with >95% accuracy
- Figuring out what each gene does 20,000x100,000,000 comparisons

Biting Off Too Much

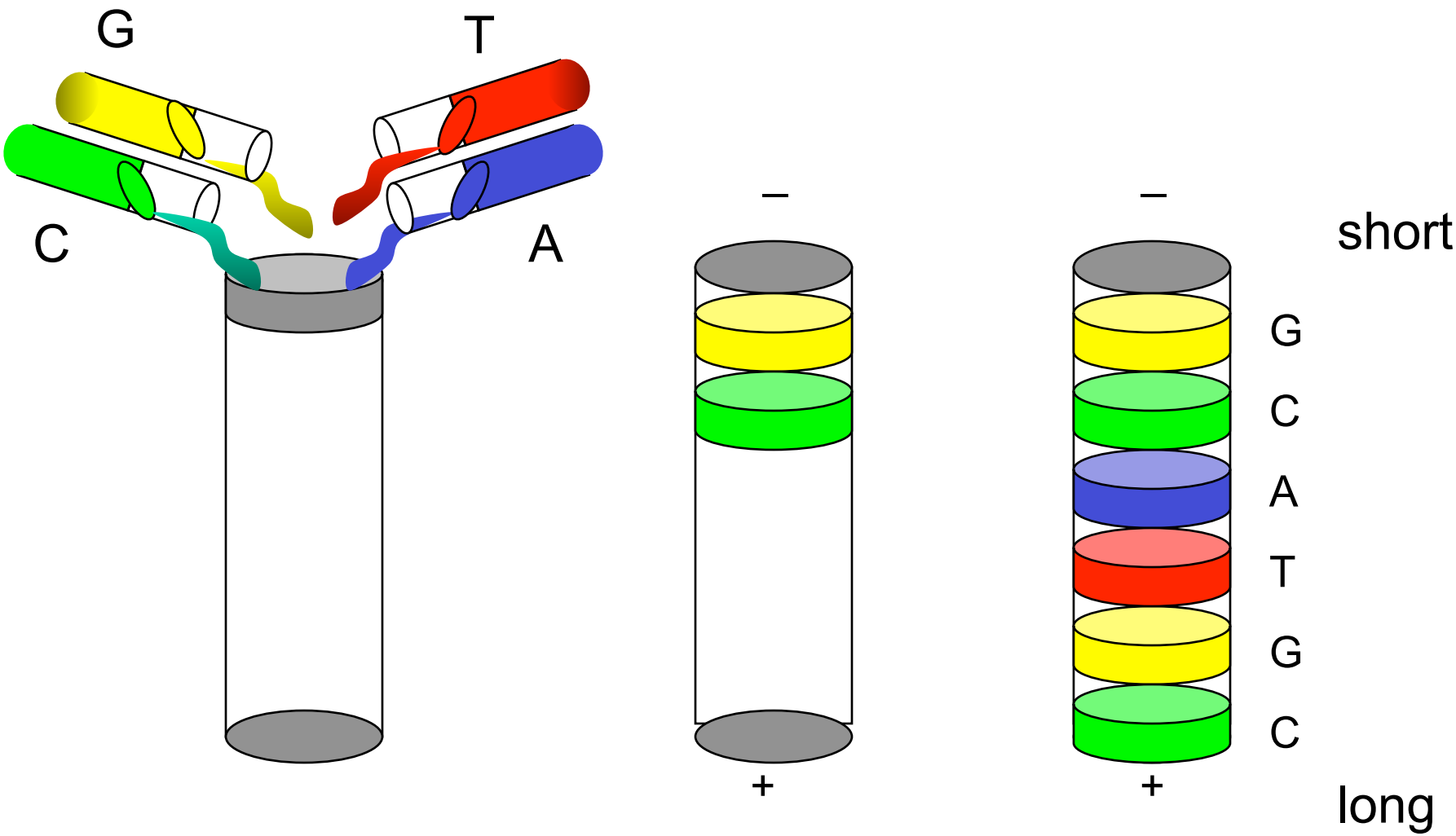
Computational Challenges

- Reading the DNA sequencer chromatograms 35 billion base calls
- Putting millions of short "reads" together to assemble the genome piecing 35 million reads together
- Identifying the genes from the DNA sequence Finding 1% signal with >95% accuracy
- Figuring out what each gene does 20,000x100,000,000 comparisons

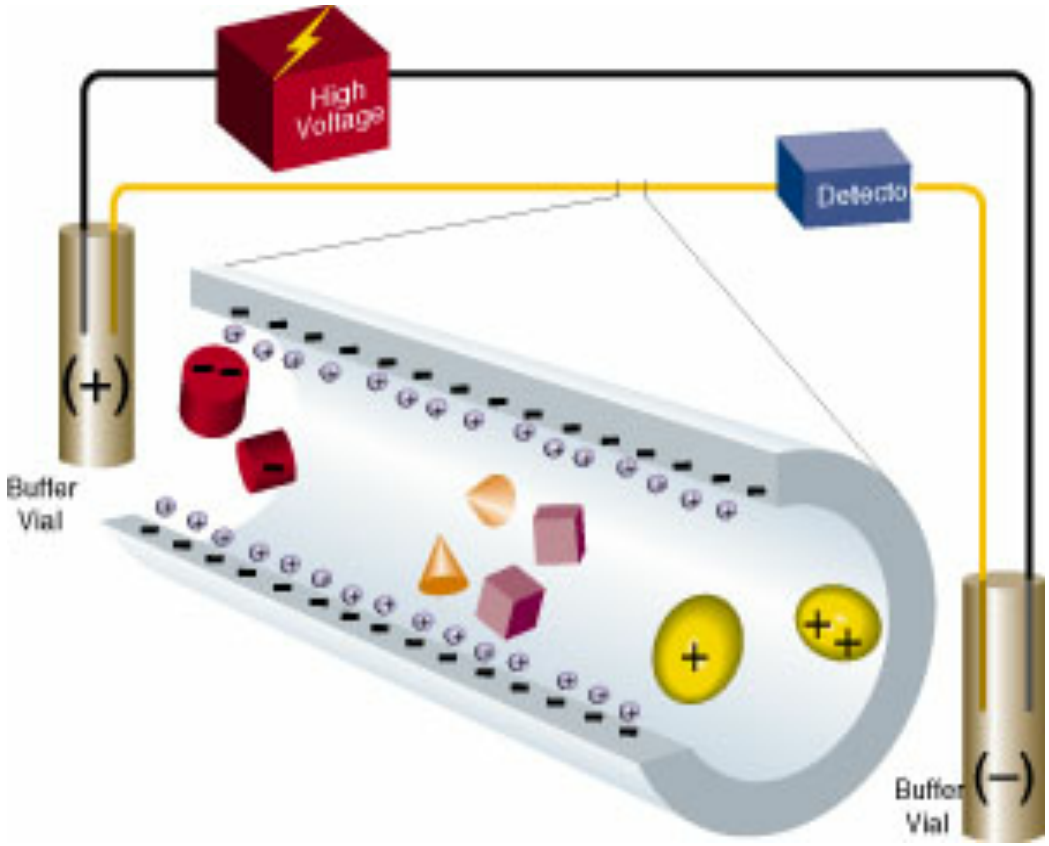
Principles of DNA Sequencing



Principles of DNA Sequencing



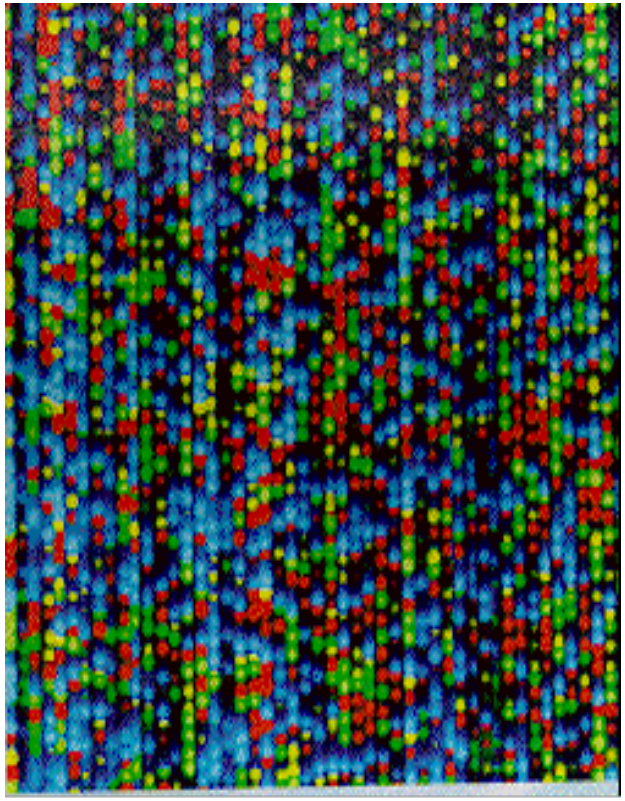
Capillary Electrophoresis



Separation by Electro-osmotic Flow

Multiplexed CE with Fluorescent detection

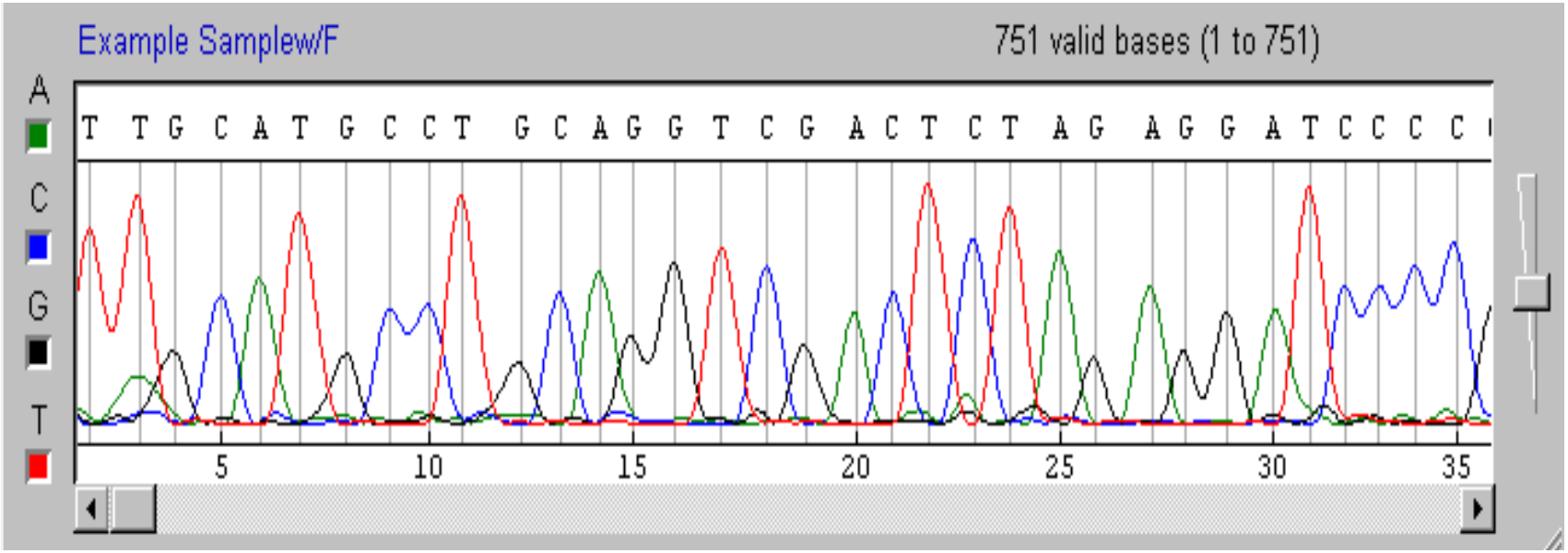
ABI 3700



96x700 bases

Base Calling

- Image processing
- Peak detection
- De-noising
- Peak deconvolution
- Signal analysis
- Reliability assessment

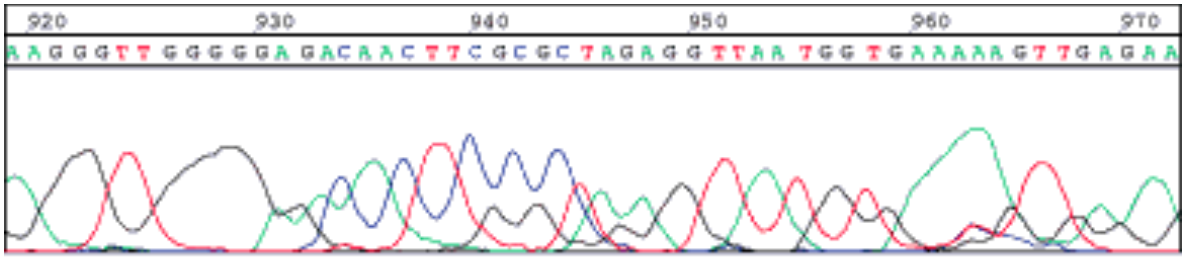


99.99% accurate for 35 billion base calls

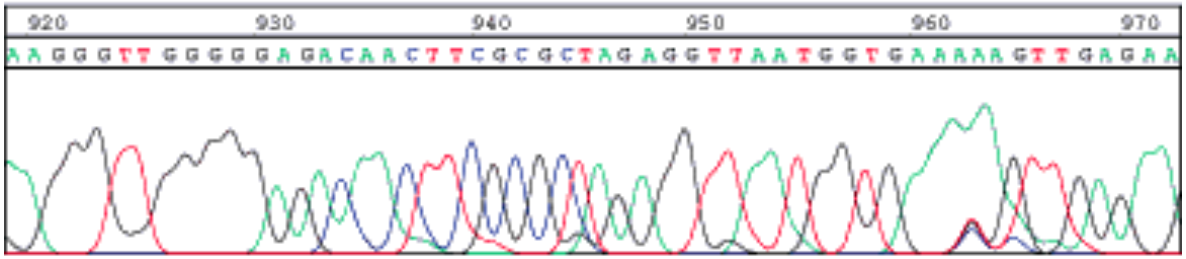
Base Calling With Phred*

READ QUALITY

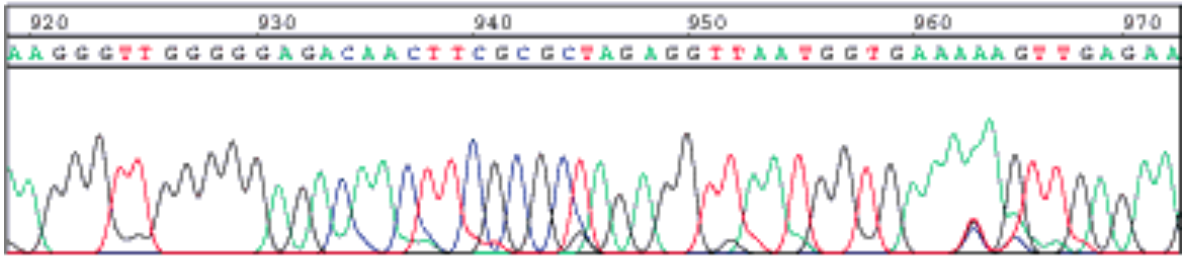
Bad



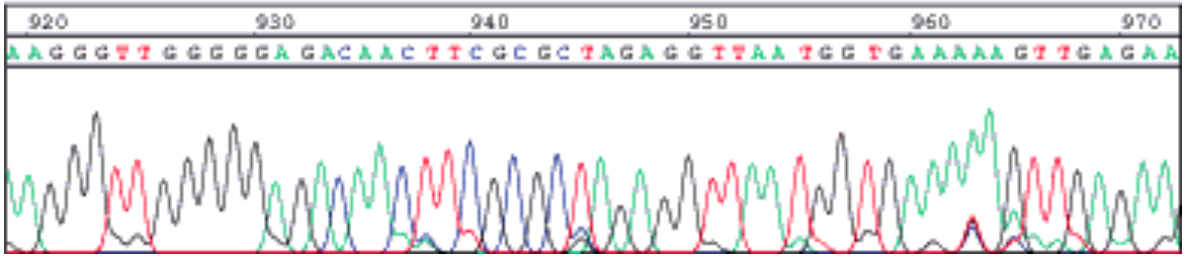
Better



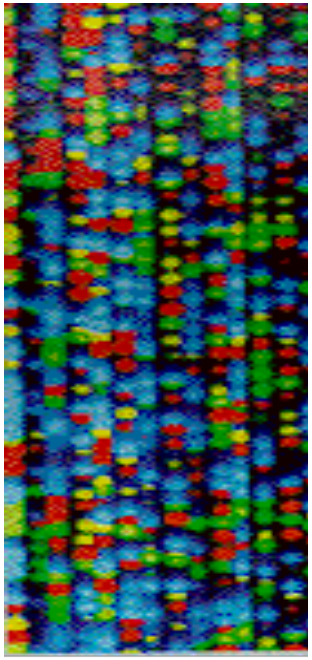
Good



Excellent

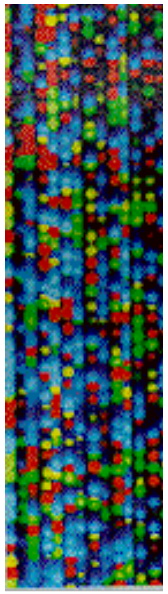


Base Calling - Result



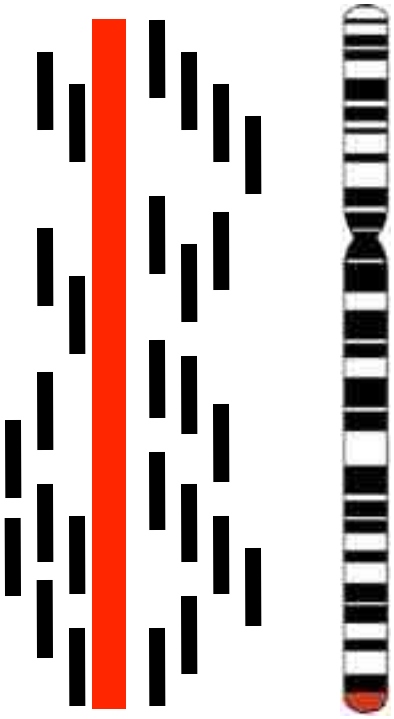
ATGTCACTGCAATTGATGTATAAATGGA
GTTAGACACTAGATCACATAGGAGTTTA
CGCTAAATGACAGATAGACA
GGGATATCTATAGATAGACACATAGCTCTCT
AATGACGACTAGCTGAGTAGATT
TTACGATCGATCGATATTACCGCGCGAAATAT
AGCTATGATGTCGAT
AGACTAGCTAGCTTCTCGGATATTAGA

Shotgun Sequencing



Sequence Chromatogram

Send to Computer



Assembled Sequence

Computational Challenges

- Reading the DNA sequencer chromatograms 35 billion base calls
- Putting millions of short "reads" together to assemble the genome piecing 35 million reads together
- Identifying the genes from the DNA sequence Finding 1% signal with >95% accuracy
- Figuring out what each gene does 20,000x100,000,000 comparisons

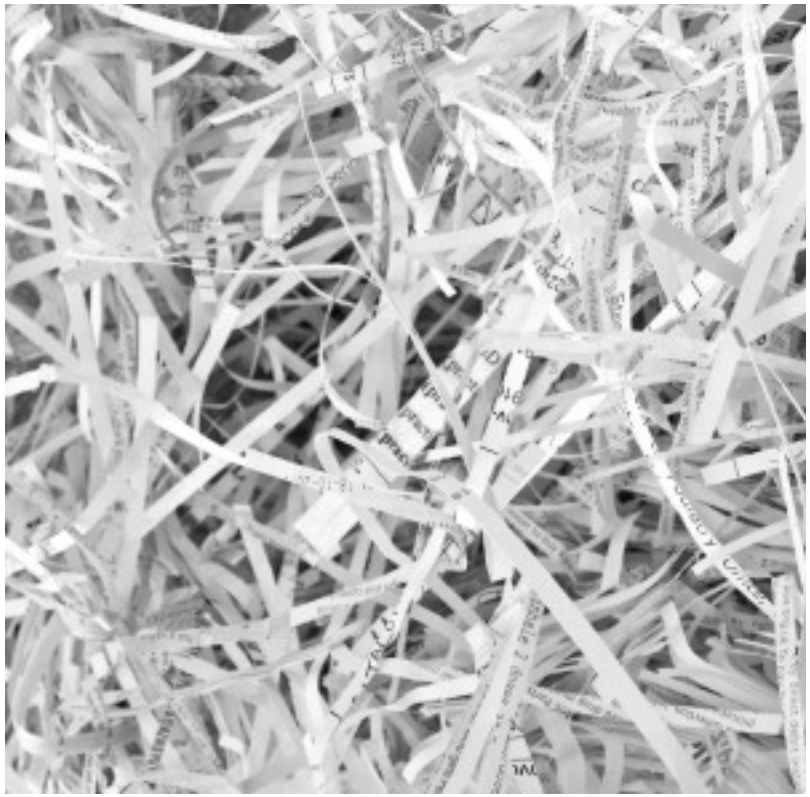
Sequence Assembly

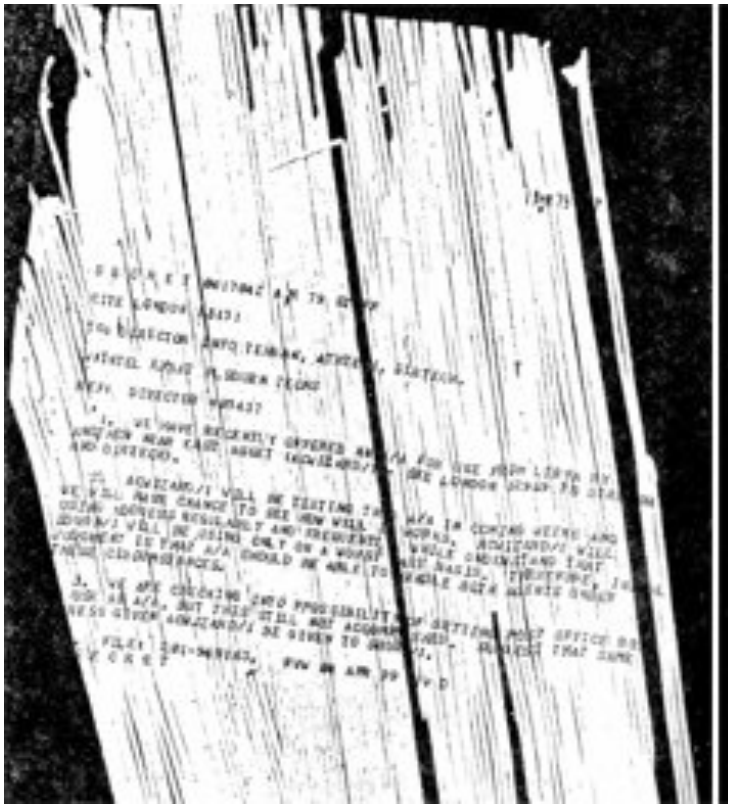
Reads

```
ATGGCATTGCAA
  TGGCATTGCAATTTG
AGATGGTATTG
 GATGGCATTGCAA
      GCATTGCAATTTGAC
  ATGGCATTGCAATTT
AGATGGTATTGCAATTTG
```

Consensus AGATGGCATTGCAATTTGAC

Sequence Assembly (An Analogy)





The DARPA Shredder Challenge

Dynamic Programming

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6

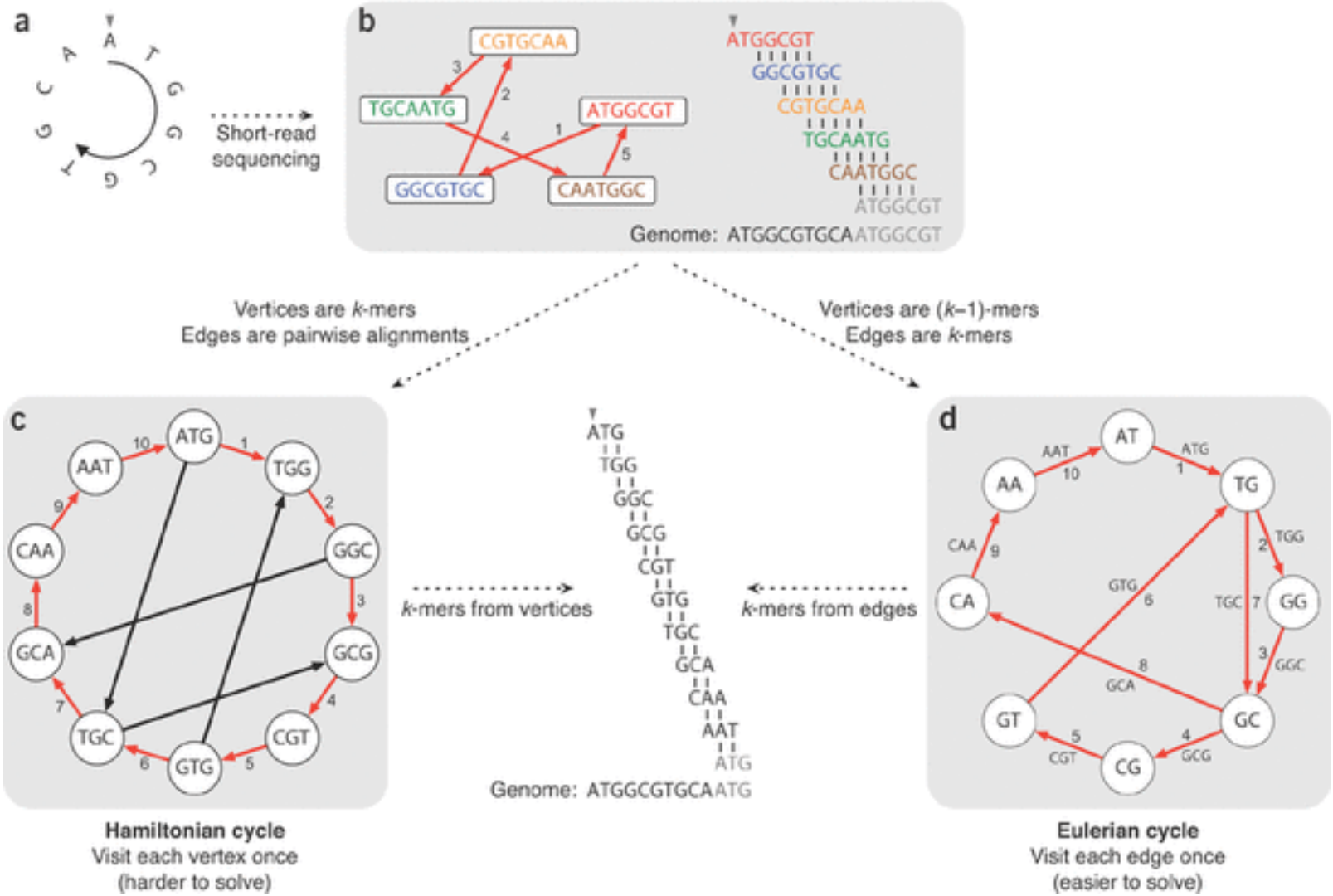
GAATTCAGTTA
GGATCGA

Dynamic Programming

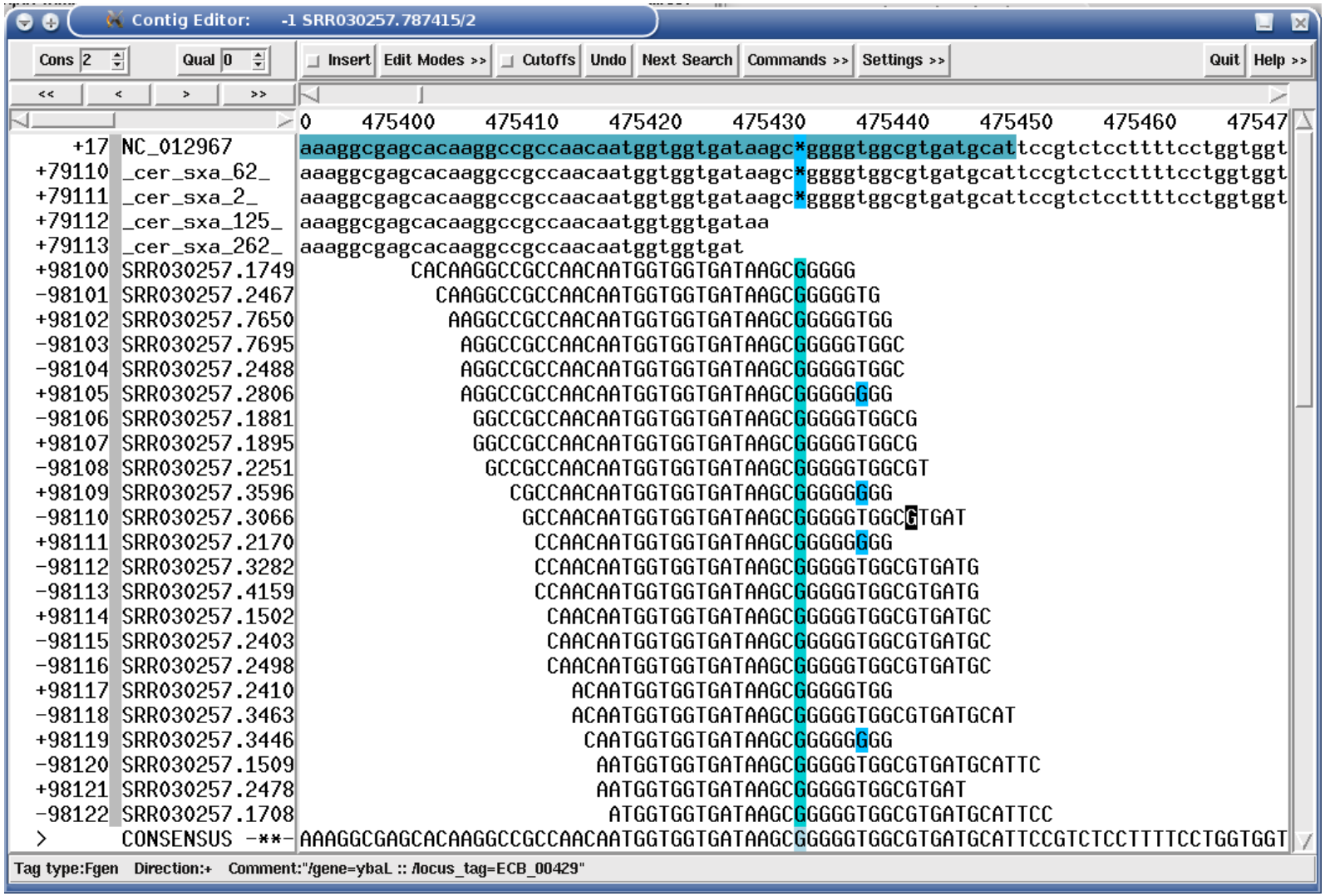
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6

GAATTCAGTTA
GGAT-C-G--A

De Bruijn Graphs & Assembly



A Real Assembler



The Result

>P12345 Human chromosome1

GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TACAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATTAC
AGATTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TTGGC…. **And on for 150,000,000 bases**

How Perl Saved the Human Genome Project *

Lincoln D. Stein

Perl remains the savior of the genome project now more than ever. Just a few weeks ago I found myself sitting in an auditorium listening to Jim Mullikin of the Wellcome Trust Sanger Institute describe how he had solved a problem that was once thought insurmountable: to assemble an entire genome (the mouse, in this case) in a single shot, without the tedious experimental mapping and subcloning that was previously thought to be critical to make the problem soluble. His genome assembly software, named Phusion, is a pipeline of Perl scripts wrapped around a nugget of high-performance C code. As Jim put it, "Perl and 70 gigabytes of main memory is all you need!"

February, 2002

Computational Challenges

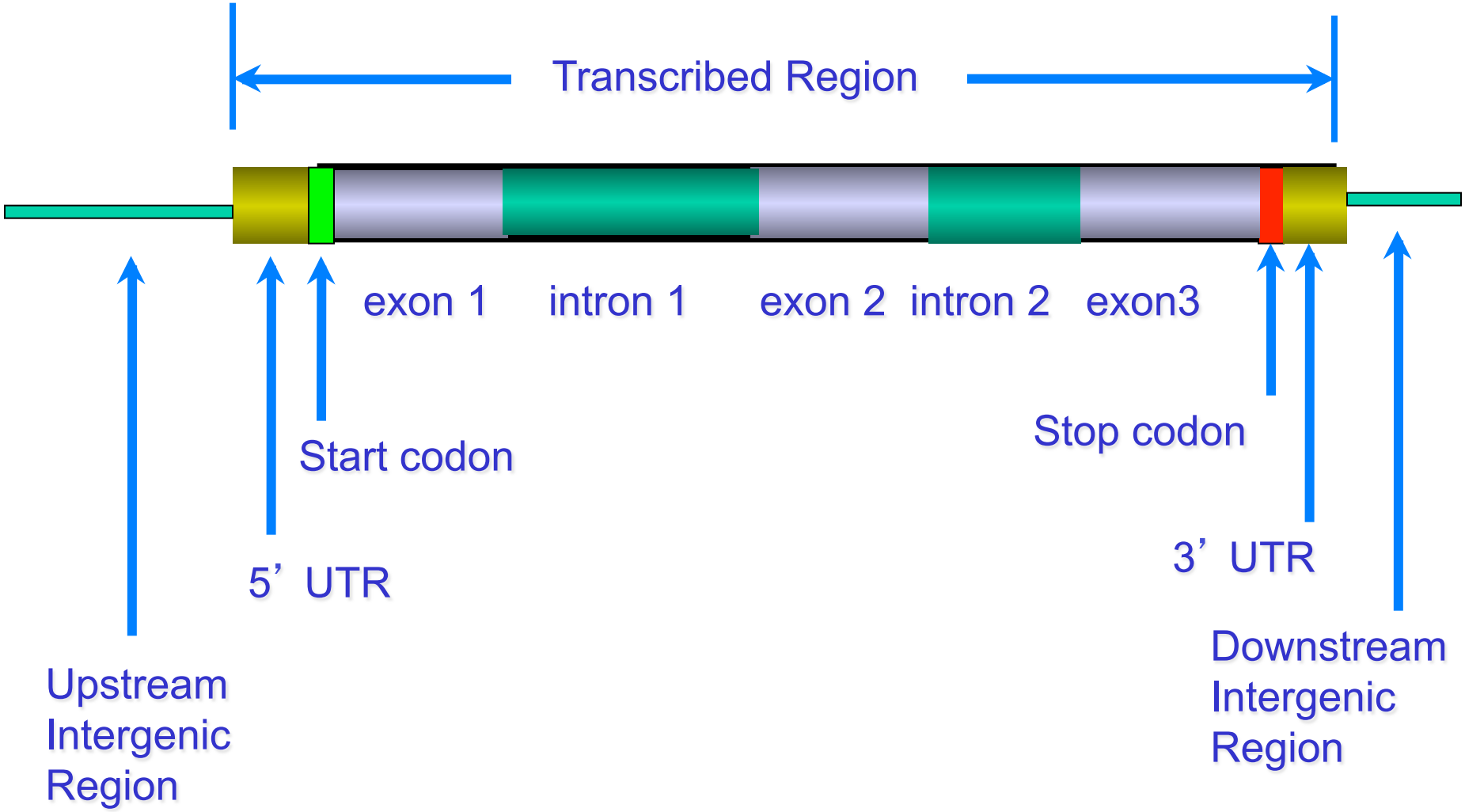
- Reading the DNA sequencer chromatograms 35 billion base calls
- Putting millions of short "reads" together to assemble the genome piecing 35 million reads together
- Identifying the genes from the DNA sequence Finding 1% signal with >95% accuracy
- Figuring out what each gene does 20,000x100,000,000 comparisons

Genome Sequence

>P12345 Human chromosome1

GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TACAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATTAC
AGATTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TTGGC…. **And on for 150,000,000 bases**

Eukaryotic Gene Structure



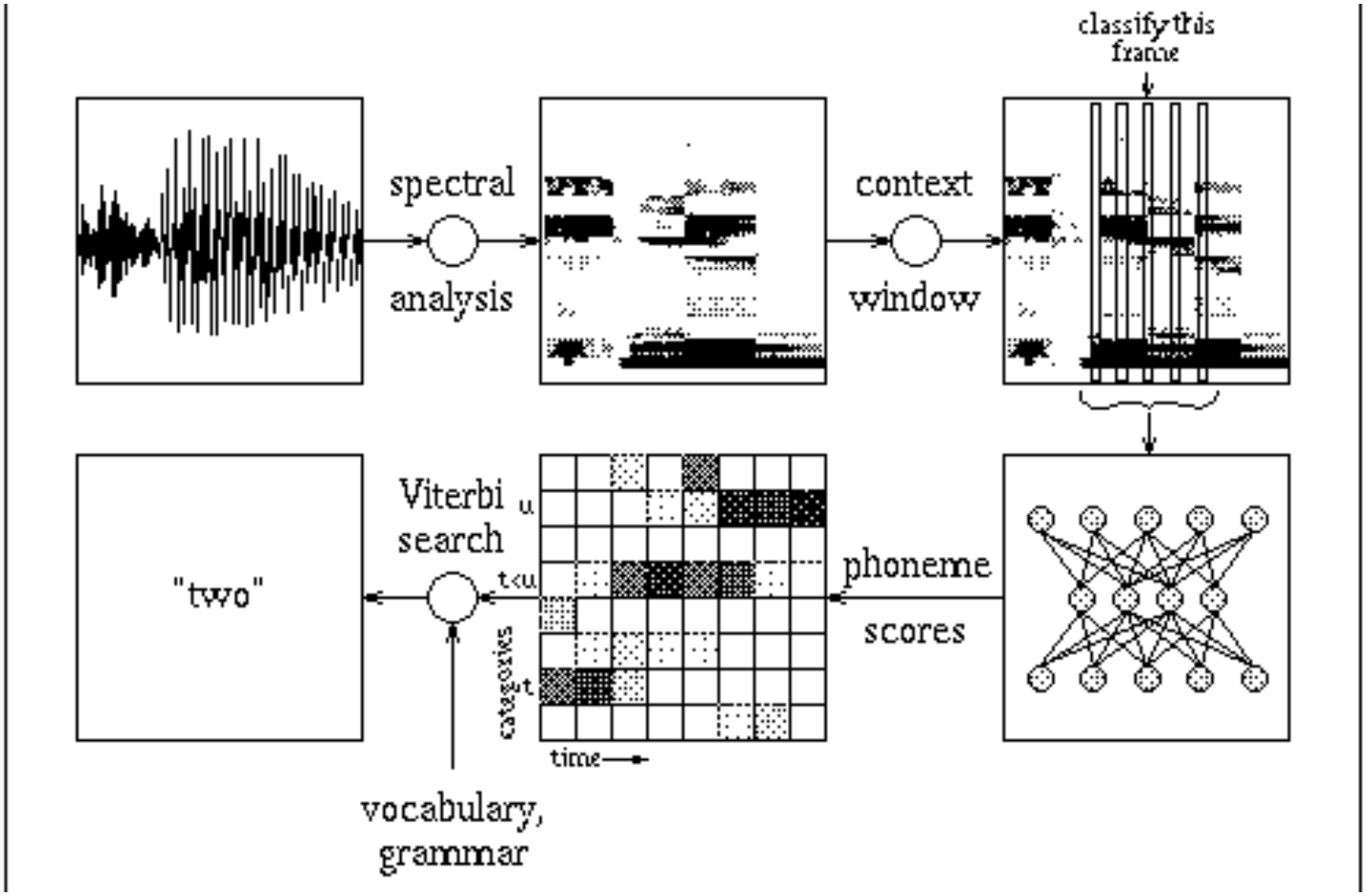
Genome Sequence

GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGATTACAGAT
TACAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATTAC
AGATTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA

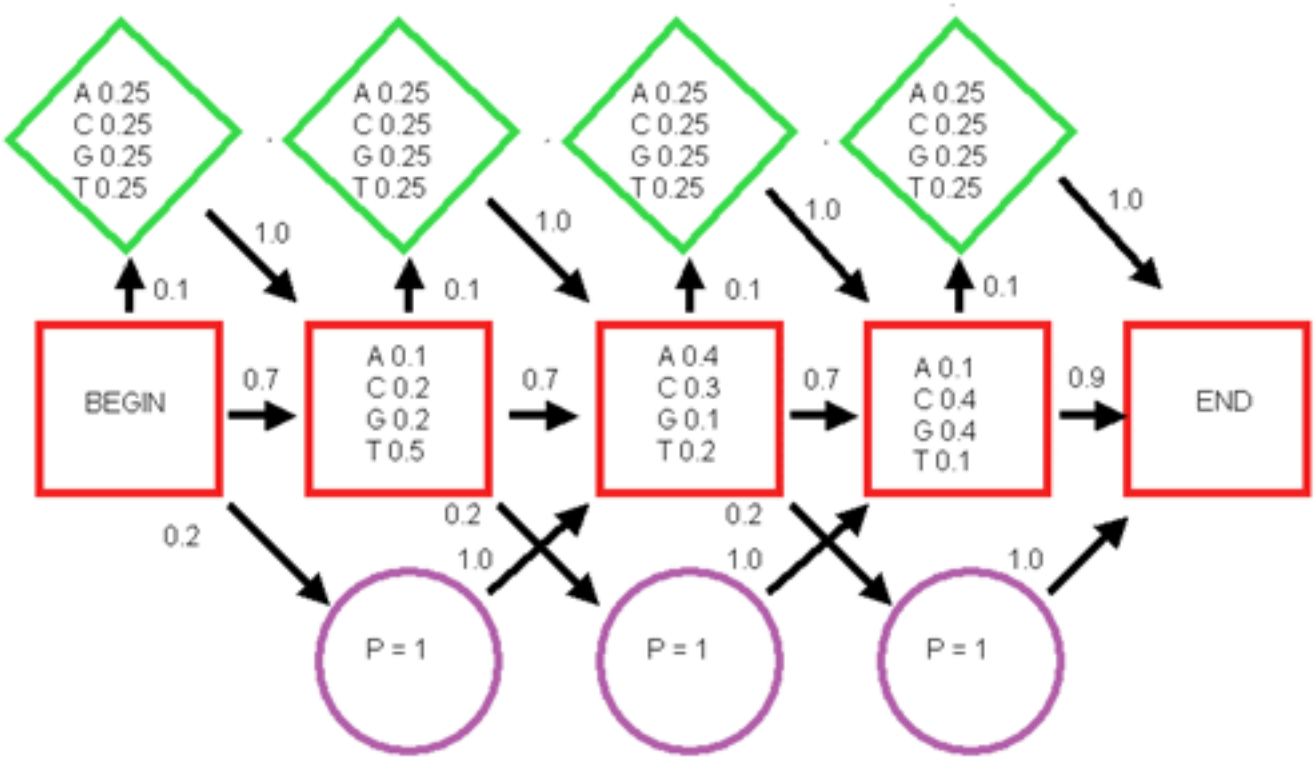
↓

ATTACAGATTACAGATTACAATTAGAGATTACAGAT
TACAGATTACAGATTACAGATTACAGATTACAGATT
ACCAGATTACAGA

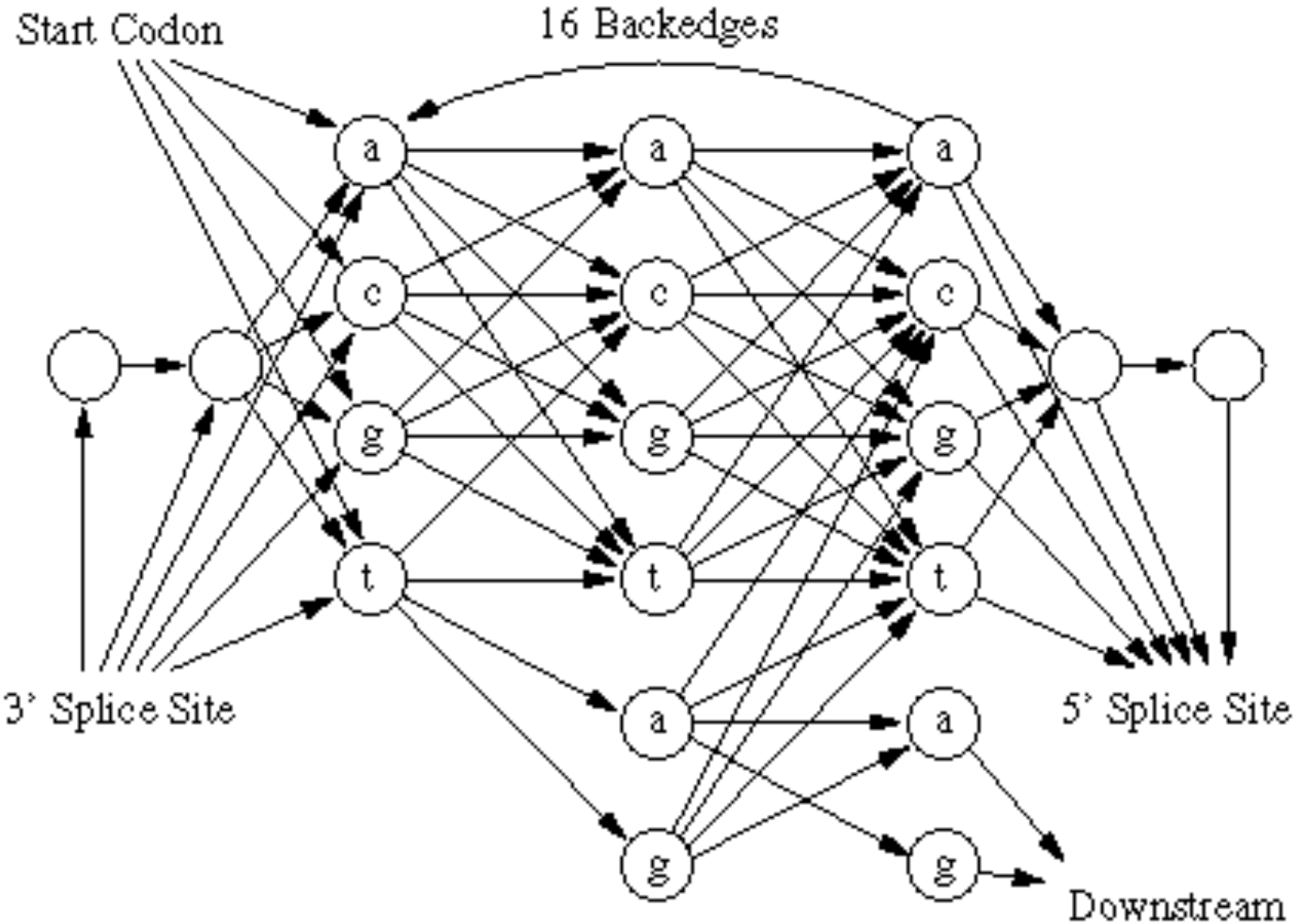
Problem Similar to Speech or Text Recognition



Hidden Markov Models



HMM for Gene Finding



Genscan – The Ultimate Gene Finder

New GENSCAN Web Server at MIT

http://genes.mit.edu/GENSCAN.html

Department o...ell Biology | Login- Depar... of Alberta | Audiobaba Music Search | Bioinformati... the U of A! | Coilgun Basics 2 | Pathguide: t...esource list

For information about Genscan, click here

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored): Choose File no file selected

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

To have the results mailed to you, enter your email address here (optional):

Run GENSCAN Clear Input

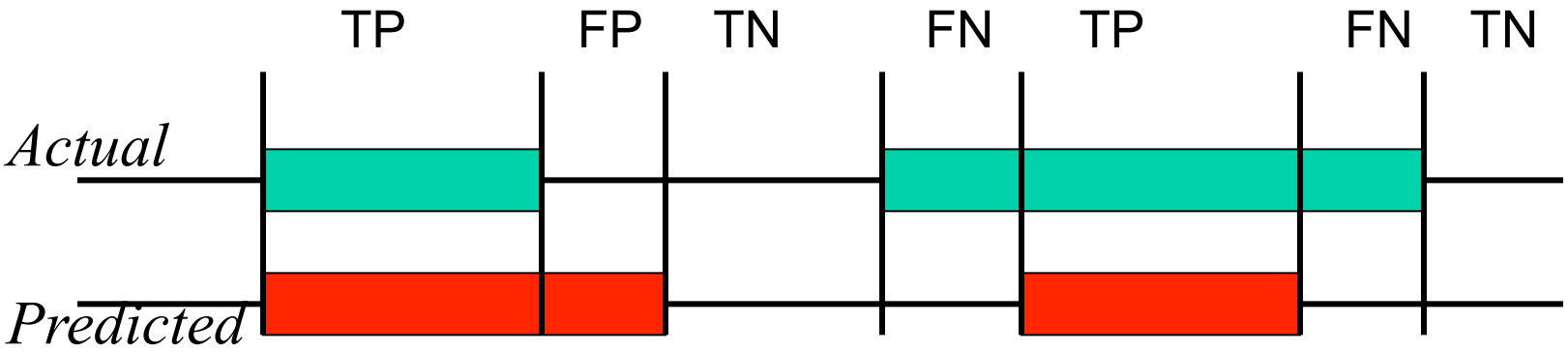
Back to the top

How Well Do They Do?

Programs	*# of seq*	*Nucleotide accuracy*				*Exon accuracy*							
		Sn	*Sp*	*AC*	*CC*	*ESn*	*ESp*	*(ESn+ESp)/2*	*ME*	*WE*	*PCa*	*PCp*	*OL*
FGENES	195(5)	0.86	0.88	0.84	0.83	0.67	0.67	0.69	0.12	0.09	0.20	0.17	0.02
GeneMark	195(0)	0.87	0.89	0.84	0.83	0.53	0.54	0.54	0.13	0.11	0.29	0.27	0.09
Genie	195(15)	0.91	0.90	0.89	0.88	0.71	0.70	0.71	0.19	0.11	0.15	0.15	0.02
Genscan	195(3)	0.95	0.90	0.91	0.91	0.70	0.70	0.71	0.08	0.09	0.21	0.19	0.02
HMMgene	195(5)	0.93	0.93	0.91	0.91	0.76	0.77	0.76	0.12	0.07	0.14	0.14	0.02
Morgan	127(0)	0.75	0.74	0.70	0.69	0.46	0.41	0.43	0.20	0.28	0.28	0.25	0.07
MZEF	119(8)	0.70	0.73	0.68	0.66	0.58	0.59	0.59	0.32	0.23	0.08	0.16	0.01

"Evaluation of gene finding programs" S. Rogic, A. K. Mackworth and B. F. F. Ouellette. Genome Research, 11: 817-832 (2001).

Gene Prediction (Evaluation)



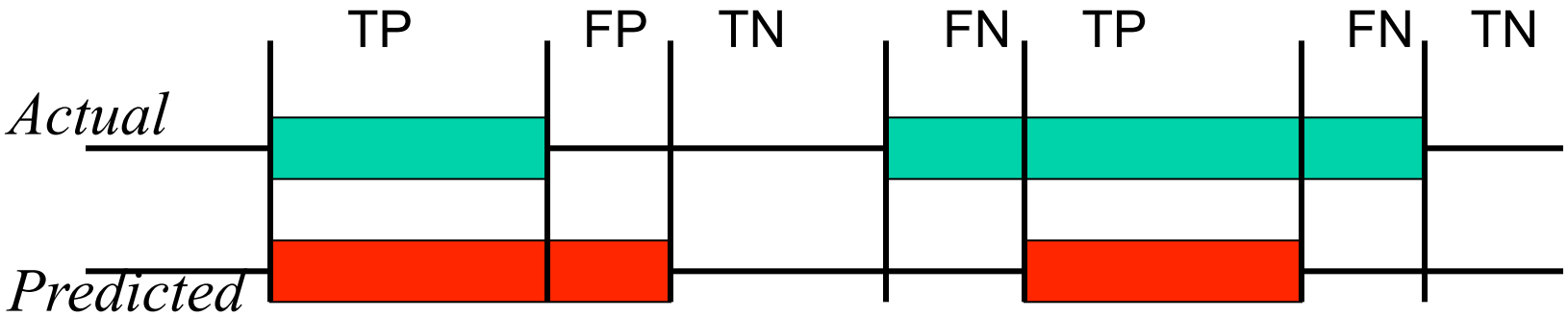
Sensitivity — *Measure of the % of false negative results (sn = 0.996 means 0.4% false negatives)*

Specificity — *Measure of the % of false positive results*

Precision — *Measure of the % positive results*

Correlation — *Combined measure of sensitivity and specificity*

Gene Prediction (Evaluation)



Sensitivity or Recall $Sn=TP/(TP + FN)$

Specificity $Sp=TN/(TN + FP)$

Precision $Pr=TP/(TP + FP)$

Correlation

$$CC=(TP*TN-FP*FN)/[(TP+FP)(TN+FN)(TP+FN)(TN+FP)]^{0.5}$$

This is a better way of evaluating

Computational Challenges

- Reading the DNA sequencer chromatograms 35 billion base calls
- Putting millions of short "reads" together to assemble the genome piecing 35 million reads together
- Identifying the genes from the DNA sequence Finding 1% signal with >95% accuracy
- Figuring out what each gene does 20,000x100,000,000 comparisons

A List of Genes

>P12346 Gene 1
ATGTACAGATTACAGATTACAGATTACAGATTACAG
ATTACAGATTACAGATTACAGATTACAGATTACAGA
TTACAGATTACAGATTACAGATTACAGAT

>P12347 Gene 2
ATGAGATTAGAGATTACAGATTACAGATTACAGATT
ACAGATTACAGATTACAGATTACAGATTACAGATTA
CAGATTACAGATTACAGATTACAGATTACAGATT

>P12348 Gene 3
ATGTTACAGATTACAGATTACAGATTACAGATTACA
GATTACAGATTACAGATTACAGATTACA...

What Biologists Want

>P12346 Gene 1

Human hemoglobin alpha chain, transports oxygen, located on chromosome 14 p.12.1

>P12347 Gene 2

Human super oxide disumutase, removes oxygen radicals and prevents rapid aging, located on chromosome 14 p.12.21

>P12348 Gene 3

Human hemoglobin beta chain, transports oxygen, located on chromosome 14 p.12.23

What Biologists Want

- Trick is to use sequence similarity or sequence matching and prior knowledge
- By 2005 millions of genes had already been characterized from other organisms
- Find the human genes that are similar to the already-characterized genes and assume they are pretty much the same
- Annotation by sequence homology
- Key is to do rapid sequence comparisons

Definitions by Similarity

Query: Bananas

Database

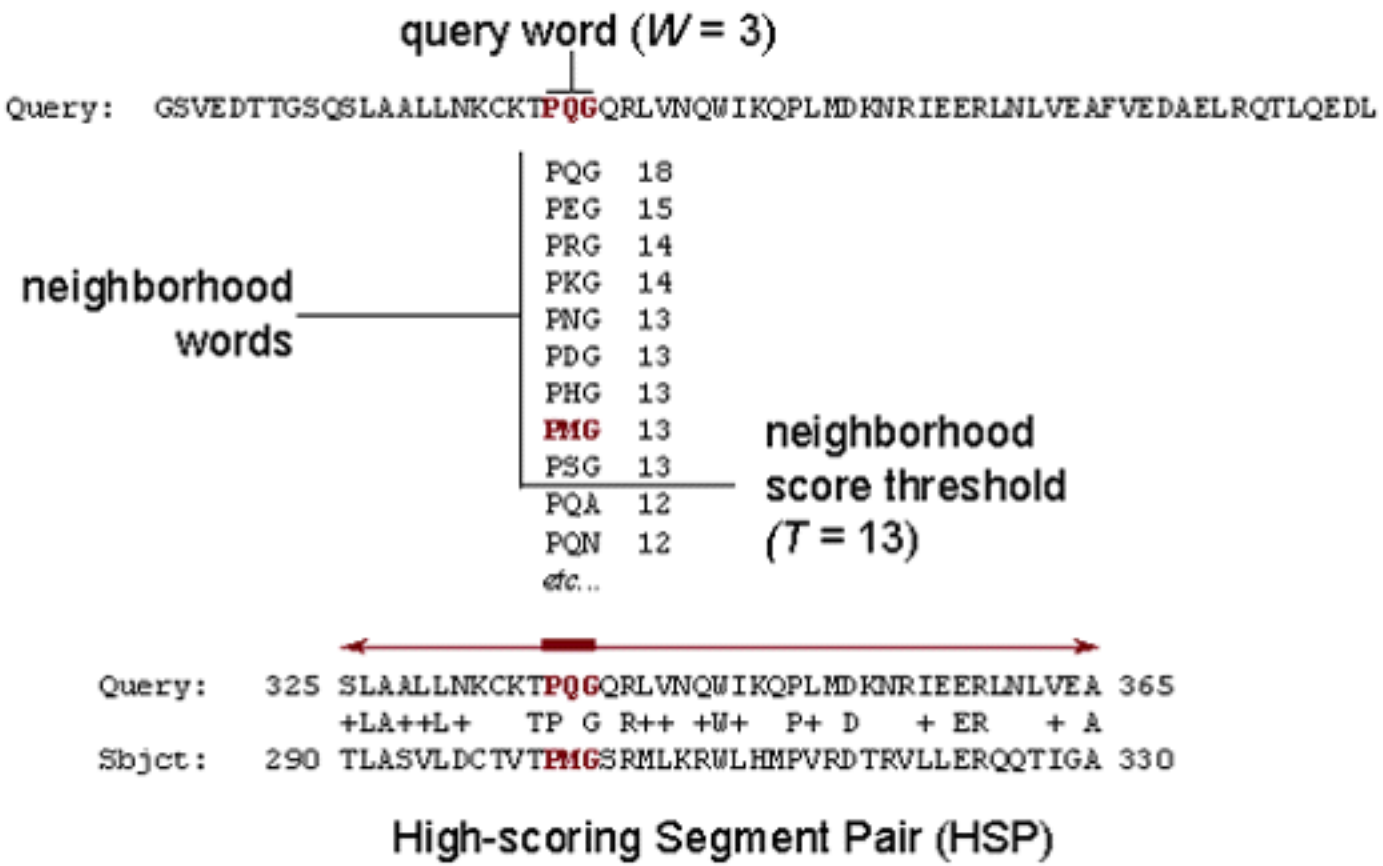
- **Banana – a yellow curved fruit**
- **Bandana – a colorful kerchief**
- **Banal – boring and obvious**
- **Banyan – a fig that starts as an epiphyte**
- **Ananas – genus name for pineapple**

Dynamic Programming – Too Slow

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	2	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6

GAATTCAGTTA
GGATCGA

The BLAST Search Algorithm



1000-10,000X faster than DP methods

The BLAST Server

Nucleotide BLAST: Search nucleotide databases using a nucleotide query

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch

Mass Spectro...uantitation | Biomarkers —...blic Portal | SOLiD Softw...axy > Home | Tutorial:Bas...enTutorials | OpenHelix: L...ee tutorials | Tutorial:Int...penTutorials

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) — Clear

```
>123
AGATTCGATGCTCTACTAATTCTCTTCTCTCGATATAGATATAGAAG
```

Query subrange: From ____ To ____

Or, upload file: Choose File — no file selected

Job Title: ____

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ◉ Others (nr etc.): Nucleotide collection (nr/nt)

Organism (Optional): Enter organism name or id--completions will be suggested ☐ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query (Optional): ____

Enter an Entrez query to limit search

Computational Challenges

- Reading the DNA sequencer chromatograms Solved with Phred
- Putting millions of short "reads" together to assemble the genome Solved with Phusion
- Identifying the genes from the DNA sequence Solved with Genscan
- Figuring out what each gene does Solved with BLAST

Who Were the Real Heroes of The Human Genome Project?

J. Craig Venter
Francis Collins

Bill Clinton
Tony Blair

Questions?

david.wishart@ualberta.ca
3-41 Athabasca Hall