CSCI 1810
Computational Molecular Biology

Of Sea Urchins, Birds and Humans

Overview

1. The Human Genome
2. The Molecular Biology Dogma: DNA, RNA and Protein
3. **Beautiful Algorithms: Rigorous, Practical, Elegant code**
4. Chapter 1: Sequence Alignment Algorithms
5. Chapter 2: Combinatorial Pattern Matching Algorithms
6. Chapter 3: Phylogenetic Trees Algorithms
7. Chapter 4: Machine Learning Methods: Hidden Markov Models Algorithms
8. Chapter 5: Genome Assembly Algorithms (Introduction)
9. Chapter 6: Genomic Privacy (Introduction)
10. The Bioinformatician as a Detective – two puzzles:

    The Adventures of the Dancing Men code, by Sherlock Holmes/Arthur Conan Doyle

    The Prison code, a code used in a prison in California

Beautiful Algorithms

- Rigorous: state-of-the-art, mathematical analysis of their accuracy
- Practical: very efficient, work on large data sets
- Elegant code: "simplicity is the ultimate sophistication"
- von Neumann's "esthetic criteria"

  many applications to different areas

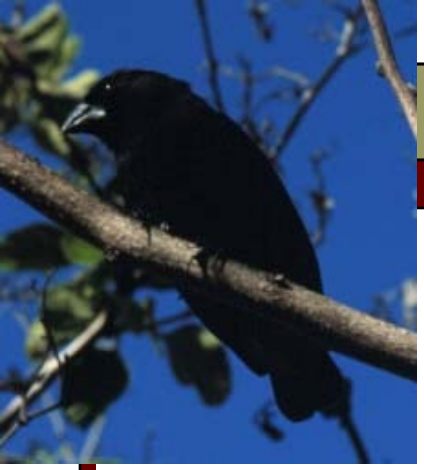
- John von Neumann's **"Beautiful" criteria**
- *"By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes the observed phenomena... Furthermore, it must satisfy certain esthetic criteria – that is, in relation to how much it describes, it must be rather simple ...One cannot tell exactly how "simple" simple is. ...Simplicity is largely a matter of historical background, of previous conditioning, of antecedents, of customary procedures, and it is very much a function of what is explained by it."*
- – John von Neumann

Evolution

Evolution

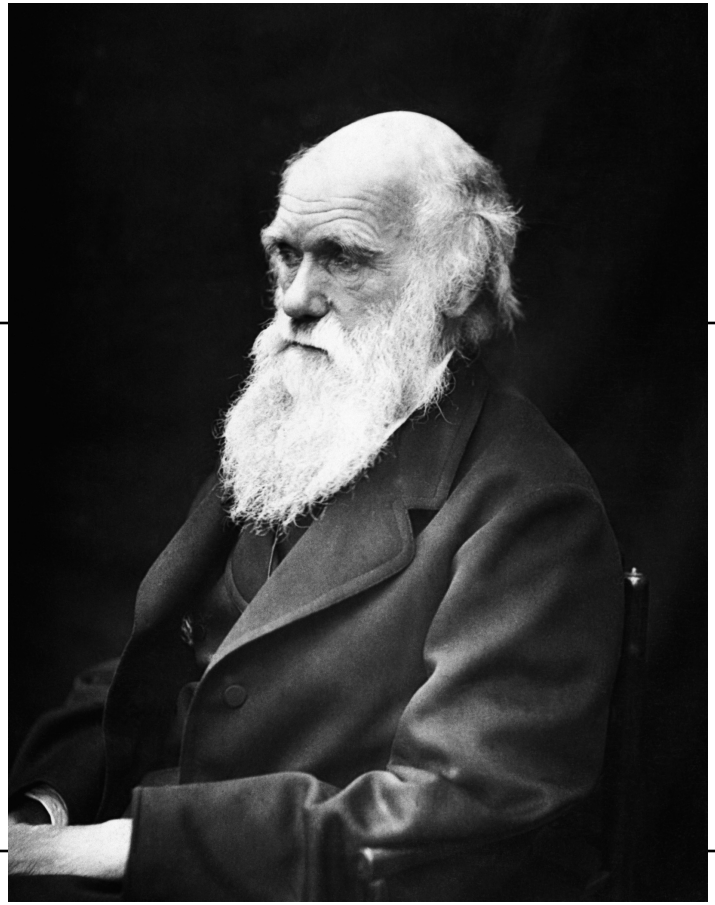
Theodosius Dobzhansky (1900-1975)

Nothing in Biology Makes Sense Except in the Light of Evolution



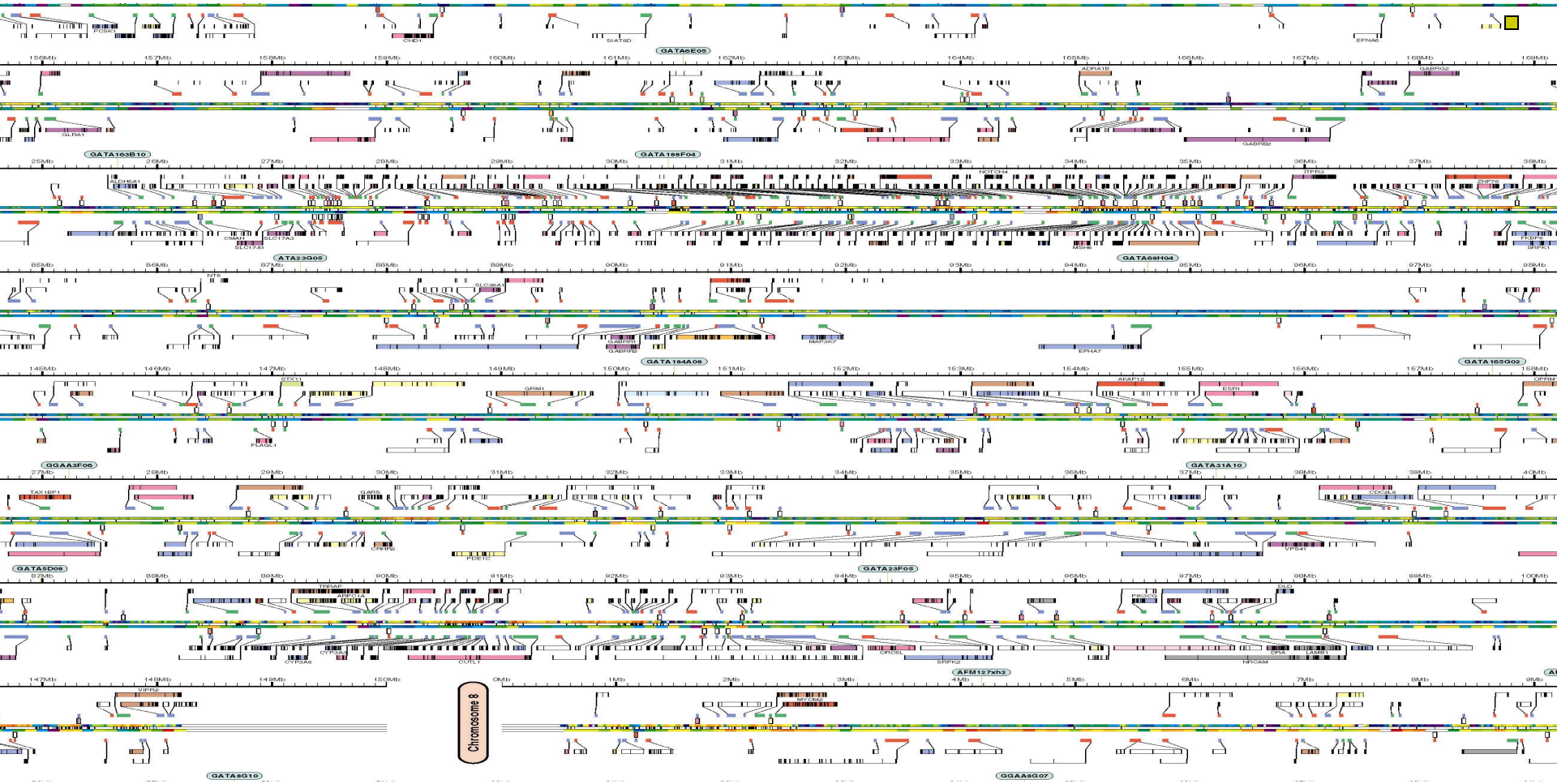
Darwin's Finches

and Coco



The Genome

The Sequence of the Human Genome

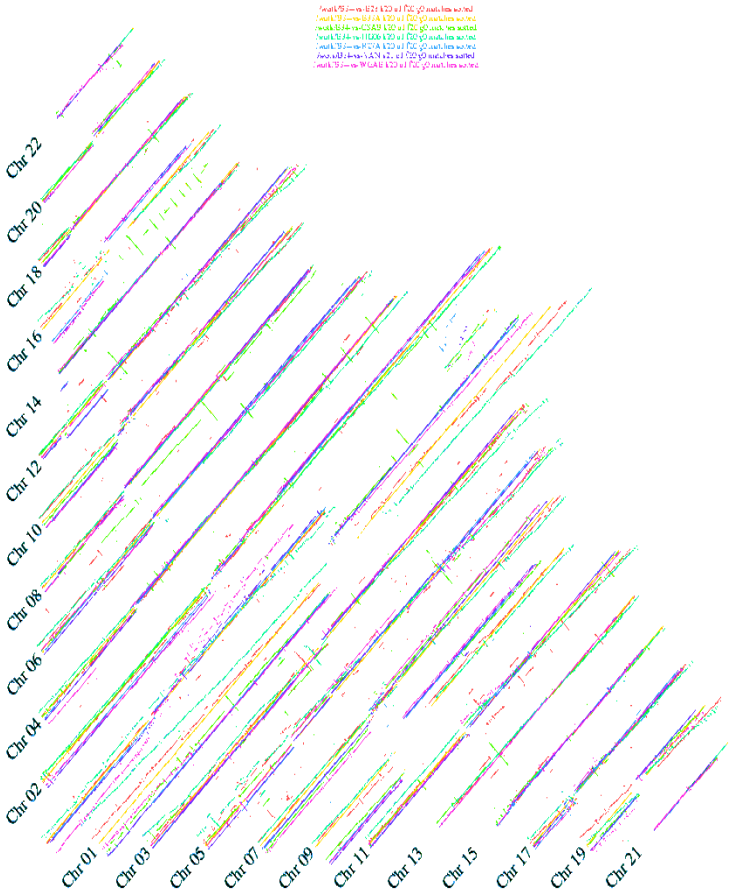


J. Craig Venter, … , Sorin Istrail, …, **"The Sequence of the Human Genome"** *Science*, 2001


16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE
HUMAN
GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

The Genome

J. Craig Venter, … , Sorin Istrail, …,
"The Sequence of the Human Genome" Science, 2001



Sorin Istrail, … , J. Craig Venter,
"Whole-genome shotgun assembly and comparison of human genome assemblies" Proc. Nat. Acad. Sci. USA, 2004

Biomolecular Data

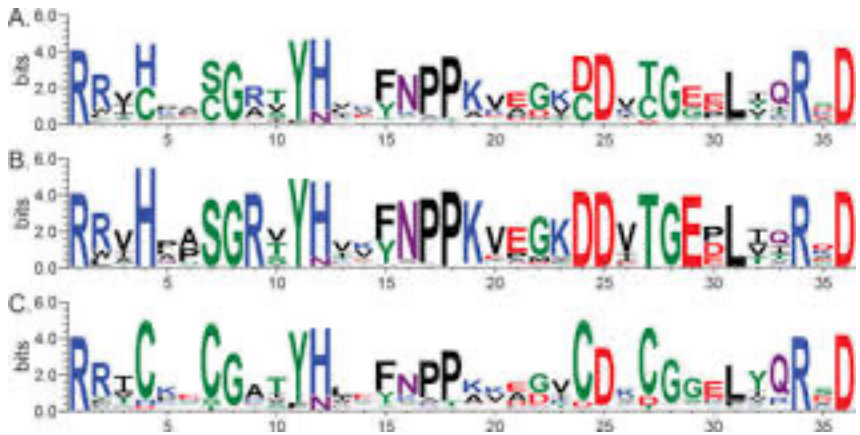
Biomolecular Data

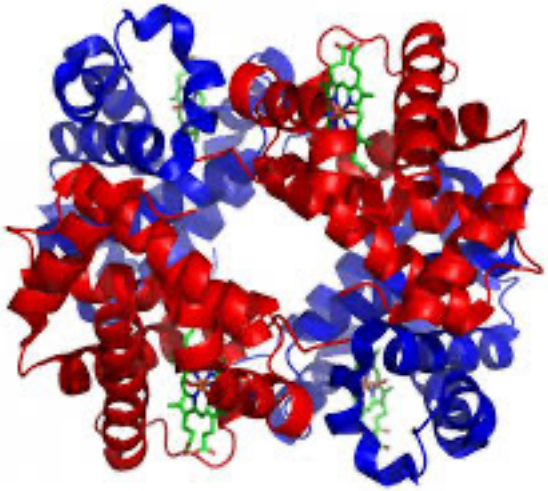
"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay."

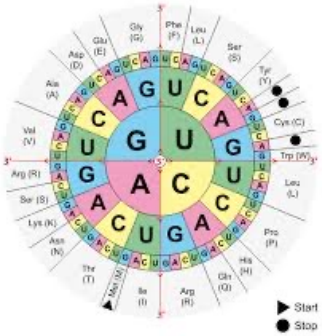
Sherlock Holmes

"The more I see the less I know for sure."

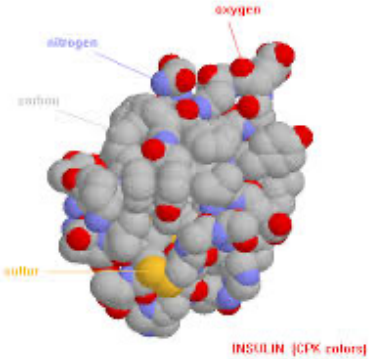
John Lennon

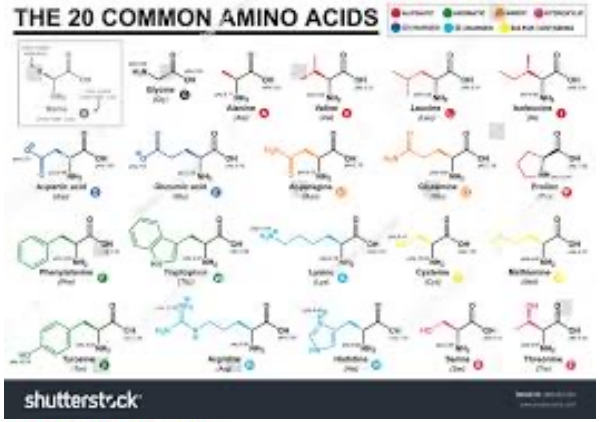


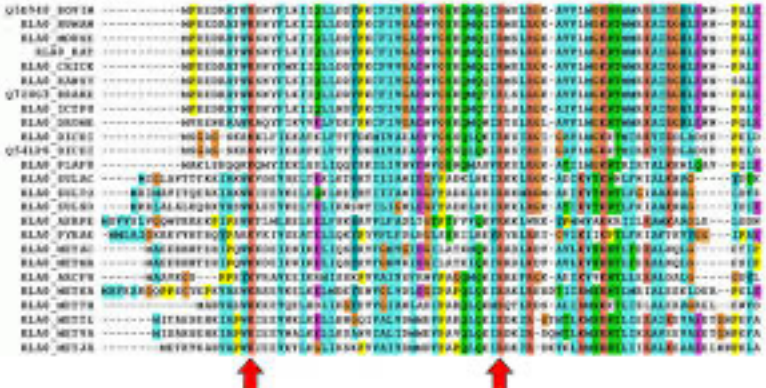


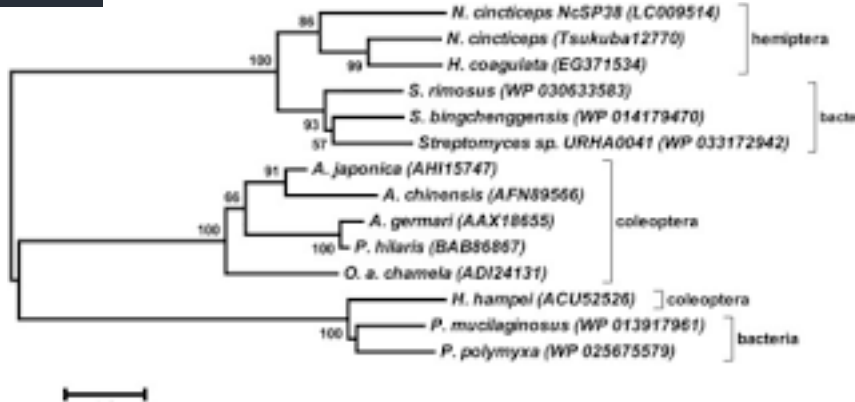


CCTTGATGCTCCTGGCACAAATGAGCAGAATCTCT
GACATGACTTTGGATTTCCCCAGGAGGAGTTTGAT
CAGCCATCTCTGTCCTCCATGAGCTGATCCAGCAG
ATCTTCAACCTCTTTACCACAAAAGATTCATCTGCTGCTTGGGATGAGGACCTCCTA
GACAAATTCTGCACCGAACTCTACCAGCAGCTGAATGACTTGGAAGCCTGTGTGATG
CAGGAGGAGAGGGTGGGAGAAACTCCCCTGATGAATGCGGACTCCATCTTGGCTGTG
AAGAAATACTTCCGAAGAATCACTCTCTATCTGACAGAGAAGAAATACAGCCCTTGT
GCCTGGGAGGTTGTCAGAGCAGAAATCATGAGATCCTCTCTTTATCAACAAACTTGC
AAGAAAGATTAAGGAGGAAGGAATAA, TGTGATCTCCCTGAGACCCACAGCCTGGA
TAACAGGAGGACCTTGATGCTCCTGGCACAAATGAGCAGAATCTCTCCTTCCTCCTG
TCTGATGGACAGACATGACTTTGGATTTCCCCAGGAGGAGTTTGATGGCAACCAGTT
CCAGAAGGCTCCAGCCATCTCTGTCCTCCATGAGCTGATCCAGCAGATCTTCAACCT

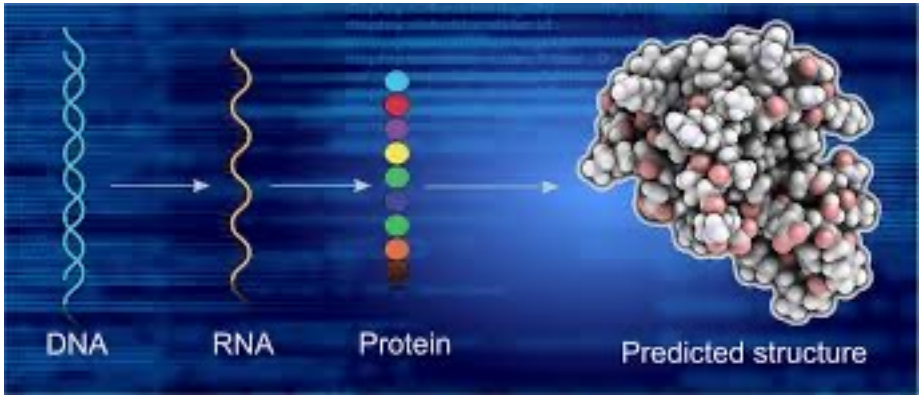




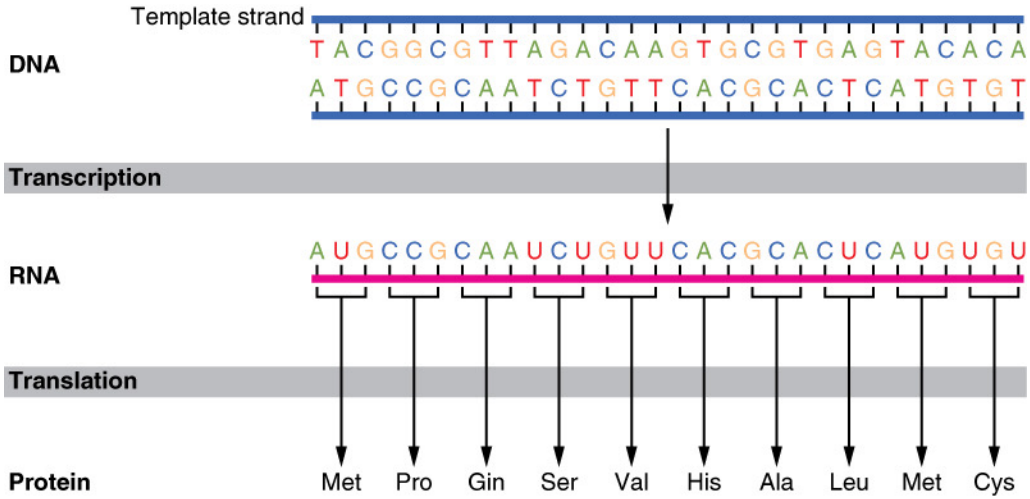




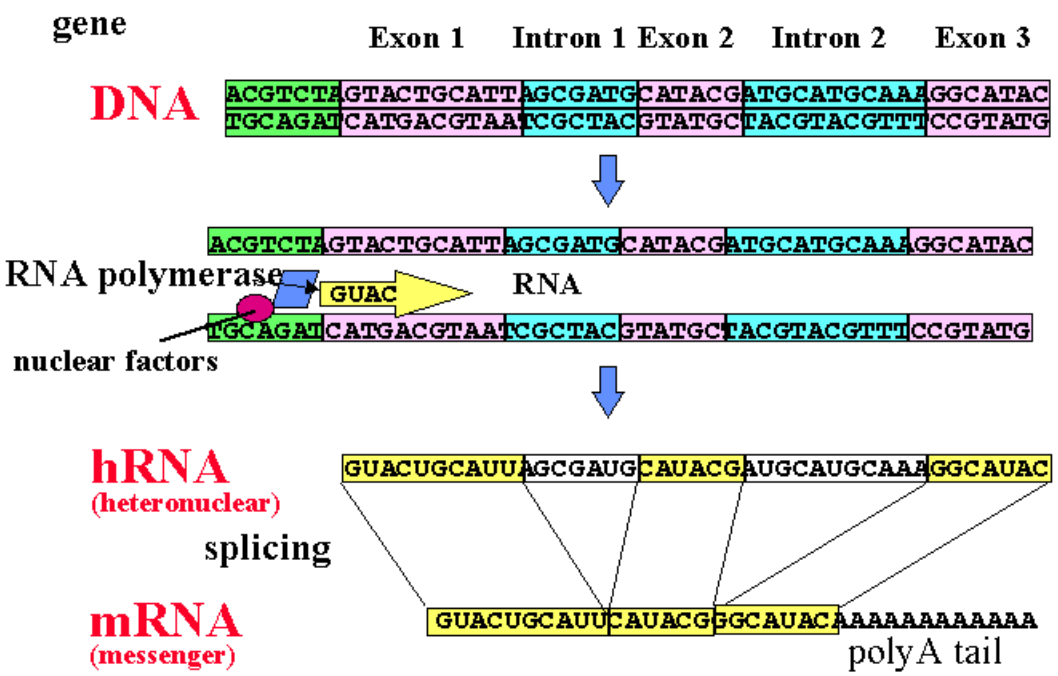
The Dogma of Molecular Biology

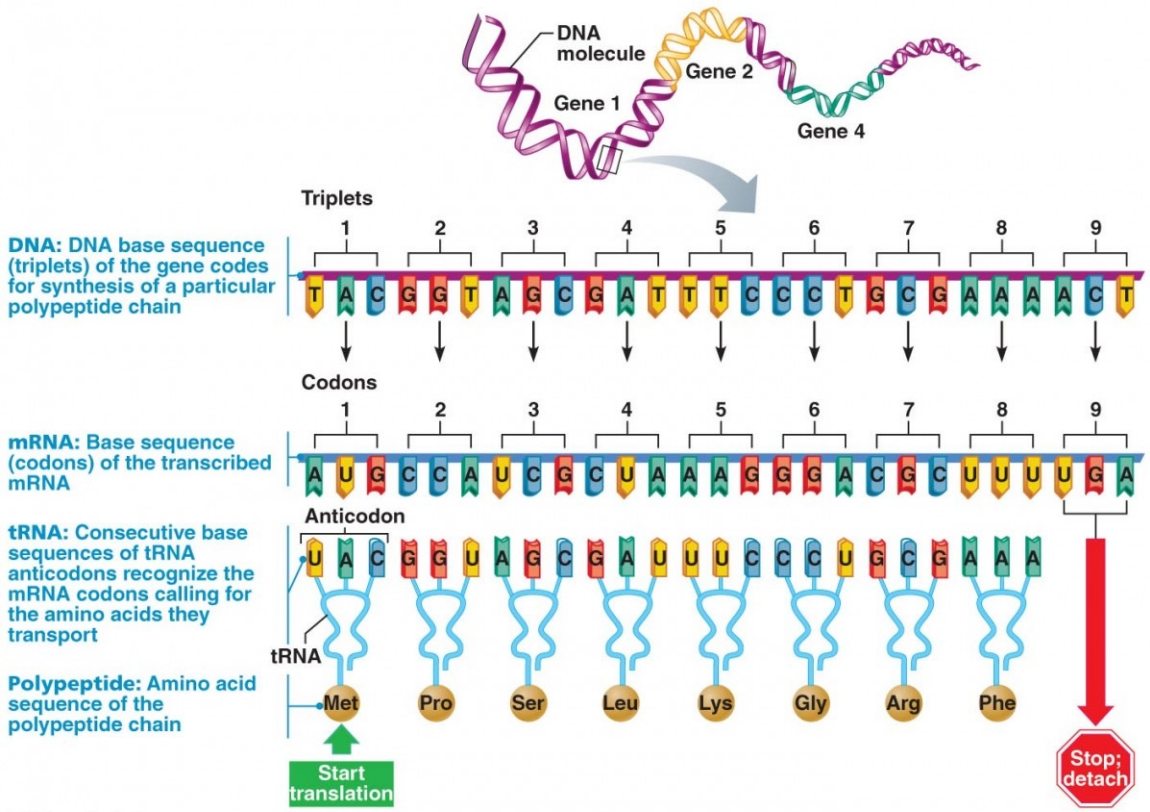


The Molecular Biology Dogma



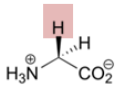
Transcription

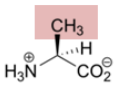




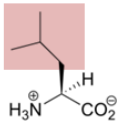
The set of 20 amino acids

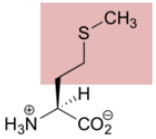
Nonpolar, aliphatic side groups

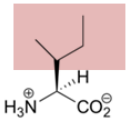

Glycine
Gly, G


Alanine
Ala, A

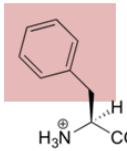

Valine
Val, V

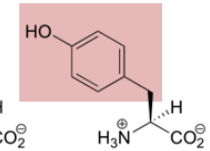

Leucine
Leu, L

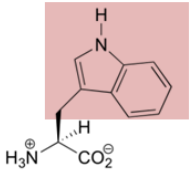

Methionine
Met, M


Isoleucine
Ile, I

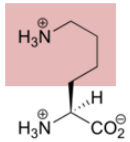
Aromatic side groups

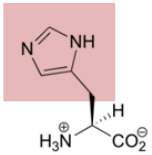

Phenylalanine
Phe, F

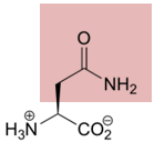

Tyrosine
Tyr, Y


Tryptophan
Trp, W

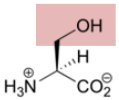
Positively charged side groups

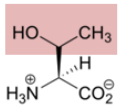

Lysine
Lys, K

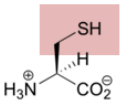

Histidine
His, H

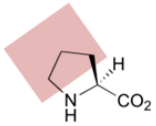

Asparagine
Asn, N

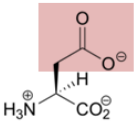
Polar, uncharged side groups

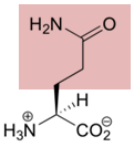

Serine
Ser, S


Threonine
Thr, T

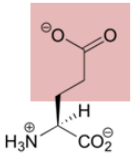

Cysteine
Cys, C

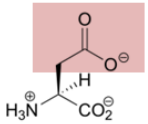

Proline
Pro, P


Aspartate
Asp, D

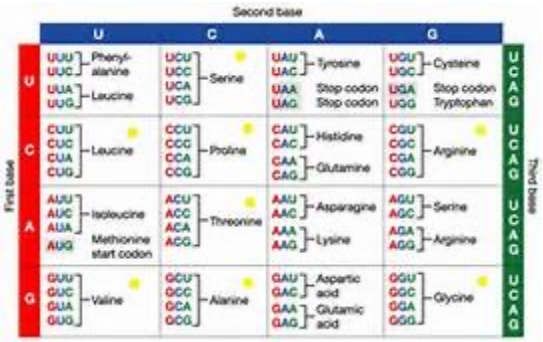

Glutamine
Gln, Q

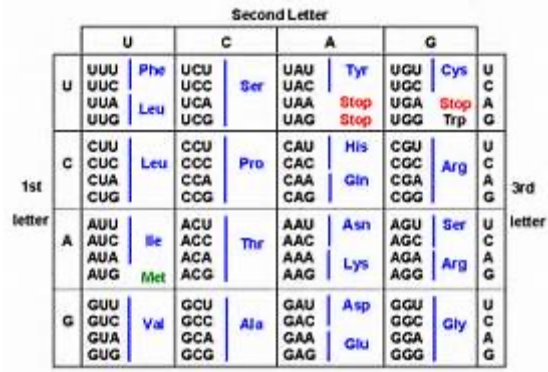
Negatively charged side groups

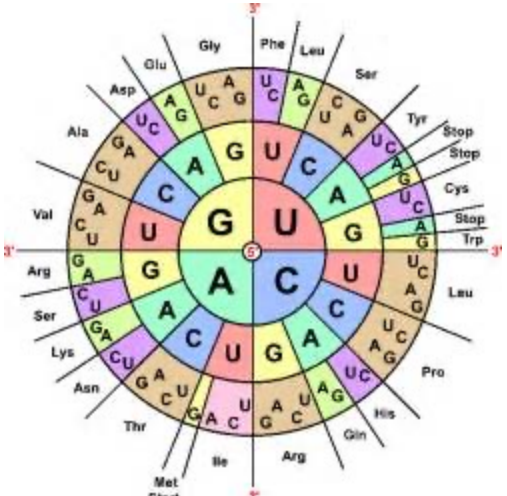

Glutamate
Glu, E

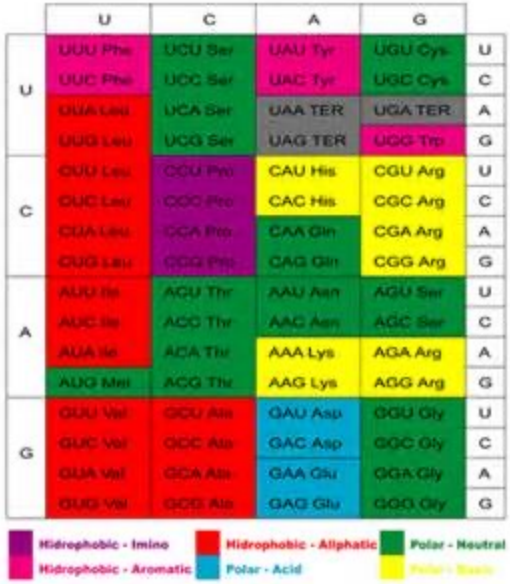

Aspartate
Asp, D

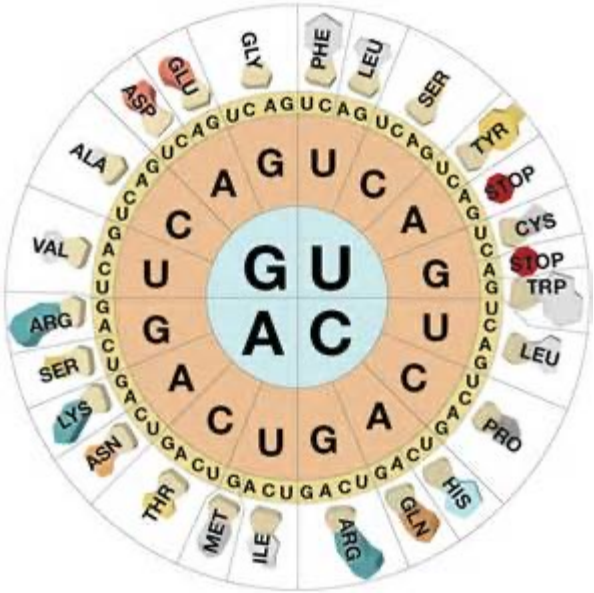
The Genetic Code











Genetic Variation

SNPs & HAPLOTYPES

Single Nucleotide Polymorphism (SNP)

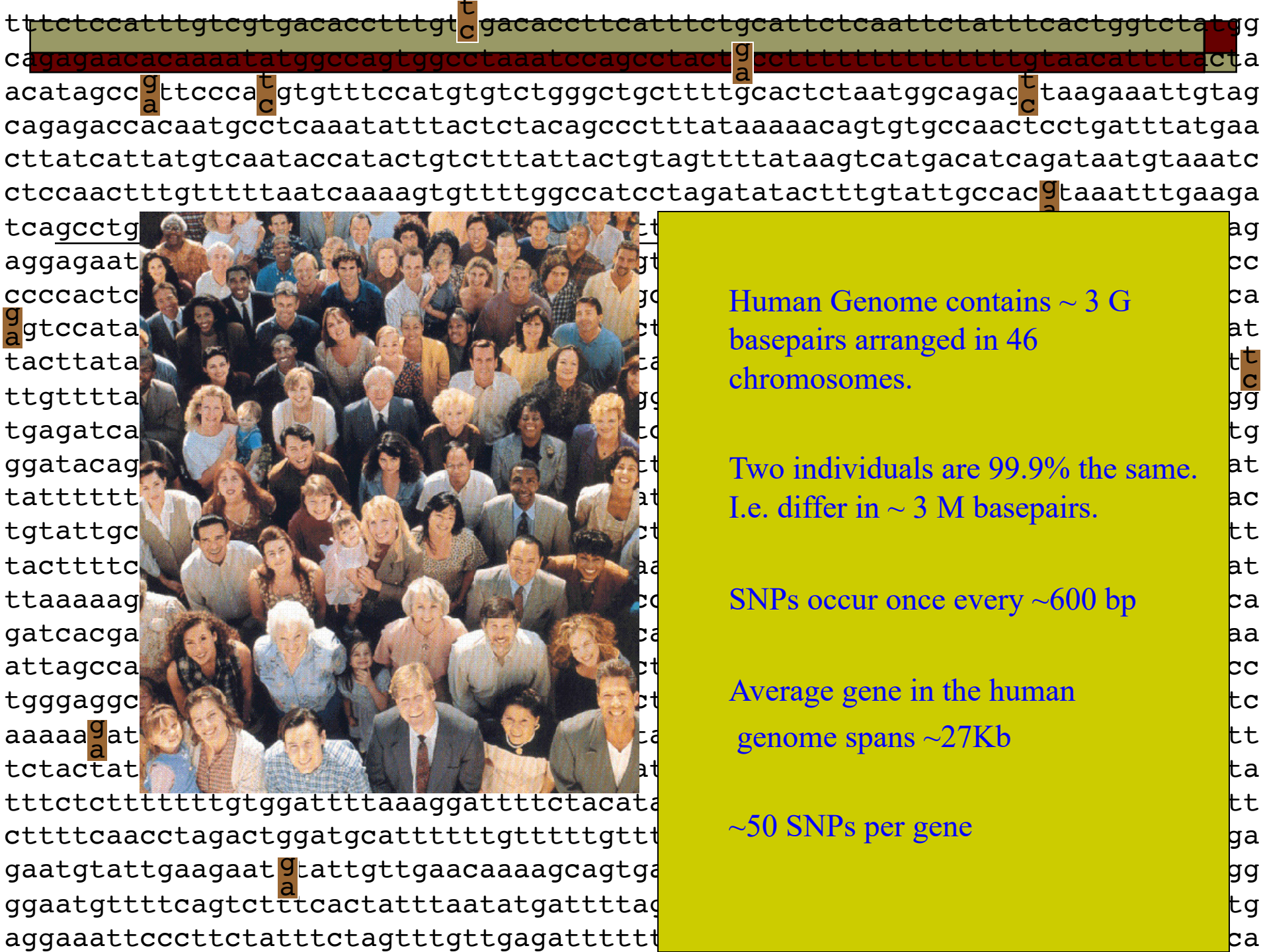
GATTTAGATCGCGATAGAG

GATTTAGATCTCGATAGAG

A SNP is a position in a genome at which two or more different bases occur in the population, each with a frequency >1%.

The two alleles at the site are **G** and **T**

- The most abundant type of polymorphism



Human Genome contains ~ 3 G basepairs arranged in 46 chromosomes.

Two individuals are 99.9% the same. I.e. differ in ~ 3 M basepairs.

SNPs occur once every ~600 bp

Average gene in the human genome spans ~27Kb

~50 SNPs per gene

Haplotype

C A G
T T G

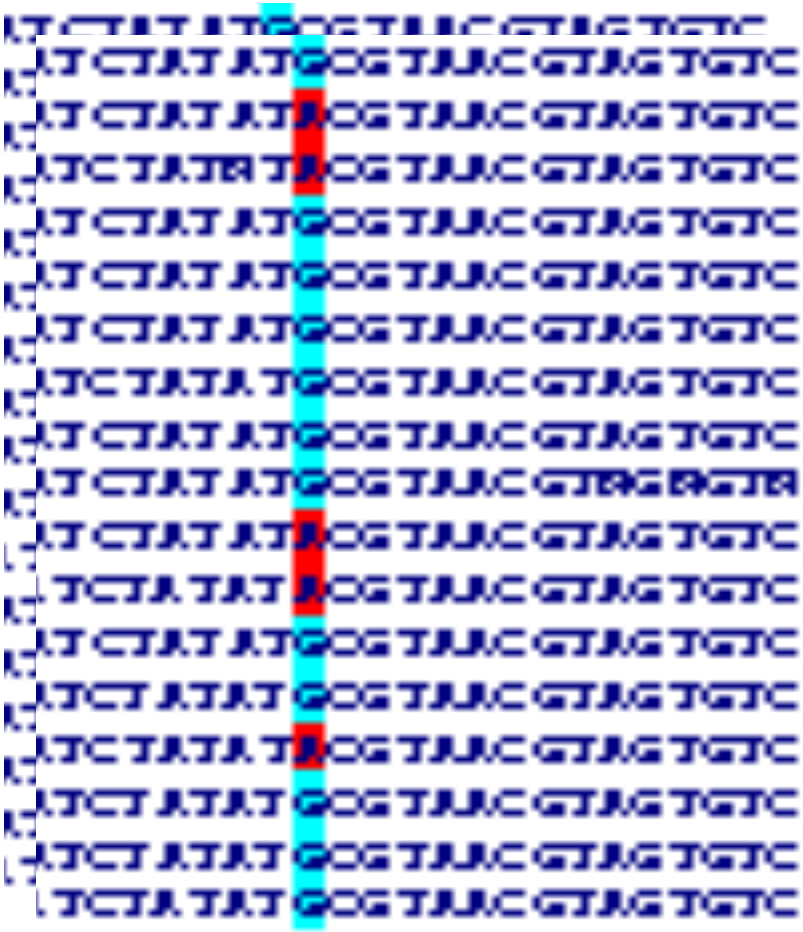
Haplotypes

G	C	T	C	G	A	C	A	A	C	A	G
G	T	T	C	G	T	C	A	A	C	A	G

SNP SNP SNP

Two individuals

Mutations



Infinite Sites Assumption:

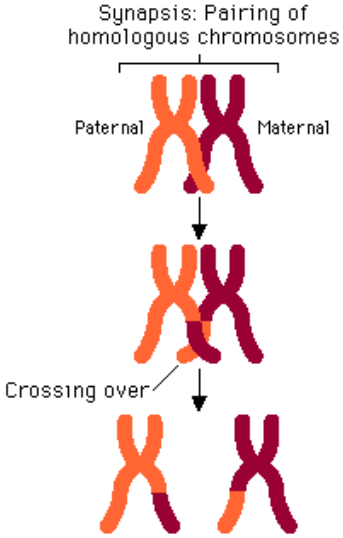
Each site mutates at most once

Haplotype Pattern

C	A	G	T	0	0	0	0
T	T	G	A	1	1	0	1
C	A	T	G	0	0	1	0
C	T	G	T	0	1	0	1

At each SNP site label the two alleles as 0 and 1. The choice which allele is 0 and which one is 1 is arbitrary.

Recombination

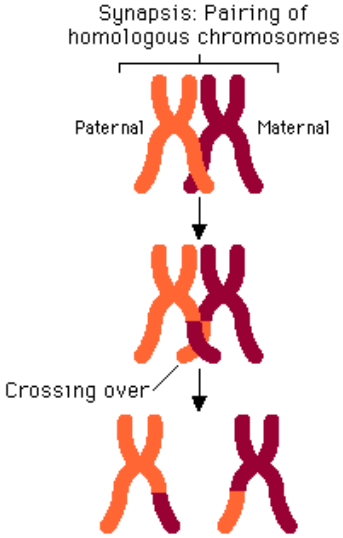


G T T C G A C A A C A T

A C G T A T C T A T T A

G T T C G A C T A T T A

Recombination



The two alleles are linked, I.e., they are "**traveling together**"

G T T C G A C A A C A T

A C G T A T C T A T T A

Recombination disrupts the linkage

?

G T T C G A C T A T T A

CACAGCCTGGATAACAGGAGGACCTTGATGCTCCTGGCACAAATGAGCAGAATCTCT
CCTTCCTCCTGTCTGATGGACAGACATGACTTTGGATTTCCCCAGGAGGAGTTTGAT
GGCAACCAGTTCCAGAAGGCTCCAGCCATCTCTGTCCTCCATGAGCTGATCCAGCAG
ATCTTCAACCTCTTTACCACAAAAGATTCATCTGCTGCTTGGGATGAGGACCTCCTA
GACAAATTCTGCACCGAACTCTACCAGCAGCTGAATGACTTGGAAGCCTGTGTGATG
CAGGAGGAGAGGGTGGGAGAAACTCCCCTGATGAATGCGGACTCCATCTTGGCTGTG
AAGAAATACTTCCGAAGAATCACTCTCTATCTGACAGAGAAGAAATACAGCCCTTGT
GCCTGGGAGGTTGTCAGAGCAGAAATCATGAGATCCTCTCTTTATCAACAAACTTGC
AAGAAAGATTAAGGAGGAAGGAATAA, TGTGATCTCCCTGAGACCCACAGCCTGGA
TAACAGGAGGACCTTGATGCTCCTGGCACAAATGAGCAGAATCTCTCCTTCCTCCTG
TCTGATGGACAGACATGACTTTGGATTTCCCCAGGAGGAGTTTGATGGCAACCAGTT
CCAGAAGGCTCCAGCCATCTCTGTCCTCCATGAGCTGATCCAGCAGATCTTCAACCT

What is the meaning of this DNA sequence?

A code to break!

Can you break this code?

Chapter 1: Sequence Alignment

Image, courtesy of Vincent van Gogh Museum

Avrilla Xinyue Qian

Chapter 1: Sequence Alignment Algorithms

Pairwise Sequence Alignment

Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

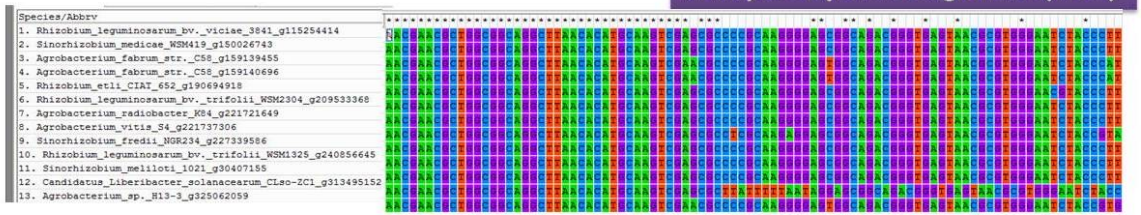
Global Alignment

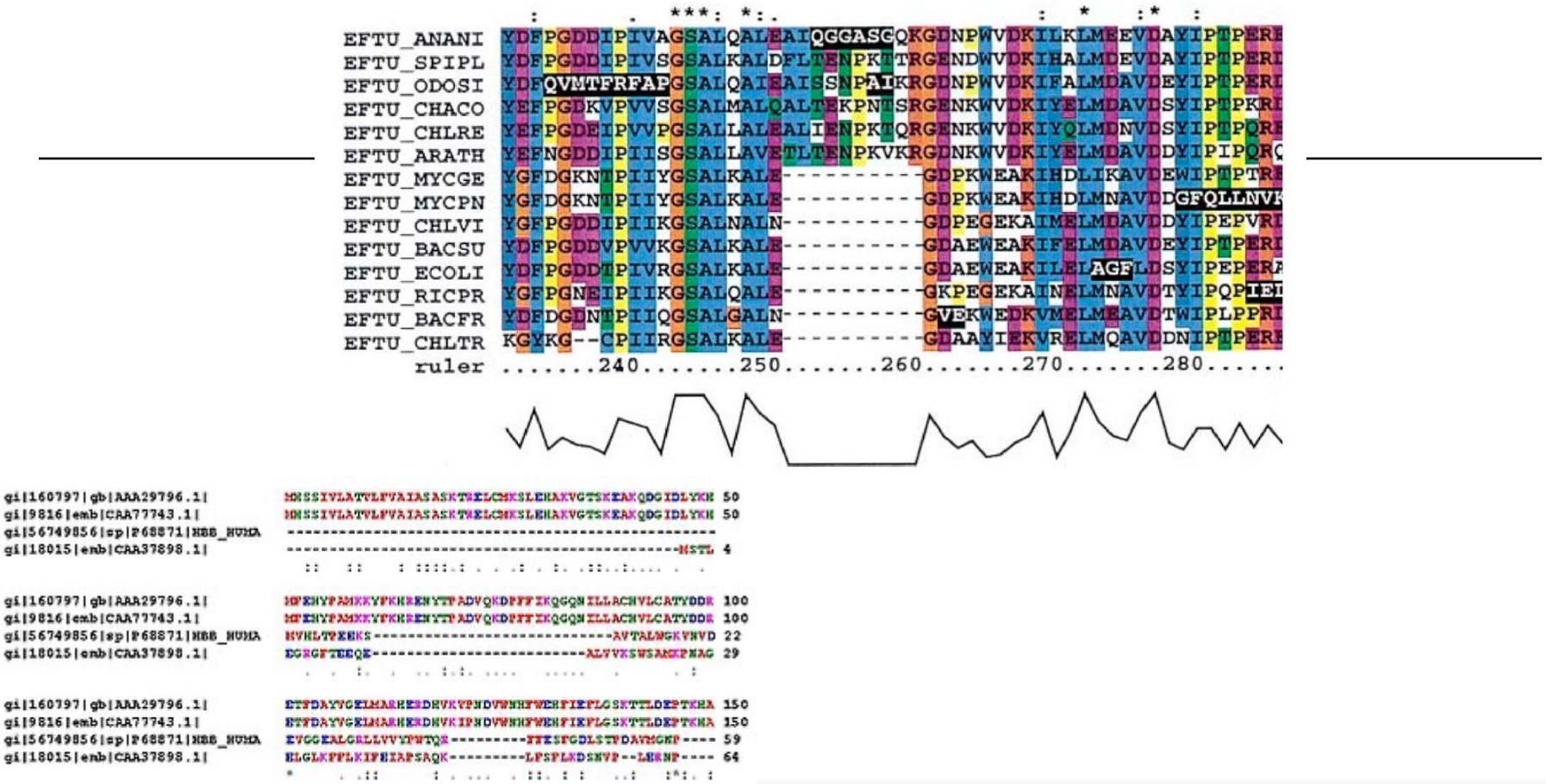
Target Sequence

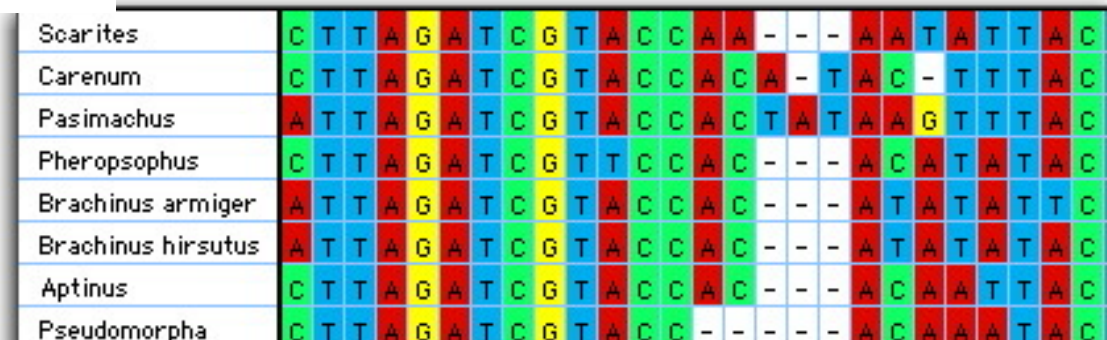
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

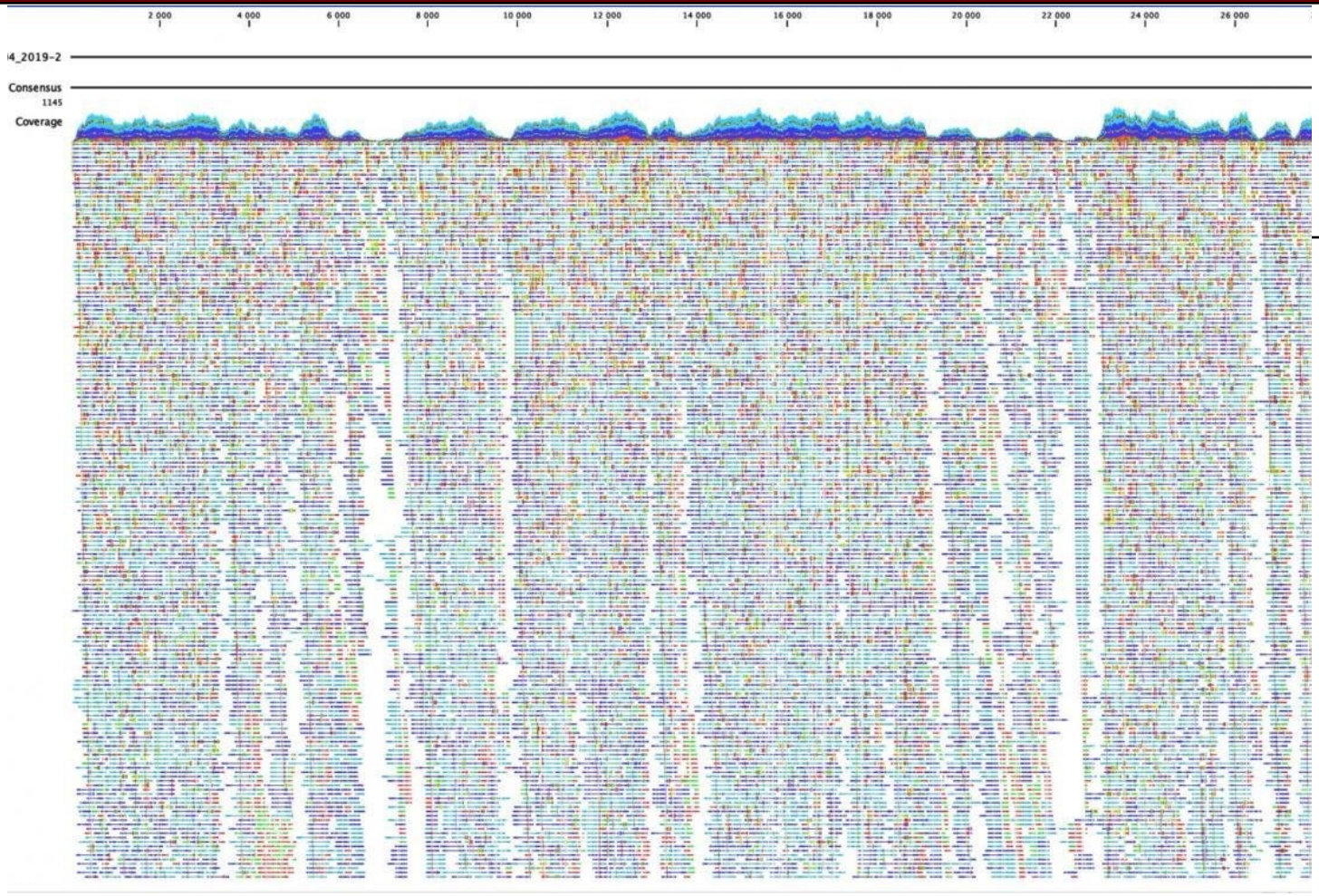
Query Sequence

Multiple Sequence Alignment (MSA)




EFTU_ANANI
EFTU_SPIPL
EFTU_ODOSI
EFTU_CHACO
EFTU_CHLRE
EFTU_ARATH
EFTU_MYCGE
EFTU_MYCPN
EFTU_CHLVI
EFTU_BACSU
EFTU_ECOLI
EFTU_RICPR
EFTU_BACFR
EFTU_CHLTR
ruler .....240.....250.....260.....270.....280.....
gi|160797|gb|AAA29796.1|
gi|9816|emb|CAA77743.1|
gi|56749856|sp|P68871|HBB_HUMA
gi|18015|emb|CAA37898.1|


Scarites
Carenum
Pasimachus
Pheropsophus
Brachinus armiger
Brachinus hirsutus
Aptinus
Pseudomorpha

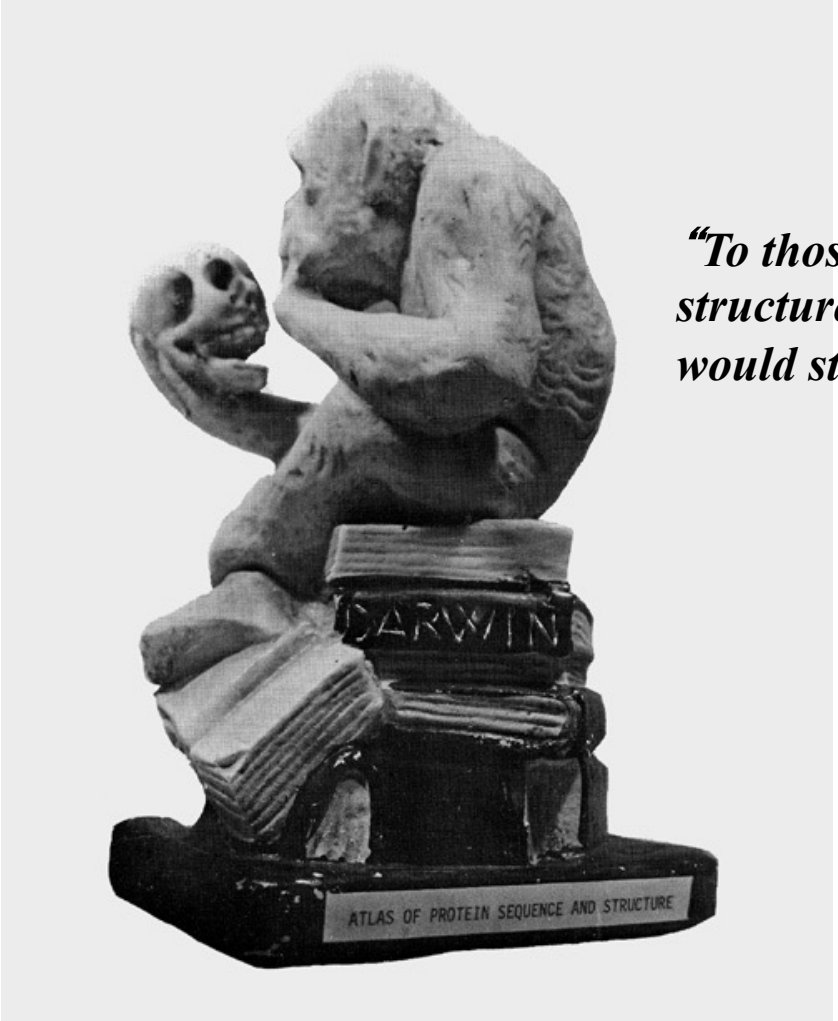


Whole genome sequence of the 2019-nCoV coronavirus, in one of the first French cases, made at the Institut Pasteur (Paris), using a unique Platform (P2M), open to all French National Reference Centers. Credit: Institut Pasteur/CNR of respiratory infection viruses.

Margaret Dayhoff & PAM Similarity Matrices

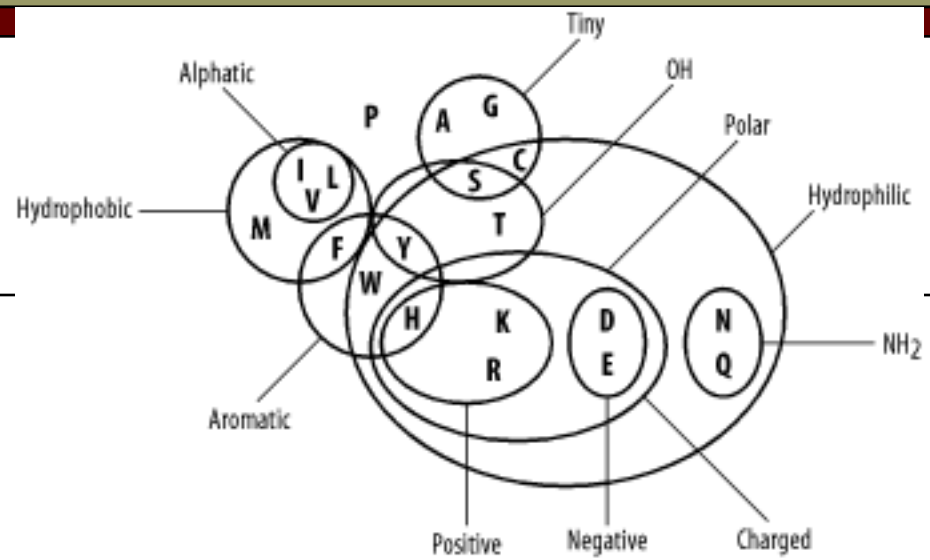
Dr. Margaret Oakley Dayhoff
The Mother & Father of Bioinformatics

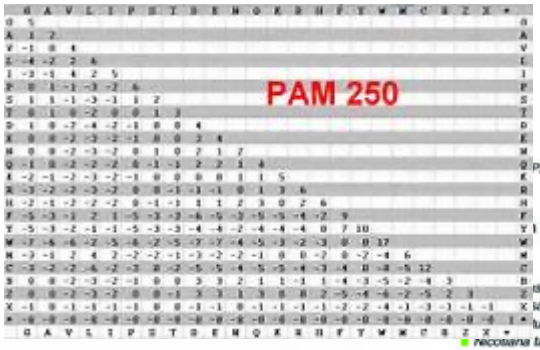
The Atlas of Protein Sequence and Structure 1972



"To those who would know the biochemical structure, function and origin of man and would strive to improve his lot."

Group & Subgroup Names	Amino Acid Residue	Group Properties
Hydrophilic -Small Aliphatic	Alanine Proline Glycine	**Small, Simple, Hydrophilic** Not hydrophobic, smallest
-Acid amide	Glutamine Asparagine	Slightly basic, amide, carbonyl
-Acid	Glutamic Acid Aspartic Acid	Acid, carbonyl
-Hydroxyl	Serine Threonine	Hydroxyl, small
Sulfhydryl	Cysteine	**Uniquely Reactive, Small**
Aliphatic	Valine Isoleucine Methionine Leucine	**Hydrophobic** Similarly branched
Basic	Lysine Arginine Histidine	**Basic, Nitrogen, Large**
Aromatic	Phenylalanine Tyrosine Tryptophan	**Aromatic Rings, Hydrophobic, Large**
Special	Histidine Tryptophan	Heterocyclic rings
	Cysteine Serine	Close similarity in shape
	Phenylalanine Leucie Isolenine Methionine	Hydrophobic; similar size

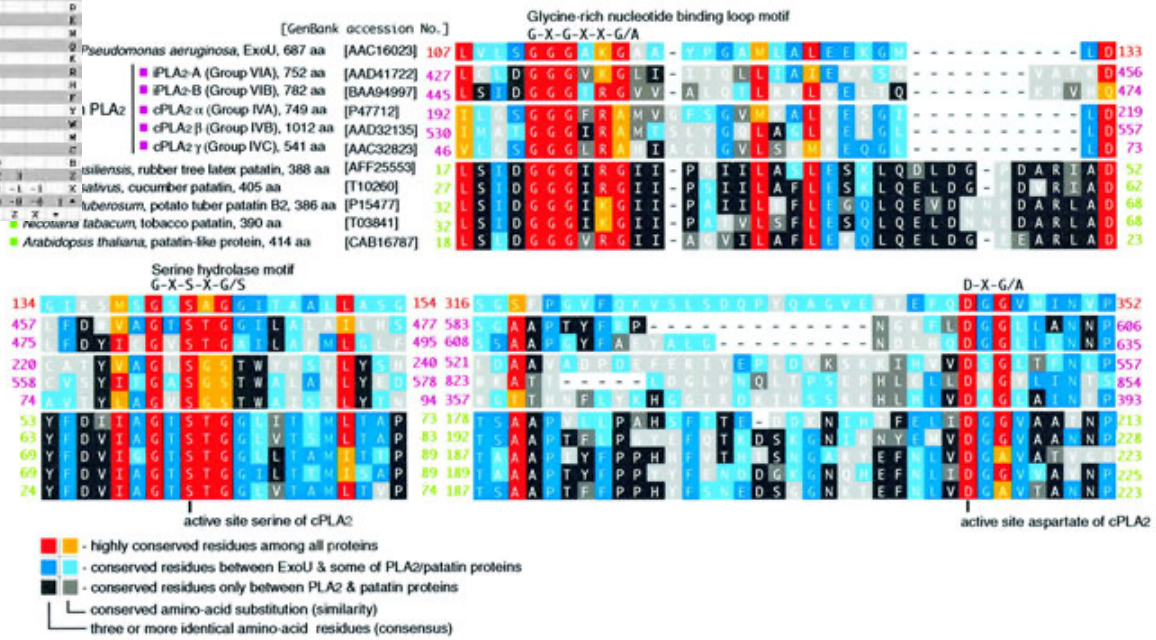




Dayhoff's PAM matrix

	A	R	N	D	C
A	9867	2	9	10	3
R	1	9913	1	0	1
N	4	1	9822	36	0
D	6	0	42	9859	0
C	1	1	0	0	9973

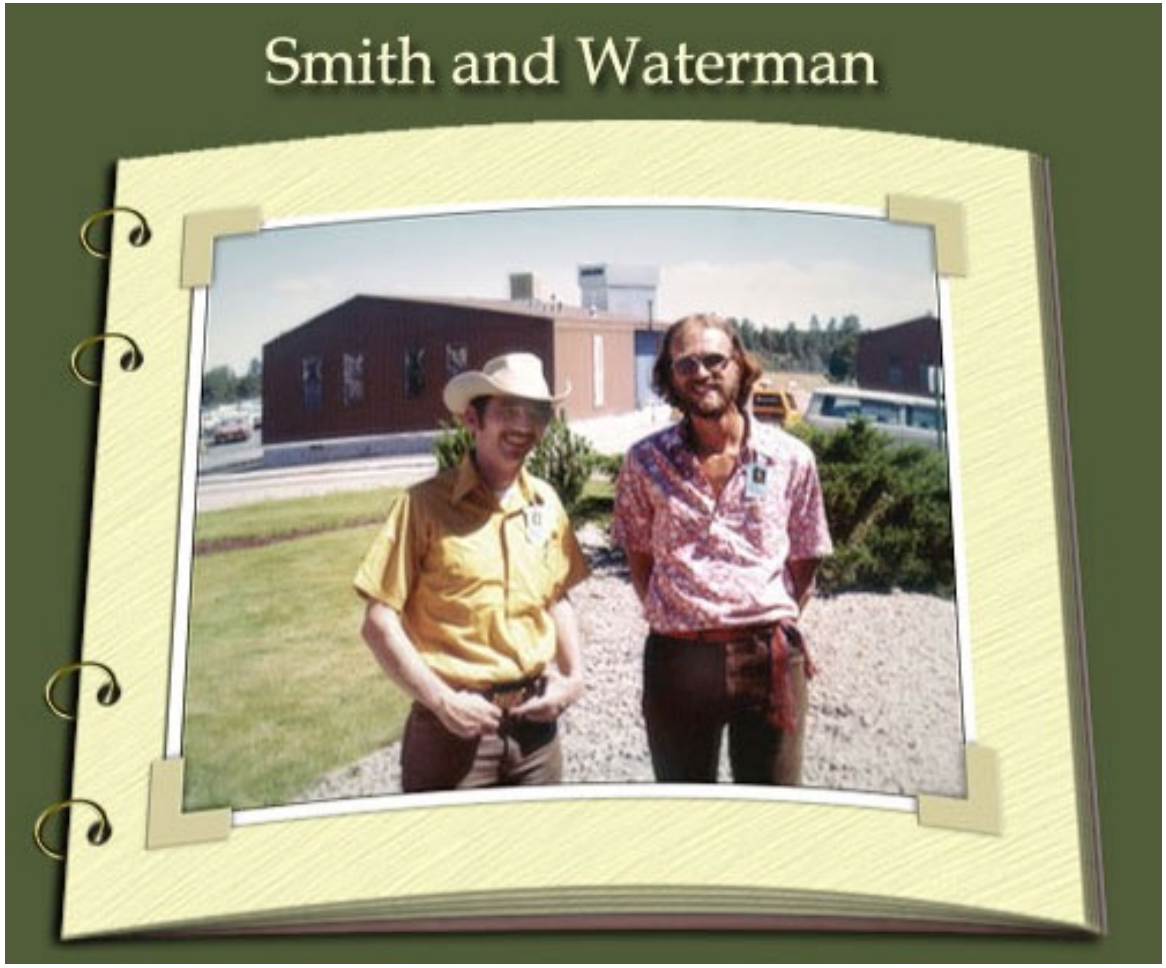
All entries × 10^4



The Smith Waterman Algorithm

Smith and Waterman at Los Alamos, New Mexico

Photo by David Lipman, Taken Summer of 1980



Viral src gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase.

AUTHORS

W. C. Barker and M. O. Dayhoff

ABSTRACT

The transforming protein sequences translated from the Rous avian and Moloney murine sarcoma virus src genes are shown to be related to the catalytic chain of bovine cAMP-dependent protein kinase (ATP:protein phosphotransferase, EC 2.7.1.37). The avian transforming protein, also a protein kinase, shows greatest homology with the bovine protein kinase in the carboxyl-terminal half, where the protein kinase activity is localized. Moreover, lysine occurs in the inferred transforming protein sequences at the position homologous with the proposed ATP-binding lysine of the bovine protein kinase. This relationship is consistent with the hypothesis that the src genes originated in the host genomes, in which they are members of a superfamily of distantly related protein kinases that are normal constituents of mammalian cells. In the host, these sequences are much more highly conserved than in the viruses.

Viral *src* gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase

```
1 BOV-PK   QIEHTL NEKRI - -LQAV NFPF LVKLEFSF KDNSNLYMVMEYV PGGE MFSH
2 MMSV     SQRSFWA ELNI AGLR HDNIVR VVAASTRT PEDSNS LG TI IME FGGNV TLH
3 RSV-PC   SPEAFL QEAQV - -MKKL RHEKLVQ L-YAVVSEEP I Y I V I EYMSKGS LLD F

           S  E  F L  E   I   L     N     LV  L          SN  Y  VI EY   GG     H
                      *                   *                              *

1 BOV-PK   - --- - -- - - - - -- LR - R I GR F- SE PHAR F YAAQI V LT F EY LHS LD LI YRDL
2 MMSV     QVIYDATRSPEPLS CR - - KQLSLGKC LKY S LDVVNG L LF LHS QS I LH LDL
3 RSV-PC   - --- - -- - - - - -- LKGEMGKYLRLPQ LV DMAAQI AS GMAYVE RMNYVHRDL

                             L    R   GK  LSLP       YAAQIV  G    Y  HS       HRDL
                             *                                              * *
```

Viral *src* gene products are related to the catalytic chain of mammalian cAMP-dependent protein kinase

```
1 BOV-PK   QIEHTL NEKRI - -LQAV NFPF LVKLEFSF KDNSNLYMVMEYV PGGE MFSH
2 MMSV     SQRSFWA ELNI AGLR HDNIVR VVAASTRT PEDSNS LG TI IME FGGNV TLH
3 RSV-PC   SPEAFL QEAQV - -MKKL RHEKLVQ L-YAVVSEEP I Y I V I EYMSKGS LLD F

           S  E  F L  E   I   L     N     LV  L          SN  Y  VI EY   GG     H
                      *                     *                              *

1 BOV-PK   - --- - - - - - - - - LR - R I GR F- SE PHAR F YAAQI V LT F EY LHS LD LI YRDL
2 MMSV     QVIYDATRSPEPLS CR - - KQLSLGKC LKY S LDVVNG L LF LHS QS I LH LDL
3 RSV-PC   - --- - - - - - - - - LKGEMGKYLRLPQ LV DMAAQI AS GMAYVE RMNYVHRDL

                               L    R   GK  LSLP       YAAQIV  G    Y  HS       HRDL
                               *                                                 * *
```

A = Alanine
V = Valine
F = Phenylalanine
P = Proline
M = Methionine
I = Isoleucine
L = Leucine

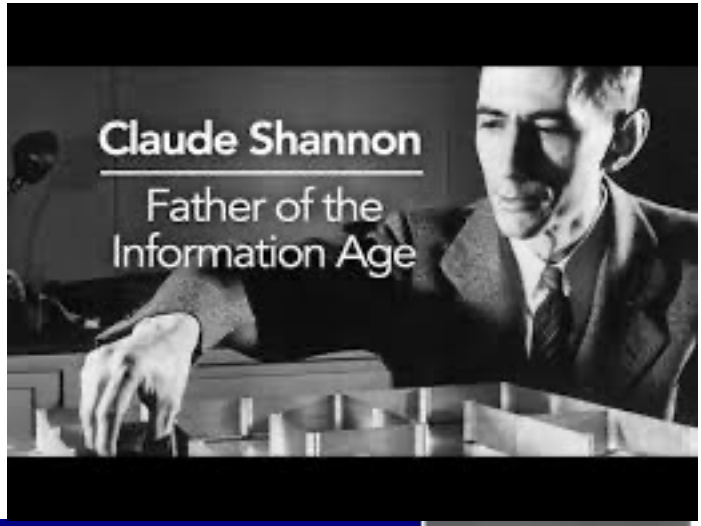
D = Asartic Acid
E = Glutamic Acid
K = Lysine
R = Arginine

S = Serine
T = Threonine
Y = Tyrosine
H = Histidine
C = Cysteine
N = Asparagine
Q = Glutamine
W = Tryptophan

G = Glycine

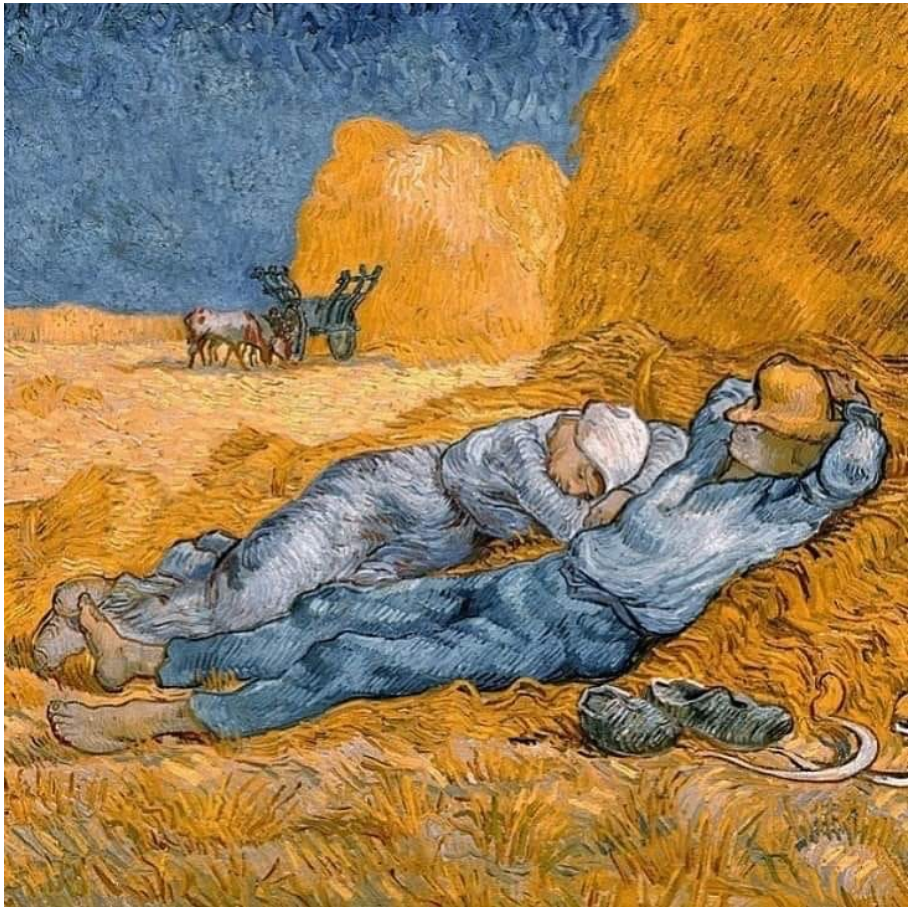
Information Theory

- How long an alignment should be to be statistically significant?



$$H = -\sum_{i=1}^{n} p_i \log p_i$$

Chapter 2: Combinatorial Pattern Matching

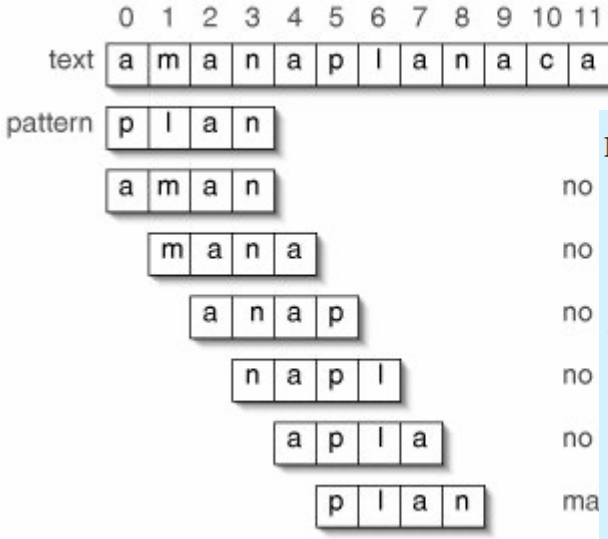


Image, courtesy of Vincent van Gogh Museum

Chapter 2

Combinatorial Pattern Matching Algorithms

Chapter 2: Combinatorial Pattern Matching Algorithms



Best Motif Found:

NAME	START	SITES	END	STRAND	MARGINAL SCORE
1	23	aaacgacgtaatgctacg	6	-	22.9
2	17	aaacgac.taat.ctacg	2	-	8.45
3	30	caacgaggta.tgcaacg	14	-	14.1
4	23	caaccacgtaatgcaacg	6	-	23.6
5	24	caaccacgtaatgcatag	7	-	17.5

Score: **51.1928**

logo with ssc

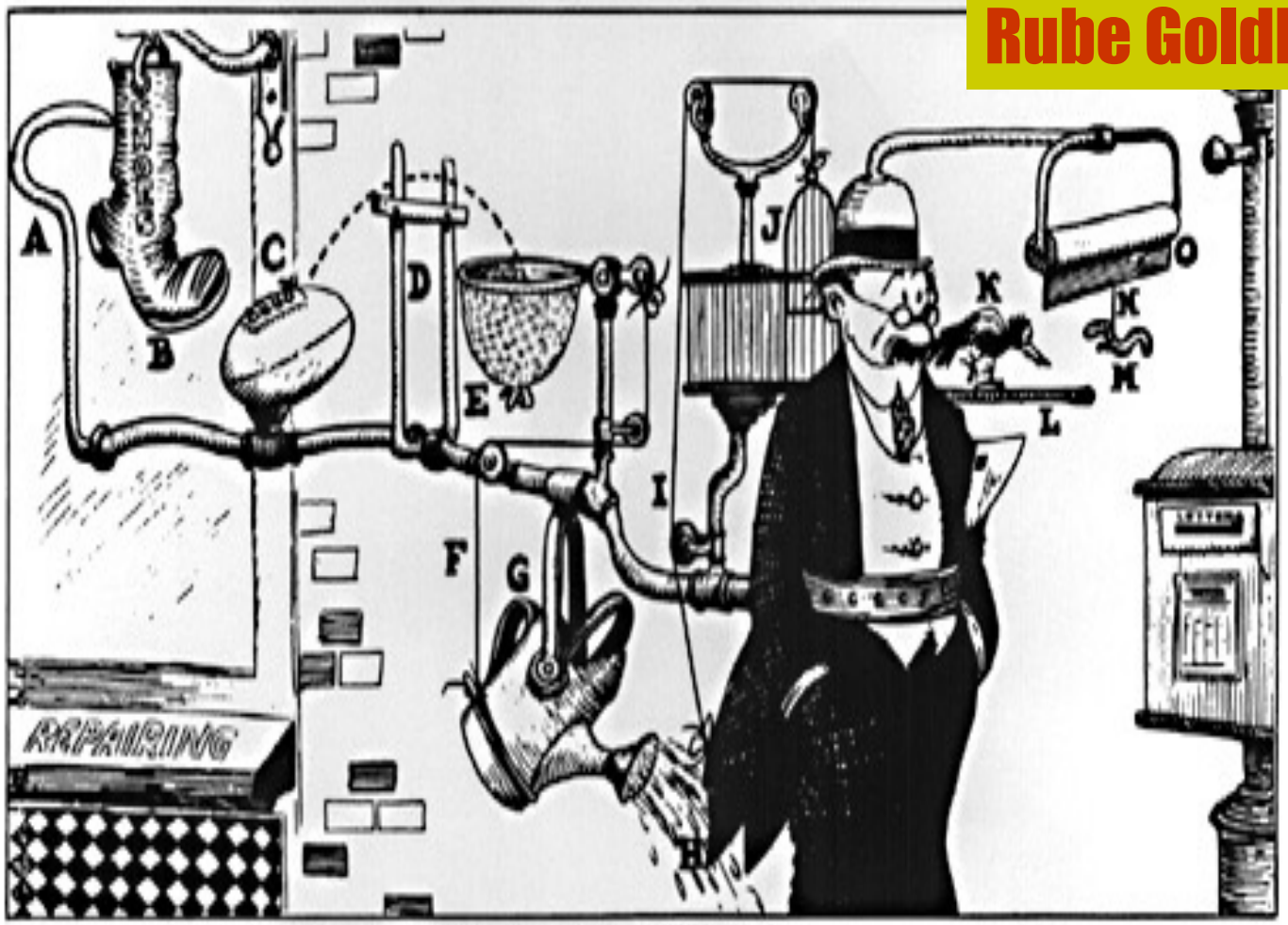
Scan alignment 1 **against sequence databases using GLAM2SCAN.**

View alignment 1 View PSPM 1

Compare PSPM 1 **to known motifs in motif databases using Tomtom.**

Regular Expression for Motif: **[ac]aac[cg]a[cg]g?taa?tg?c[at][at][ac]g**

Rube Goldberg's Innovation



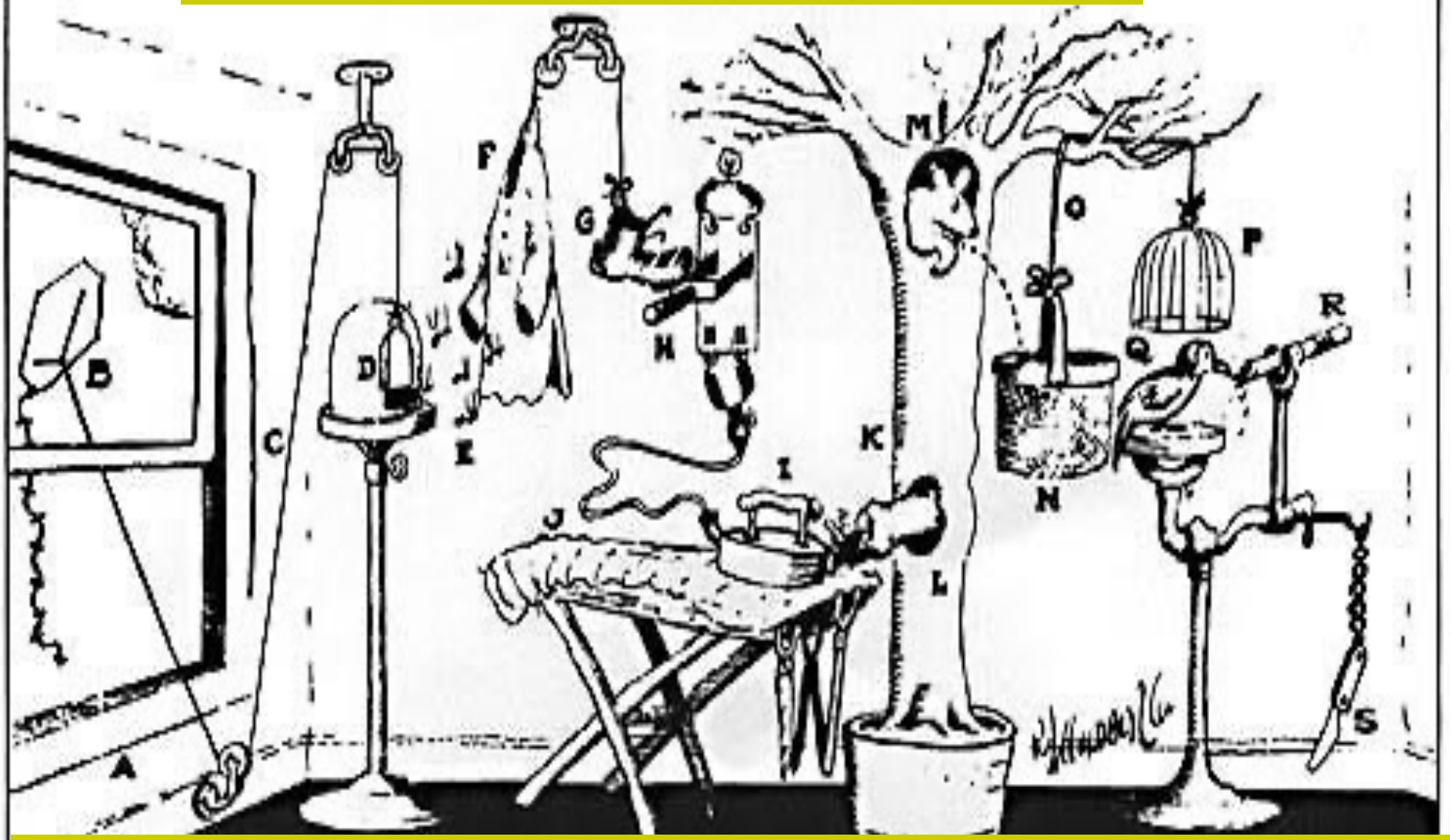
Keep You From Forgetting To Mail Your Wife's Letter RUBE GOLDBERG (tm) RGI 049

Mixed character of the problem :

continuous mathematics
discrete mathematics

GENOMIC REGULATORY SYSTEMS

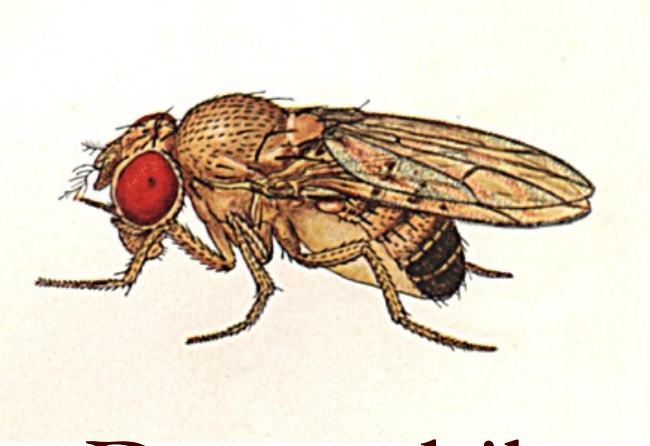
Rube Goldberg's Pencil Sharpener invention



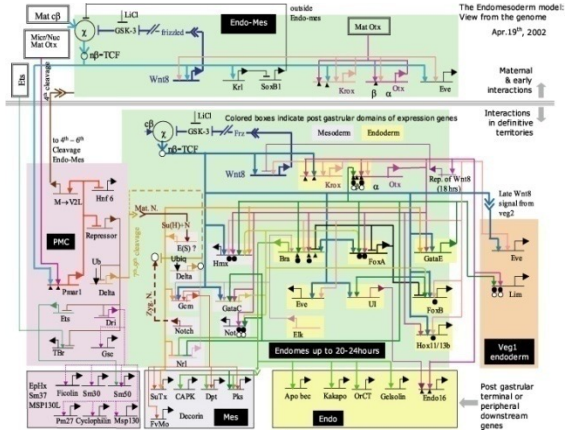
Emergency knife (S) is always handy in case opossum or the woodpecker gets sick and can't work.

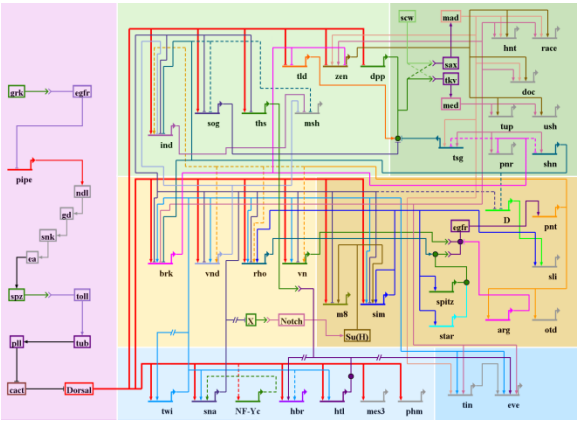
A Tale of Two Networks

Sea Urchin

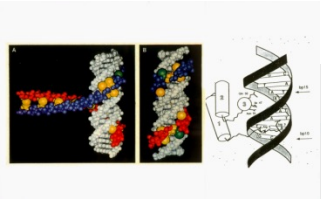


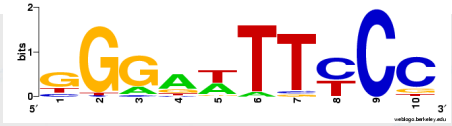
Drosophila





The Dogma





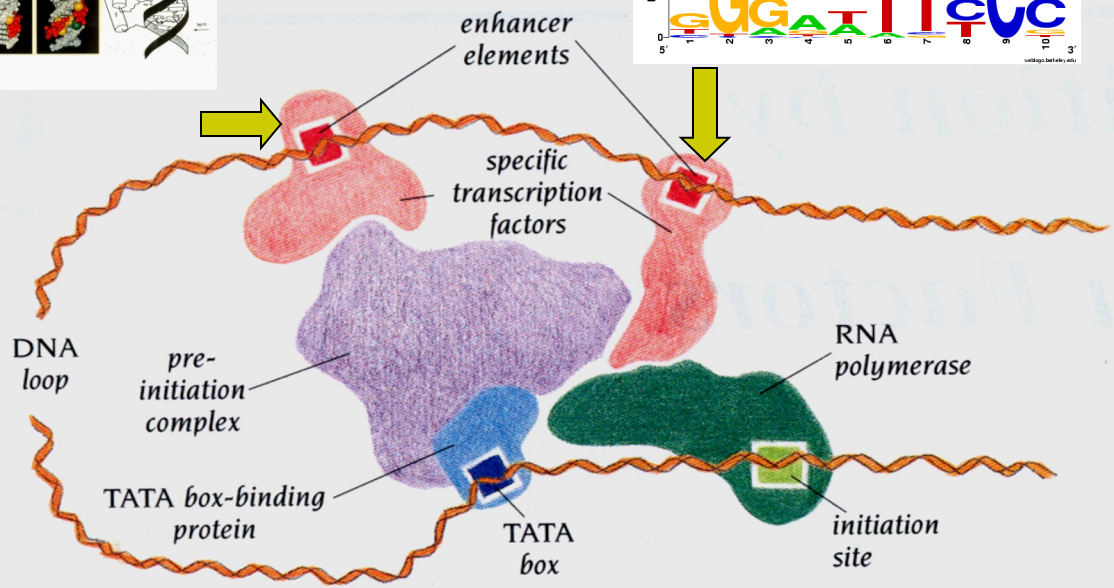
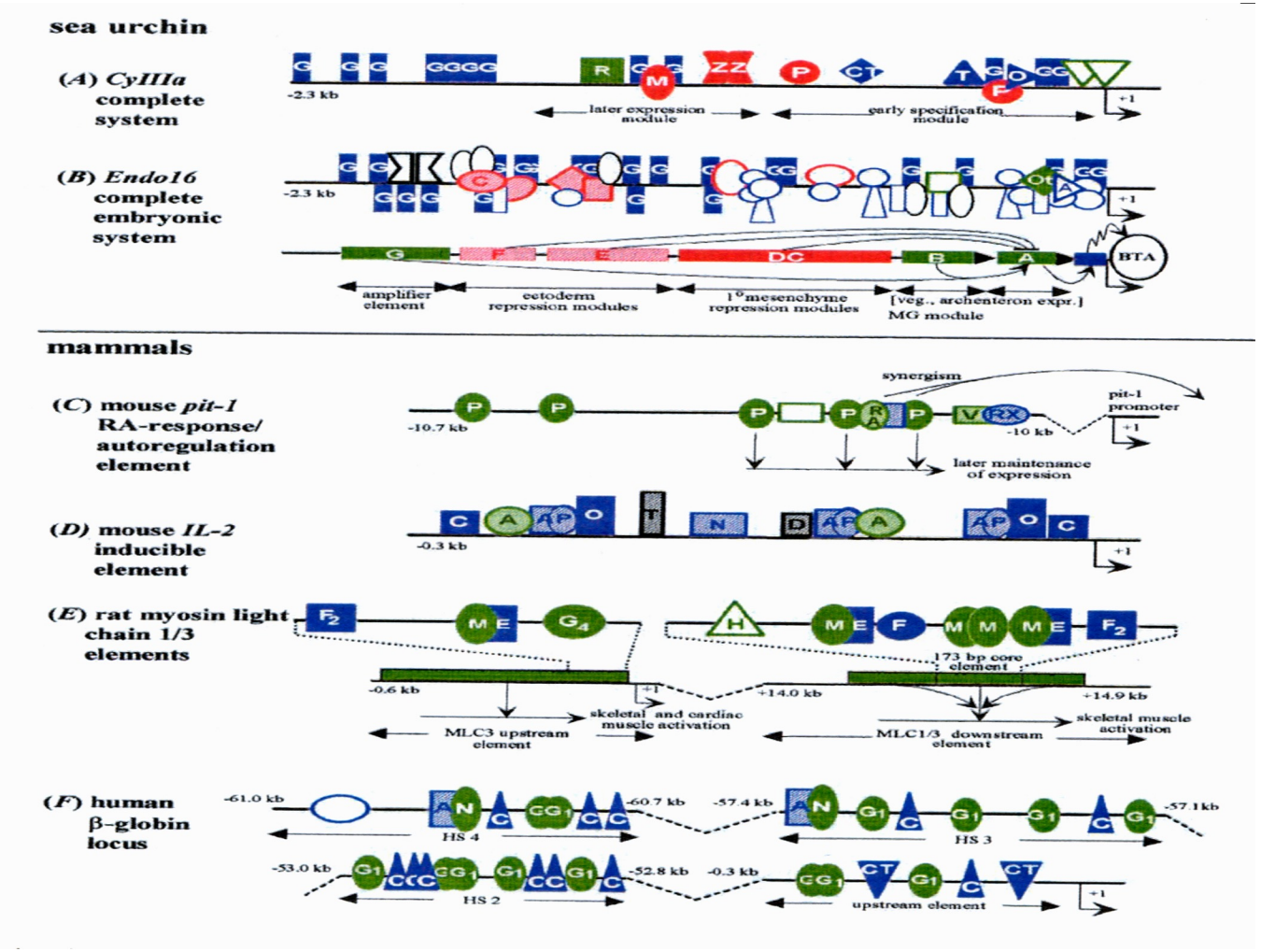


Figure 9.2 Schematic model for transcriptional activation. The TATA box-binding protein, which bends the DNA upon binding to the TATA box, binds to RNA polymerase and a number of associated proteins to form the preinitiation complex. This complex interacts with different specific transcription factors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

Genomic Regulatory Regions

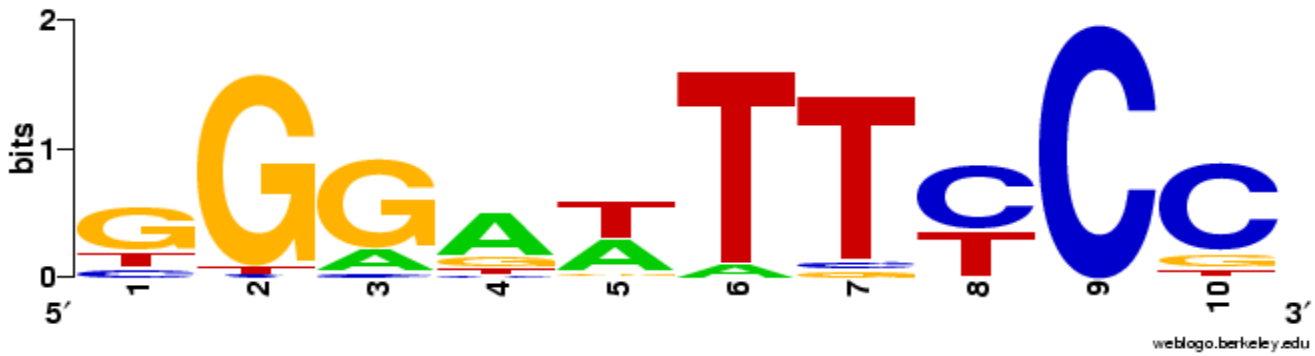


Phylogenetic Trees (Ch. 3) ???

Big open problem about what is an evolutionary model for regulatory regions of genes !!!

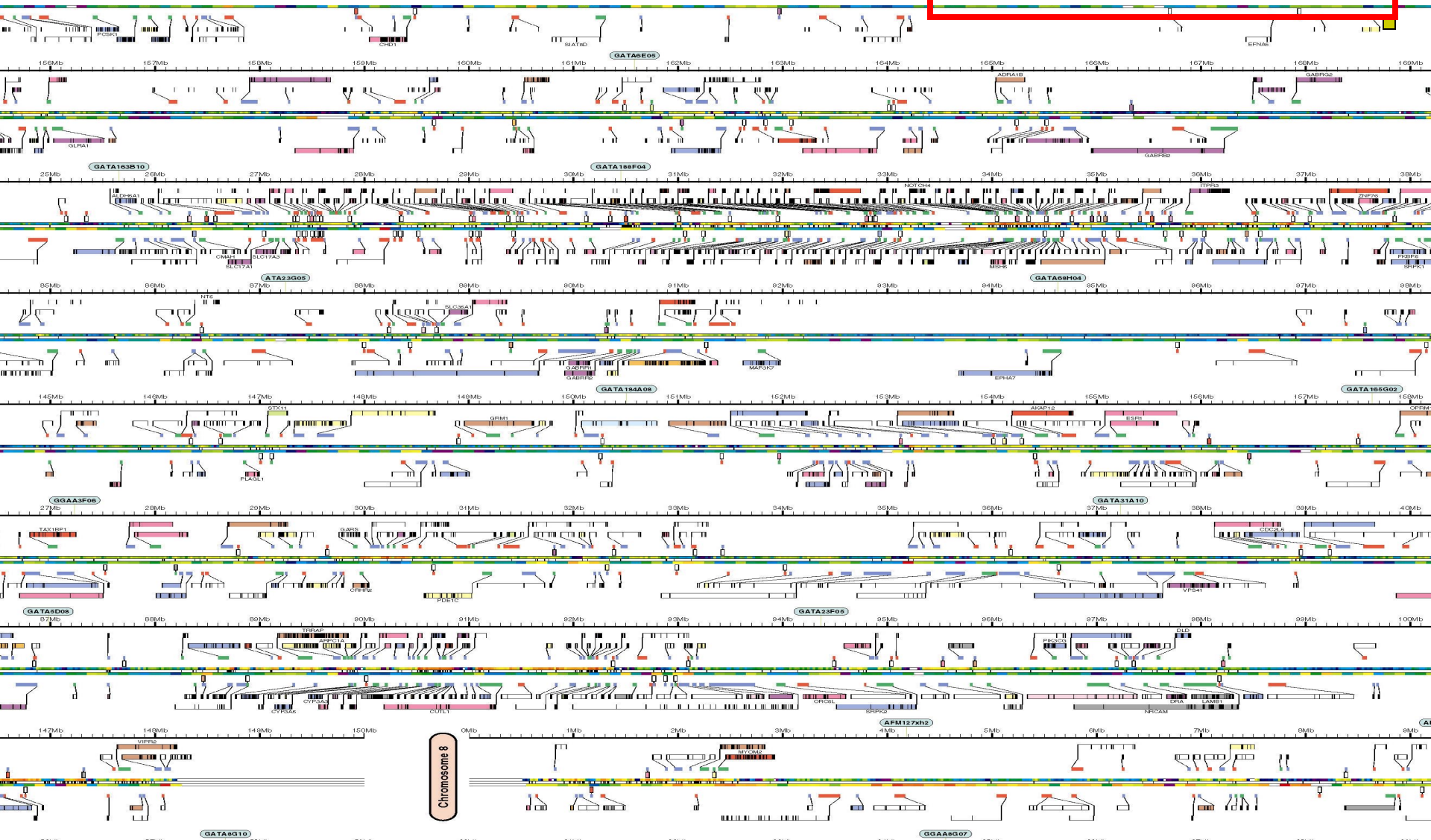
Phylogenetic trees are not good models for the Regulatory Genome

TF Binding Site Complexity

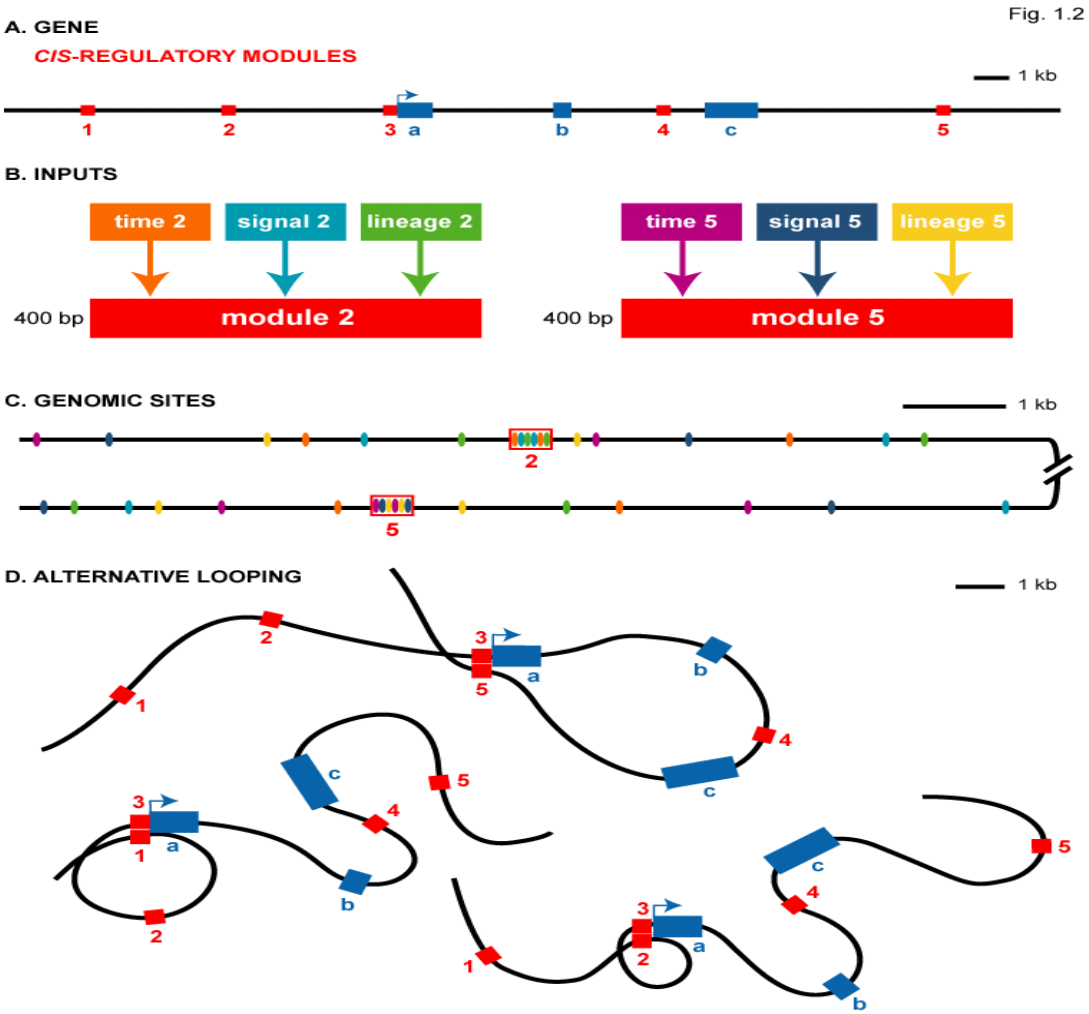


Genome Complexity

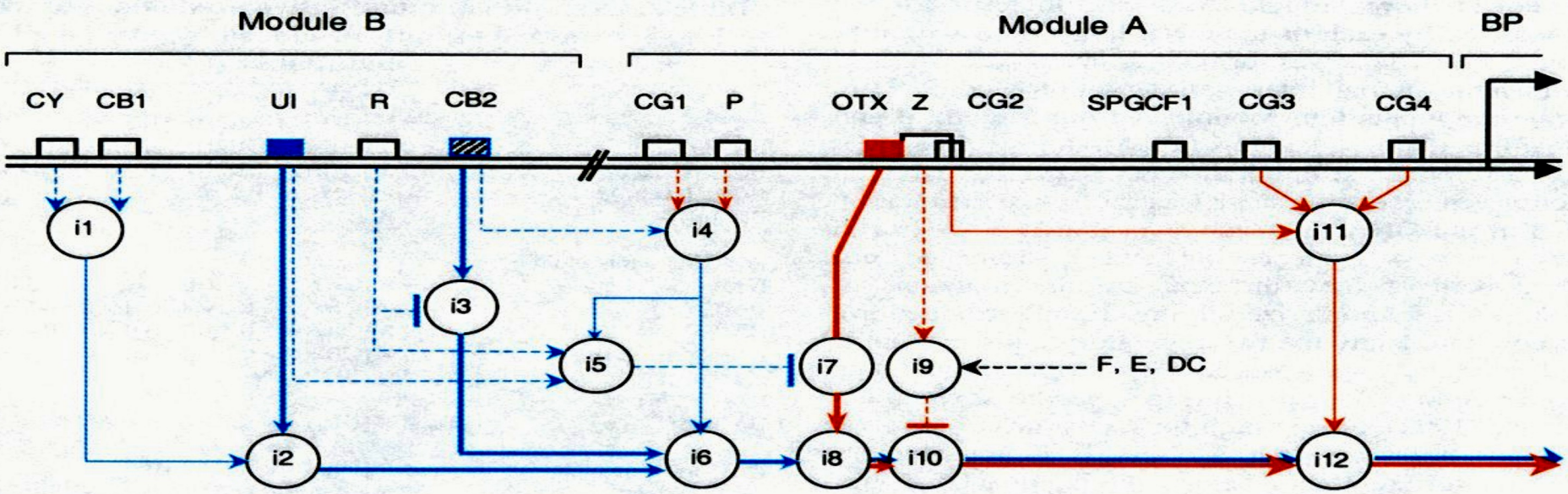
1 Billion DNA bases
20,000 Genes



cis-Regulatory Modules Complexity



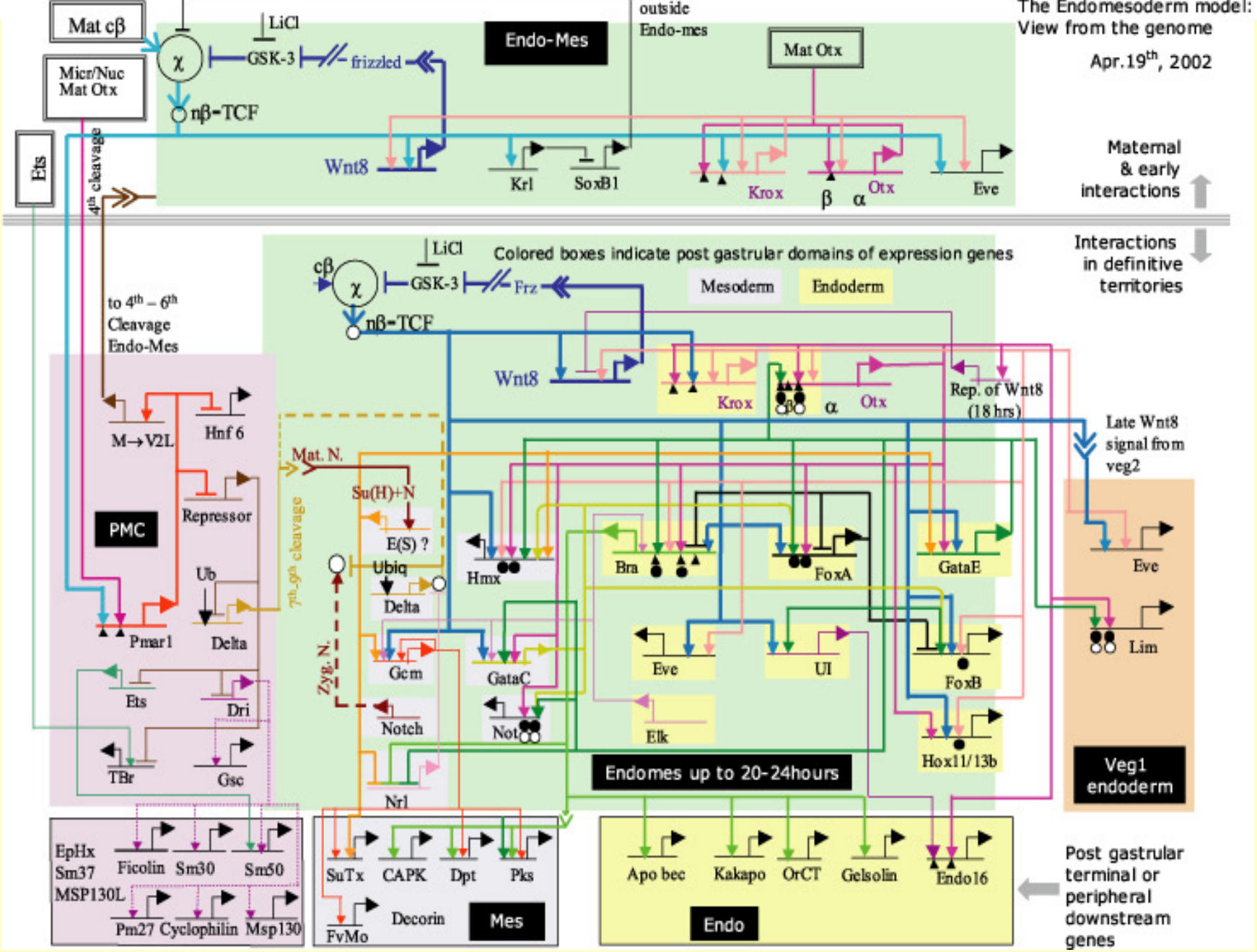
200,000 cis-Modules



if CY & CB1	i1 = 1
else	i1 = 0.5
	i2 = i1 • UI(t)
if R	i3 = CB2(t)
else	i3 = k • CB2(t) (1<k<2)
if P & CG1 & CB2	i4 = 2
else	i4 = 0
if UI(t)>threshold & R & i4≠0	i5 = 1
else	i5 = 0
	i6 = i4 • (i2+i3)
if i5=0	i7 = OTX(t)
else	i7 = 0
	i8 = i6 + i7
if (F or E or DC) & Z	i9 = 1
else	i9 = 0
if i9=1	i10 = 0
else	i10 = i8
if (CG2 & CG3 & CG4)	i11 = 2
else	i11 = 1
	i12 = i10 • i11

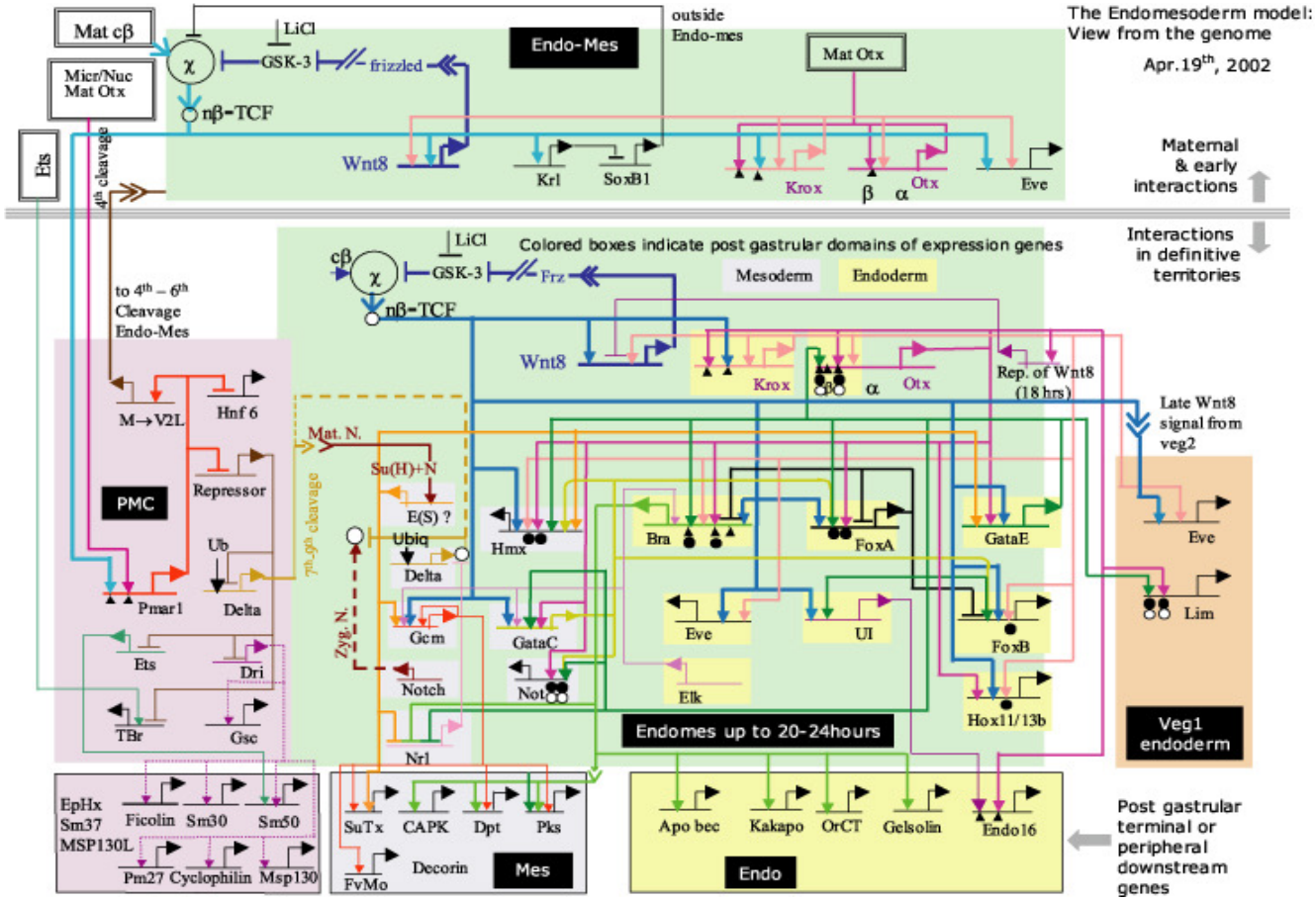
The DNA program that regulates the expression of ***endo16*** in sea urchin

▪THE FIRST GENE



THE FIRST NETWORK

The View from the Genome



A Case Study

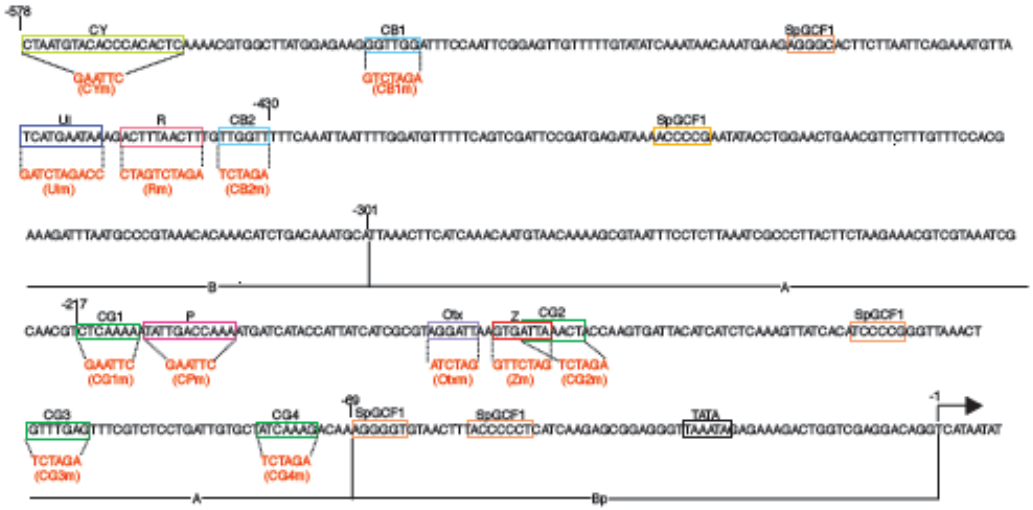


Figure 2: Quintessential diagram (from [25])

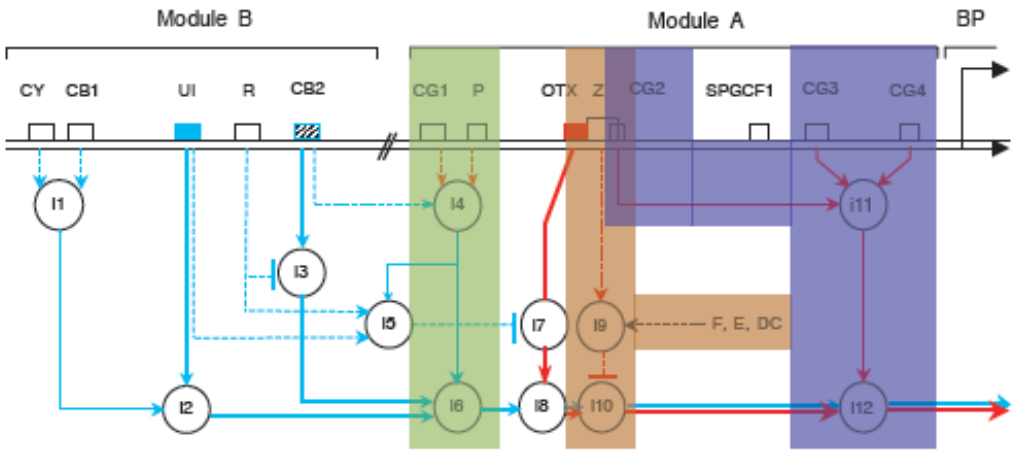
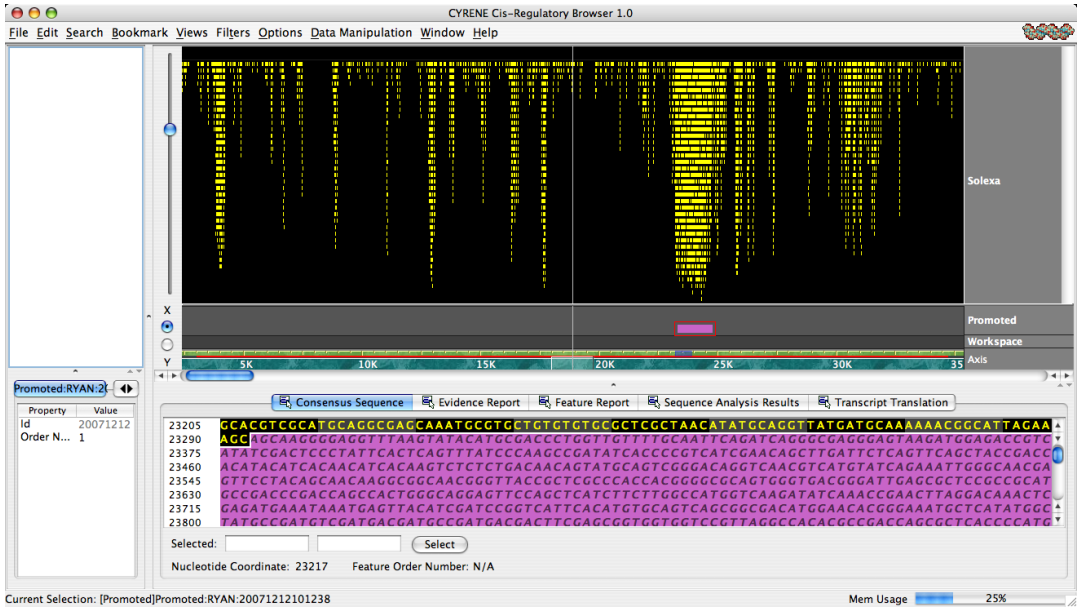


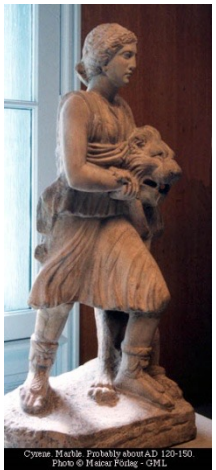
Figure 3: Computational logic model for Modules A and B of *endo16* (from [25])

Cyrene

Ryan Tarpine

The CYRENE project seeks to address the fundamental problem of determining de novo the function of regulatory sequence by developing the cis-Lexicon, a database of known cis-regulatory modules, the cis-Browser, a next-generation regulatory genome browser, and a library of tools for assisting in the annotation pipeline. The cis-Lexicon will be a comprehensive catalog of experimentally-validated gene regulatory knowledge, designed to be a foundation and benchmark for future prediction algorithms. The cis-Browser is a high-speed integrative environment for viewing and annotating all types of genomic information. It is capable of displaying data from the cis-Lexicon, public online databases, BLAST hits, and precomputed comparative genomics analyses. To aid annotators' entry of information into the cis-Lexicon, we are developing high-throughput tools for finding relevant literature and assisting in the extraction of correct information. We suggest several algorithms to analyze the cis-regulatory data as the cis-Lexicon expands. The CYRENE project is being carried out in cooperation with Eric Davidson at the California Institute of Technology.





The cis-Browser

Zoom Slider

Zoom Focus Slider

Annotation View

Selected Region

Loading Indicator

Outline View

Selected Feature

Annotation Workspace Tier

Genomic Axis

Promoted Annotation Tier

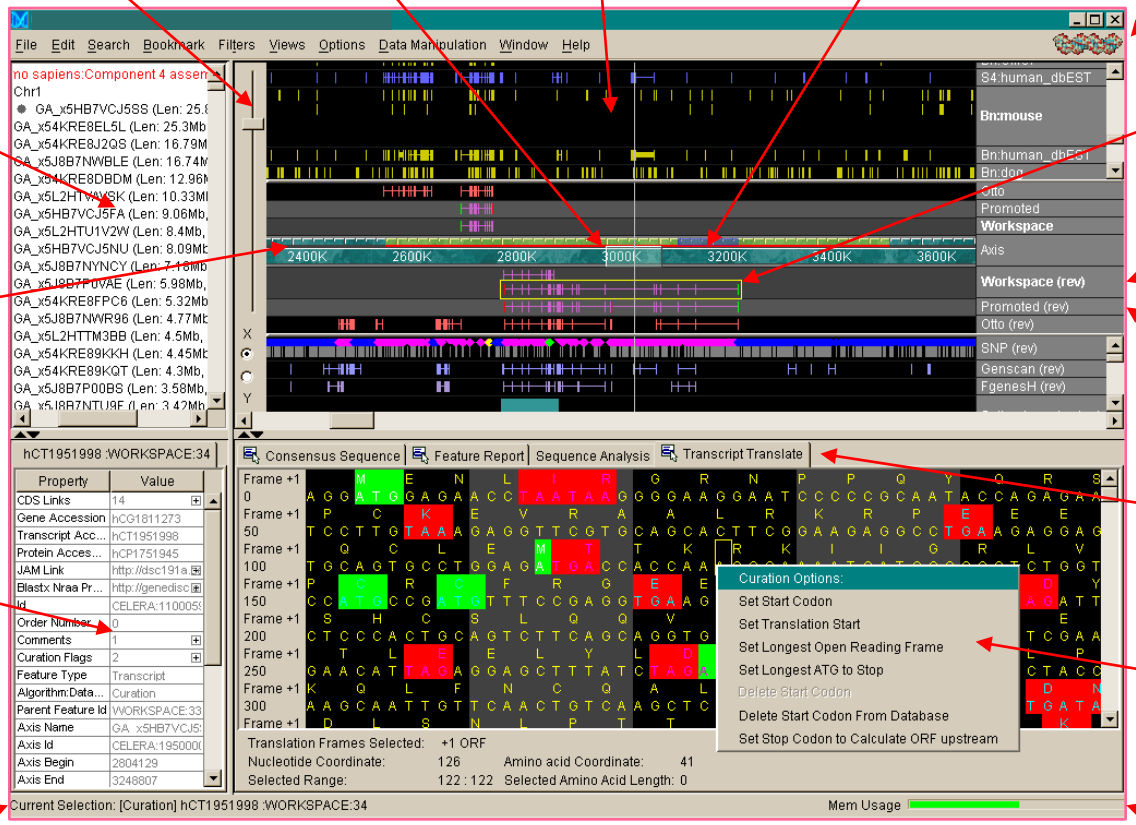
Property Inspector View

Subviews

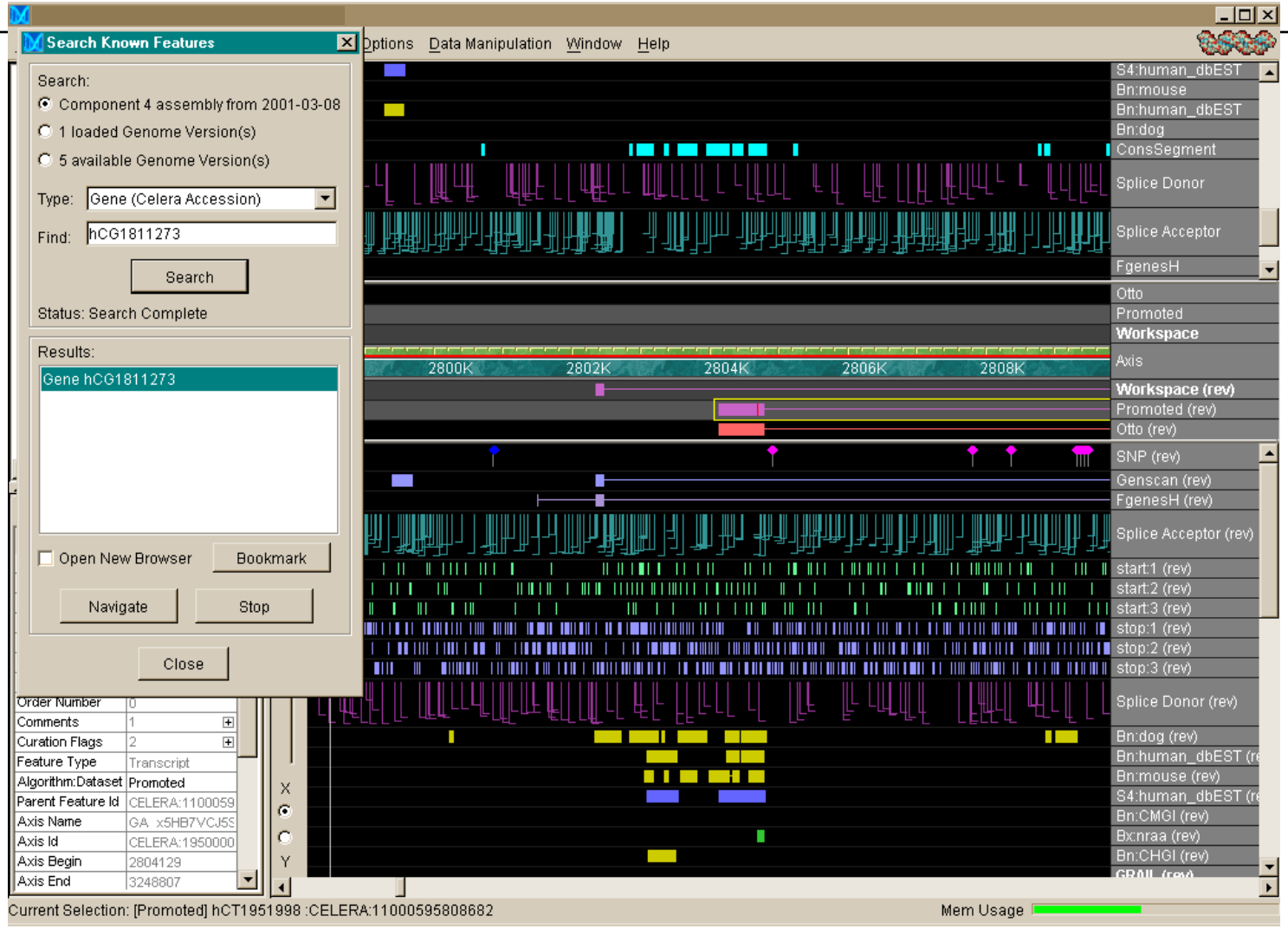
Popup Curation Menu

Description Line

Memory Usage Meter



Transcript Curation



Sequence Comparison

File Edit Search Bookmark Filters Views Options Data Manipulation Window Help

no sapiens:Component 4 asser
Chr1
GA_x5HB7VCJ5SS (Len: 25.
GA_x54KRE8EL5L (Len: 25.3Mb
GA_x54KRE8J2QS (Len: 16.79M
GA_x5J8B7NWBLE (Len: 16.74M
GA_x54KRE8DBDM (Len: 12.96I
GA_x5L2HTVAVSK (Len: 10.33M
GA_x5HB7VCJ5FA (Len: 9.06Mb
GA_x5L2HTU1V2W (Len: 8.4Mb,
GA_x5HB7VCJ5NU (Len: 8.09MI
GA_x5J8B7NYNCY (Len: 7.18Mb
GA_x5J8B7P0VAE (Len: 5.98Mb,
GA_x54KRE8FPC6 (Len: 5.32Mb
GA_x5J8B7NWR96 (Len: 4.77MI
GA_x5L2HTTM3BB (Len: 4.5Mb,
GA_x54KRE89KKH (Len: 4.45MI
GA_x54KRE89KQT (Len: 4.3Mb,
GA_x5J8B7P00BS (Len: 3.58Mb
GA_x5J8B7NTU9F (Len: 3.42Mb

GRAIL
Bn:CHGI
Bx:nraa
Bn:CMGI
S4:human_dbEST
Bn:mouse
Bn:human_dbEST
Bn:dog
Otto
Promoted
Workspace
Axis
Workspace (rev)
Promoted (rev)
Otto (rev)
SNP (rev)

260.4K 5260.6K 5260.8K 5261.0K 5261.2K 5261.4K 5261.6K 5261.8K

Bx:nraa:CELERA:5000009072895

Property	Value
Id	CELERA:500000
Aliases (Numb...	0
Order Number	1
Comments	0
Feature Type	High Scoring Pair
Algorithm:Data...	Bx:nraa
Parent Feature...	CELERA:500000
Axis Name	GA_x5HB7VCJ5
Axis Id	CELERA:195000
Axis Begin	5261006
Axis End	5261183
Entity Length	177
Entity Orientation	Forward
Is Child	true
Is Composite	false
Relative Asse...	985037830
Display Priority	Low

Consensus Sequence | Feature Report | Query Sequence Alignments | Sequence Analysis | Subject Sequence Alignments

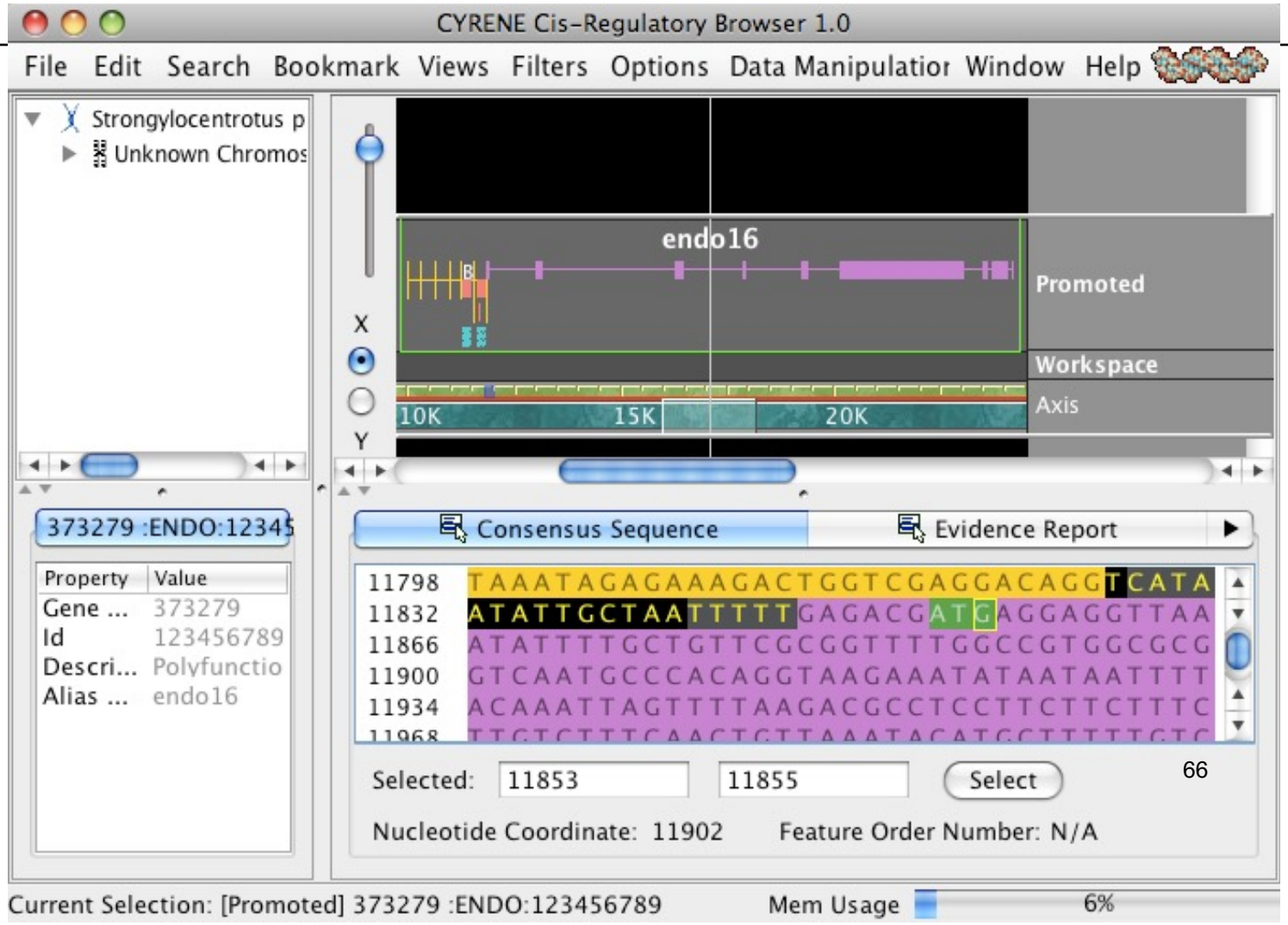
+3
+2
+1
DNA +
Query Sequence
Axis
DNA -
-1
-2
-3

GTGGGCAGTGGGGATCTTCCGAATTGAAAGGAGATGACCTTCAGGCCATAAGAAGGAATTGACCCAGATAAAACAAAA

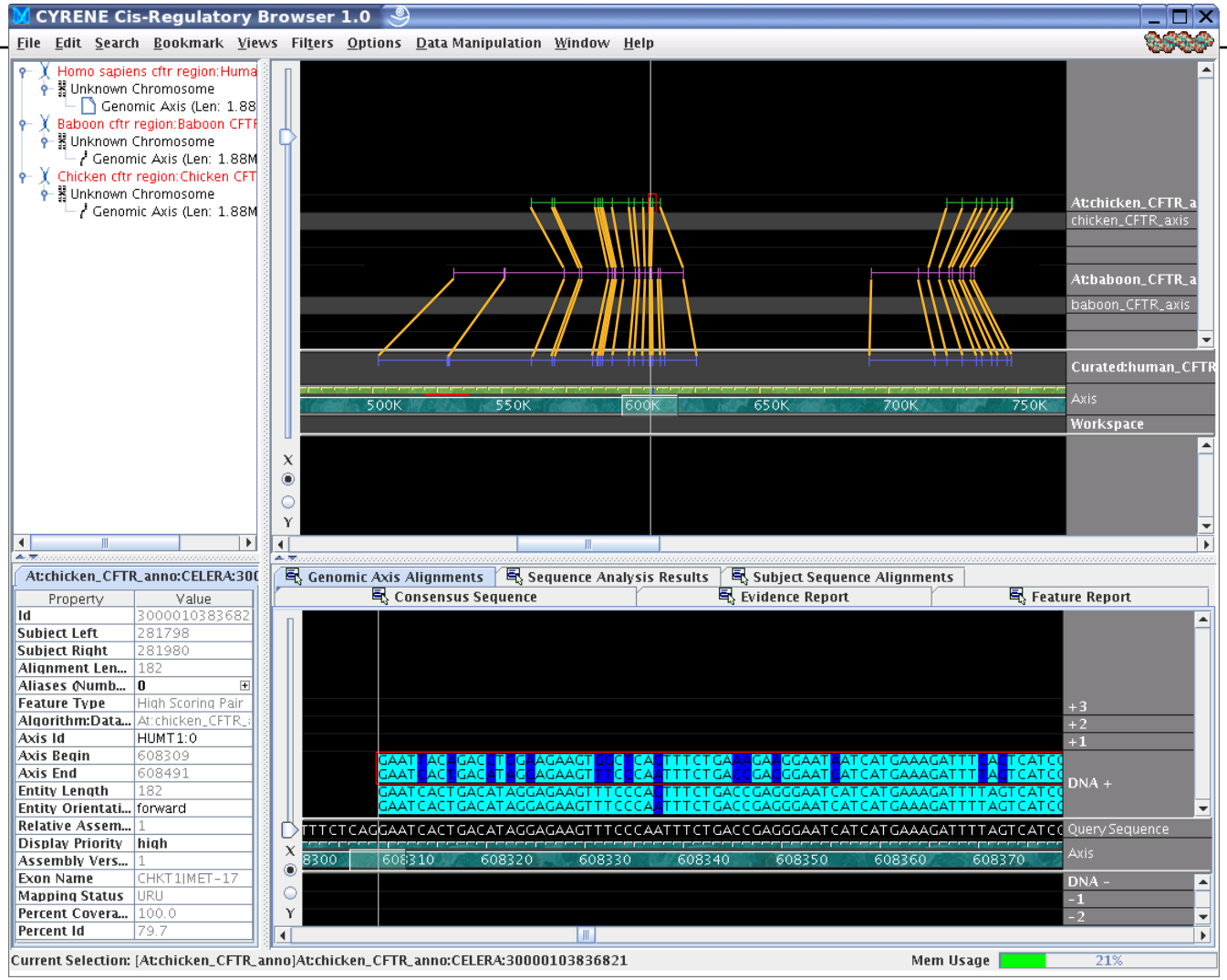
5260960 5260970 5260980 5260990 5261000 5261010 5261020 5261030

Current Selection: [Bx:nraa]Bx:nraa:CELERA:50000090728954

Mem Usage

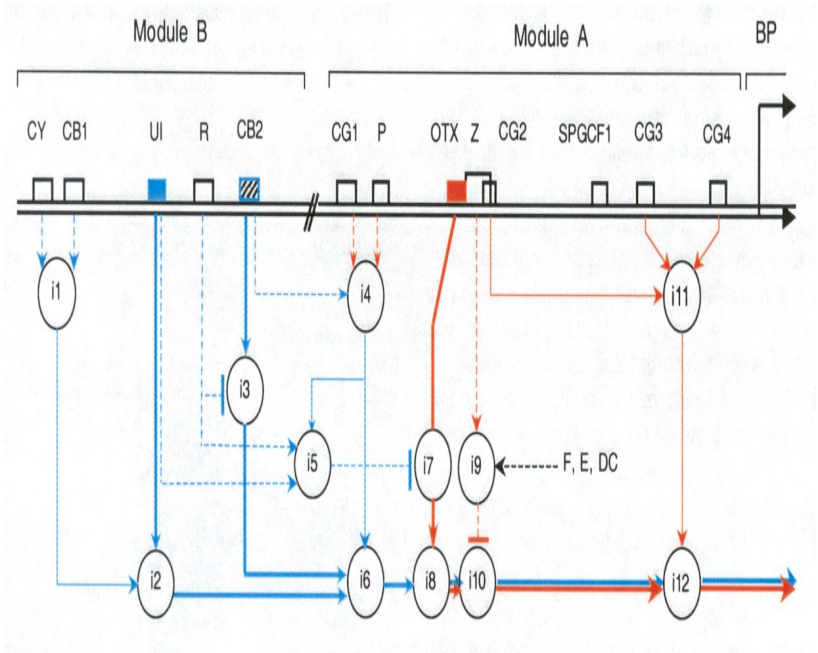

CYRENE Cis-Regulatory Browser 1.0
File Edit Search Bookmark Views Filters Options Data Manipulatior Window Help
Strongylocentrotus p
Unknown Chromos
endo16
Promoted
Workspace
Axis
10K
15K
20K
X
Y
373279 :ENDO:12345
Property Value
Gene ... 373279
Id 123456789
Descri... Polyfunctio
Alias ... endo16
Consensus Sequence
Evidence Report
11798 TAAATAGAGAAAGACTGGTCGAGGACAGGTCATA
11832 ATATTGCTAATTTTTGAGACGATGAGGAGGTTAA
11866 ATATTTTGCTGTTCGCGGTTTTGGCCGTGGCGCG
11900 GTCAATGCCCACAGGTAAGAAATATAATAATTTT
11934 ACAAATTAGTTTTAAGACGCCTCCTTCTTCTTTC
Selected: 11853 11855 Select
Nucleotide Coordinate: 11902 Feature Order Number: N/A
Current Selection: [Promoted] 373279 :ENDO:123456789
Mem Usage 6%

Inter-species comparison



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

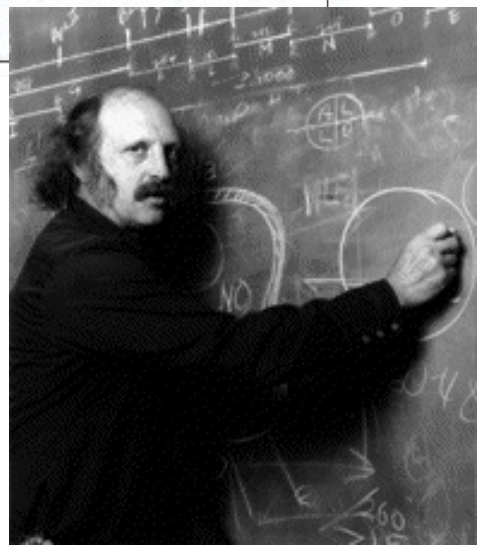
A Proposal for Nobel Prize



if CY & CB1	i1 = 1	if i5 = 0	i7 = OTX(t)
else	i1 = 0.5	else	i7 = 0
			i8 = i6 + i7
	i2 = i1 • UI(t)		
		if (F or E or DC) & Z	i9 = 1
		else	i9 = 0
if R	i3 = CB2(t)		
else	i3 = k • CB2(t) (1<k<2)	if i9=1	i10 = 0
		else	i10 = i8
if P & CG1 & CB2	i4 = 2		
else	i4 = 0	if (CG2 & CG3 & CG4)	i11 = 2
		else	i11 = 1
if UI(t)>threshold & R & i4‡0	i5 = 1		i12 = i11 • i10
else	i5 = 0		
	i6 = i4 • (i2+i3)		

"Programs built into the DNA of every animal."

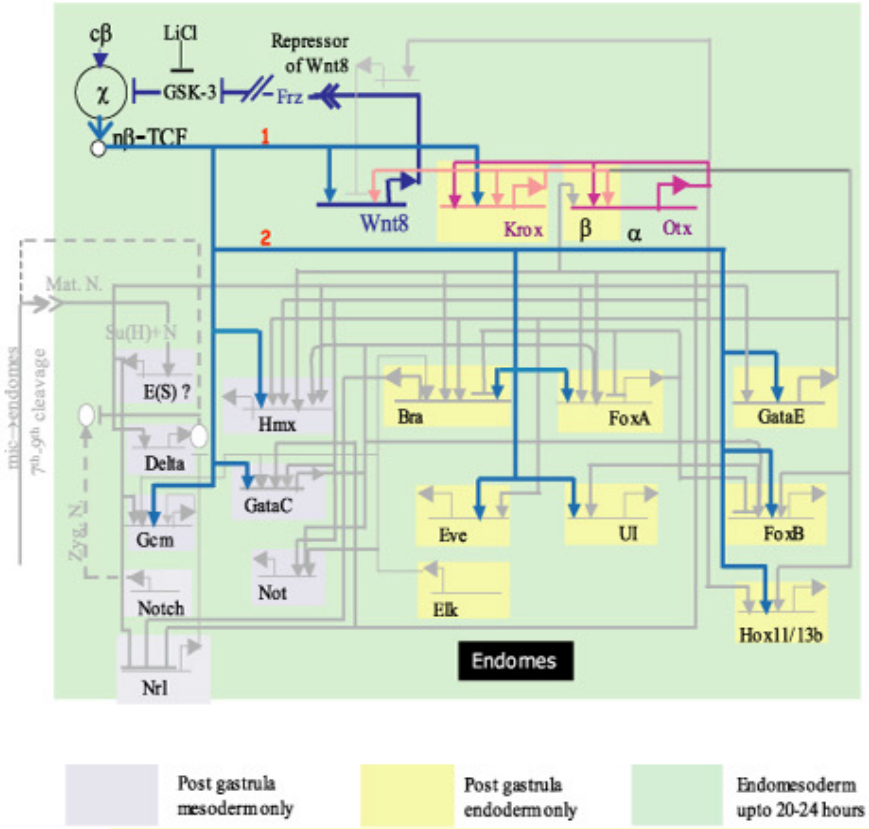
Eric H. Davidson



Genomic Regulatory Systems

The View from the Nucleus

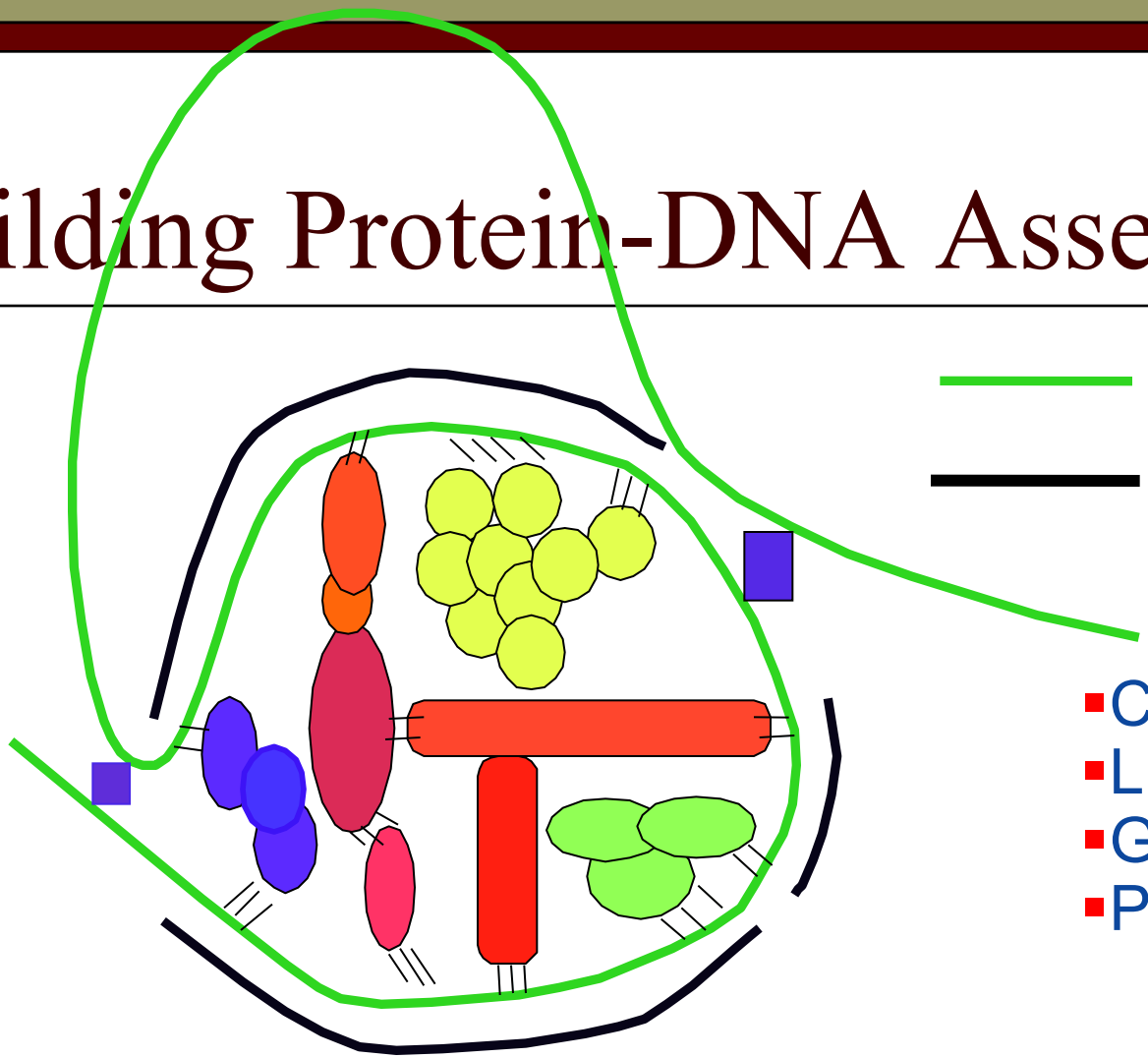
View from the nucleus: Endomesoderm nuclei to hatching blastula stage; the Wnt8/Tcf signalling loop and its genes.
Apr. 19th , 2002



Notes:

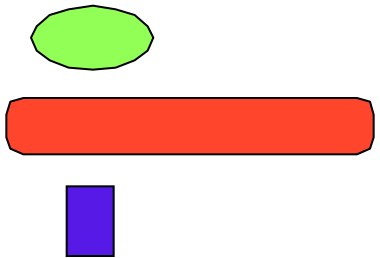
1. β-catenin/Tcf input now produced by a zygotic signaling loop driven by Wnt8 expression in endmesoderm cells.
2. β-catenin/Tcf input required for expression of many regulatory genes that become active in the veg_2 endomesodermal territory during early-mid blastula stage.

Building Protein-DNA Assemblies



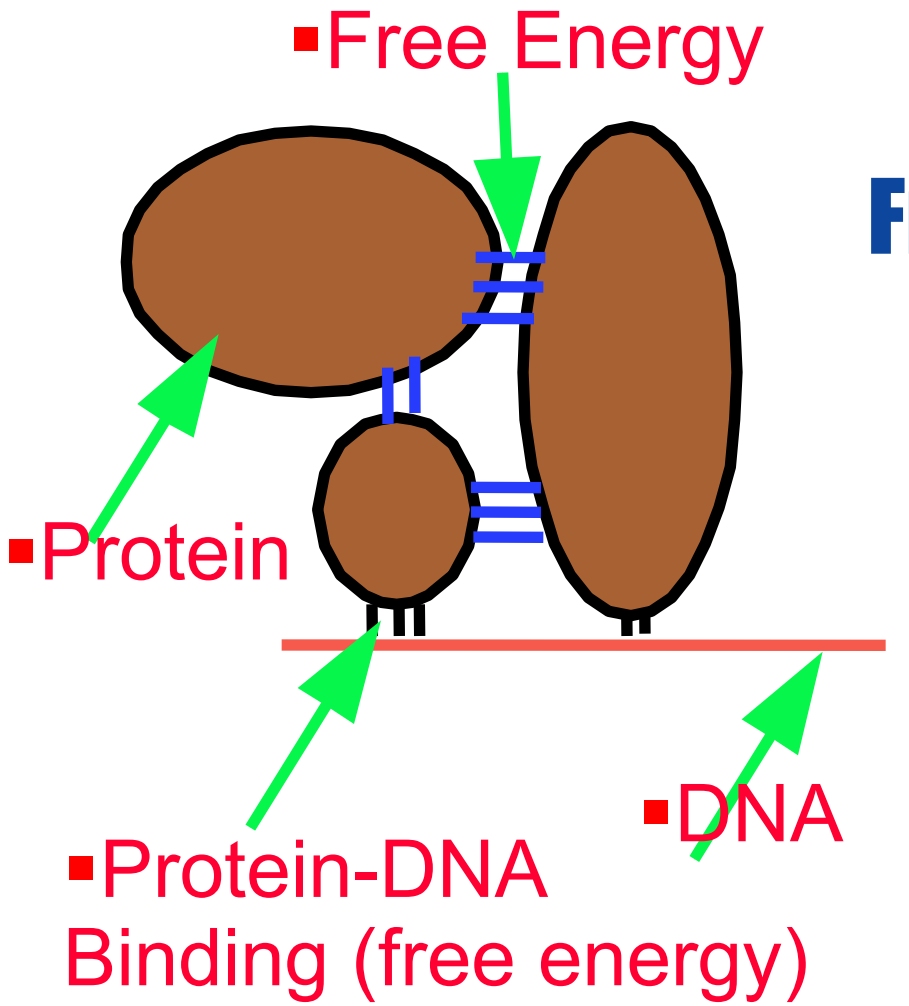
- DNA
- cismodule

- Cooperativity
- Linear-amp
- Gates
- Potentiality



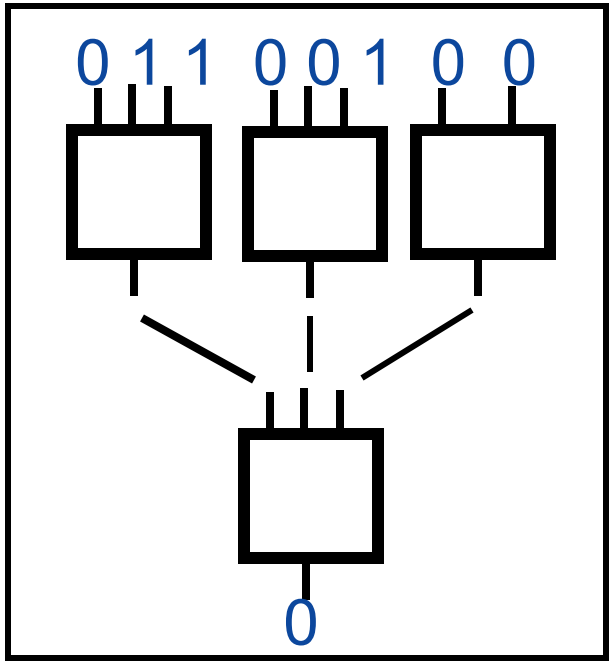
- Inter-cismodule linkage
- Insulation
- Communication

The Building Blocks

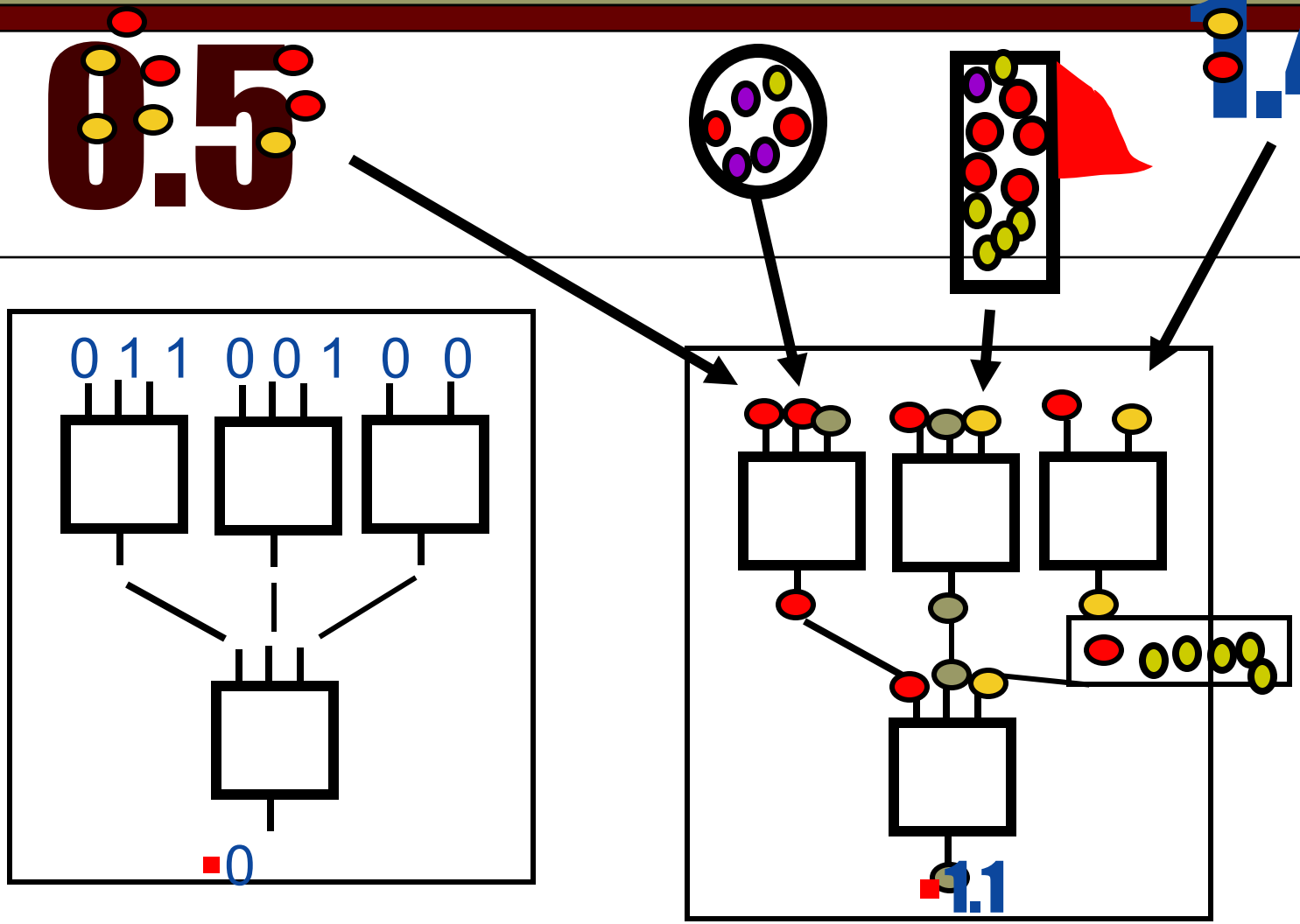


Free energy is the "GLUE"

Information Processing



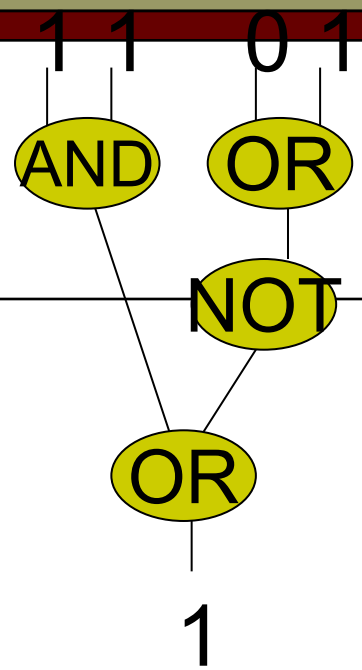
- **Boolean Circuit**
- **Synchronous** input and output
- **Completely** defined gates

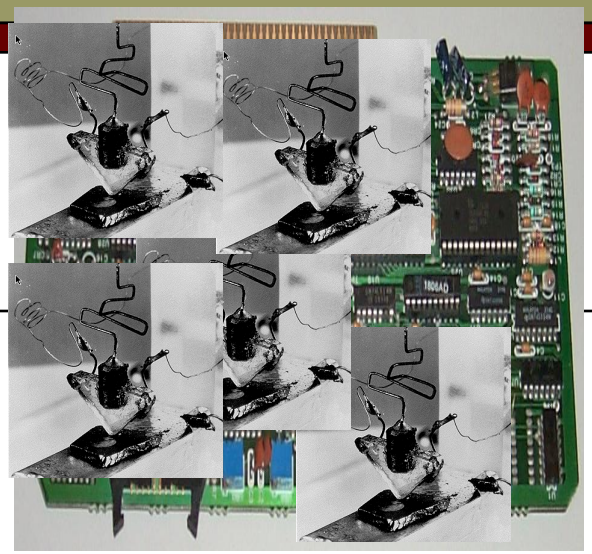


- **Boolean Circuit**
  - **Synchronous** input and output
  - **Completely** defined gates

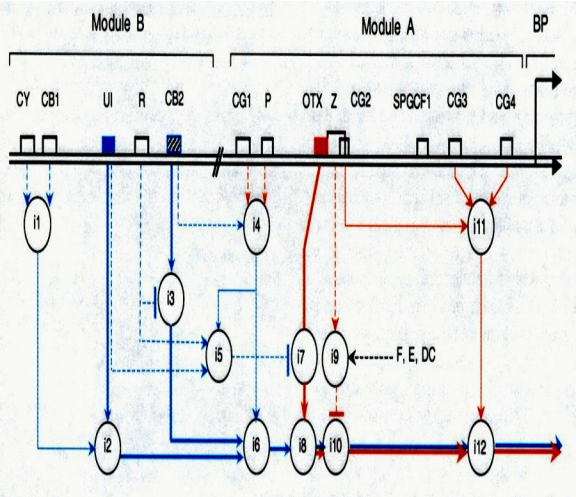
- **Boolinear Circuit**
  - **Asynchronous** input and output
  - **Incompletely** defined gates

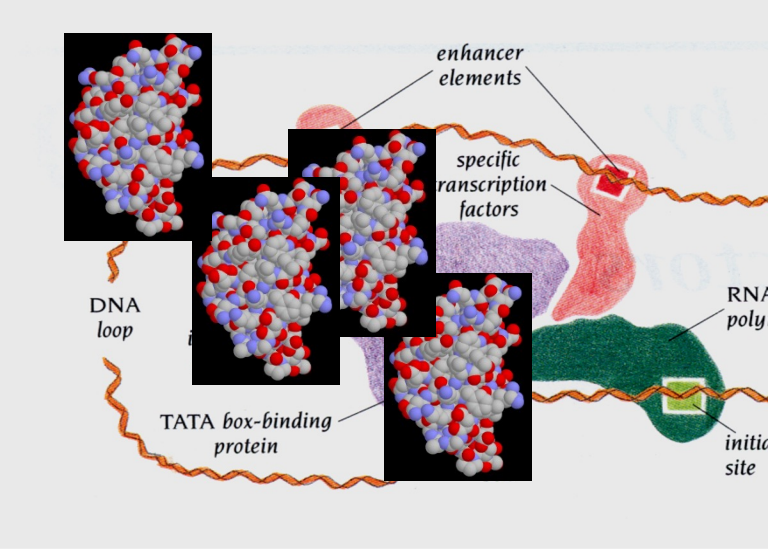
IF (x1 = 1
AND
x2= 1)
THEN
.....





GTAGGATTAAG
.....
CATCCTAATTC
.....
GTATCTAGAAG
.....





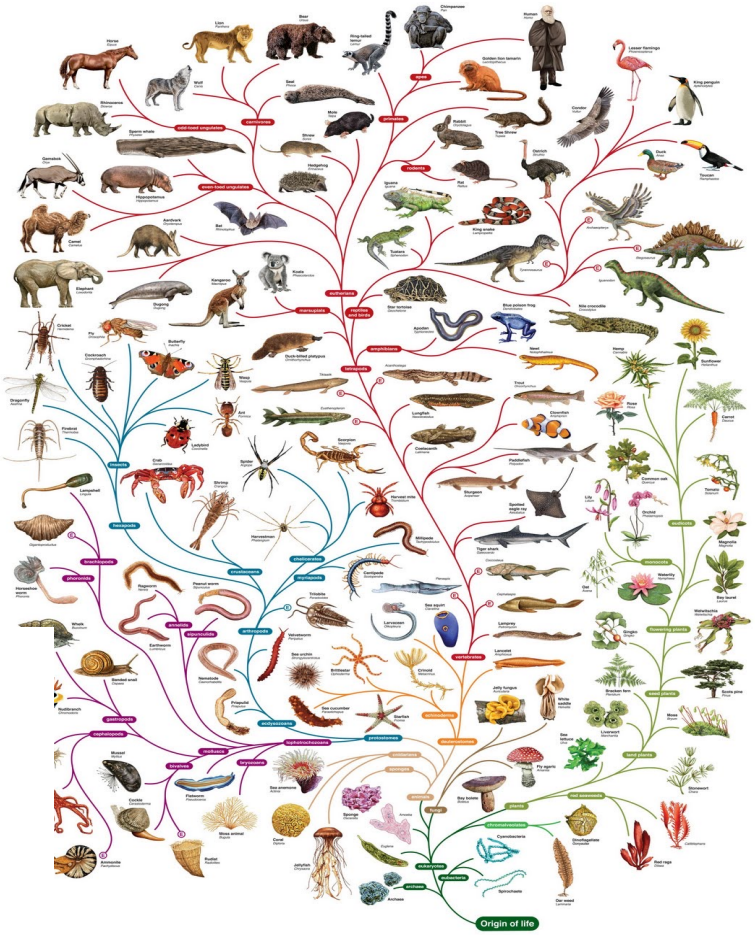
Chapter 3: Phylogenetic Trees

Image, courtesy of Vincent van Gogh Museum

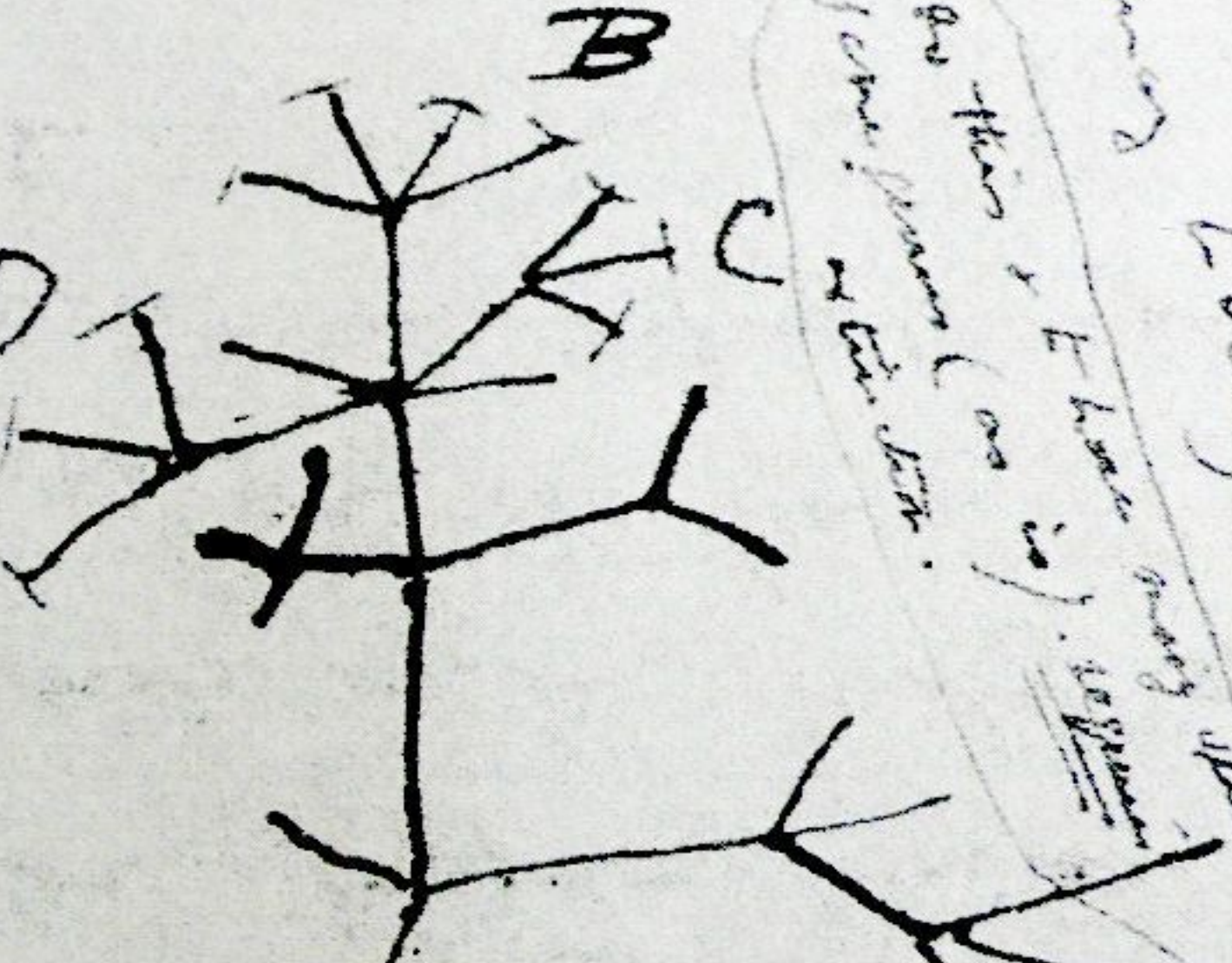
Chapter 3

Phylogenetic Trees Algorithms

Chapter 3: Phylogenetic Trees Algorithms

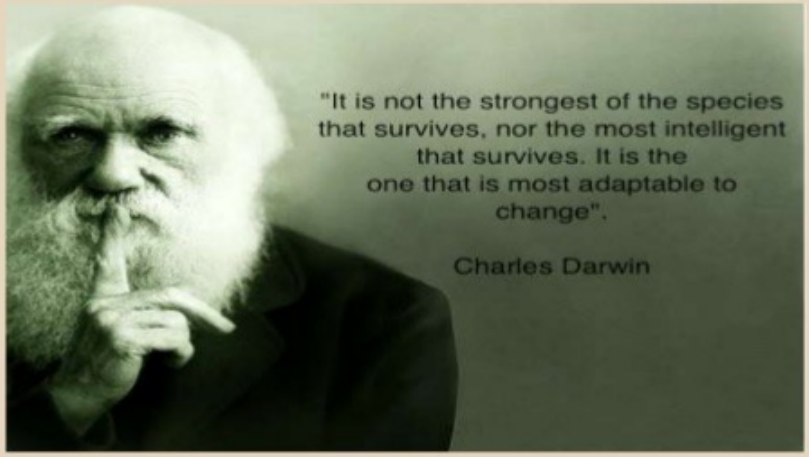


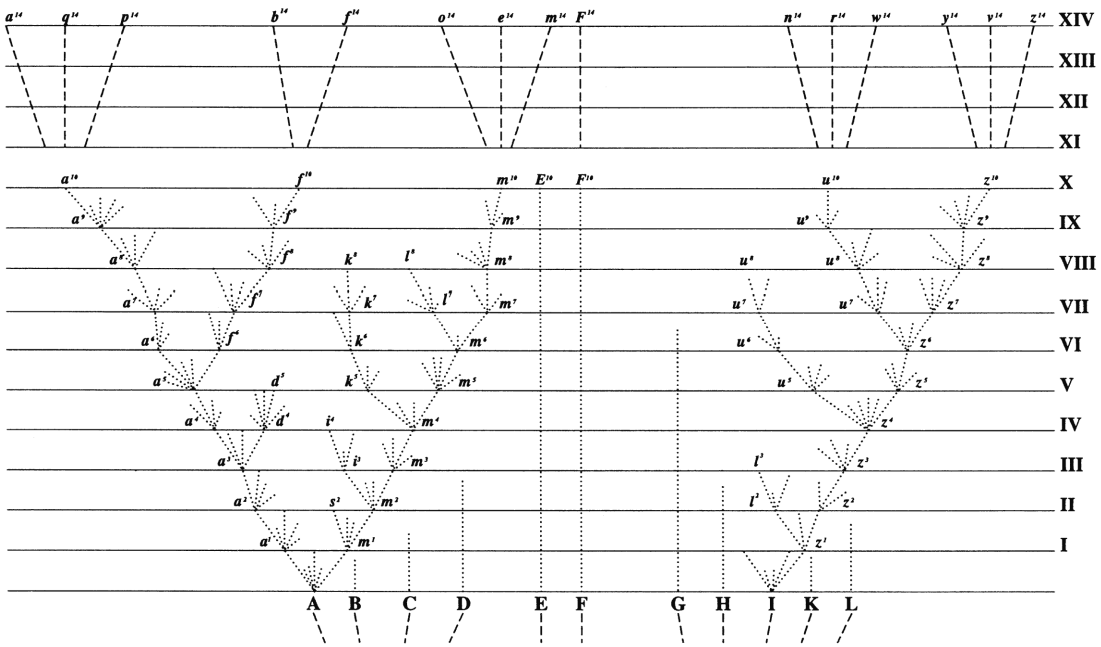
thus



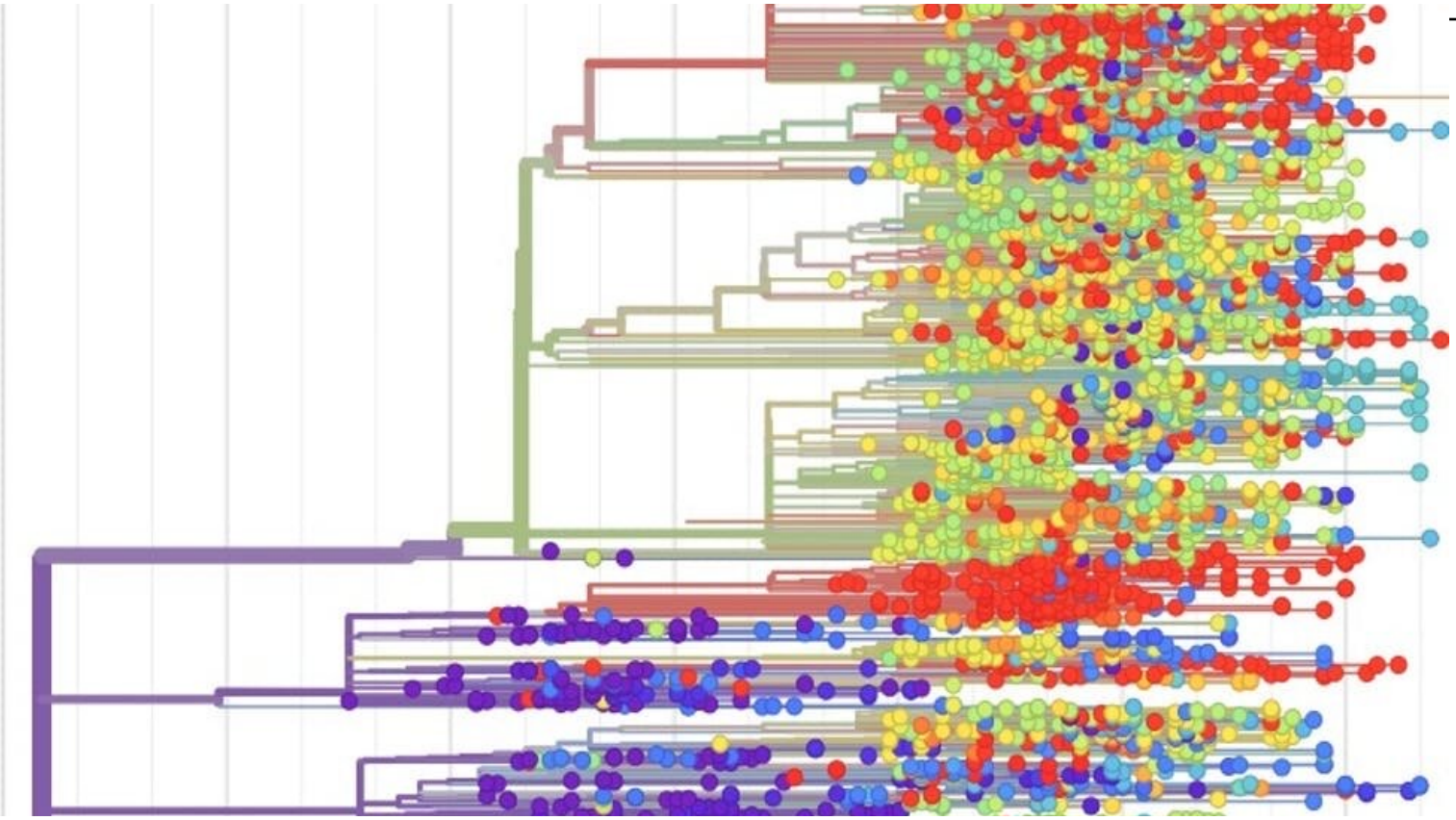
[illegible] the [illegible] on
[illegible] to
[illegible] them
[illegible] on
[illegible] lines
[illegible] on
[illegible] specimen
[illegible] to
[illegible] as
[illegible] specimen
[illegible] with

CHARLES DARWIN QUOTE





SARS-CoV-2 (COVID) phylogenetic tree

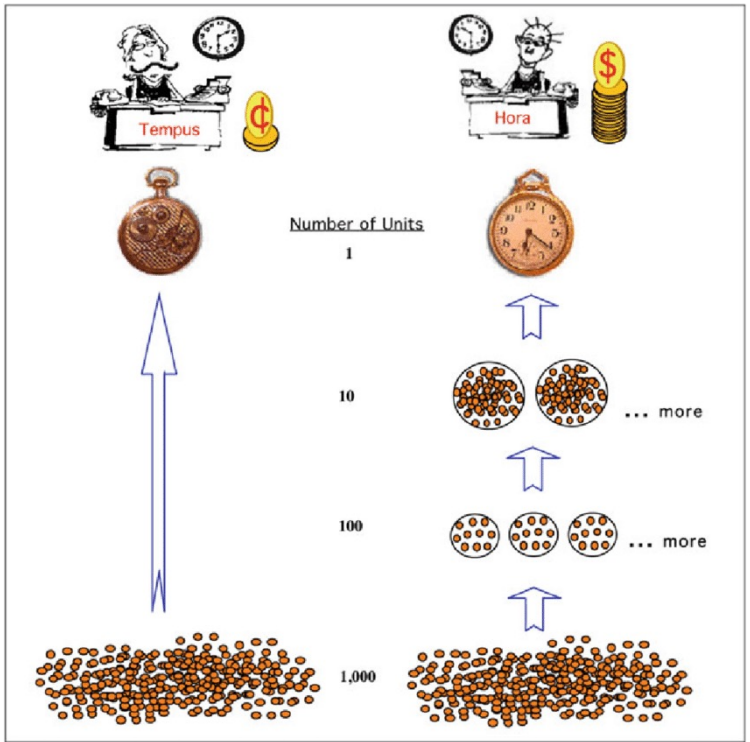


The SARS-CoV-2 phylogenetic tree – the family tree that shows the evolution of all the sequenced coronavirus samples worldwide.

Herbert Simon's Parable on Evolution "The Parable of the Two Watchmakers"

A mathematical theory of "interruptions"

How can we quantify the Speed of Evolution?



- **Two fine watchmakers**: Hora and Tempus, who were making equally fine watches, each consisting of 1000 parts. Their popularity attracted many phone calls from customers, interrupting their work and making the partly assembled watch at hand fall apart.
- **The math**: Assuming that the probability of an interrupton occurring while a part is being added to an assembly is 0.01, Hora made 111 times as many complete assemblies per watch as Tempus; Tempus had to make 20,000 as many attempts per completed assembly as Hora.
- **The outcome**: Hora became a rich man, but Tempus went bankrupt.

HORA TEMPUS

Chapter 4: Hidden Markov Models

Image, courtesy of Vincent van Gogh Museum

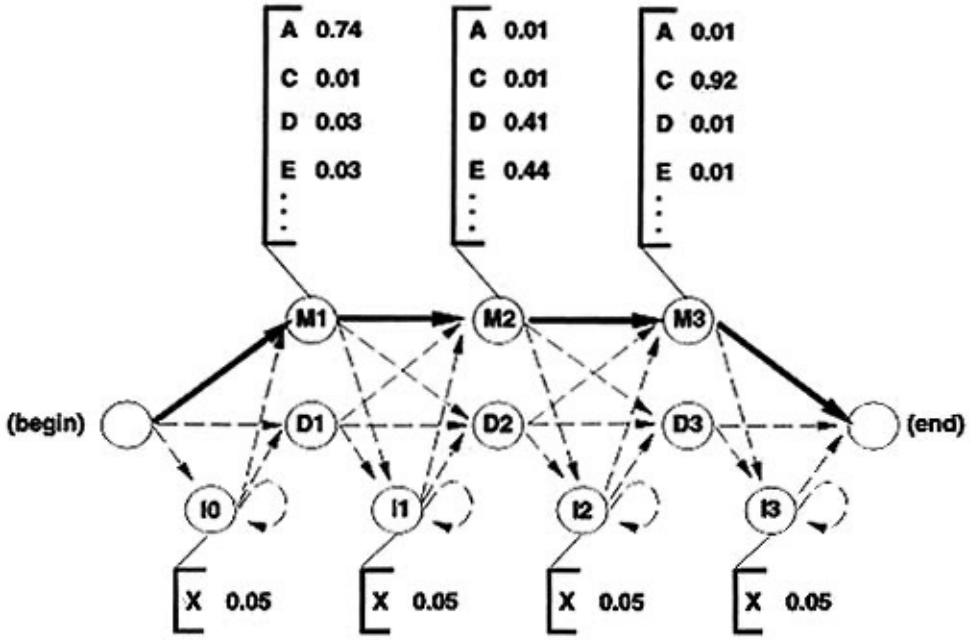
Chapter 4

Machine Learning Methods:

Hidden Markov Models Algorithms

Chapter 4: Hidden Markov Chains

Algorithms

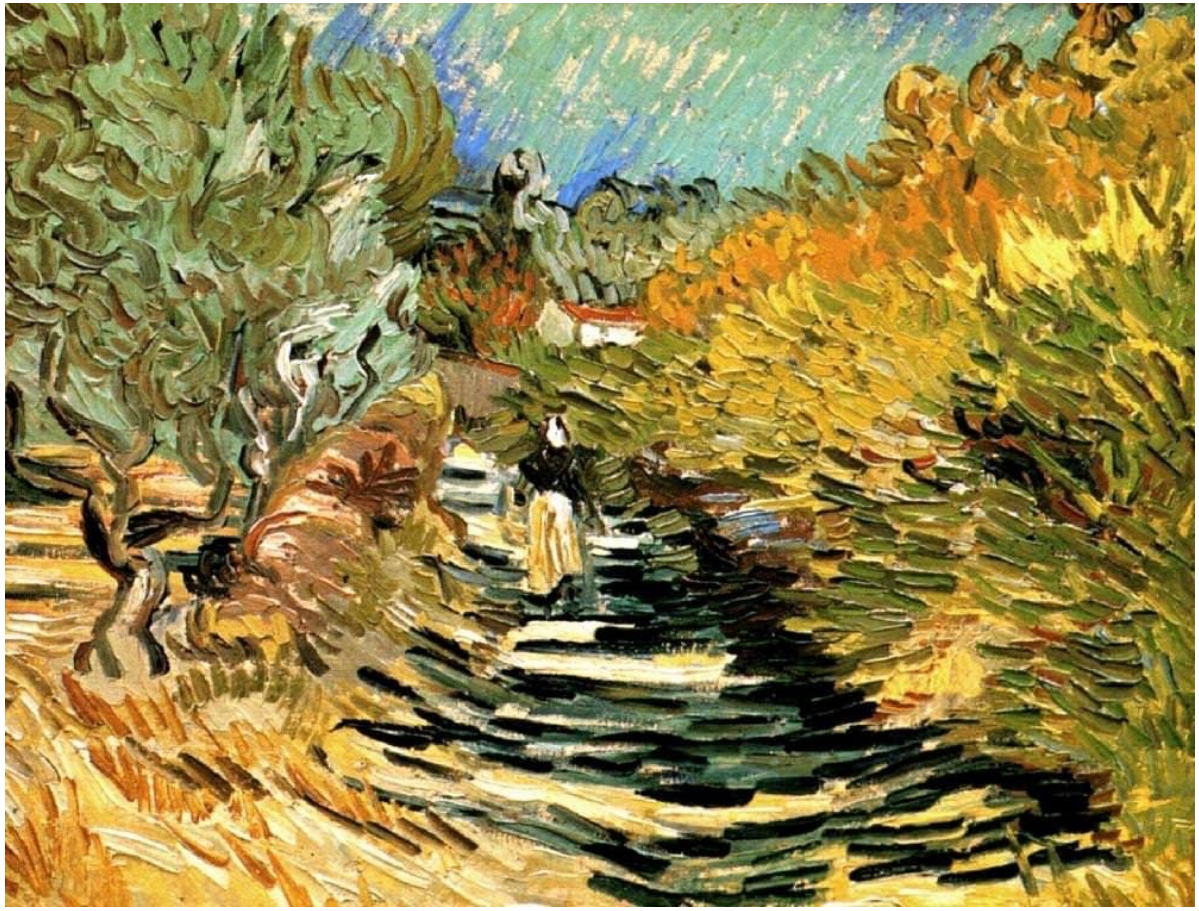


Gene finding in a genome using HMMs algorithms

"For one rational line or true sentence there are thousands of nonsense cacophonies, mountains of verbal trash and incoherencies."

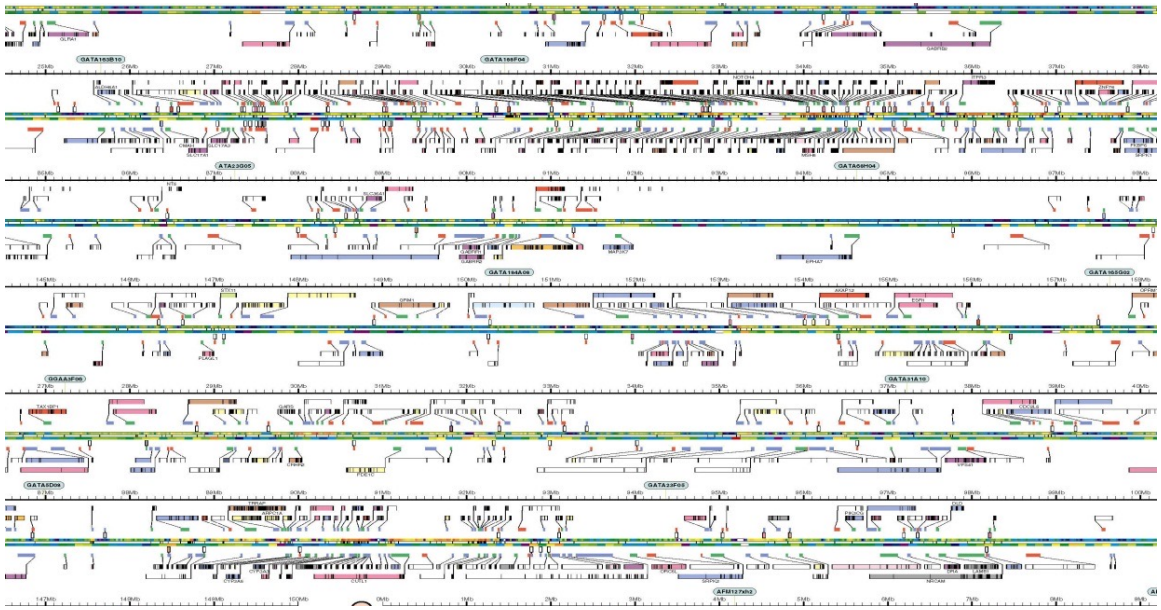
Jorge Luis Borges

Chapter 5: Genome Assembly Algorithms: An Introduction



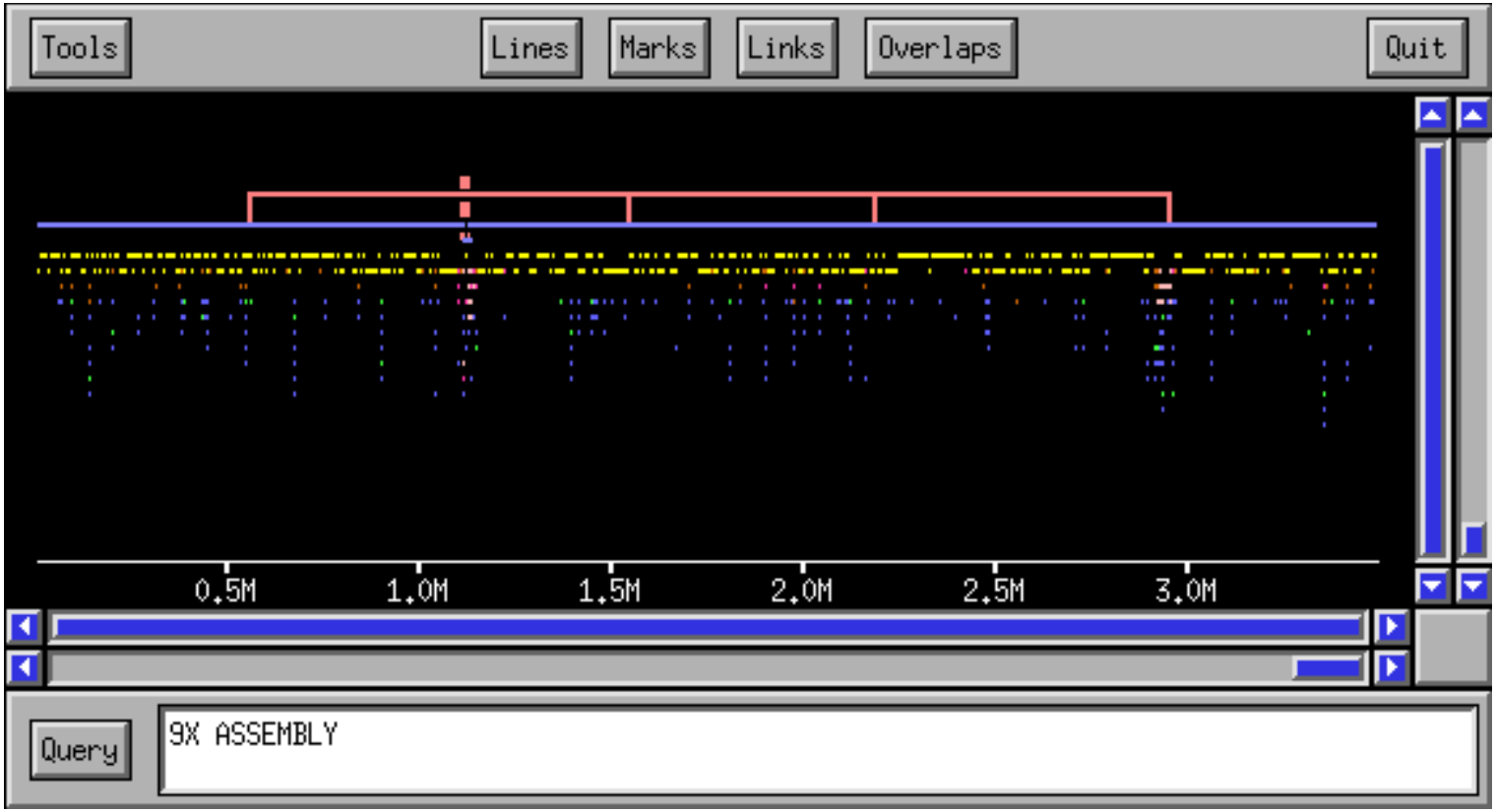
Image, courtesy of Vincent van Gogh Museum

Chapter 5: Genome Assembly Algorithms

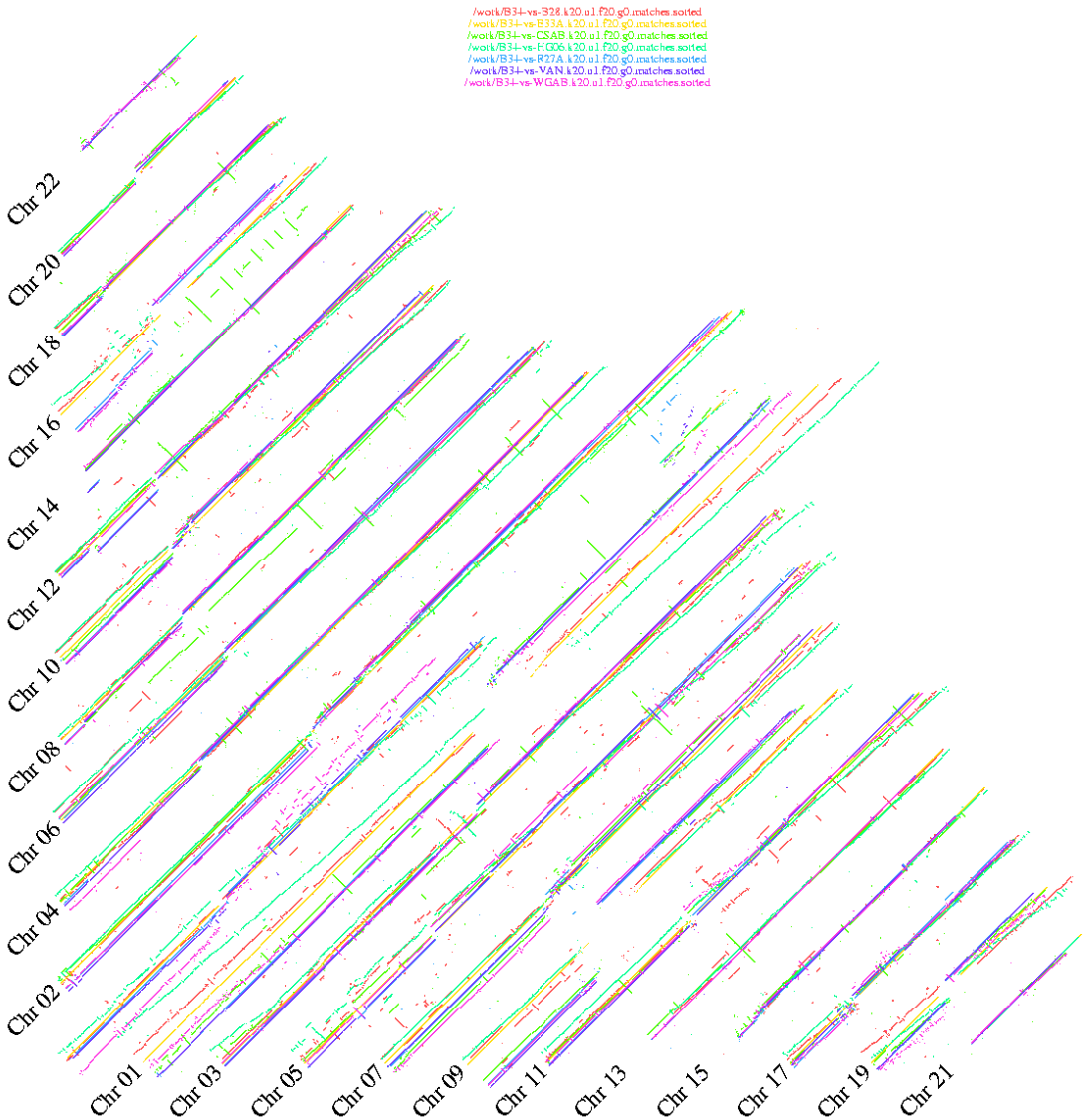


Genome Assembly Algorithm

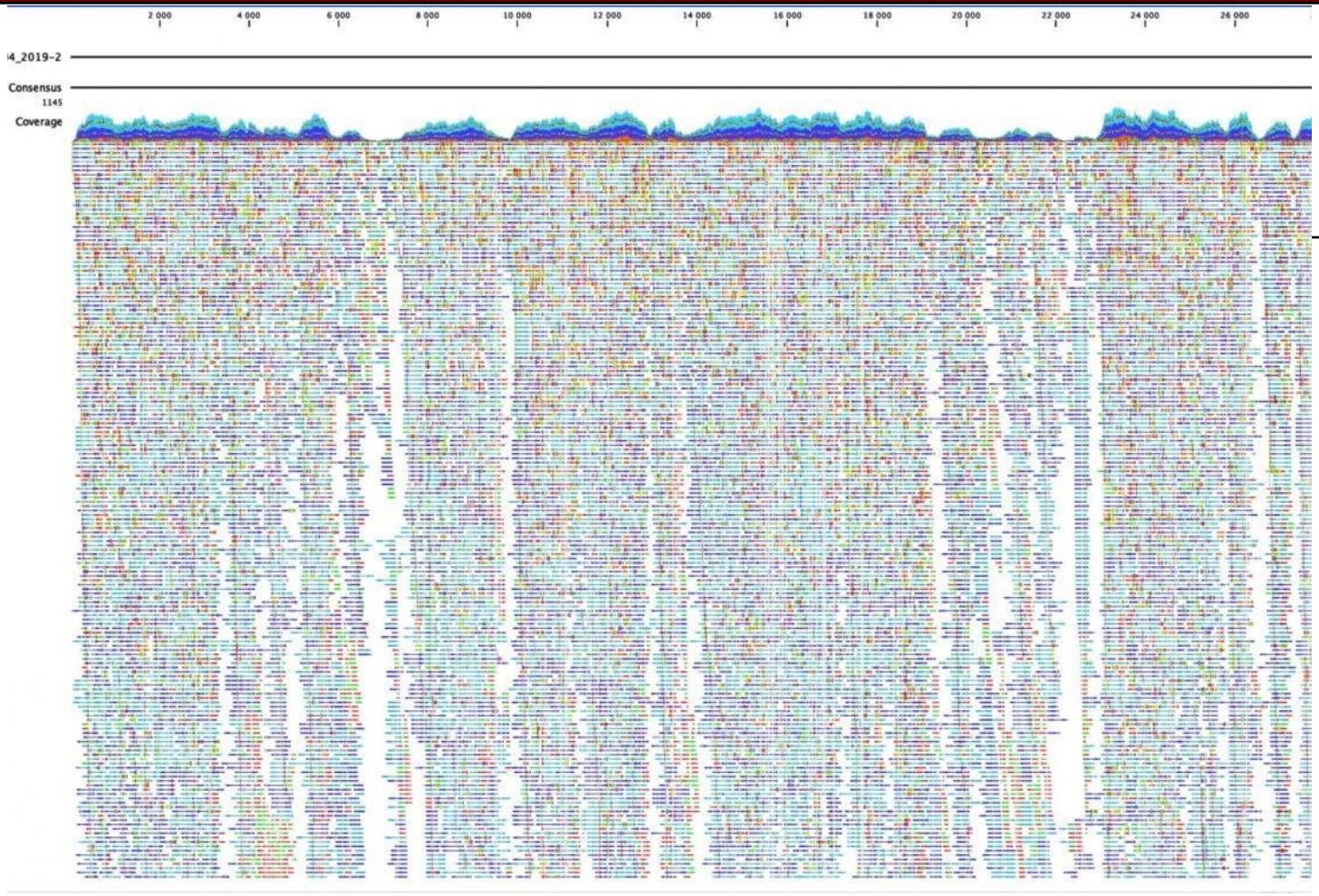
Celera Assembler



The Father of All Dot Plots



The Human Genome



Whole genome sequence of the 2019-nCoV coronavirus, in one of the first French cases, made at the Institut Pasteur (Paris), using a unique Platform (P2M), open to all French National Reference Centers. Credit: Institut Pasteur/CNR of respiratory infection viruses.

Chapter 6: Genomic Privacy

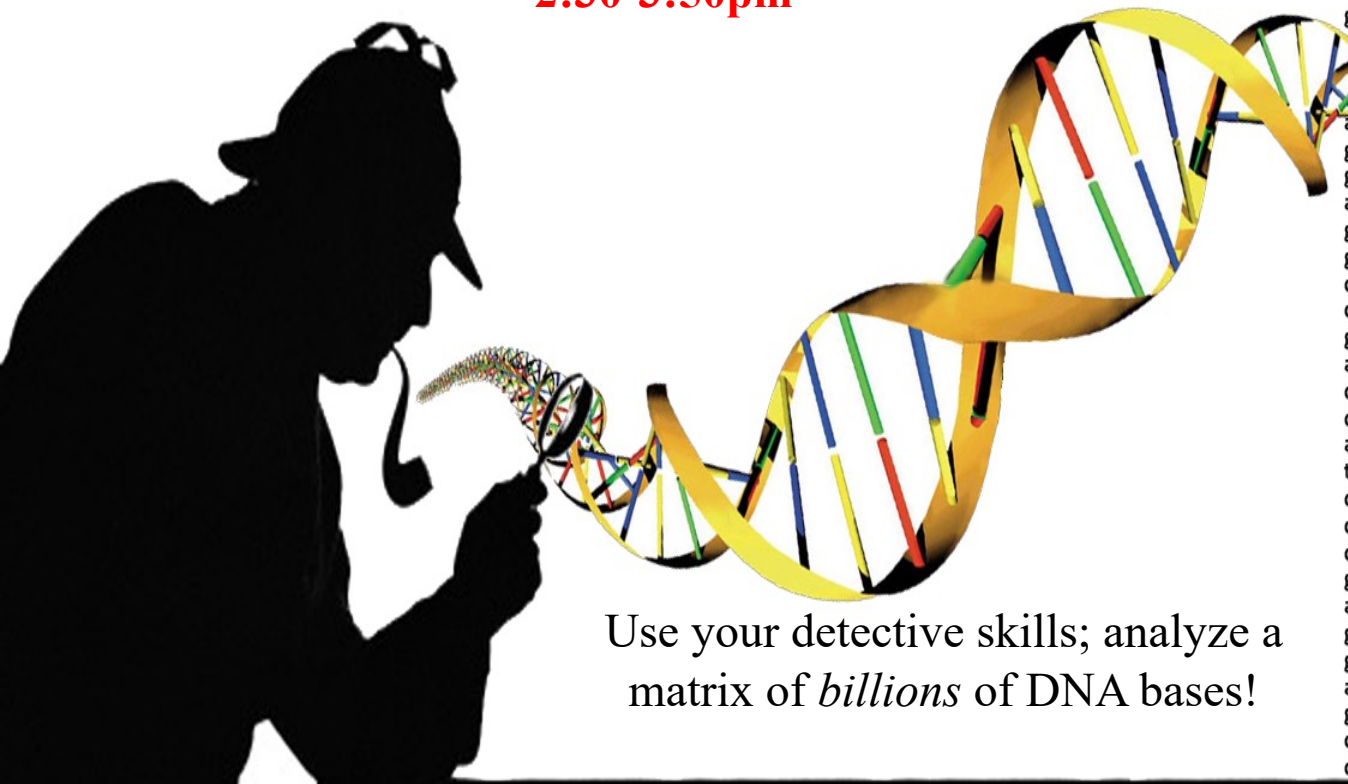
Image, courtesy of Vincent van Gogh Museum

HOMER's attack:

Genomic privacy studies on Genome-Wide Association Studies (GWAS) were first introduced as the well-known Homer's attack (2008) that showed that publicly released GWAS statistics can be used to estimate a GWAS participant's disease status from knowing her/his genotypes at certain risk factors.

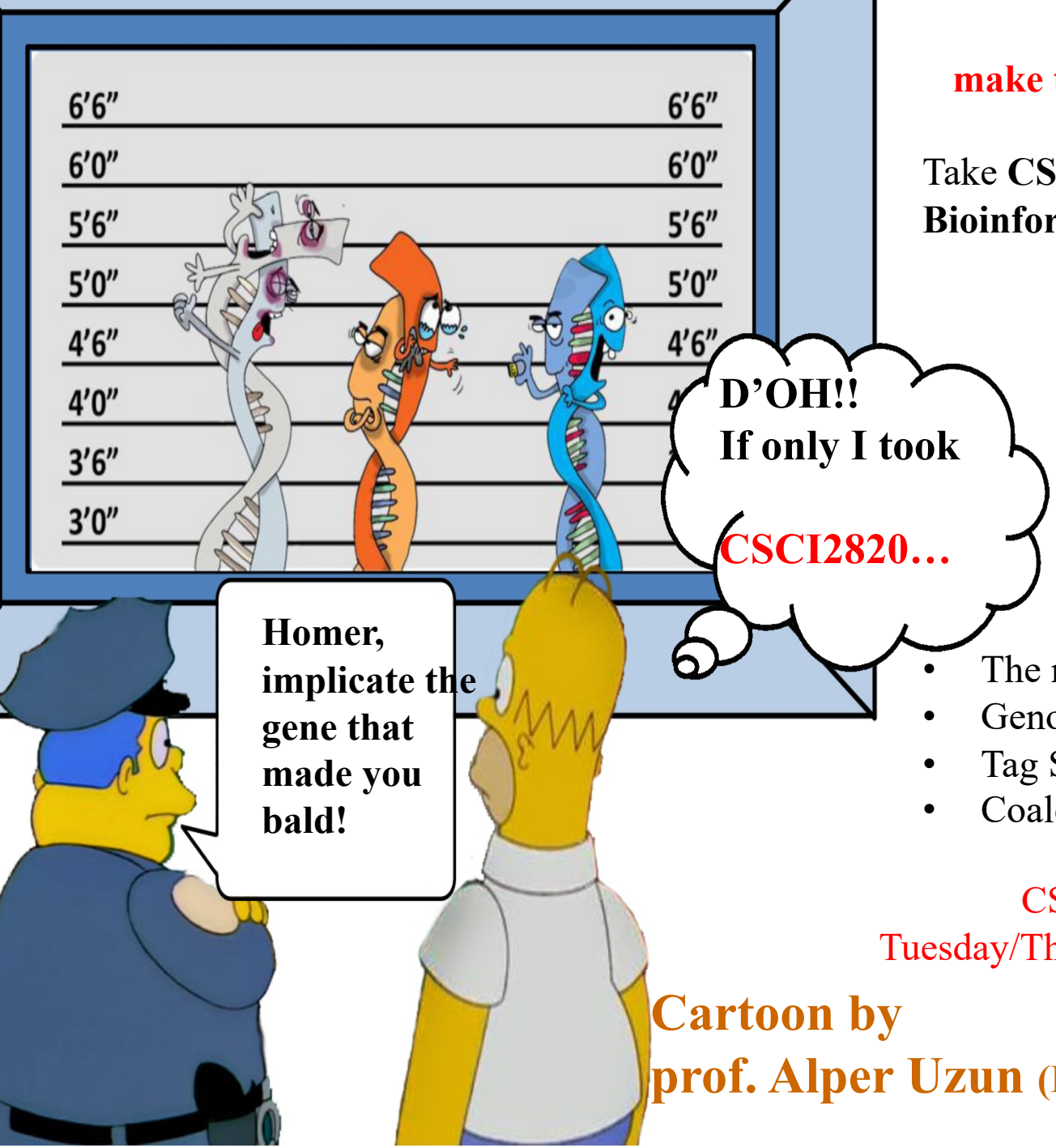
Enjoy working on large data?
Solve mysteries in your spare time?
Are you hard-boiled?

Work to uncover the mysteries of complex disease!
Take **CSCI2820: Medical Bioinformatics** in the **Fall**
Tuesday/Thursday
2:30-3:50pm



Use your detective skills; analyze a matrix of *billions* of DNA bases!

Featuring private investigator Sorin Istrail



Do not make the same mistake as Homer

Take **CSCI2820: Medical Bioinformatics** and study the genetics of complex disease.

Topics include

- Hypothesis testing
- Haplotype phasing
- The missing heritability problem
- Genome-wide Association Studies
- Tag SNP selection
- Coalescent Theory

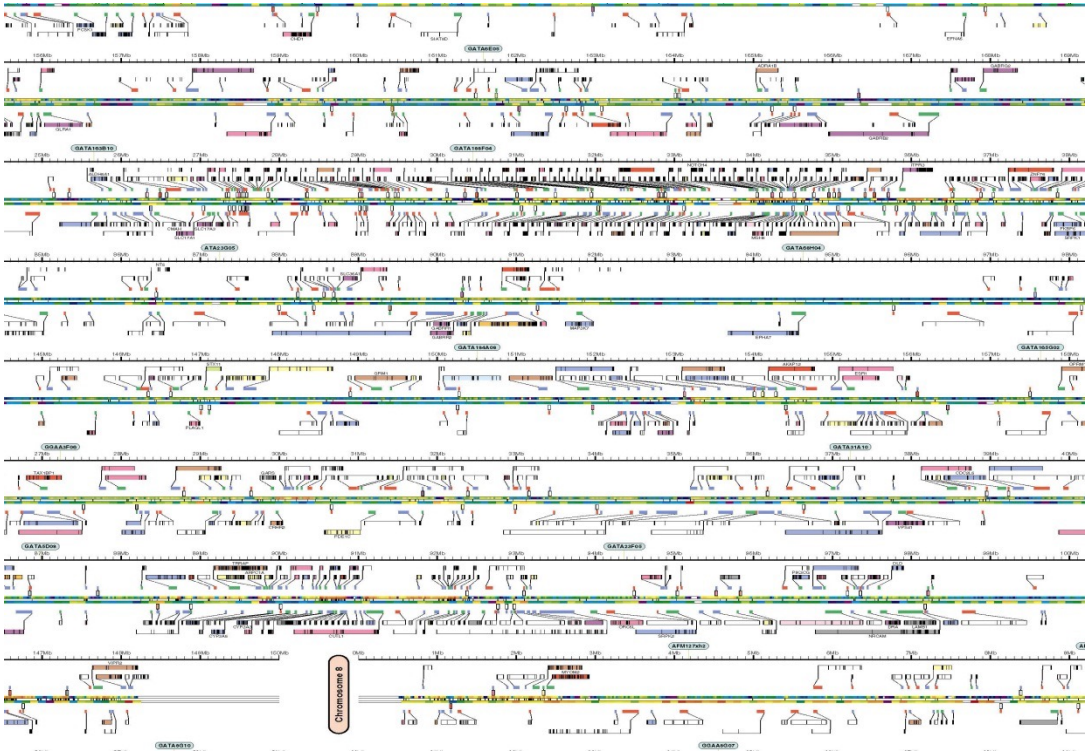
CSCI2820
Tuesday/Thursday 2:30-3:50

Cartoon by prof. Alper Uzun (Brown Medical School)

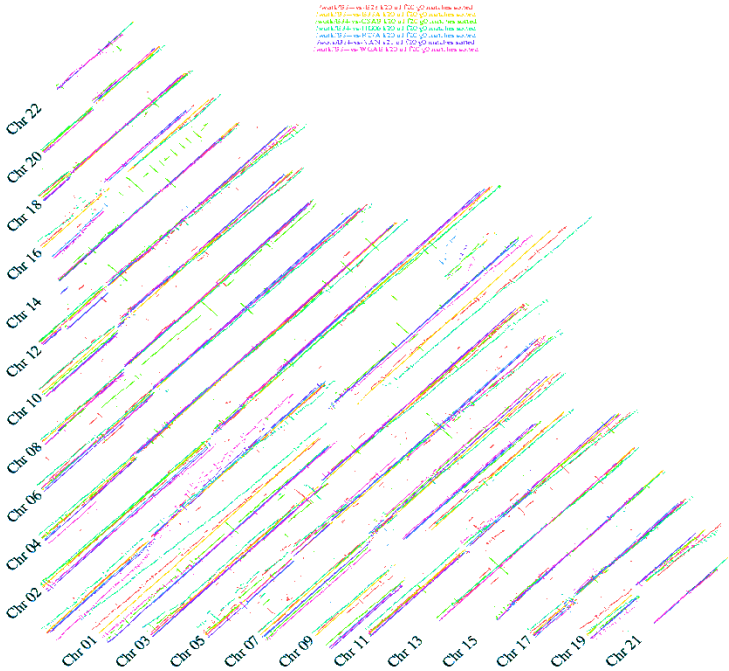
CSCI1820 Algorithmic Foundations of Computational Biology

http://www.cs.brown.edu/courses/csci1820/

Prof. Sorin Istrail



"The Sequence of the Human Genome" Science, 2001



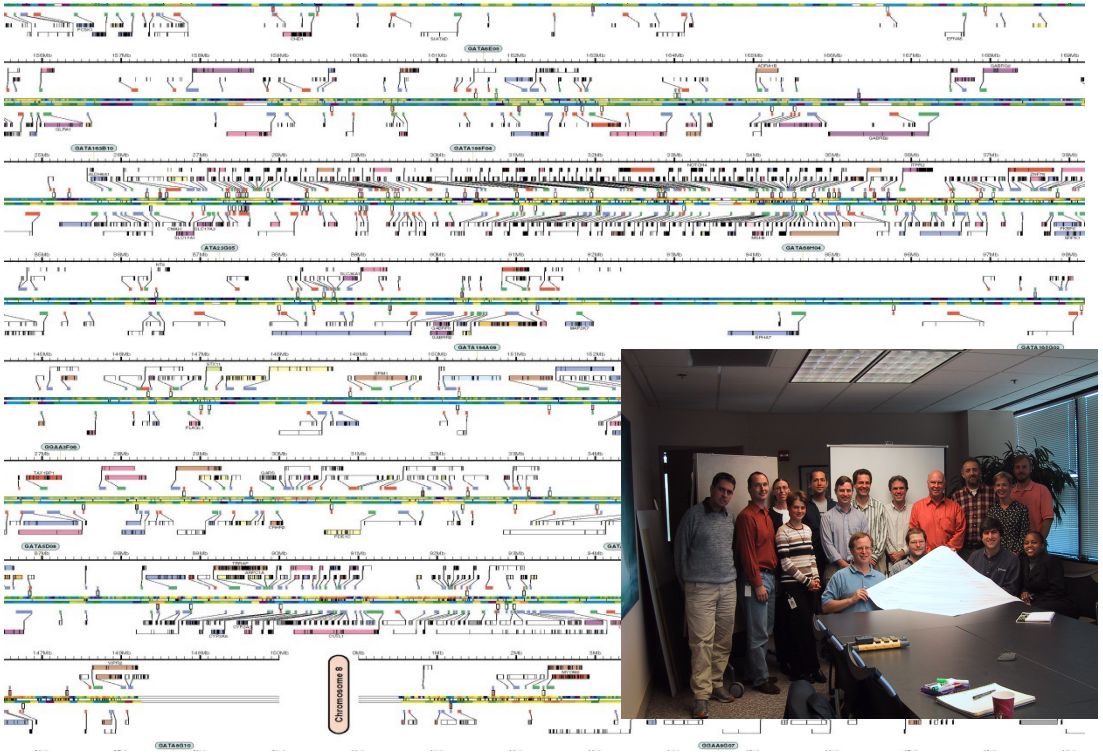
"Whole-genome shotgun assembly and comparison of human genome assemblies" Proc. Nat. Acad. Sci. USA, 2004


Science
16 February 2001
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

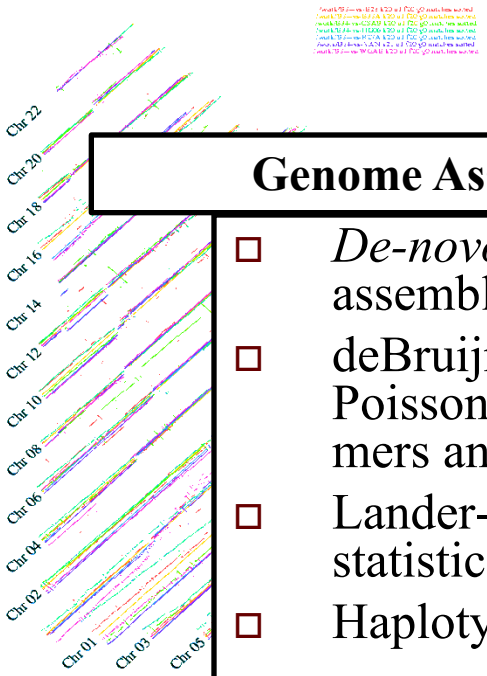
CSCI1820 Algorithmic Foundations of Computational Biology

http://www.cs.brown.edu/courses/csci1820/

Prof. Sorin Istrail



"The Sequence of the Human Genome" Science, 2001



Genome Assembly

- *De-novo* genome assembly algorithms
- deBruijn graphs and Poisson theory of k-mers and NGS
- Lander-Waterman statistical theory
- Haplotype assembly

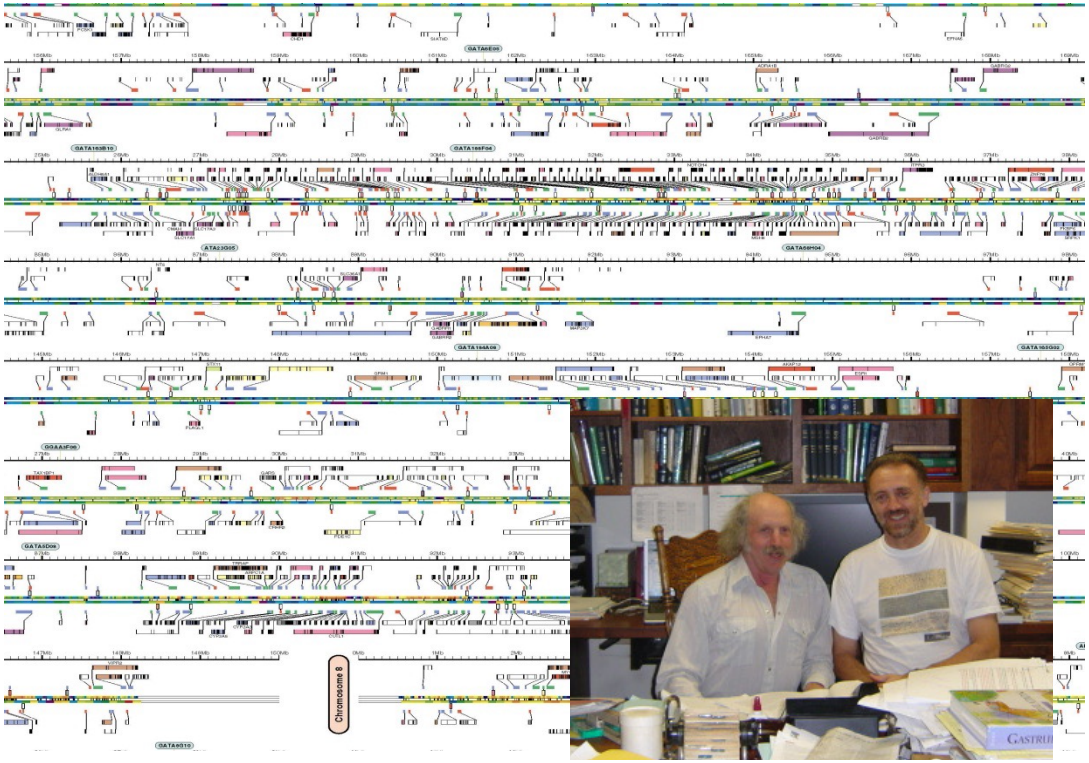
"Whole-genome shotgun assembly and comparison of human genome assemblies" Proc. Nat. Acad. Sci. USA, 2004

CSCI1820 Algorithmic Foundations of Computational Biology

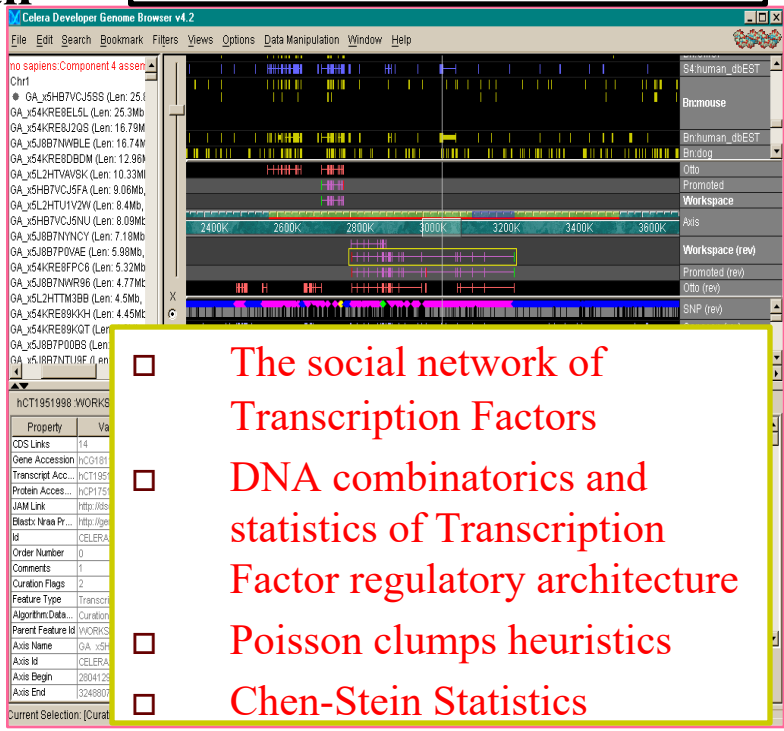
http://www.cs.brown.edu/courses/csci1820/

Prof. Sorin Istrail

The Regulatory Genome



Eric Davidson and Sorin working on "Logic functions of the genetic cis-regulatory code"



- The social network of Transcription Factors
- DNA combinatorics and statistics of Transcription Factor regulatory architecture
- Poisson clumps heuristics
- Chen-Stein Statistics

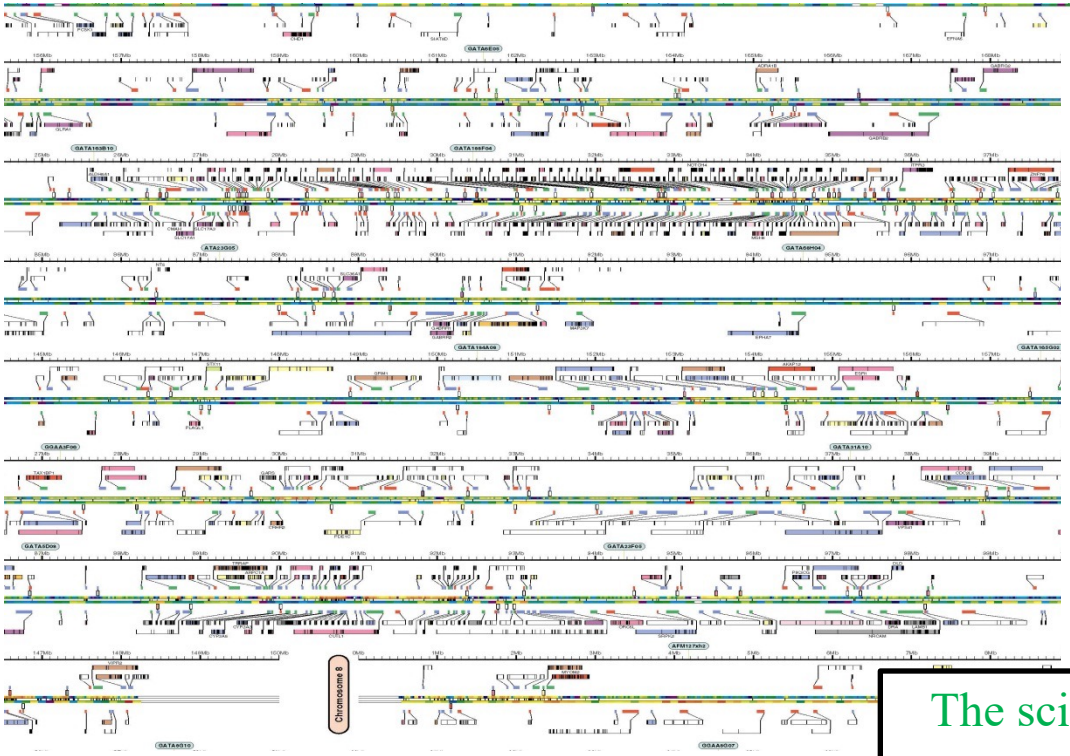
The cis-Regulatory CYRENE Genome Browser

CSCI1820 Algorithmic Foundations of Computational Biology

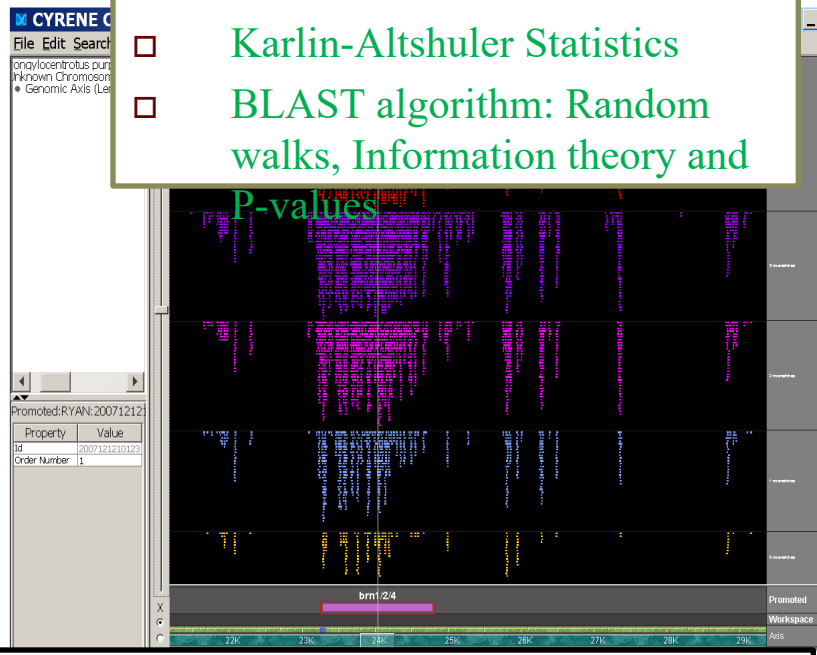
http://www.cs.brown.edu/courses/csci1820/

Prof. Sorin Istrail

- Suffix trees in linear time
- Burrows-Wheeler transform
- Karlin-Altshuler Statistics
- BLAST algorithm: Random walks, Information theory and P-values



"The Sequence of the Human Genome" Science, 2001



The science and art of mapping DNA fragments, genes, and genomes to genomes

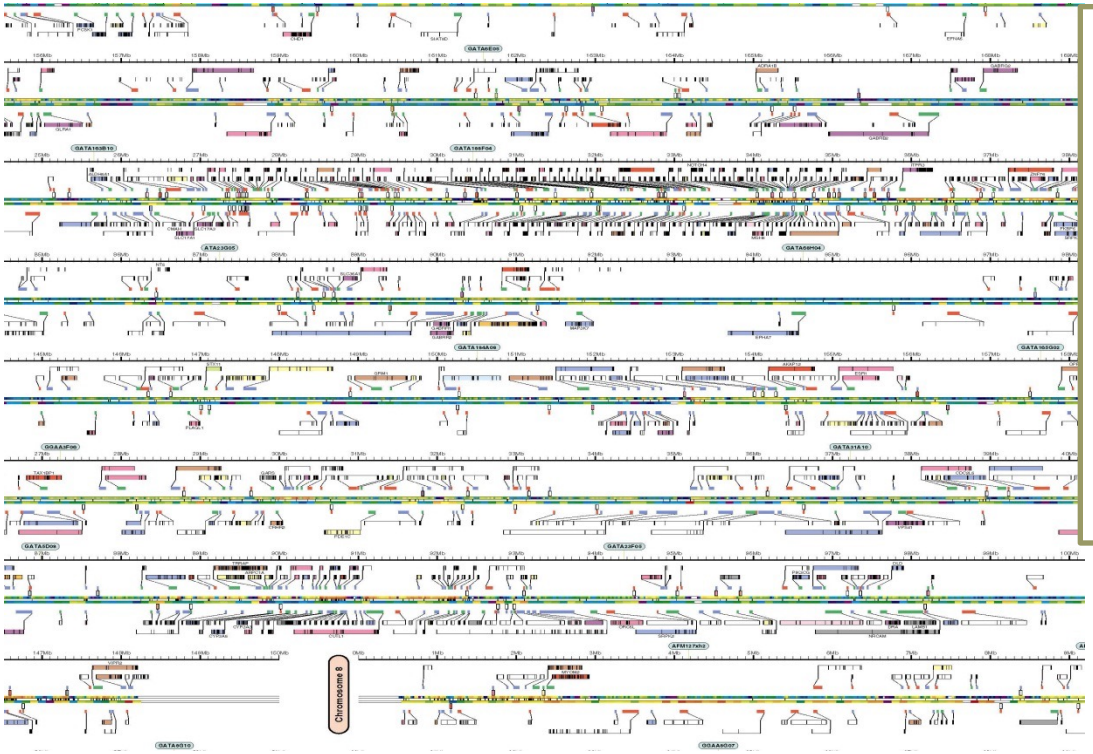
CSCI1820 Algorithmic Foundations of Computational Biology

http://www.cs.brown.edu/courses/csci1820/

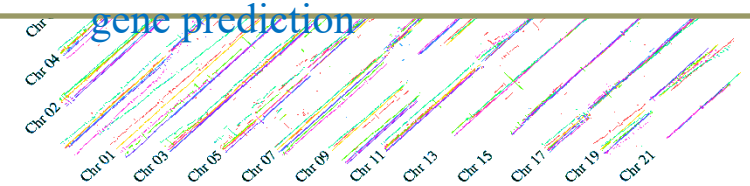
Prof. Sorin Istrail

Topics include

- Genome sequencing and assembly: algorithms and statistical theory
- BLAST algorithms and statistical theory of alignment and searching
- Mapping reads and genomes to genomes
- DNA combinatorics and statistical theory of regulatory regions of genes
- Hidden Markov model algorithms and gene prediction



"The Sequence of the Human Genome" Science, 2001



"Whole-genome shotgun assembly and comparison of human genome assemblies" Proc. Nat. Acad. Sci. USA, 2004

CSCI2840 Advanced Algorithms in Computational Biology and Medical Bioinformatics

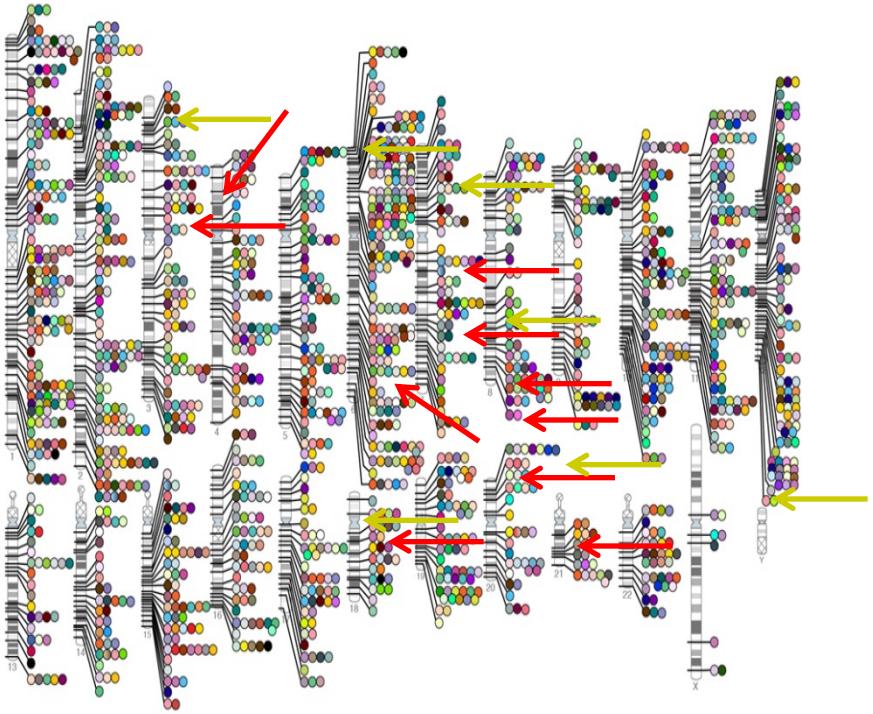
Genome-wide Association Studies (GWAS)

Prof. Sorin Istrail

Published Genome-Wide Associations through 2011

1,617 published GWA at $p \leq 5X10^{-8}$ for 249 traits

The GWAS Human Genome



Autism marker

Multiple Sclerosis Marker

- Abdominal aortic aneurysm
- Acute lymphoblastic leukemia
- Adhesion molecules
- Adiponectin levels
- Age-related macular degeneration
- AIDS progression
- Alcohol dependence
- Alopecia areata
- Alzheimer disease
- Amyloid A levels
- Amyotrophic lateral sclerosis
- Angiotensin-converting enzyme activity
- Ankylosing spondylitis
- Arterial stiffness
- Asparagus anosmia
- Asthma
- Atherosclerosis in HIV
- Atrial fibrillation
- Attention deficit hyperactivity disorder
- Autism
- Basal cell cancer
- Behcet's disease
- Bipolar disorder
- Biliary atresia
- Bilirubin
- Bitter taste response
- Birth weight
- Bladder cancer
- Bleomycin sensitivity
- Blond or brown hair
- Blood pressure
- Blue or green eyes
- BMI, waist circumference
- Bone density
- Breast cancer
- Butyrylcholinesterase levels
- C-reactive protein
- Calcium levels
- Cardiac structure/function
- Cardiovascular risk factors
- Carnitine levels
- Carotenoid/tocopherol levels
- Carotid atherosclerosis
- Celiac disease
- Celiac disease and rheumatoid arthritis
- Cerebral atrophy measures
- Chronic lymphocytic leukemia
- Chronic myeloid leukemia
- Cleft lip/palate
- Coffee consumption
- Cognitive function
- Conduct disorder
- Colorectal cancer
- Corneal thickness
- Coronary disease
- Cortical thickness
- Creutzfeldt-Jakob disease
- Crohn's disease
- Crohn's disease and celiac disease
- Cutaneous nevi
- Cystic fibrosis severity
- Dermatitis
- DHEA-s levels
- Diabetic retinopathy
- Dilated cardiomyopathy
- Drug-induced liver injury
- Drug-induced liver injury (amoxicillin-clavulanate)
- Endometrial cancer
- Endometriosis
- Eosinophil count
- Eosinophilic esophagitis
- Epirubicin-induced leukopenia
- Erectile dysfunction and prostate cancer treatment
- Erythrocyte parameters
- Esophageal cancer
- Essential tremor
- Exfoliation glaucoma
- Eye color traits
- F cell distribution
- Fibrinogen levels
- Folate pathway vitamins
- Follicular lymphoma
- Fuch's corneal dystrophy
- Freckles and burning
- Gallstones
- Gastric cancer
- Glioma
- Glycemic traits
- Graves disease
- Hair color
- Hair morphology
- Handedness in dyslexia
- HDL cholesterol
- Heart failure
- Heart rate
- Height
- Hemostasis parameters
- Hepatic steatosis
- Hepatitis
- Hepatitis B vaccine response
- Hepatocellular carcinoma
- Hirschsprung's disease
- HIV-1 control
- Hodgkin's lymphoma
- Homocysteine levels
- HPV seropositivity
- Hypospadias
- Idiopathic pulmonary fibrosis
- IFN-related cytopeni
- IgA levels
- IgE levels
- Inflammatory bowel disease
- Insulin-like growth factors
- Intracranial aneurysm
- Iris color
- Iron status markers
- Ischemic stroke
- Juvenile idiopathic arthritis
- Keloid
- Kidney stones
- LDL cholesterol
- Leprosy
- Leptin receptor levels
- Liver enzymes
- Longevity
- LP (a) levels
- LpPLA(2) activity and mass
- Lung cancer
- Magnesium levels
- Major mood disorders
- Malaria
- Male pattern baldness
- Mammographic density
- Matrix metalloproteinase levels
- MCP-1
- Melanoma
- Menarche & menopause
- Meningioma
- Meningococcal disease
- Metabolic syndrome
- Migraine
- Moyamoya disease
- Multiple sclerosis
- Myeloproliferative neoplasms
- Myopia (pathological)
- N-glycan levels
- Narcolepsy
- Nasopharyngeal cancer
- Natriuretic peptide levels
- Neuroblastoma
- Nicotine dependence
- Obesity
- Open angle glaucoma
- Open personality
- Optic disc parameters
- Osteoarthritis
- Osteoporosis
- Otosclerosis
- Other metabolic traits
- Ovarian cancer
- Pancreatic cancer
- Pain
- Paget's disease
- Panic disorder
- Parkinson's disease
- Periodontitis
- Peripheral arterial disease
- Personality dimensions
- Phosphatidylcholine levels
- Phosphorus levels
- Photic sneeze
- Phytosterol levels
- Platelet count
- Polycystic ovary syndrome
- Primary biliary cirrhosis
- Primary sclerosing cholangitis
- PR interval
- Progranulin levels
- Progressive supranuclear palsy
- Prostate cancer
- Protein levels
- PSA levels
- Psoriasis
- Psoriatic arthritis
- Pulmonary funct. COPD
- QRS interval
- QT interval
- Quantitative traits
- Recombination rate
- Red vs.non-red hair
- Refractive error
- Renal cell carcinoma
- Renal function
- Response to antidepressants
- Response to antipsychotic therapy
- Response to carbamazepine
- Response to clopidogrel therapy
- Response to hepatitis C treat
- Response to interferon beta therapy
- Response to metaformin
- Response to statin therapy
- Restless legs syndrome
- Retinal vascular caliber
- Retinol levels
- Rheumatoid arthritis
- Ribavirin-induced anemia
- Schizophrenia
- Serum metabolites
- Skin pigmentation
- Smoking behavior
- Speech perception
- Sphingolipid levels
- Statin-induced myopathy
- Stevens-Johnson syndrome
- Stroke
- Sudden cardiac arrest
- Suicide attempts
- Systemic lupus erythematosus
- Systemic sclerosis
- T-tau levels
- Tau AB1-42 levels
- Telomere length
- Testicular germ cell tumor
- Thyroid cancer
- Thyroid volume
- Tooth development
- Total cholesterol
- Triglycerides
- Tuberculosis
- Type 1 diabetes
- Type 2 diabetes
- Ulcerative colitis
- Urate
- Urinary albumin excretion
- Urinary metabolites
- Uterine fibroids
- Venous thromboembolism
- Ventricular conduction
- VEGF levels
- Vertical cup-disc ratio
- Vitamin B12 levels
- Vitamin D insufficency
- Vitamin E levels
- Vitiligo
- Warfarin dose
- Weight
- White cell count
- White matter hyperintensity
- YKL-40 levels

CSCI2820 Advanced Algorithms in Computational Biology and Medical Bioinformatics

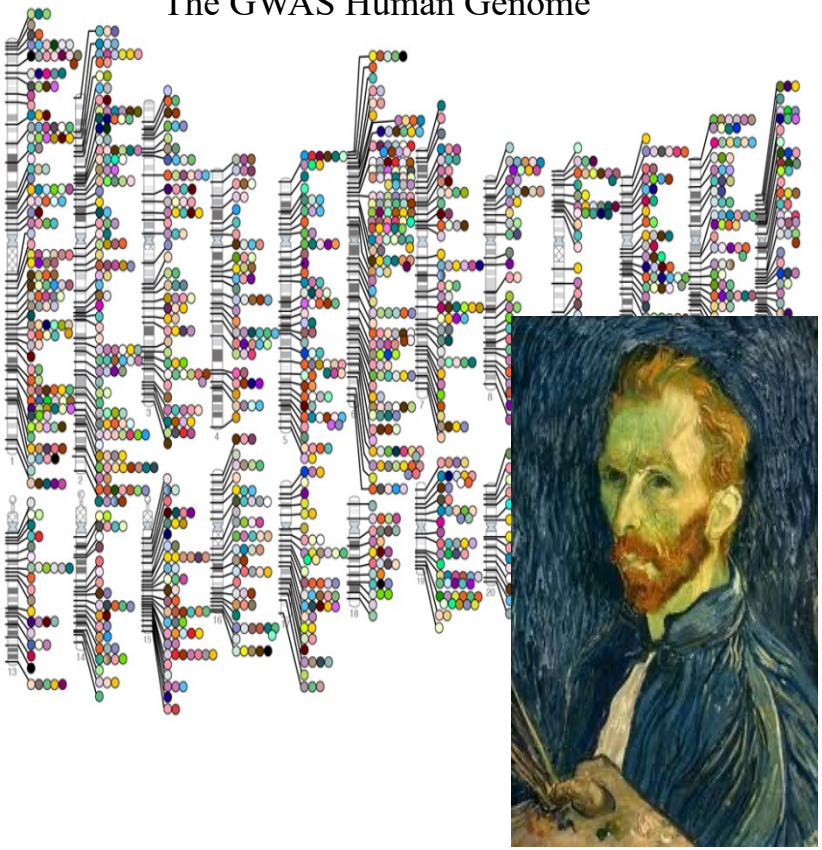
Genome-wide Association Studies (GWAS)

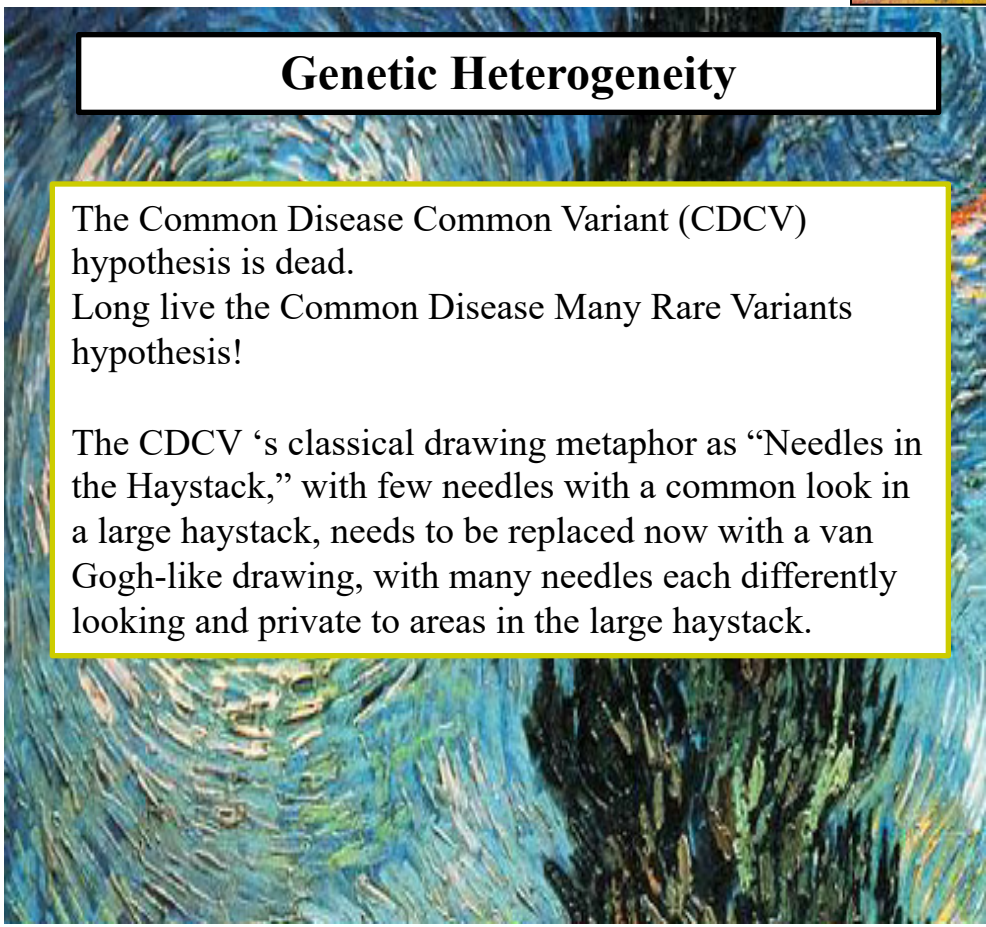
Prof. Sorin Istrail

Published Genome-Wide Associations through 2011

1,617 published GWA at $p \leq 5X10^{-8}$ for 249 traits

The GWAS Human Genome





Genetic Heterogeneity

The Common Disease Common Variant (CDCV) hypothesis is dead.
Long live the Common Disease Many Rare Variants hypothesis!

The CDCV 's classical drawing metaphor as "Needles in the Haystack," with few needles with a common look in a large haystack, needs to be replaced now with a van Gogh-like drawing, with many needles each differently looking and private to areas in the large haystack.

Vincent

CSCI2820 Advanced Algorithms in Computational Biology and Medical Bioinformatics

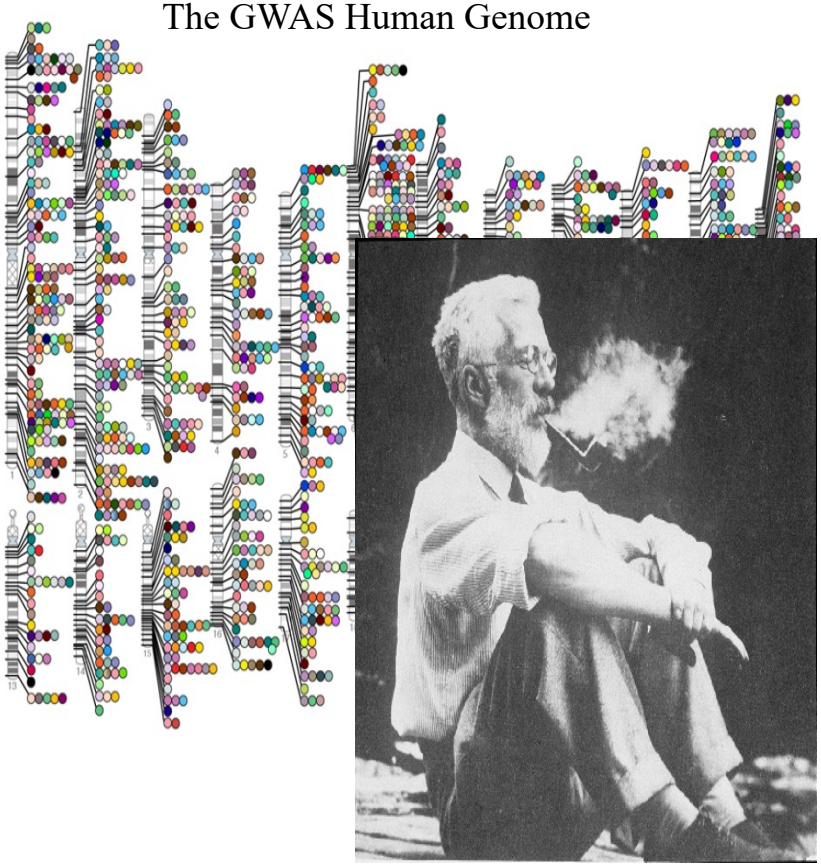
Genome-wide Association Studies (GWAS)

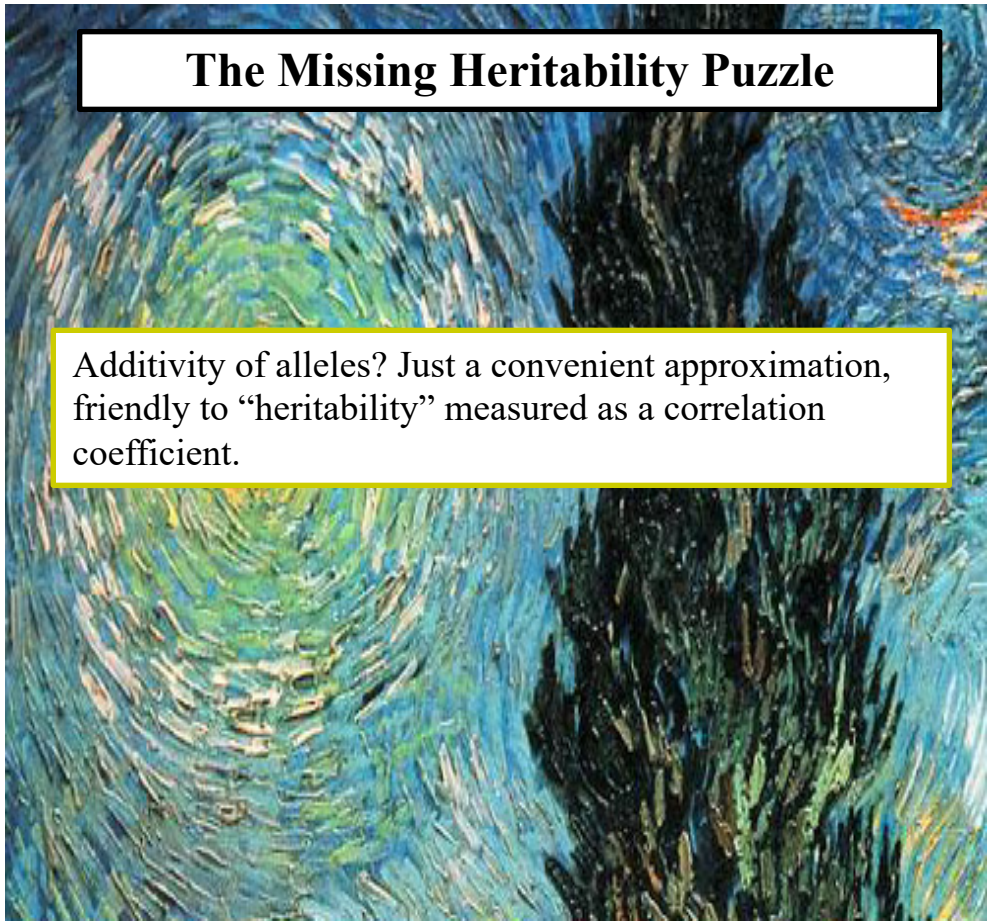
Prof. Sorin Istrail

Published Genome-Wide Associations through 2011

1,617 published GWA at $p \leq 5X10^{-8}$ for 249 traits

The GWAS Human Genome





Ronald

CSCI2820 Advanced Algorithms in Computational Biology and Medical Bioinformatics

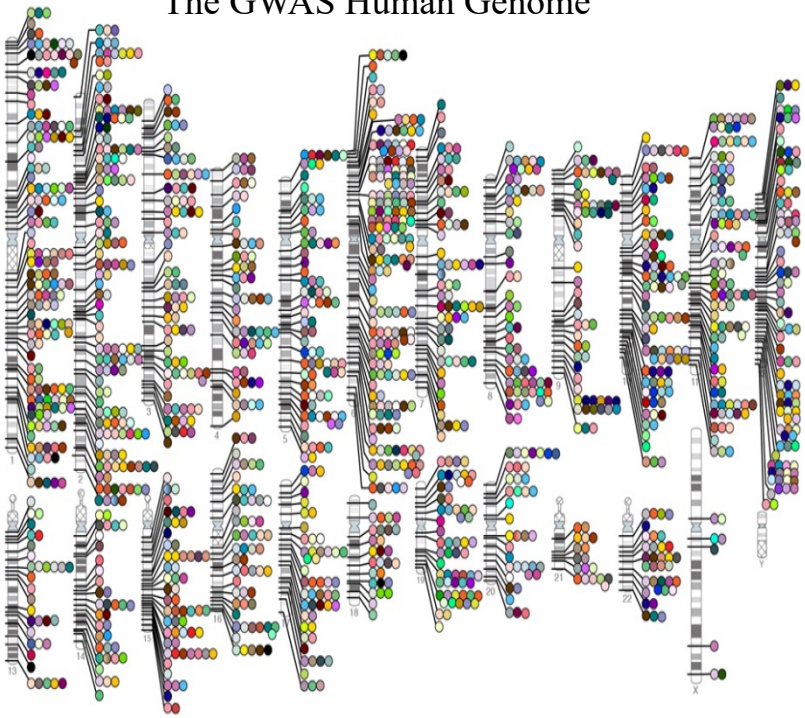
Genome-wide Association Studies (GWAS)

Prof. Sorin Istrail

Published Genome-Wide Associations through 2011

1,617 published GWA at $p \leq 5X10^{-8}$ for 249 traits

The GWAS Human Genome



Topics include

- haplotype phasing, linkage disequilibrium, tagging SNPs, identical by descent (IBD), pedigrees, trios
- coalescent theory, Polya urn game, Ewens sampling lemma, genome-wide graph theory algorithms
- the genetic heterogeneity problem, the missing heritability problem
- statistical models of disease, association tests and multiple hypothesis testing
- autism, multiple sclerosis, type 2 diabetes

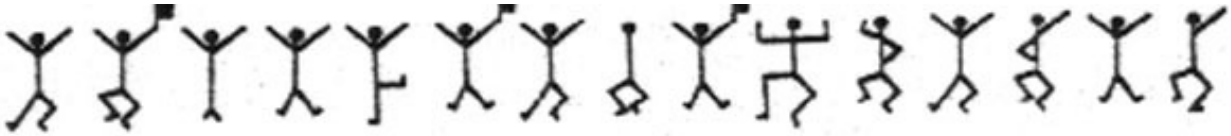
Bioinformatics is detective work

- **The Dancing Men code**, Sherlock Holmes
- **The Prison code**, a real life code used in CA

The Adventures of the Dancing Men

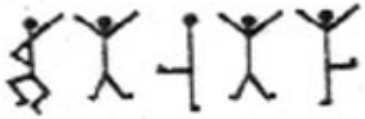
by Sir Arthur Conan Doyle

"Sherlock Holmes"

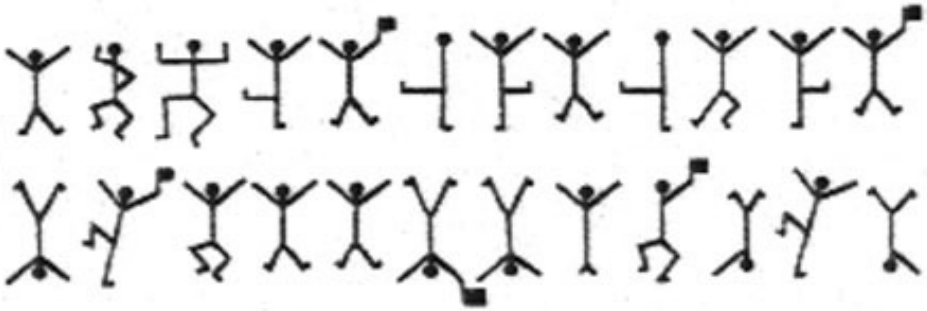


criminal's message (1)

criminal's message (2)



Elsie's reply



criminal's message (3)

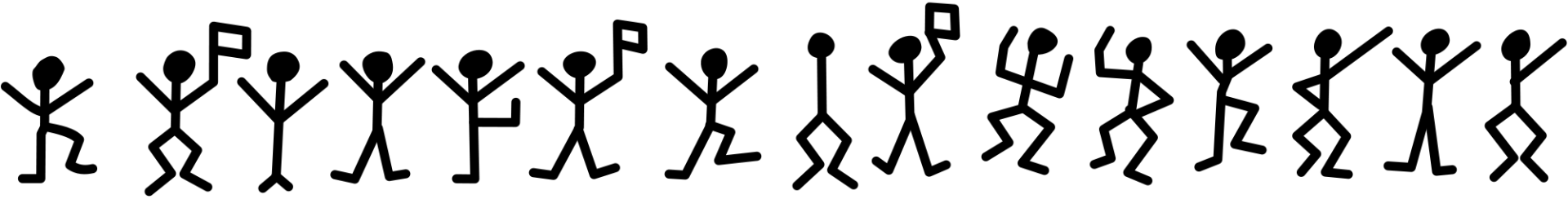
The Adventures of the Dancing Men

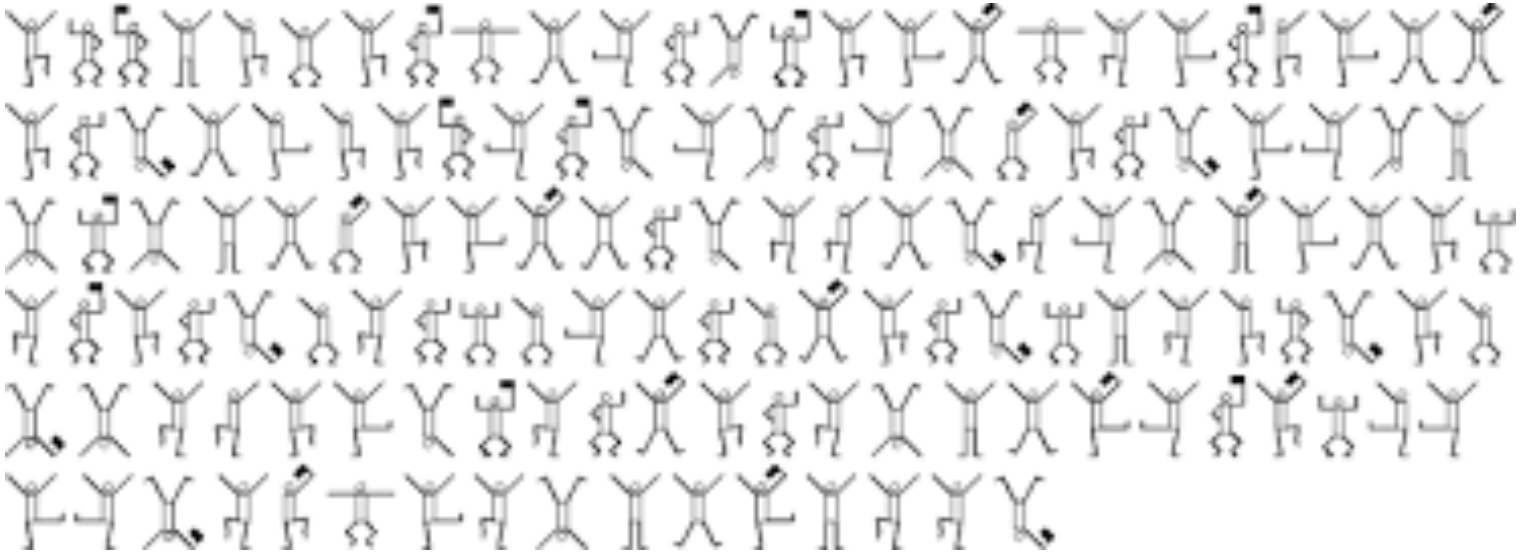
by Sir Arthur Conan Doyle

"Sherlock Holmes"

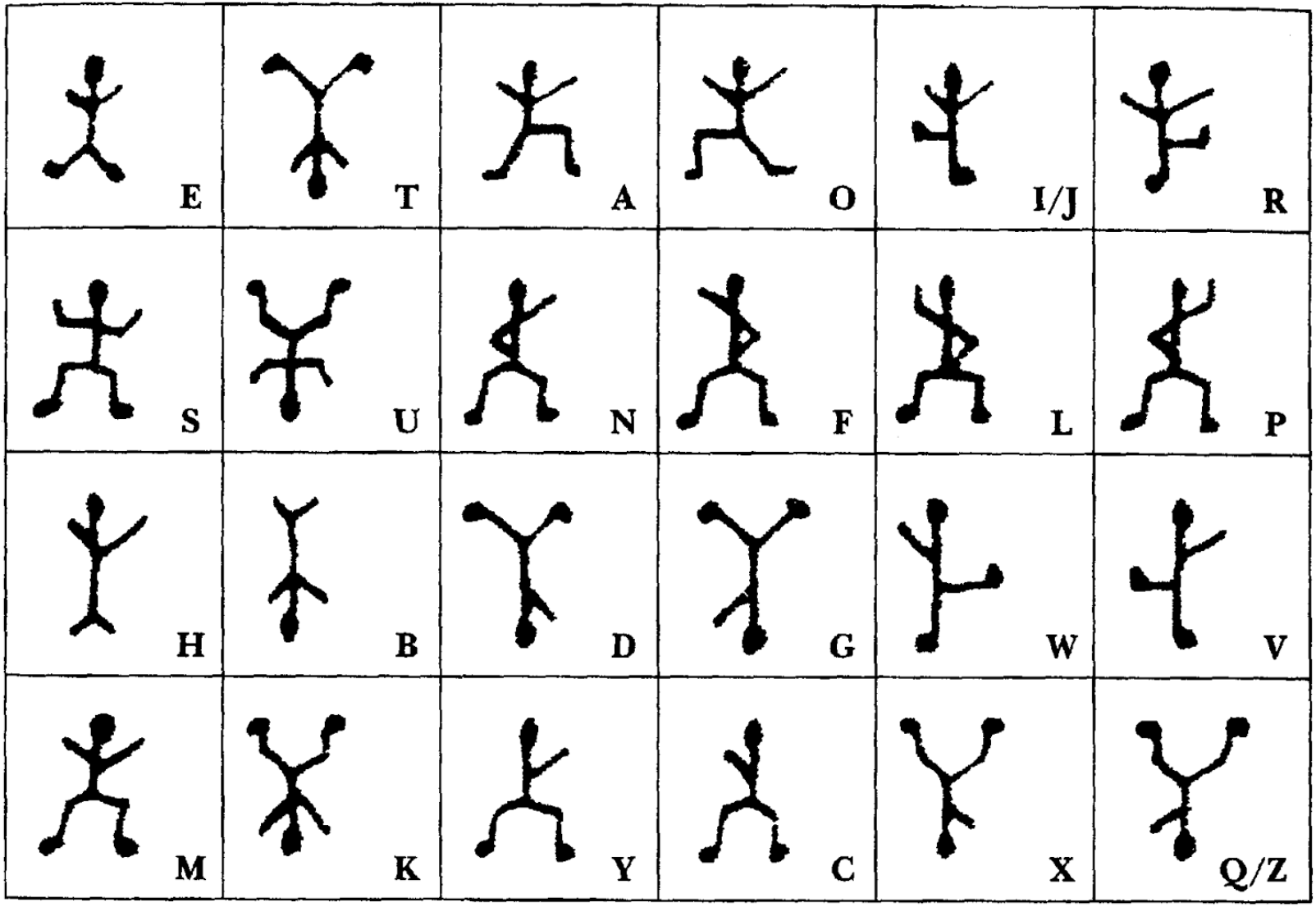
The Dancing Men code

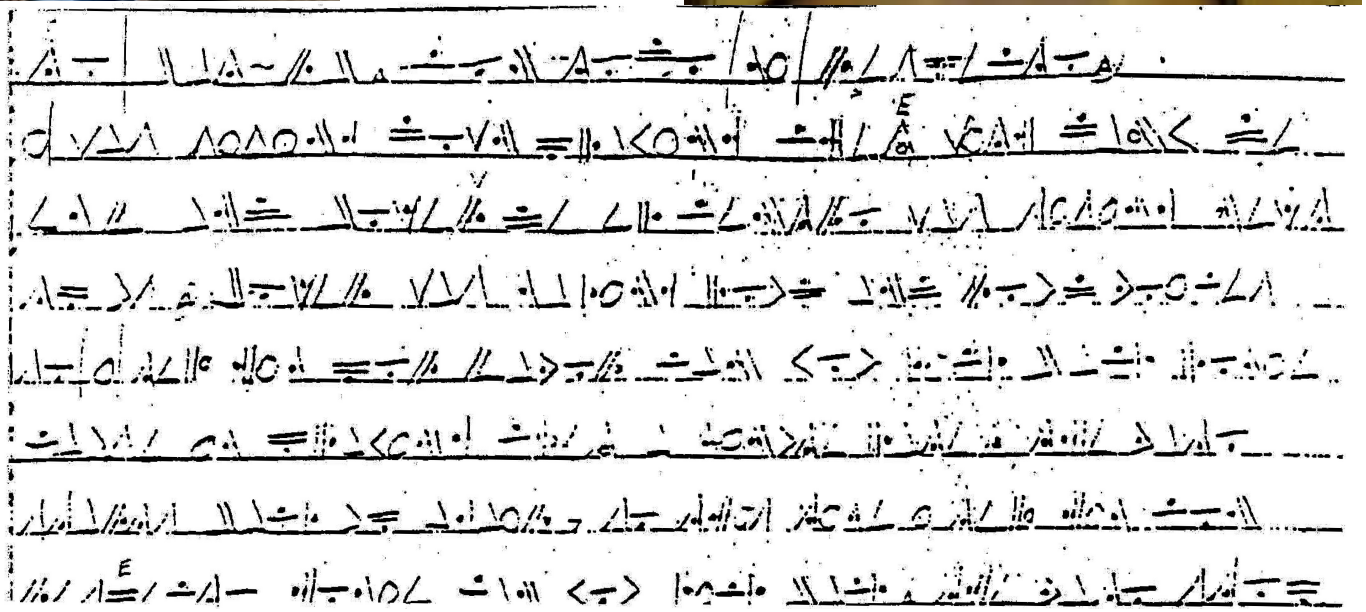
by Arthur Conan Doyle: “The Adventures of the Dancing Men”






Hello, how are you?
Good morning!
How was your weekend?
What the heck!
I've seen that movie!
THE CODE OF THE DANCING MEN
sherlockholmes-fan.com


E
T
A
O
I/J
R
S
U
N
F
L
P
H
B
D
G
W
V
M
K
Y
C
X
Q/Z



The Prison code

Solution: An Algorithm based on Markov Chain Monte Carlo


Otx-beta
Ul-Novel-POU
Otx-beta3
CELERA
Caltech, Davidson Lab
October 2004

Eric Davidson

– in memoriam