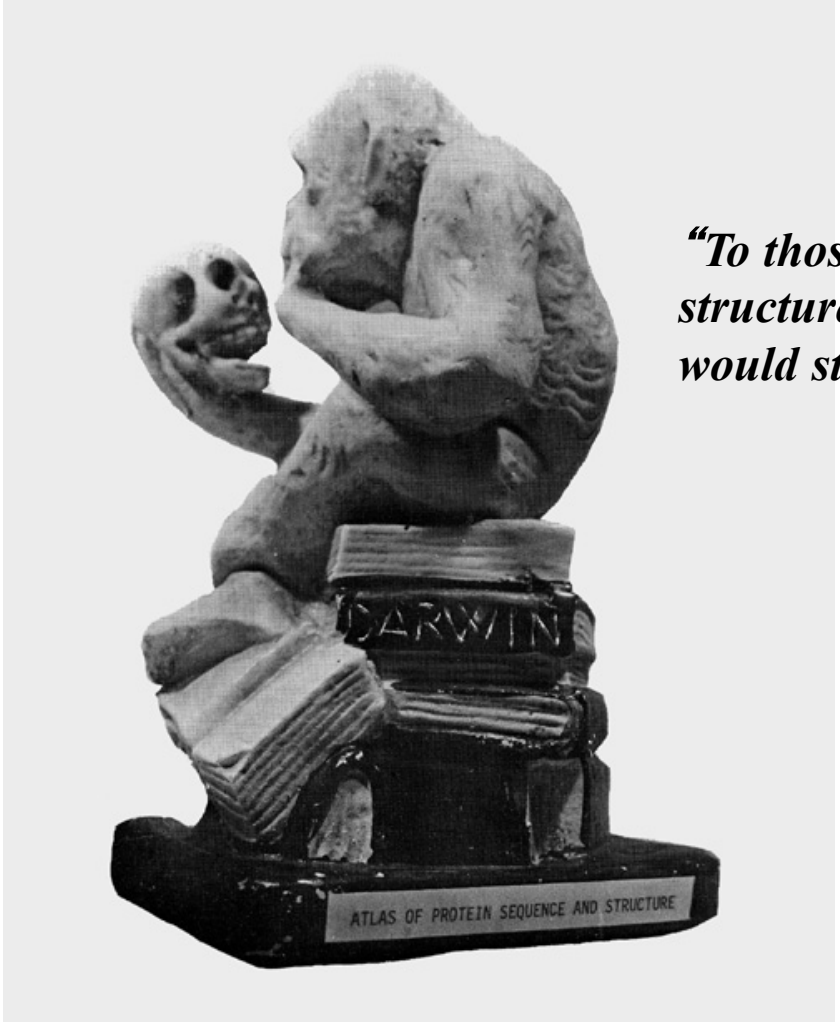
Margaret Dayhoff "mother and father of Bioinformatics"

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."

Mutation probability matrix for the evolutionary distance of 2 PAMs

normalized probabilities multiplied by 10000

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Mutation probability matrix for the evolutionary distance of 2 PAMs

Hydrophobic Amino Acids
Charged Amino Acids
Polar Amino Acids
Glycine

normalized probabilities multiplied by 10000

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

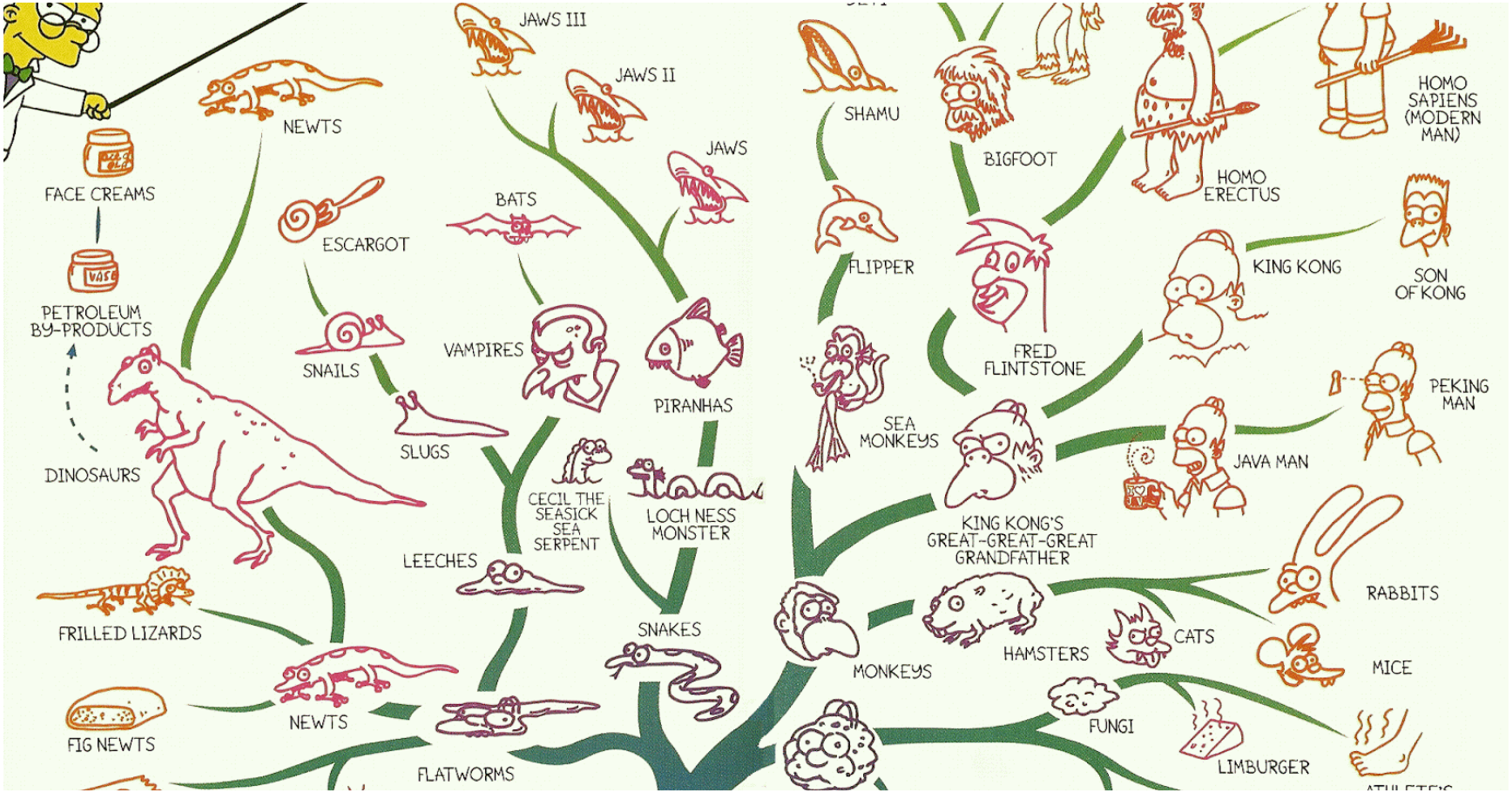
Mutation probability matrix for the evolutionary distance of 2 PAMs (Dayhoff Color Scheme)

Hydrophilic Amino Acids
Sulfhydryl
Aliphatic
Basic
Aromatic
Special

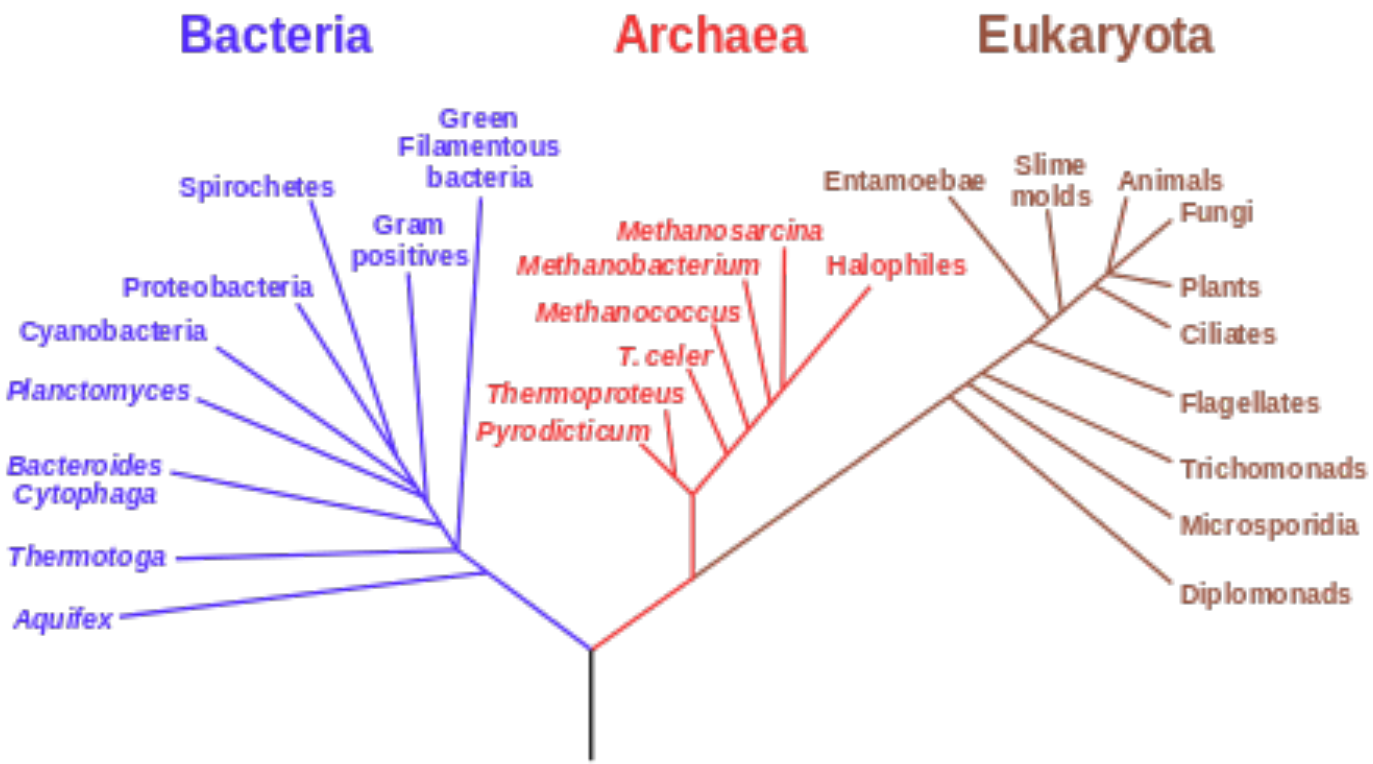
normalized probabilities multiplied by 10000

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Phylogenetic Trees (Ch. 3)



Phylogenetic Tree of Life



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 Cysteine Serine Phenylalanine Leucie Isolenine Methionine	Heterocyclic rings Close similarity in shape Hydrophobic; similar size

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

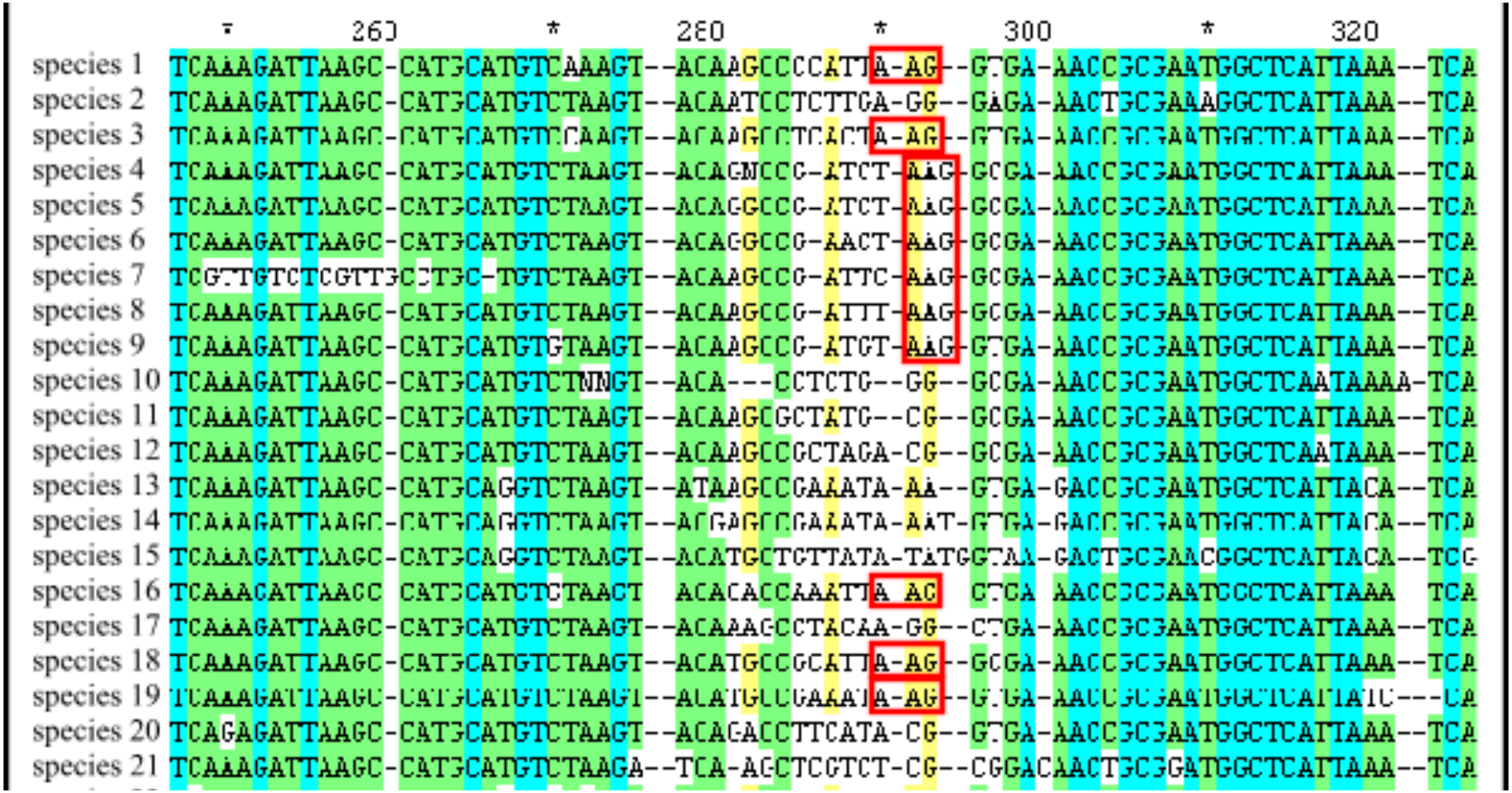
AUTHORS

W. C. Barker and M. O. Dayhoff

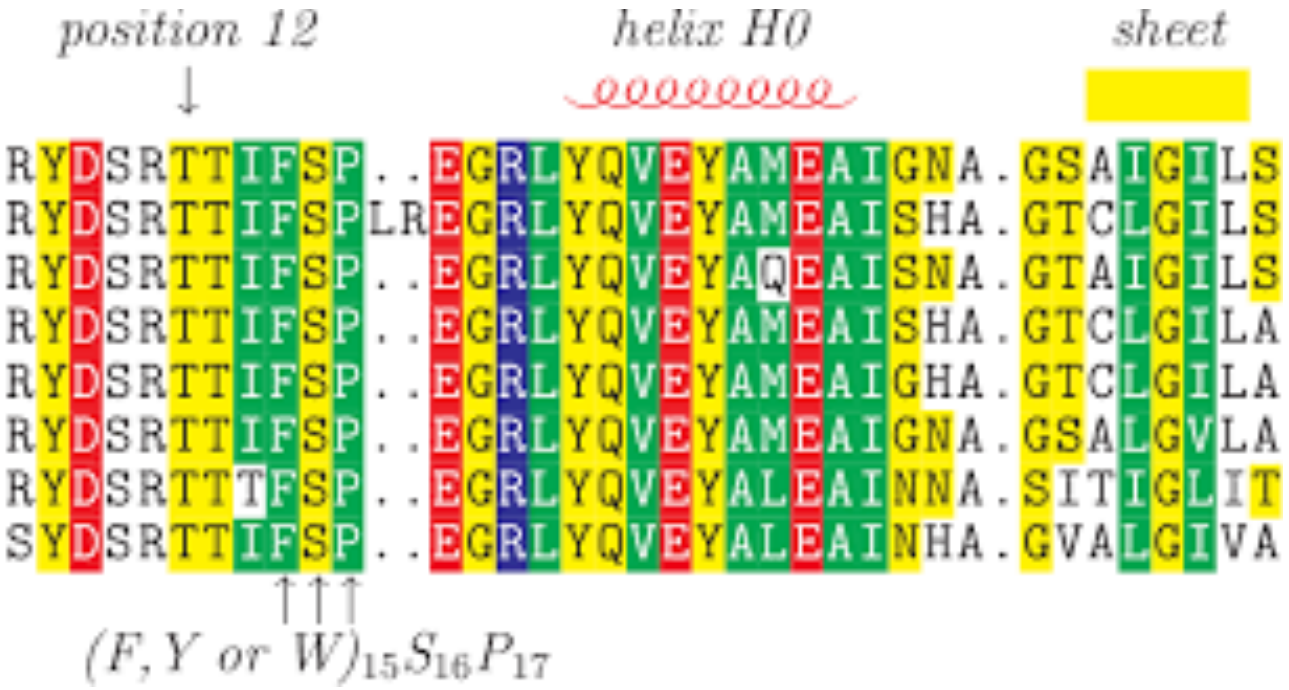
ABSTRACT

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.

Sequence Alignment (Ch. 1) DNA



Protein sequence alingment



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 ELN I AGLR HDNIVR VVAASTRT PEDSNS LG T I 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   Q I E H T L  N E K R I  - - L Q A V N F P F  L V K L E F S F  K D N S N L Y M V M E Y V  P G G E  M F S H
2 MMSV     S Q R S F W A  E L N  I  A G L R  H D N I V R  V V A A S T R T  P E D S N S  L G  T  I  I M E  F G G N V  T L H
3 RSV-PC   S P E A F L  Q E A Q V -  - M K K L  R H E K L V Q  L - Y A V V S E E P  I  Y I  V  I  E Y M S K G S  L L D  F

           S   E   F L   E    I    L    N    L V   L    S N   Y   V I  E Y   G G    H
                         *                  *                                *

1 BOV-PK   -  - - -  - - -  - - - - - L R - R I G R F - S E  P H A R  F Y A A Q I V L T F  E Y  L H S L D L I Y R D L
2 MMSV     Q V I Y D A T R S P E P L S  C R - - K Q L S L G K C  L K Y S  L D V V N G L  L F  L H S Q S I L H  L D L
3 RSV-PC   -  - - -  - - -  - - - - - L K G E M G K Y L R L P Q  L V  D M A A Q I  A S  G M A Y V E  R M N Y V H R D L

                                      L    R   G K  L S L P    Y A A Q I V   G    Y  H S    H R D L
                                      *                                                     *  *
```

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

Simian Sarcoma Virus *onc* Gene, v-*sis,* Is Derived from the Gene (or Genes) Encoding a Platelet-Derived Growth Factor

AUTHORS

Russell F. Doolittle, Michael W. Hunkapiller, Leroy E. Hood, Sushilkumar G. Devare, Keith C. Robbins, Stuart A. Aaronson, Harry N. Antoniades

ABSTRACT

The transforming protein of a primate sarcoma virus and a platelet derived growth factor are derived from the same or closely related cellular genes. This conclusion is based on the demonstration of extensive sequence similarity between the transforming protein derived from the simian sarcoma virus onc gene, v-sis, and a human platelet-derived growth factor. The mechanism by which v-sis transforms cells could involve the constitutive expression of a protein with functions similar or identical to those of a factor active transiently during normal cell growth.

Simian Sarcoma Virus *onc* Gene, v-*sis,* Is Derived from the Gene (or Genes) Encoding a Platelet-Derived Growth Factor

```
p28sis  1    M T L T W Q G D P I P E E L Y K M L S G H S I A S F D D L Q R L L Q G D S G K E D G A E L D L N M T

p28sis 51    A S H S G G E L E S L A R G K R S L G S L S V A E P A M I A E C K T A T E V F E I S A A L I D A T N
PDGF-2                                 S L G S L T I A E P A M I A E C K T A E E V F C I C A A L ? D A ? ?
PDGF-1                                       S I E E A V P A V C K T A I V I Y E I S A A E L D ? ? ?

p28sis       A N F L V W P P C V E V Q R C S G C C N N R N V Q C R P T Q V Q L R P V Q V R K I E I V R K K P I F
PDGF-2       ? ? ? ? ? ?P P C V E V K R C T G C C N N R N V K C R P S Q V Q L R P ?Q V R K I E I V R K [
PDGF-1       A N F L [

p28sis       K K A T V T L E D H L A C K C E I V A A A R A V T R S P G T S Q E Q R A K T T Q S R V T I R T V R V
PDGF-2
PDGF-1

P28sis       R R P P K G K H A K C K H T H D K T A L K E T L G A
PDGF-2
PDGF-1
```

Simian Sarcoma Virus *onc* Gene, v-*sis,* Is Derived from the Gene (or Genes) Encoding a Platelet-Derived Growth Factor

```
p28sis  1    M T L T W Q G D P I P E E L Y K M L S G H S I A S F D D L Q R L L Q G D S G K E D G A E L D L N M T

p28sis 51    A S H S G G E L E S L A R G K R S L G S L S V A E P A M I A E C K T A T E V F E I S A A L I D A T N
PDGF-2                                S L G S L T I A E P A M I A E C K T A E E V F C I C A A L ? D A ? ?
PDGF-1                                      S I E E A V P A V C K T A I V I Y E I S A A E L D ? ? ?

p28sis       A N F L V W P P C V E V Q R C S G C C N N R N V Q C R P T Q V Q L R P V Q V R K I E I V R K K P I F
PDGF-2       ? ? ? ? ? ?P P C V E V K R C T G C C N N R N V K C R P S Q V Q L R P ?Q V R K I E I V R K [
PDGF-1       A N F L [

p28sis       K K A T V T L E D H L A C K C E I V A A A R A V T R S P G T S Q E Q R A K T T Q S R V T I R T V R V
PDGF-2
PDGF-1

P28sis       R R P P K G K H A K C K H T H D K T A L K E T L G A
PDGF-2
PDGF-1
```

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

Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus

AUTHORS

Michael D. Waterfield, Geoffrey T. Scrace, Nigel Whittle, Paul Stroobant, Ann Johnsson, Ake Wasteson, Bengt Westermark, Carl-Henrik Heldin, Jung Sang Huang, and Thomas F. Deuel

ABSTRACT

A partial amino acid sequence of human platelet-derived growth factor, the major mitogen in serum for cells of mesenchymal origin, has been determined. A region of 104 contiguous amino acids shows virtual identity with the predicted sequence of p28sis, the putative transforming protein of simian sarcoma virus (SSV). This similarity suggests a mechanism for transformation by SSV and other agents, involving expression of growth factors.

Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus

```
V-sis M T L T W Q G D P  I P E E  L Y K M L S G H S  I  R S F D D L Q R L L Q G D S G K E D G A E L  D L N M T R S H S G G E L  E S

V-sis L  A R G K R S  L G S L S V A E P  A M I A  E C K T R T E V F E  I  S R R L  I  D R T N A N F L V W P P  C V E V Q R C S G C C N

Peptide I          S L G S L T I A E P A M I A E C K T R T E V F E I S R R L  I  D ----------------------------------------------------------
Peptide II               S I  E E A V P A V C K T R T V I Y E I P R S Q V D P T S A N F L V W P P C V E ----------------------------------
Peptide III                                                                T S A N F L V W P P C V E V Q R C S G C C N
Peptide IV                                                                 T  I A N F L V W P P C V E V Q R C S G C C N

V-sis N R N V Q  C  R P  T Q V Q L R P V Q V R K I E I V R K K P I F K K A T V T L E D H L A C K G E I V A A A R A V T R S P G T

Peptide I          ----------------------------------------------------------------------------------------------------------------
Peptide II         ----------------------------------------------------------------------------------------------------------------
Peptide III        N R N V Q C R P T Q V Q L X P V Q----------------*
Peptide IV         N R N V Q C R P T Q V Q L R P V Q V R K I E ---*
Peptide V                                                                                  K K P I F K K A X V X L E D H L A C K C X I V A A A *

V-sis S Q E Q R A K T T Q S R V T I R T V R V R R P P K G K H R K C K H T H D K T A L K E T L G A

Peptide I          ------------------------------------*
Peptide II         ----------------------------------------------------------------------------------*
```

Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus

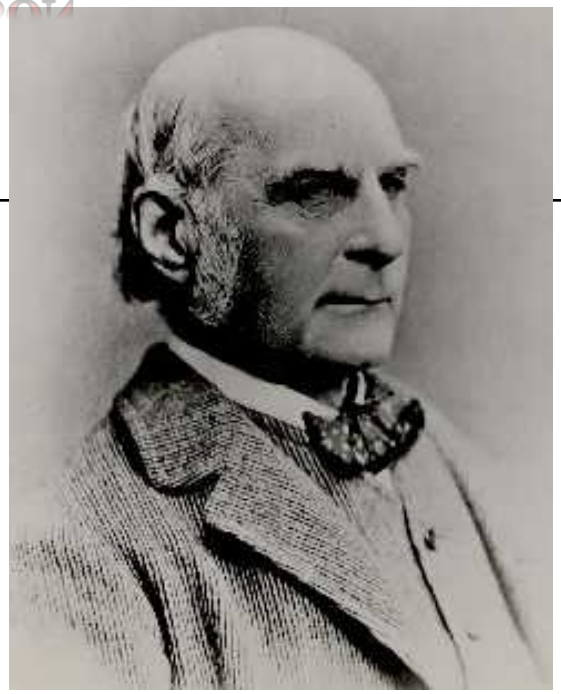
```
V-sis  MTLTWQGDP IPEE LYKMLSGHS I RSFDDLQRLLQGDSGKEDGAEL DLNMTRSHSGGEL ES

V-sis  L ARGKRS LGSLSVAEP AMIA ECKTRTEVFE I SRRL I DRTNANFLVWPP CVEVQRCSGCCN

Peptide I        SLGSLTIAEPAMIAECKTRTEVFEISRRLID----------------------------------------------------------
Peptide II             SIEEAVPAVCKTRTVIYEIPRSQVDPTSANFLVWPPCVE-----------------------------
Peptide III                                                  TSANFLVWPPCVEVQRCSGCCN
Peptide IV                                                   TIANFLVWPPCVEVQRCSGCCN

V-sis  NRNVQ C RP TQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKGEIVAAARAVTRSPGT

Peptide I      ----------------------------------------------------------------------------------------------------------------
Peptide II     ----------------------------------------------------------------------------------------------------------------
Peptide III    NRNVQCRPTQVQLXPVQ----------------*
Peptide IV     NRNVQCRPTQVQLRPVQVRKIE---*
Peptide V                                          KKPIFKKAXVXLEDHLACKCXIVAAA*

V-sis  SQEQRAKTTQSRVTIRTVRVRRPPKGKHRKCKHTHDKTALKETLGA

Peptide I      ----------------------------------*
Peptide II     ----------------------------------------------------------------------------*
```

Theodosius Dobshansky
Sir Francis Galton
Sir Ronald Fisher

Theodosius Dobzhansky (1900-1975)

Nothing in Biology Makes Sense Except in the Light of Evolution

THE SUPREME LAW OF UNREASON



Sir Francis Galton

A quotation (Galton 1889) relevant to the central limit theorem and about the normal or "Gaussian" distribution:

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

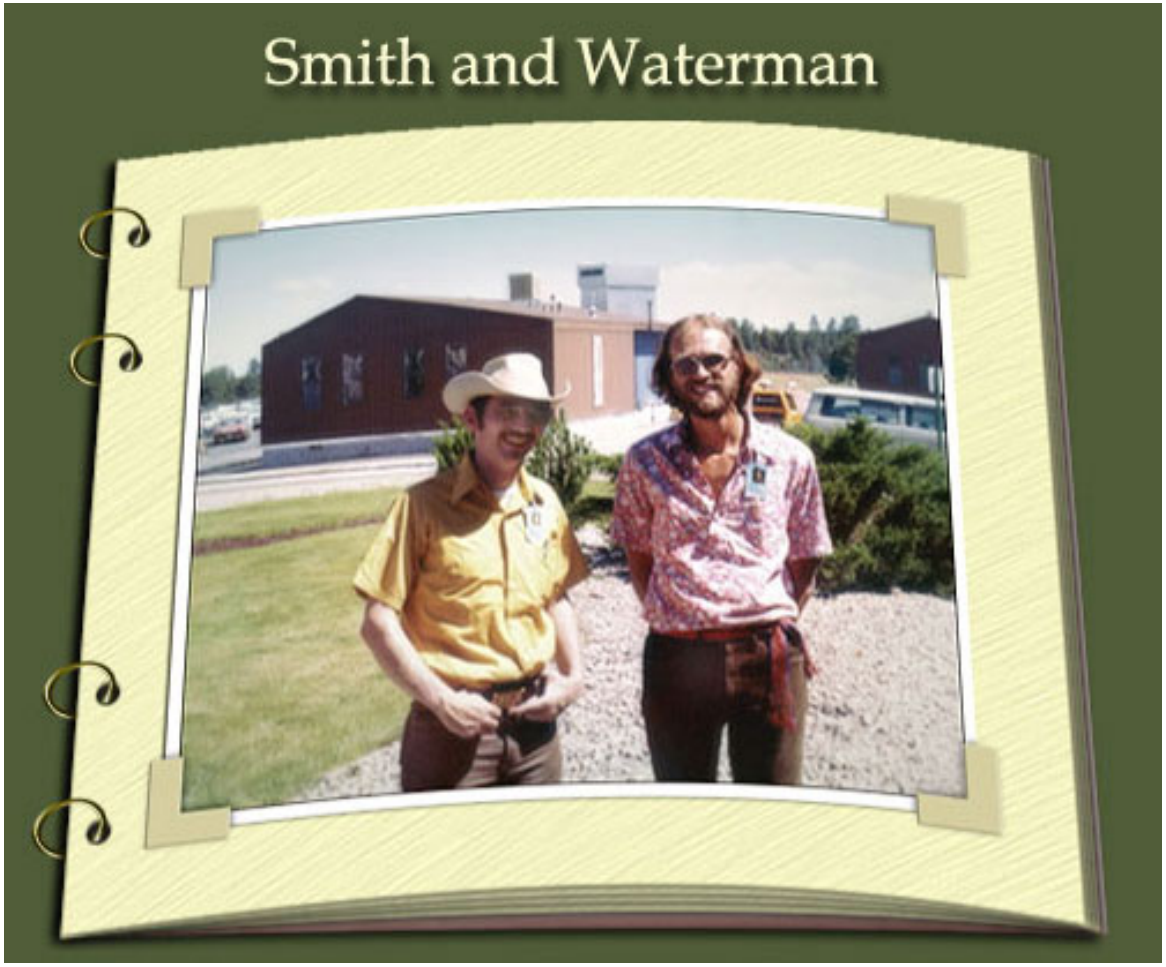
In Fisher We Trust

- Sir Ronald Aylmer Fisher

The Smith-Waterman Algorithm

Smith and Waterman at Los Alamos, New Mexico

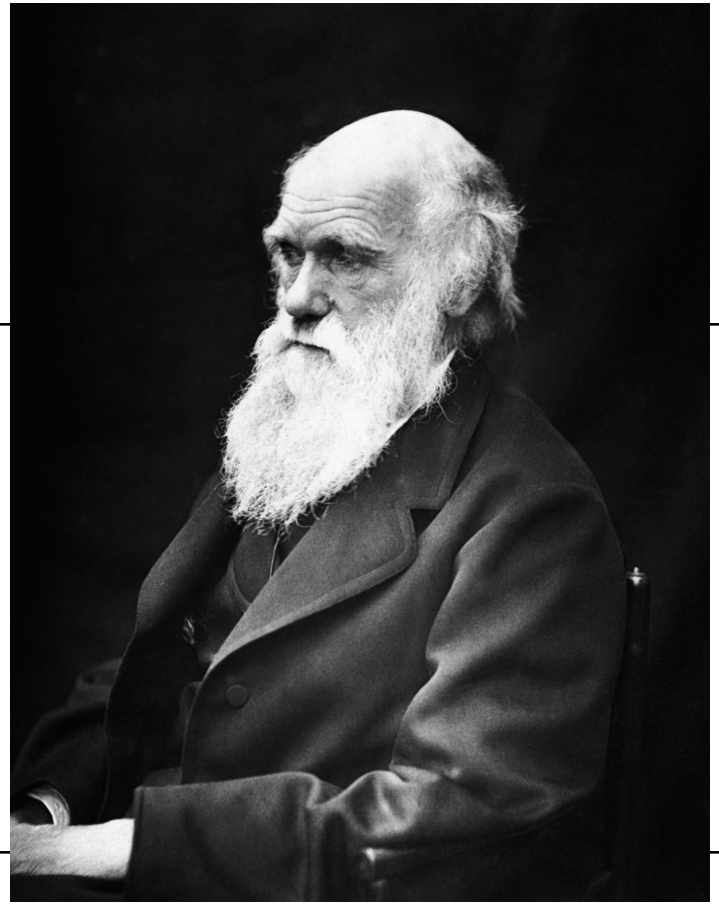
Photo by David Lipman, Taken Summer of 1980

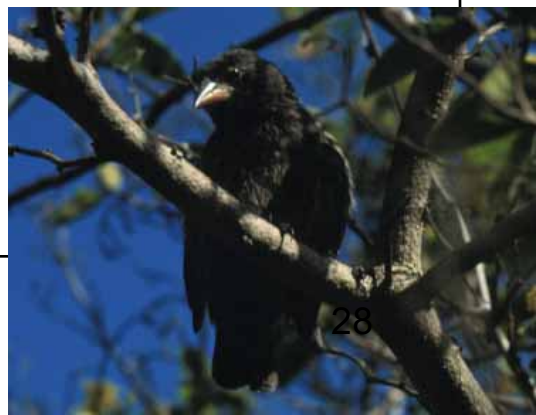


Of Sea Urchins, Birds and Men

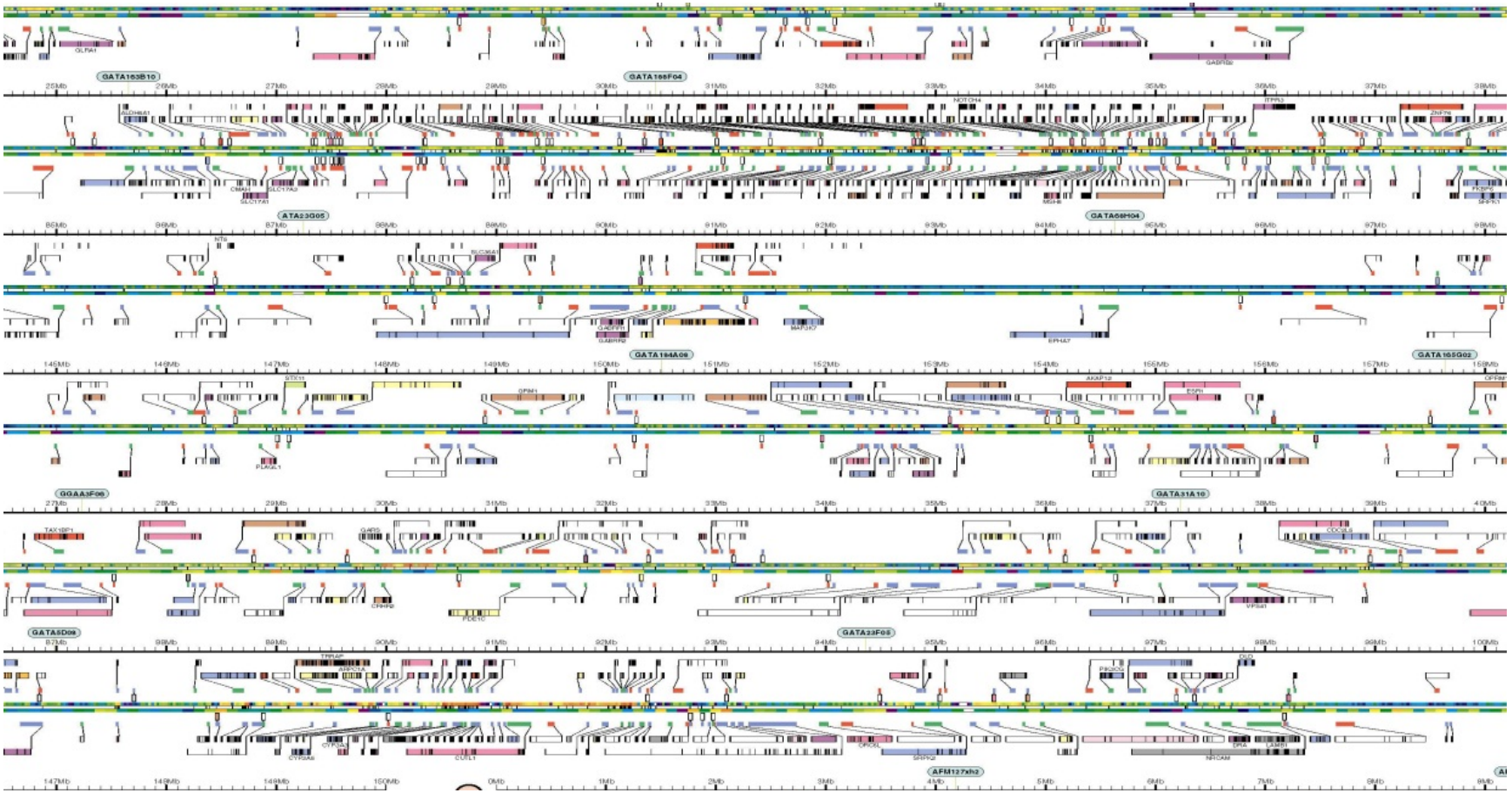
Darwin's Finches

and Coco

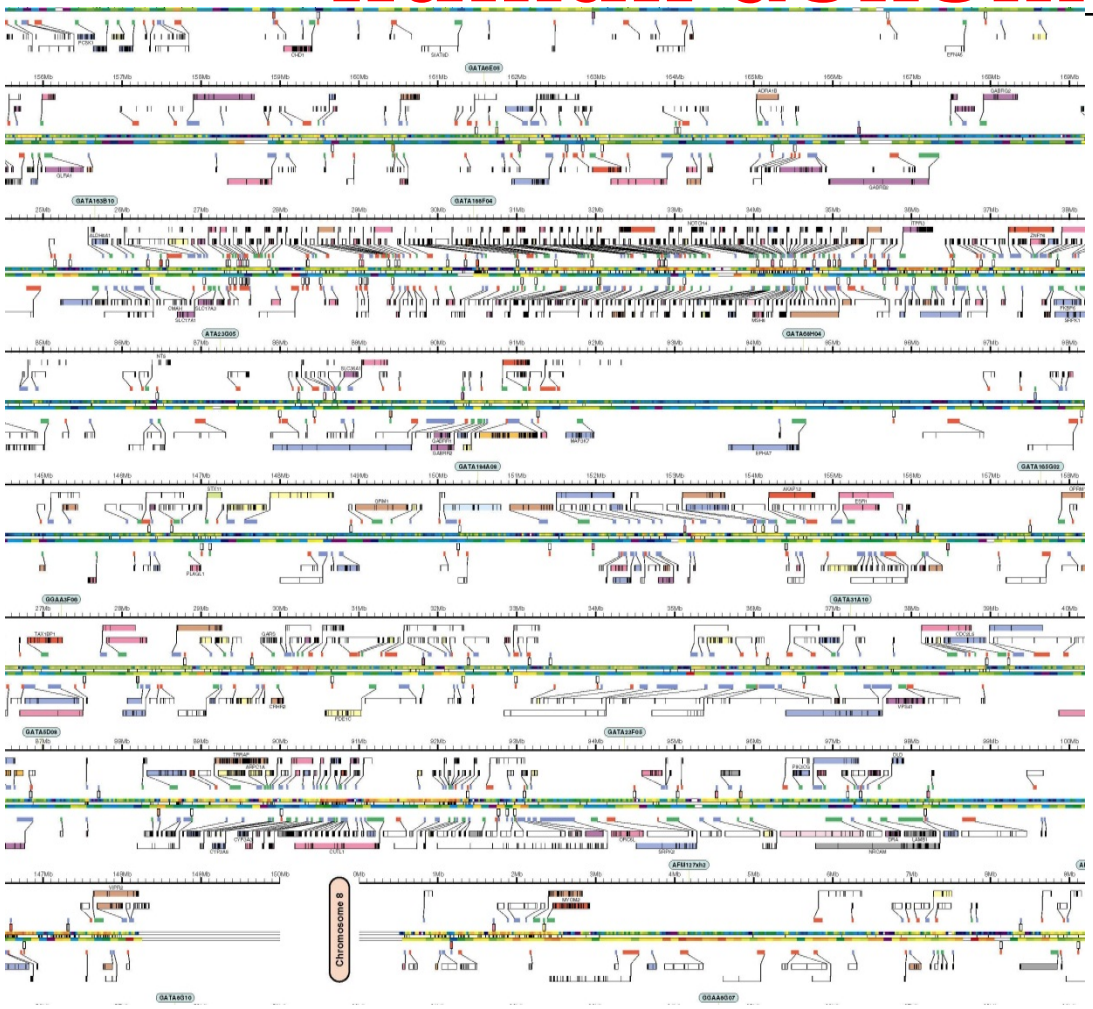




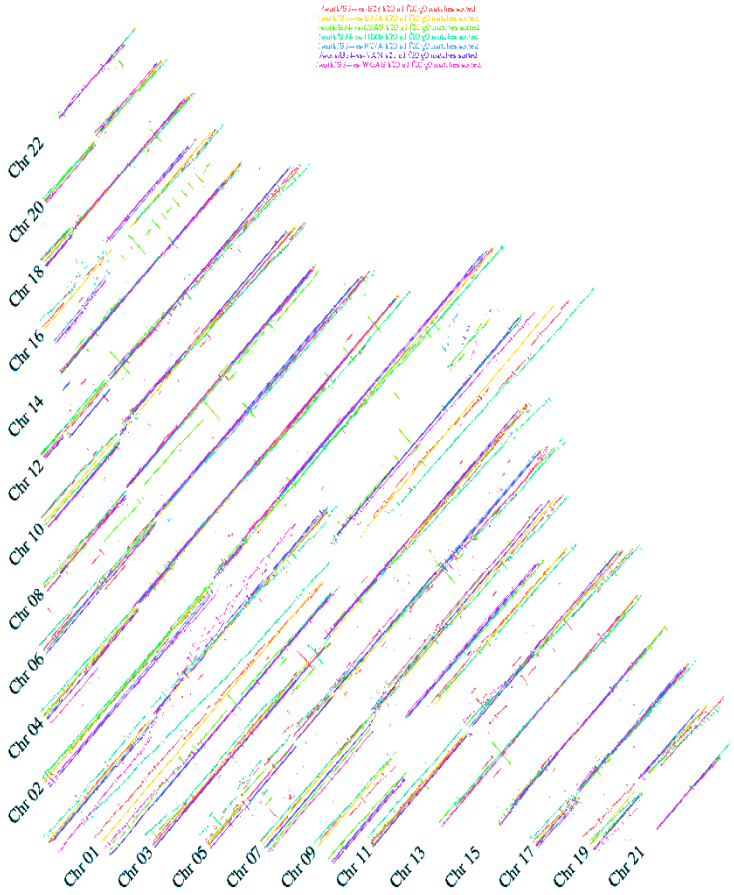
Genome Assembly (Ch. 5)



The Sequence of the Human Genome



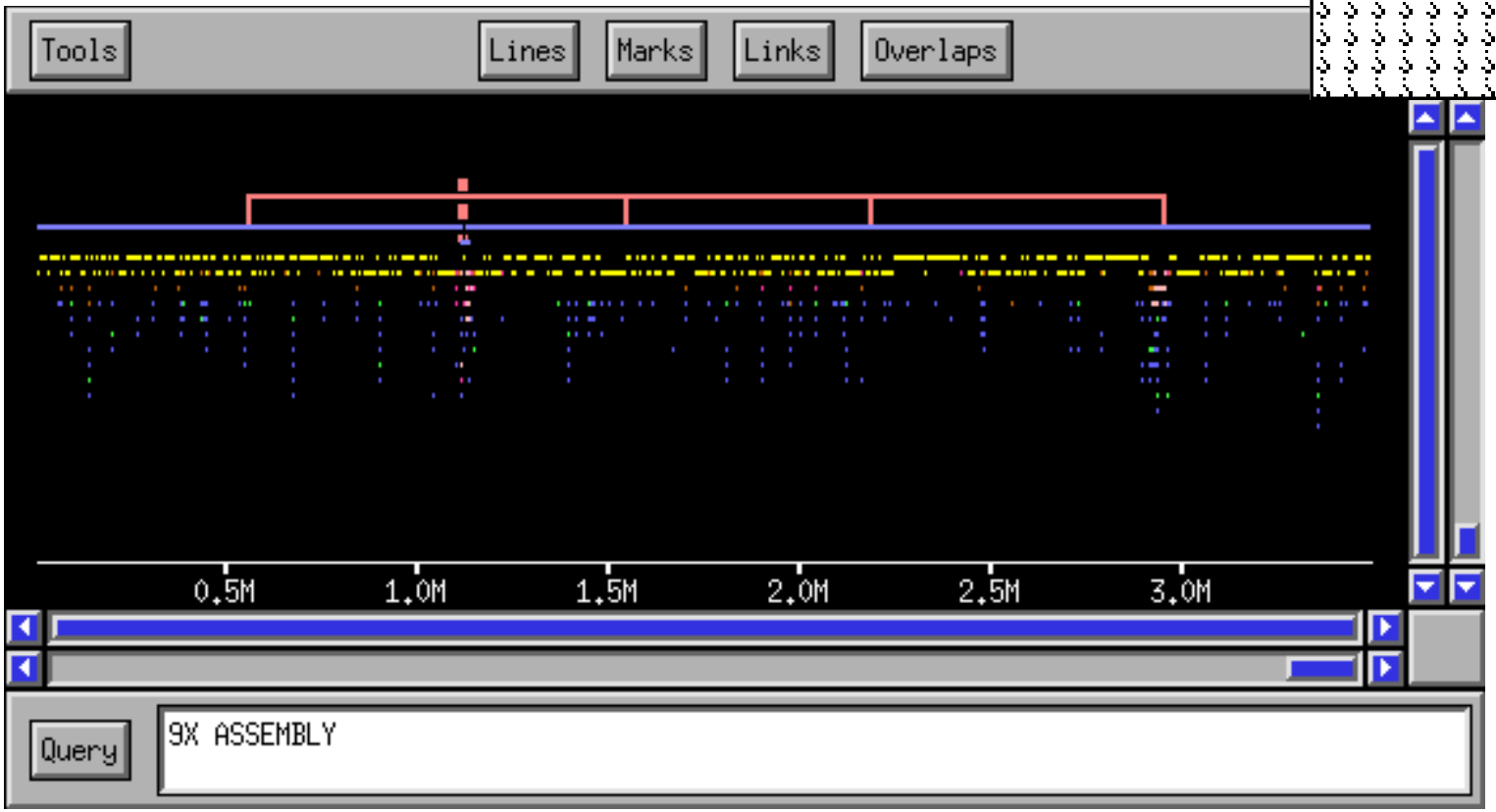
"The Sequence of the Human Genome" Science, 2001



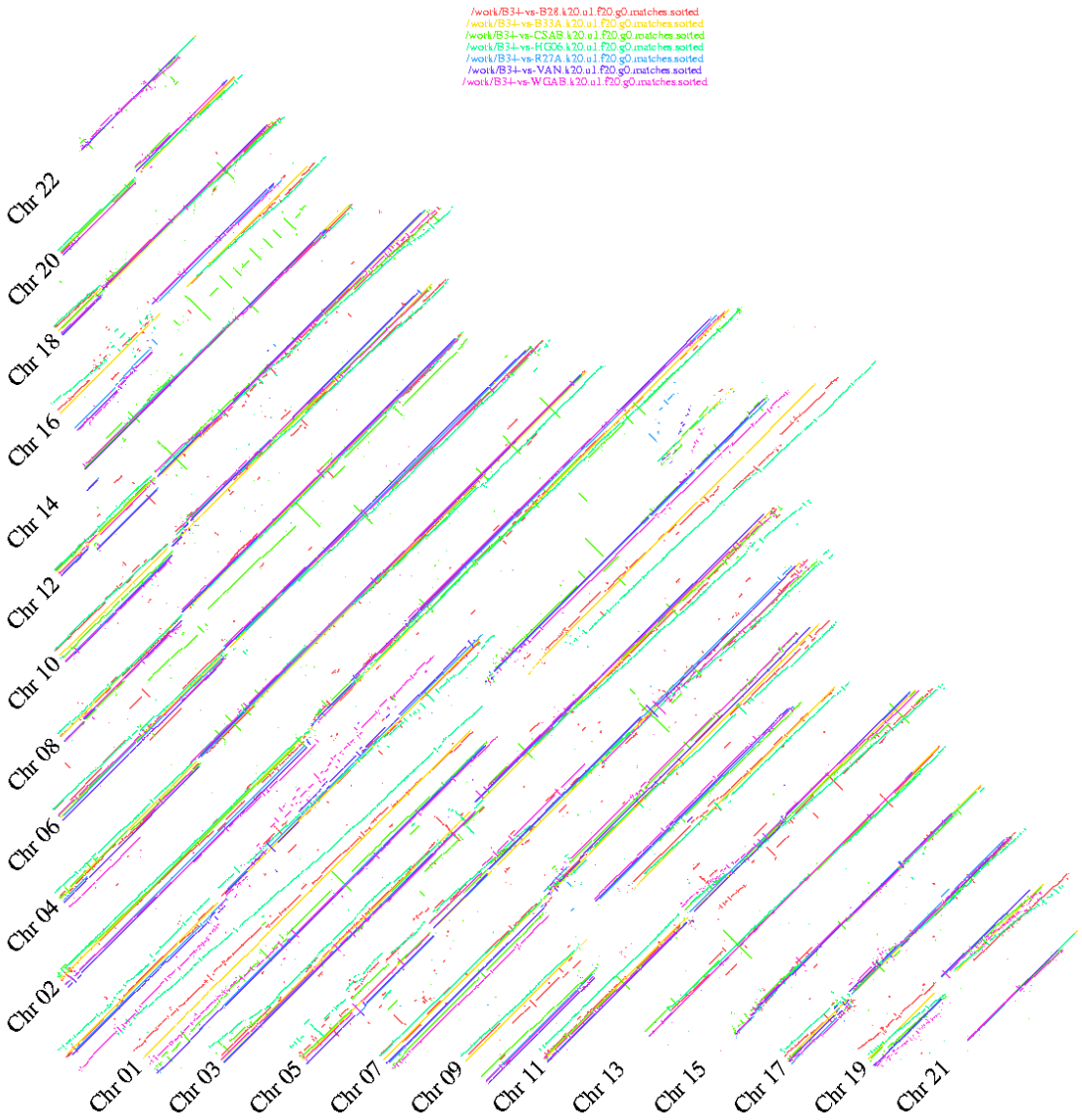
"Whole-genome shotgun assembly and comparison of human genome assemblies" Proc. Nat. Acad. Sci. USA, 2004

Genome Assembly Algorithm

Celera Assembler



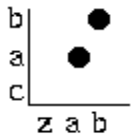
The Father of All Dot Plots

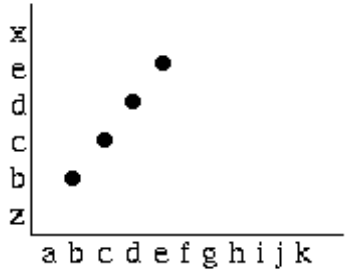


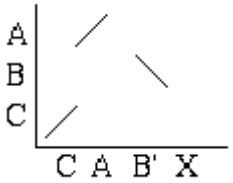
The Human Genome

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)

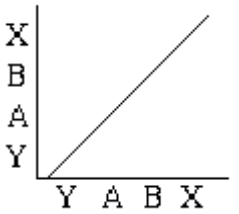


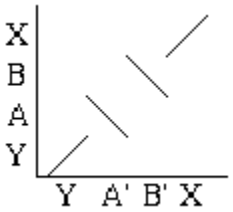


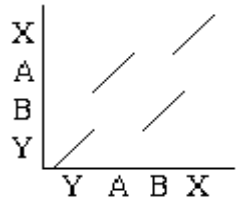


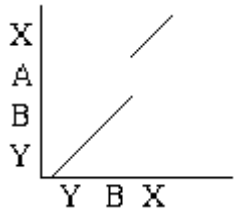
Dot Plots 101

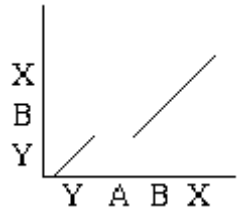
- When words line up
- Reversed
- Misplaced
- Something gained (relative to horizontal)
- Something lost (relative to horizontal)



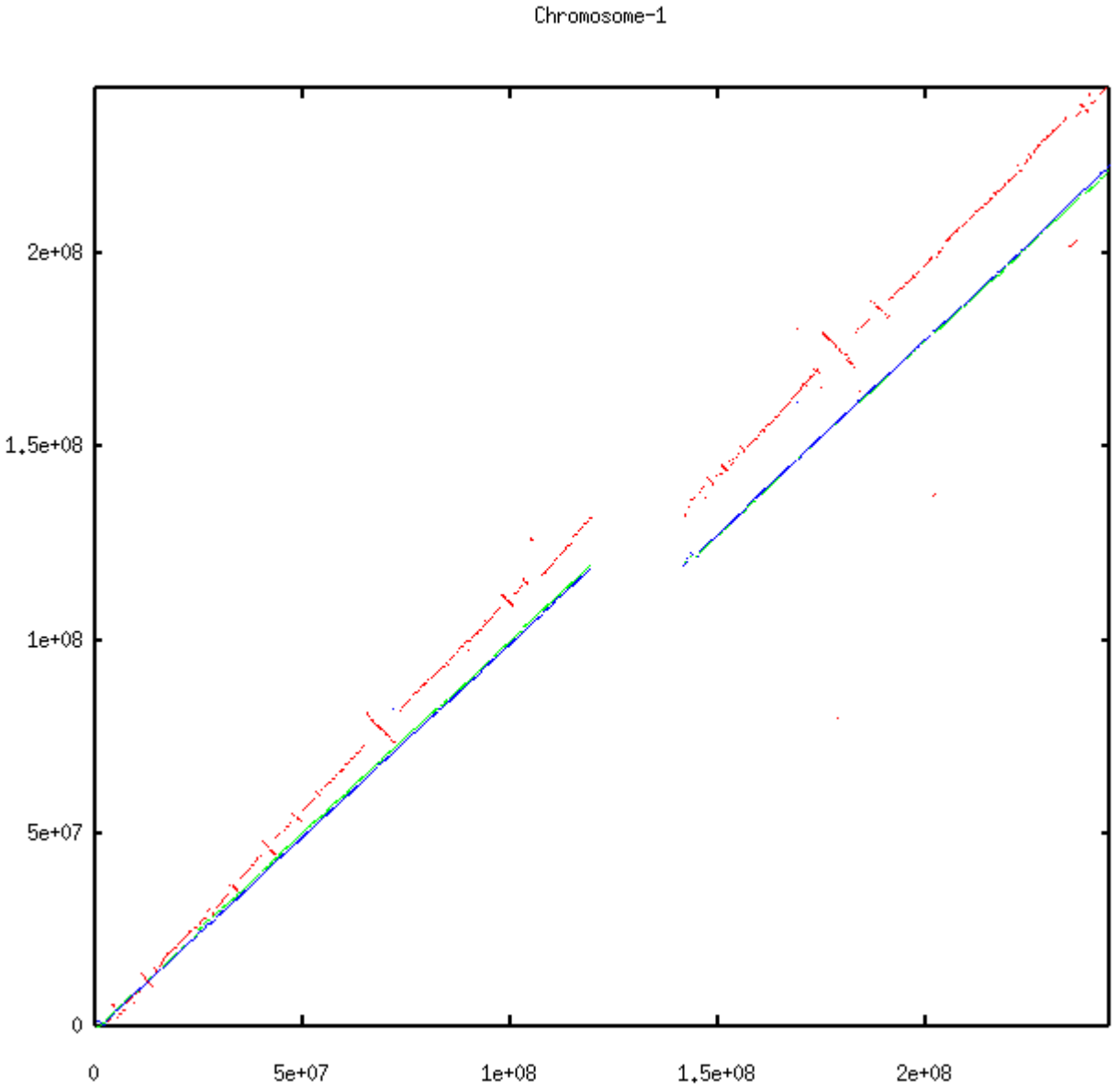




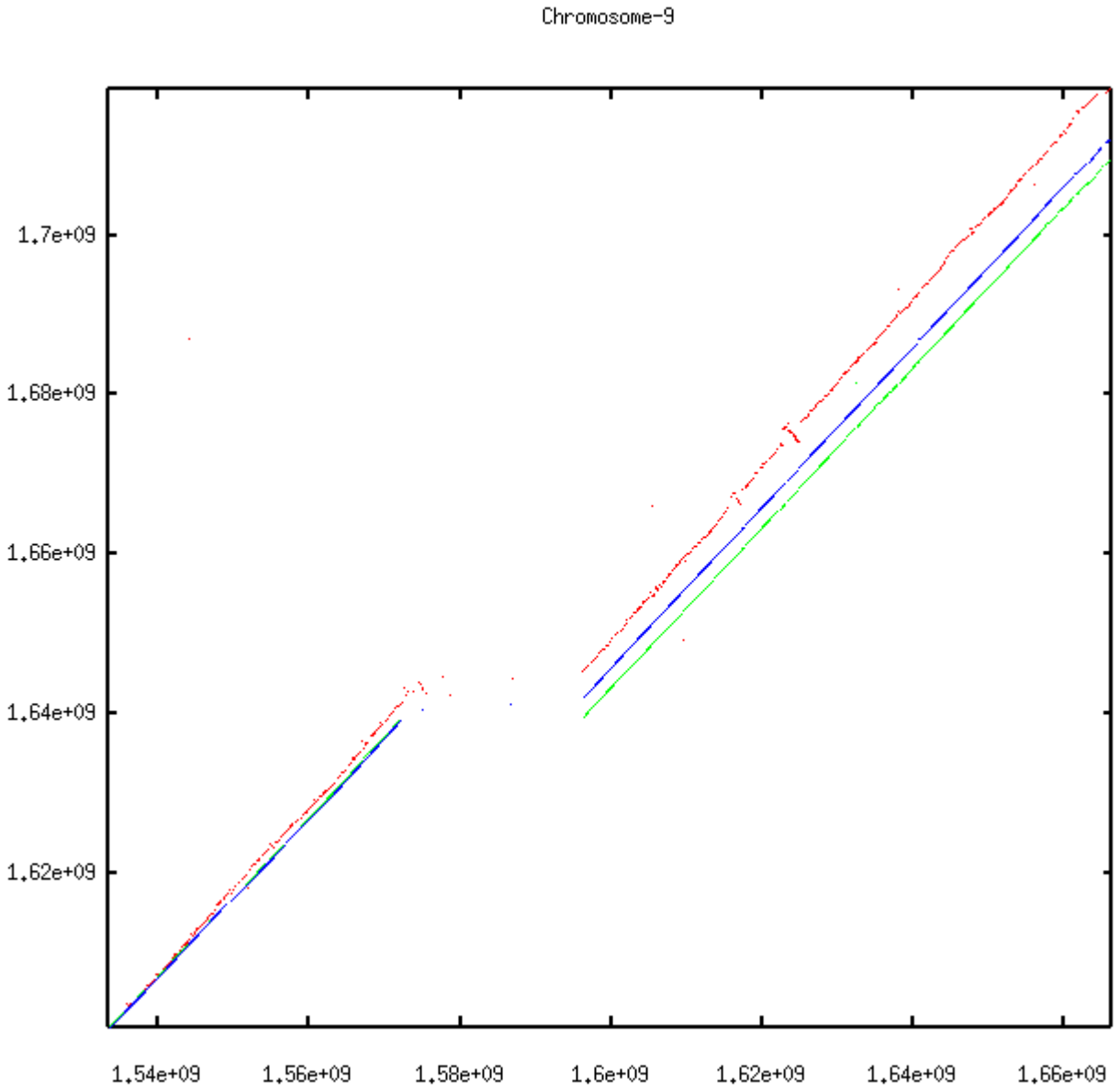




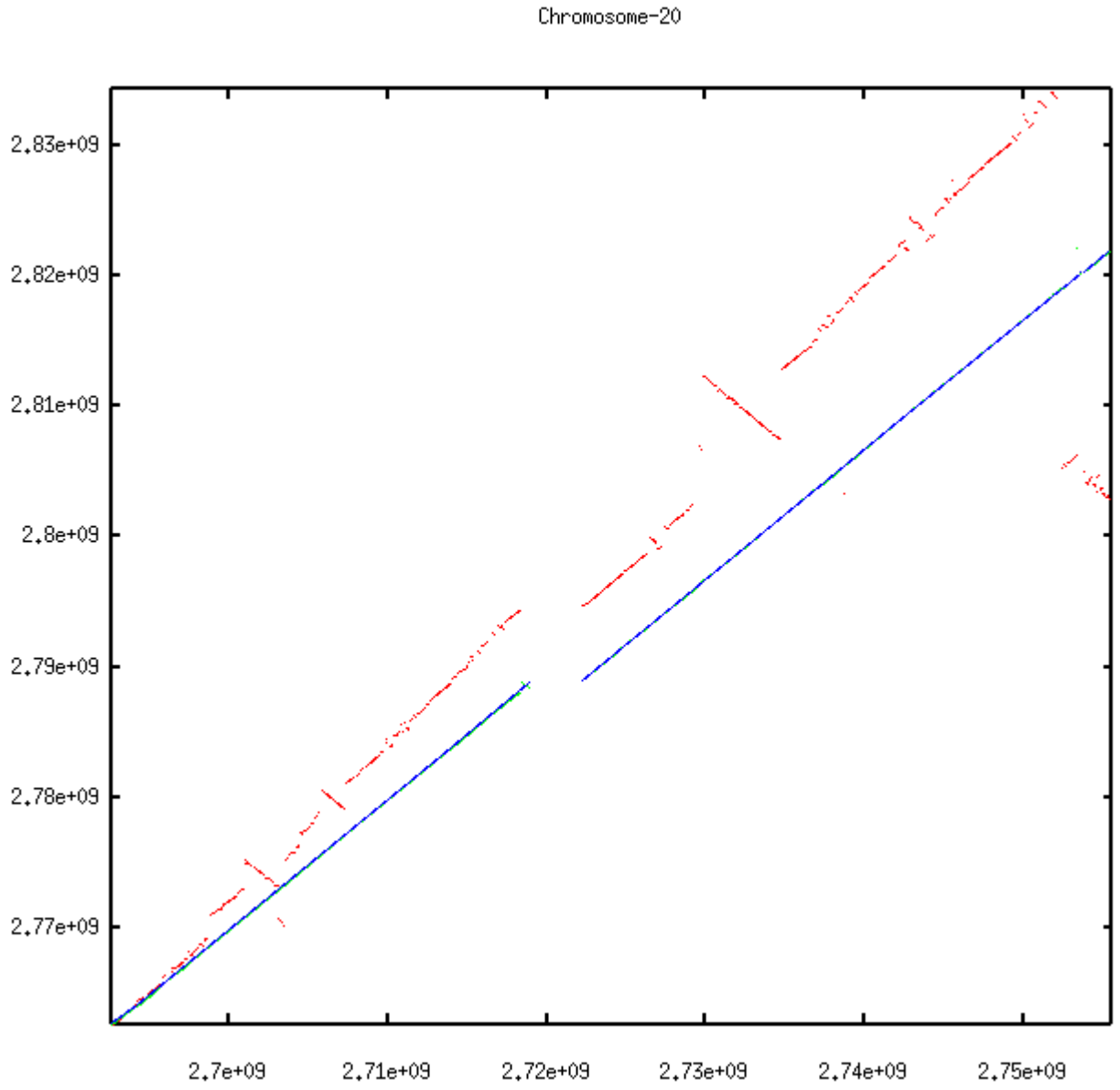
Some large reversals in GP



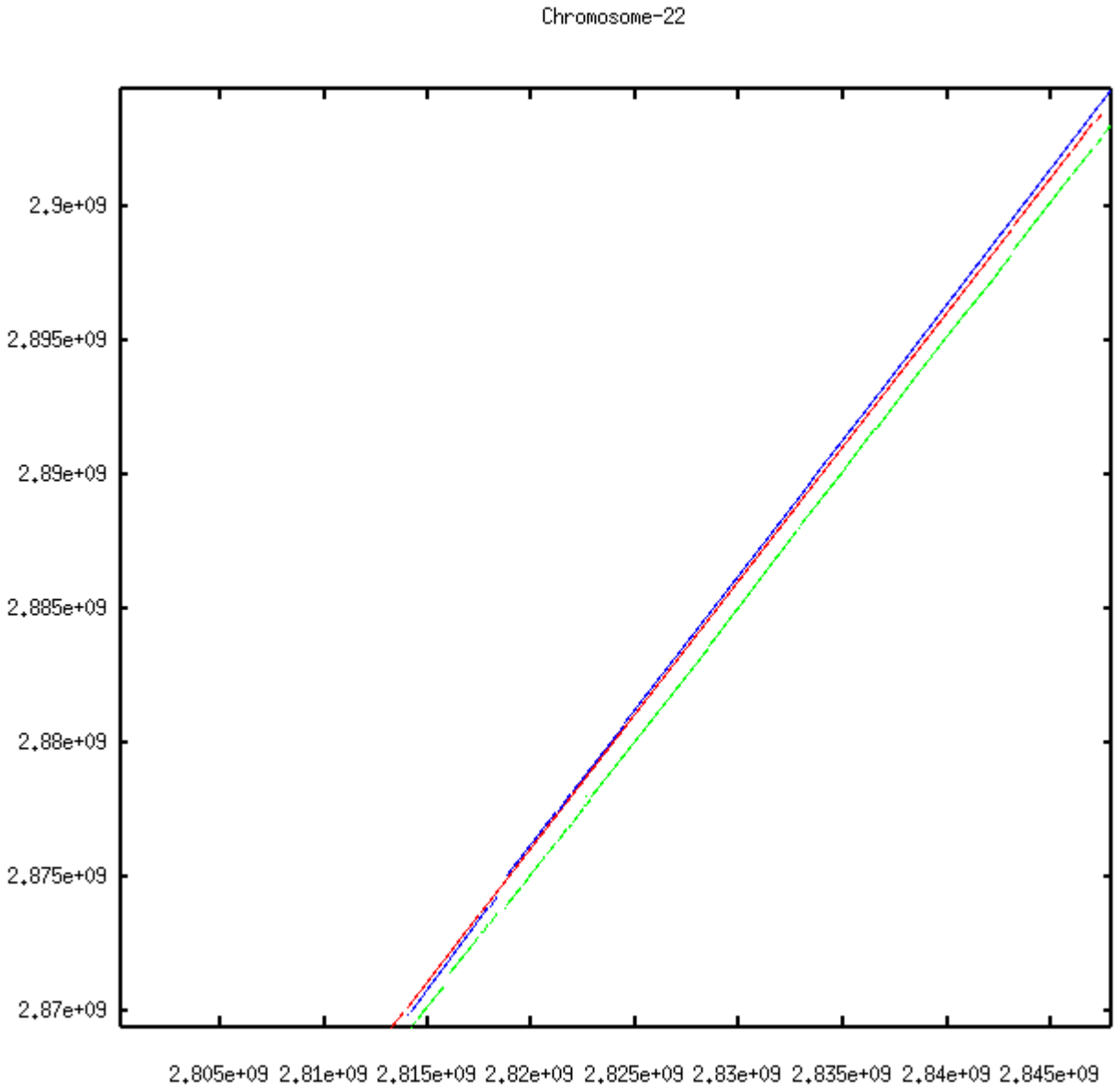
NCBI has more of the centromere than anyone else (or is that N's?)



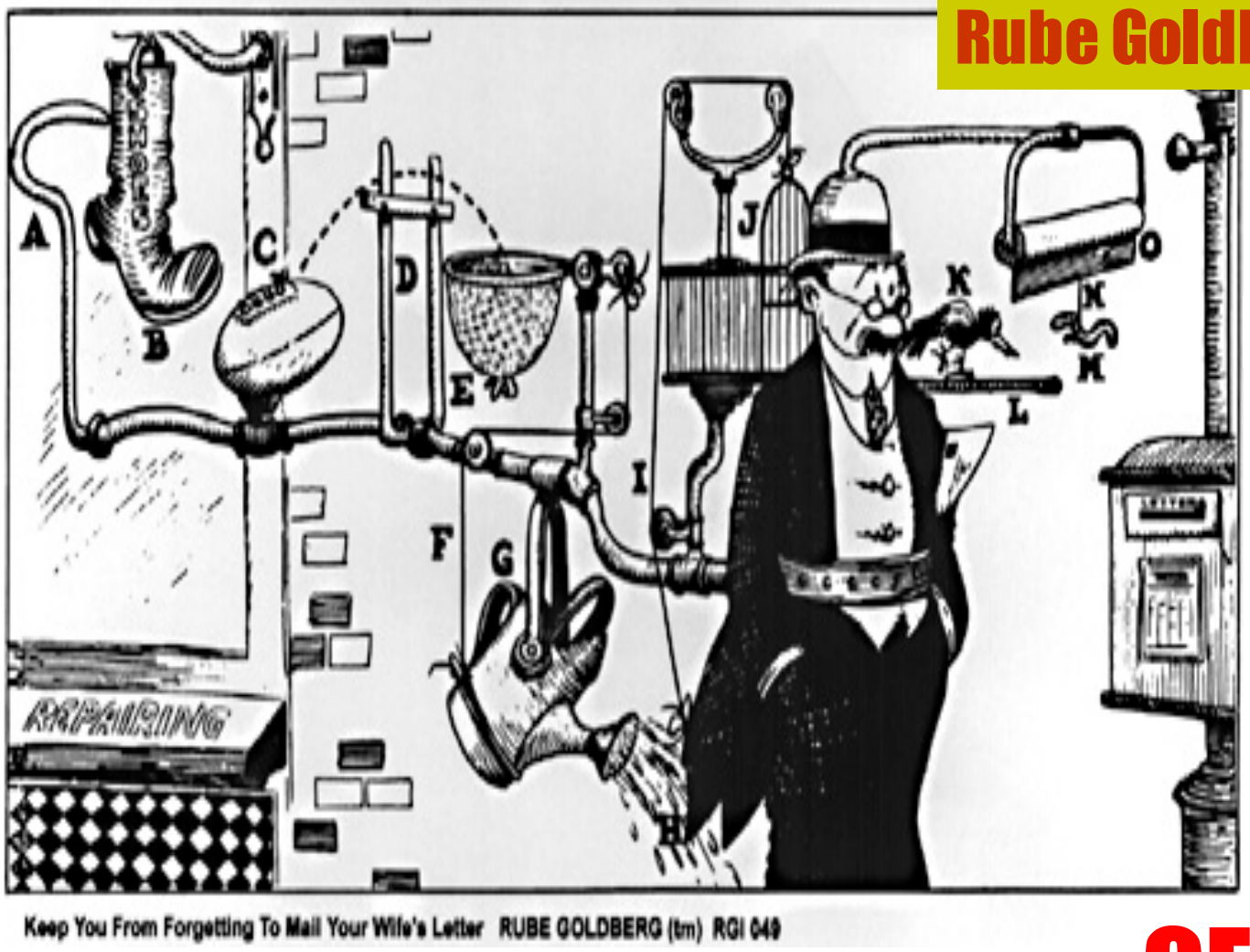
Many reversals in GP, a piece of the end is re-ordered to the middle, celera assemblies boringly good.



Again everyone misses the first 10MB (or are those N's) of NCBI31



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







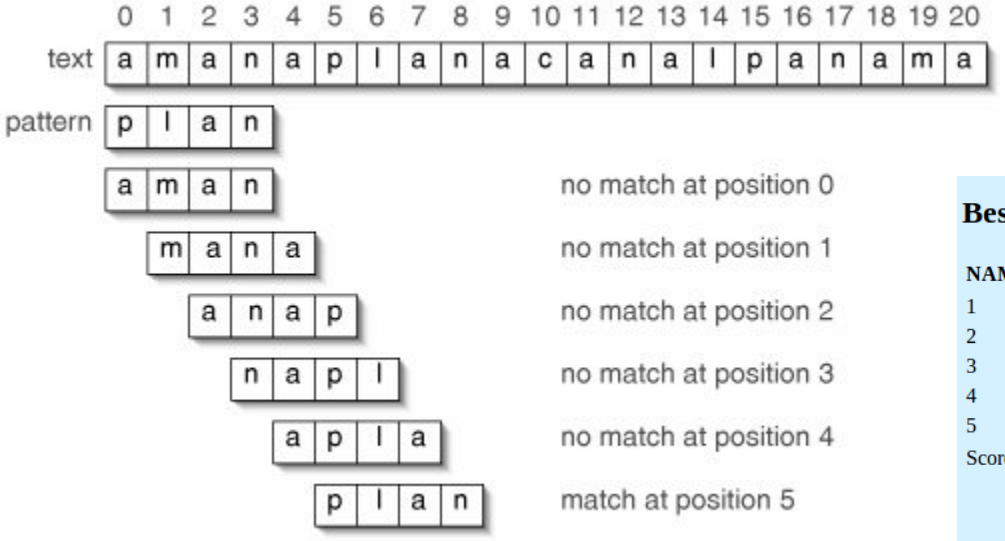




Eric Davidson

– in memoriam

Combinatorial pattern matching (Ch. 2)



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

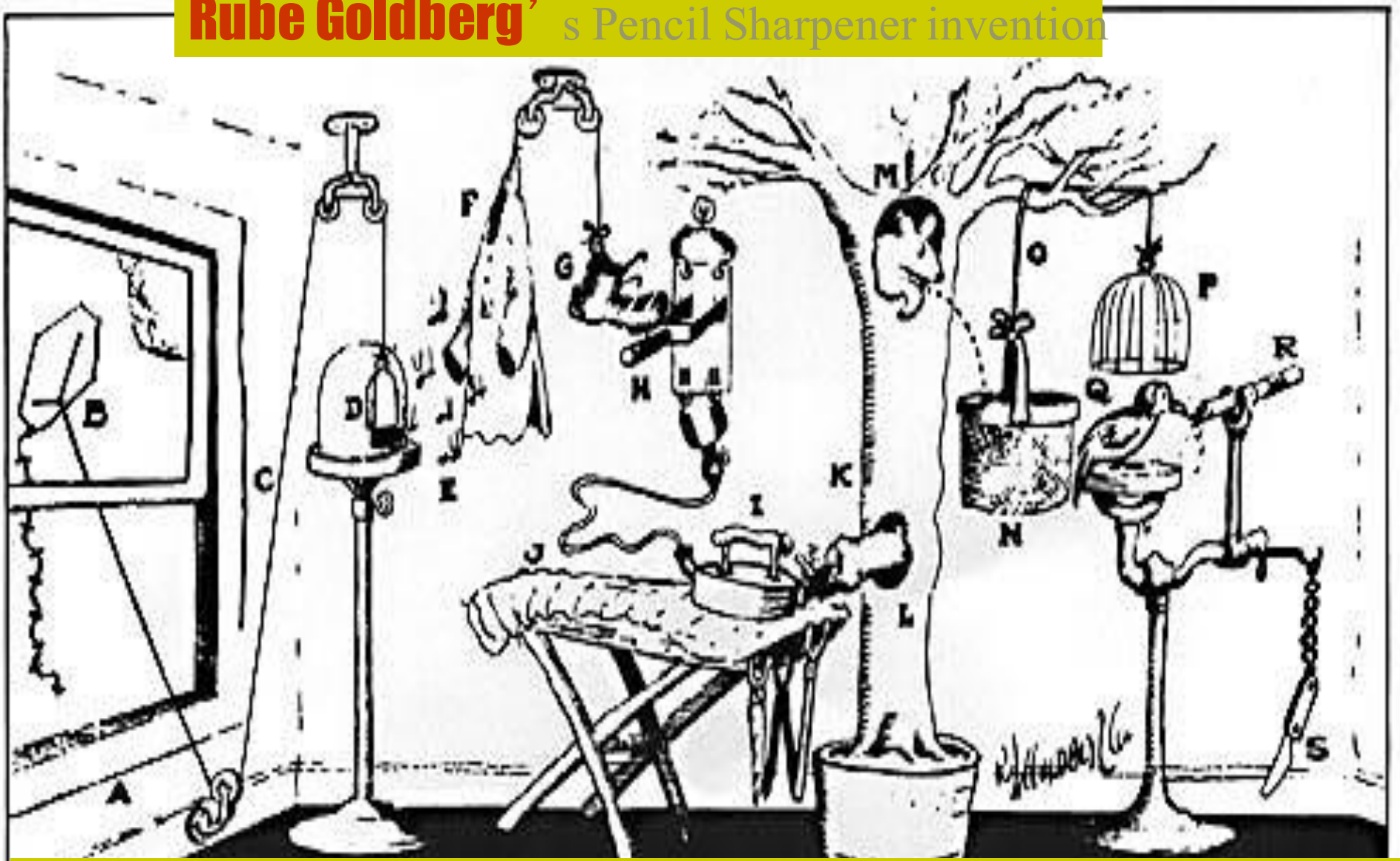
Scan alignment 1 **against sequence databases using GLAM2SCAN.**

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**

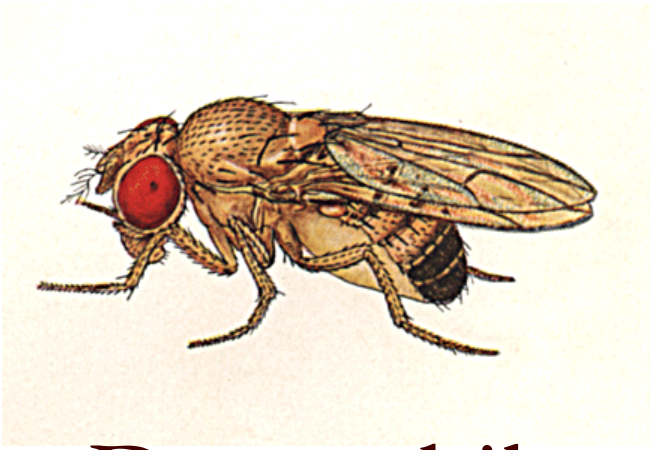
Rube Goldberg's Pencil Sharpener invention



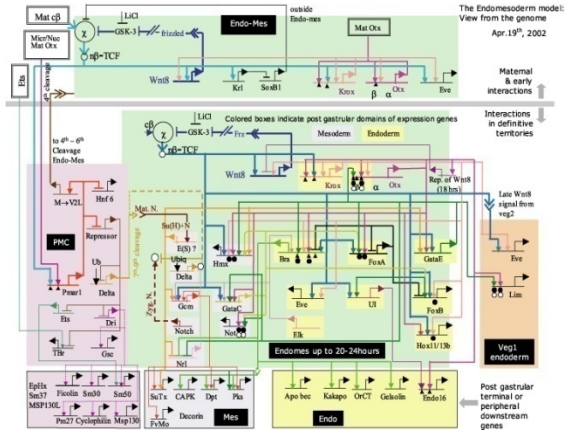
Emergency knife **(S)** is always handy in case opossum or the woodpecker gets sick and can't work.

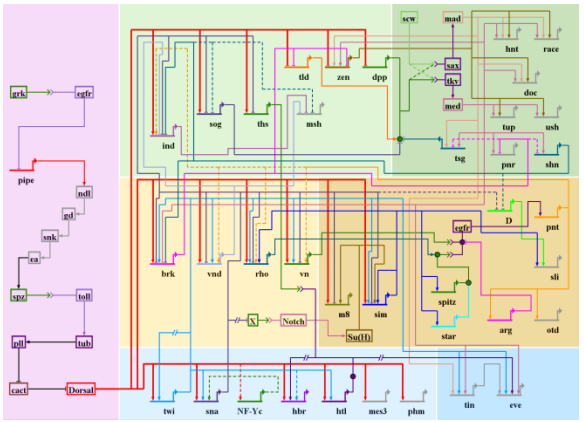
A Tale of Two Networks

Sea Urchin



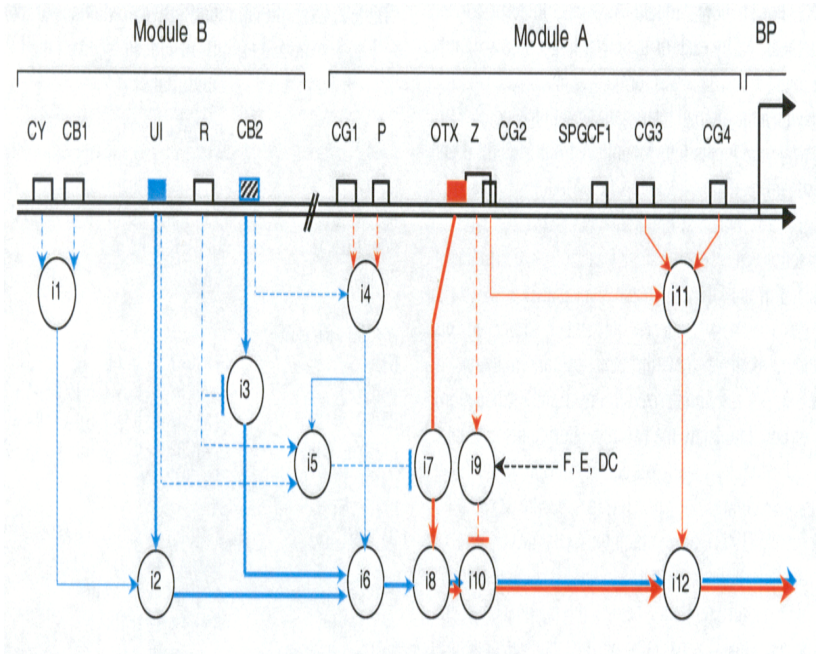
Drosophila





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

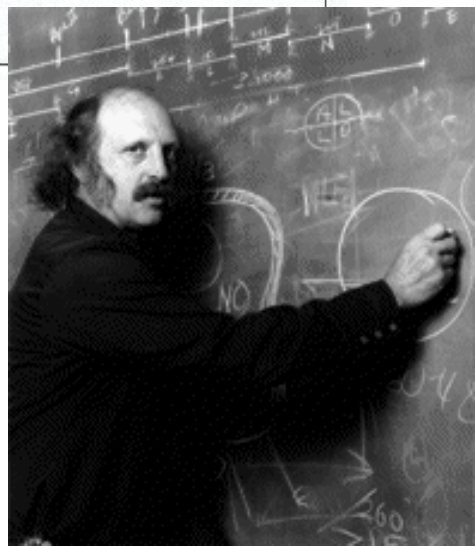
A Proposal for Nobel Prize



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)		

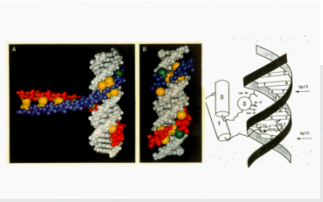
"Programs built into the DNA of every animal."

Eric H. Davidson



Genomic Regulatory Systems

The Dogma



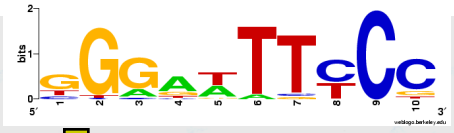
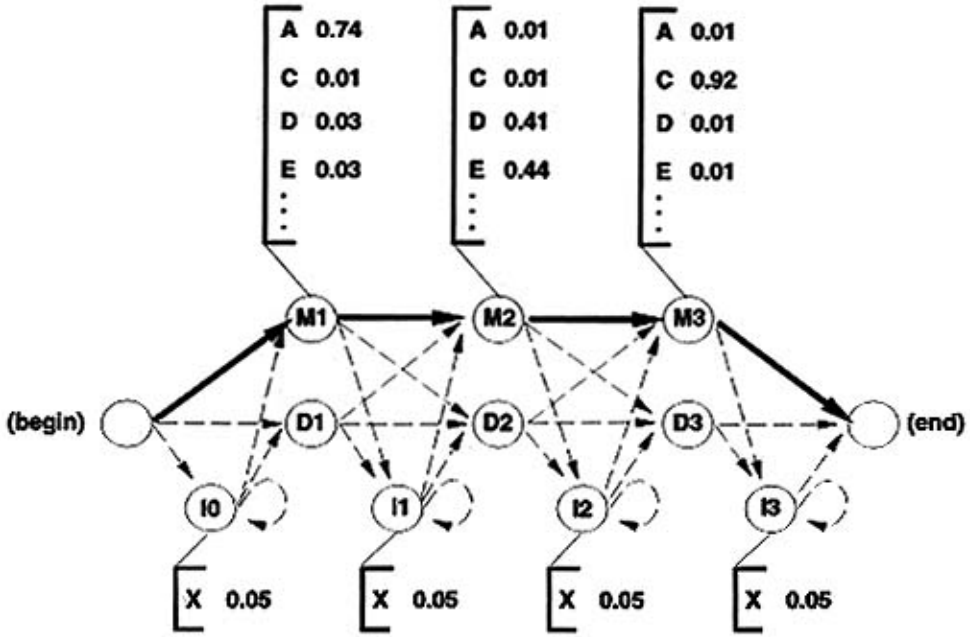
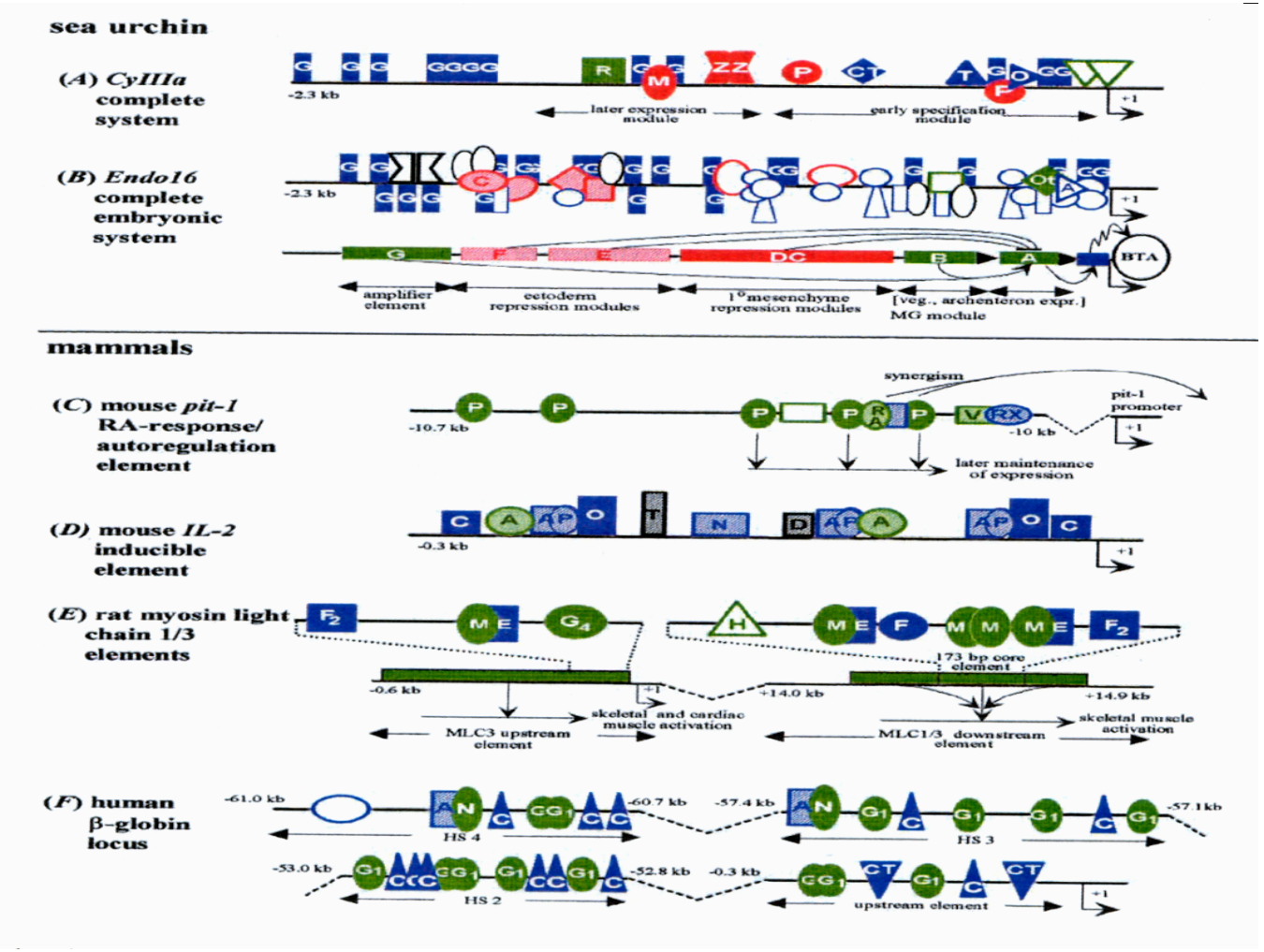


Figure 9.2 Schematic model for transcriptional activation. The TATA box-binding protein, which bends the DNA upon binding to the TATA box, binds to RNA polymerase and a number of associated proteins to form the preinitiation complex. This complex interacts with different specific transcription factors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

Hidden Markov Models (Ch. 4)



Genomic Regulatory Regions

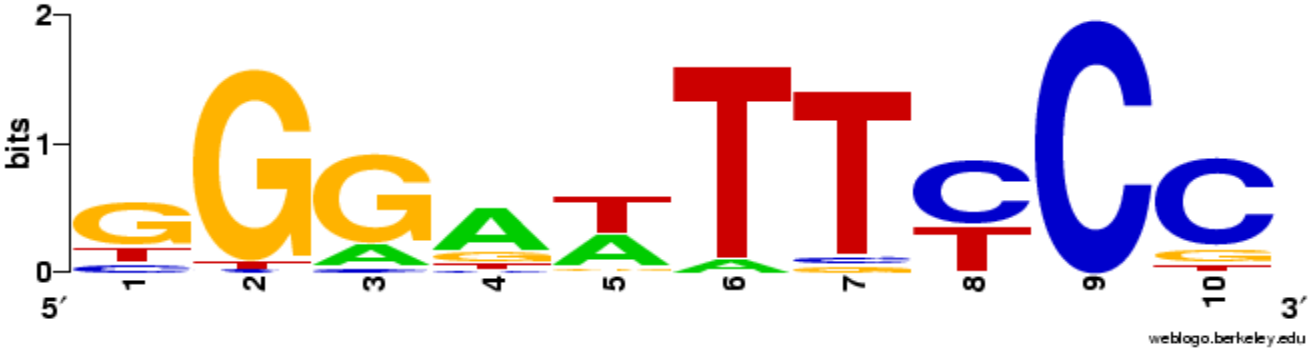


Phylogenetic Trees (Ch. 3) ???

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

Phylogenetic trees are not good models for the Regulatory Genome

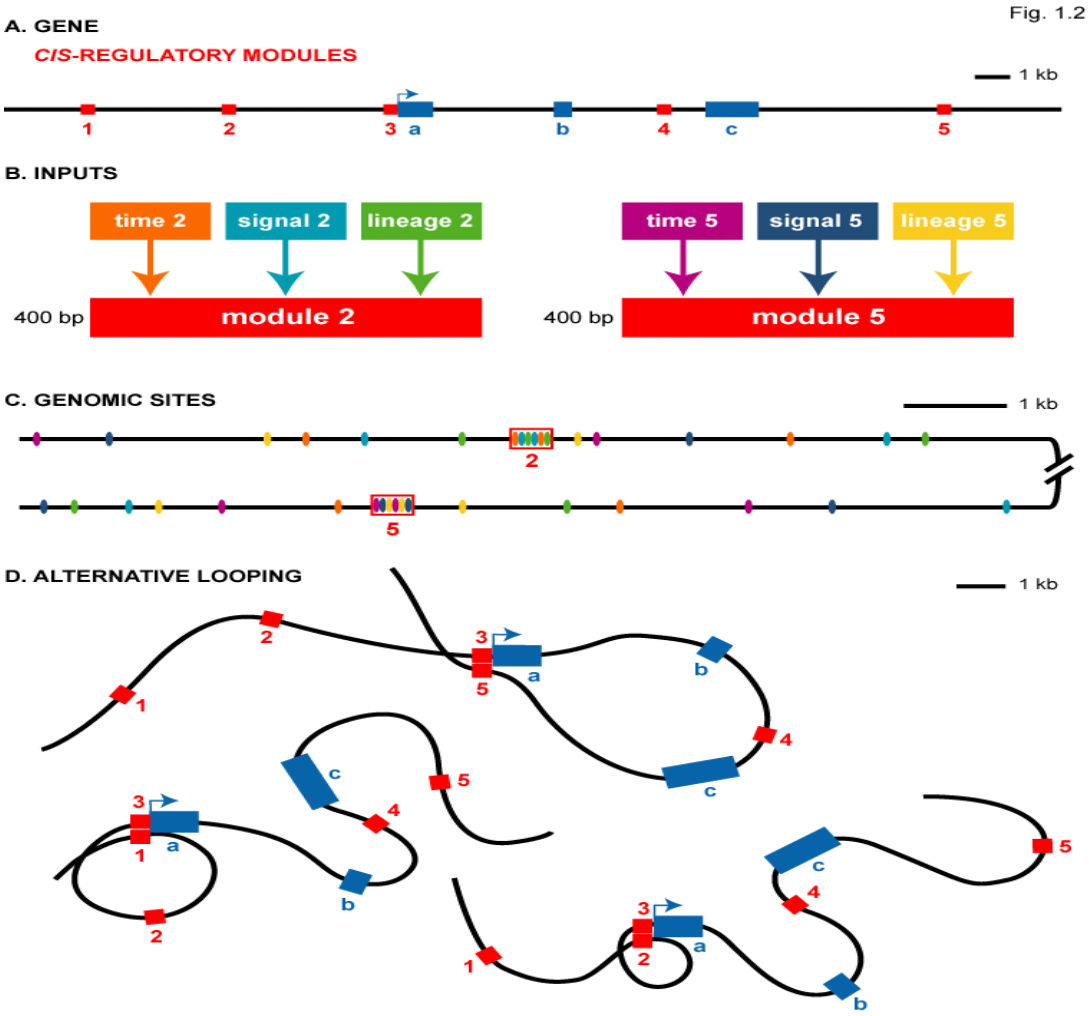
TF Binding Site Complexity



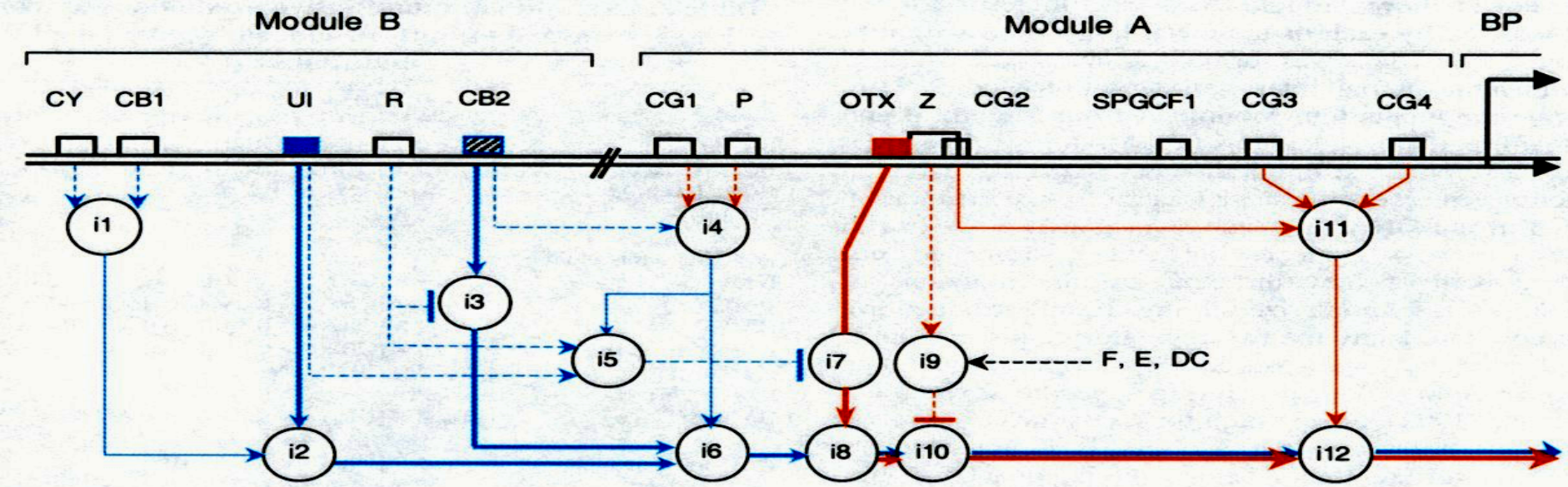
Genome Complexity

1 Billion DNA bases
20,000 Genes

cis-Regulatory Modules Complexity



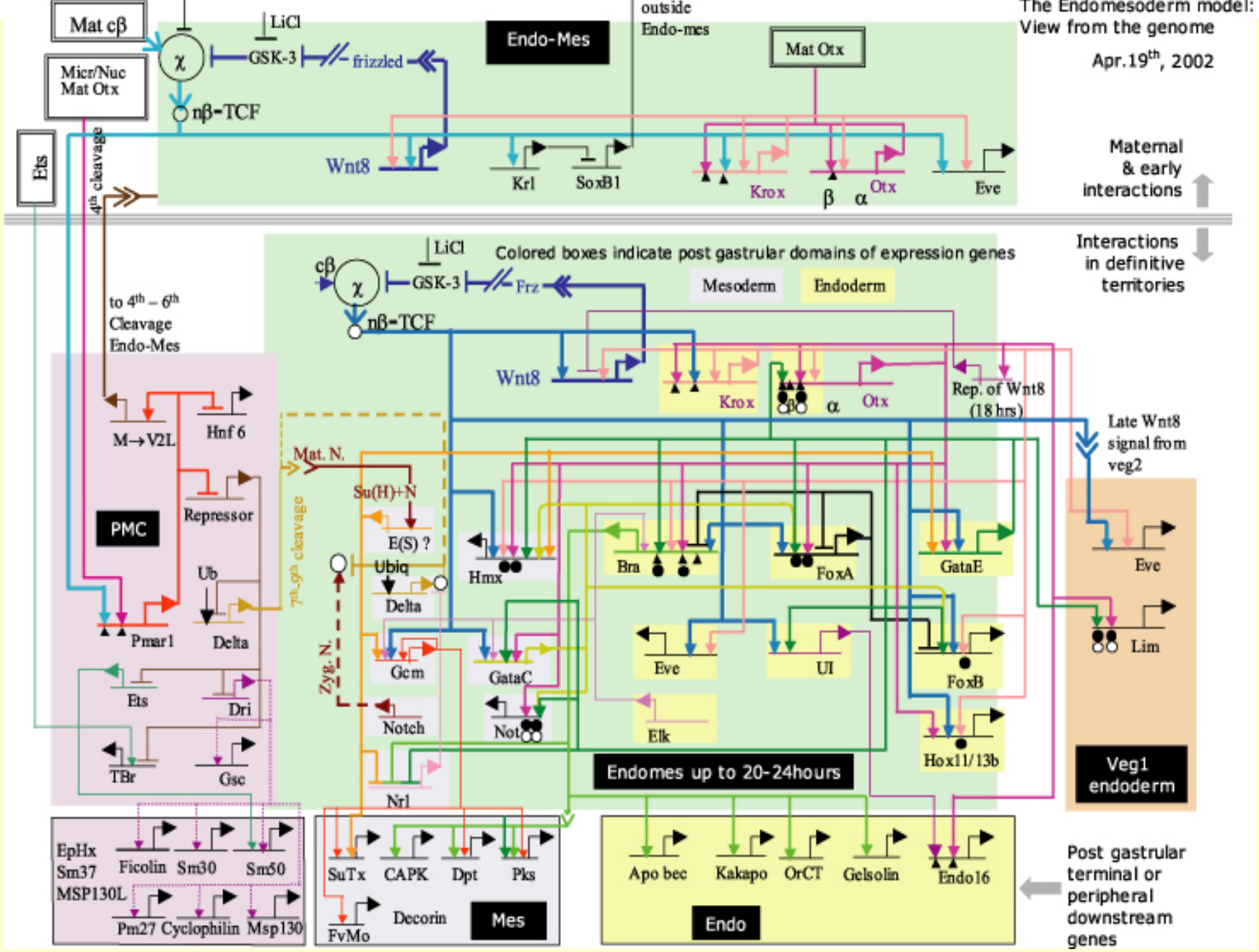
200,000 cis-Modules



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

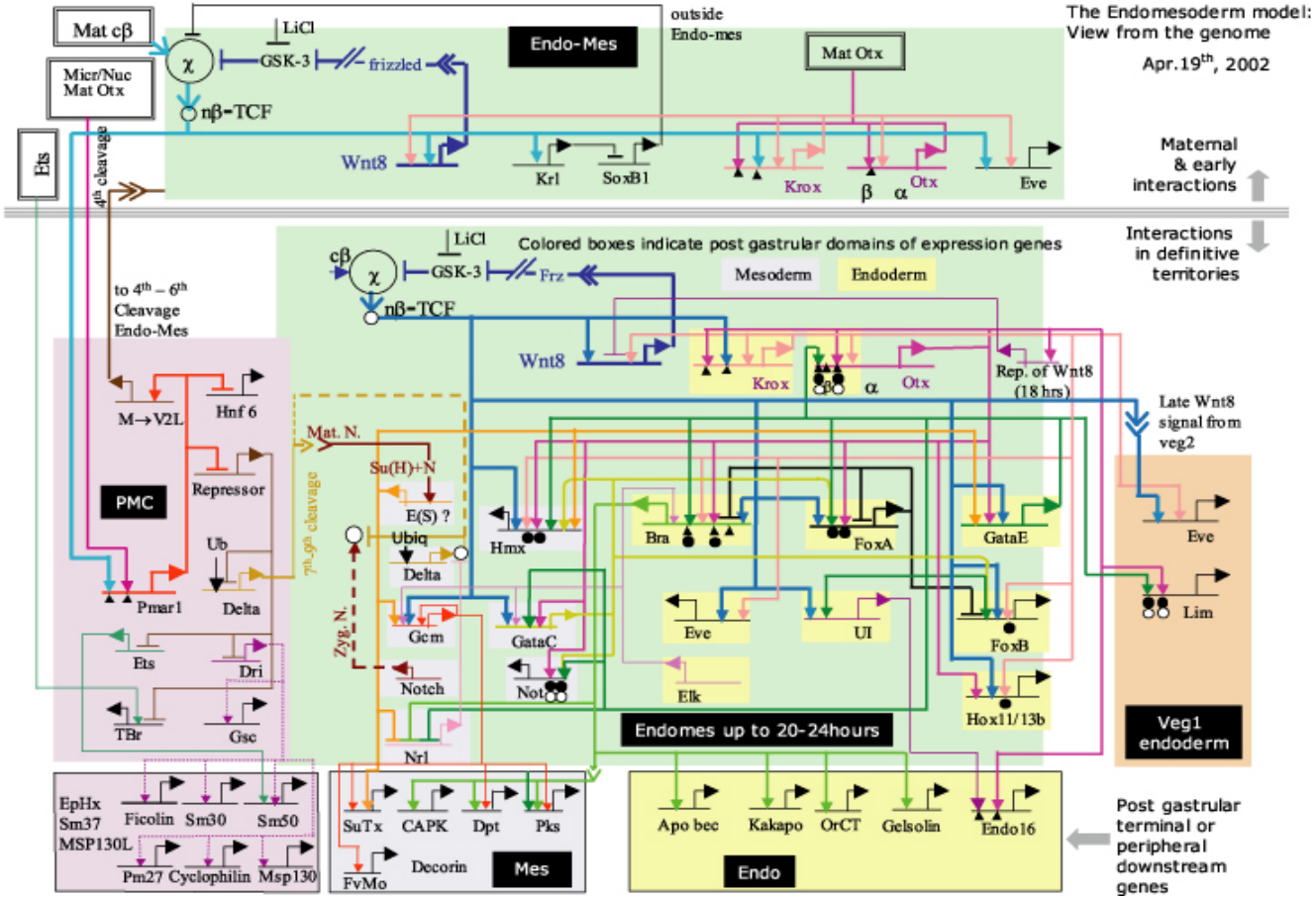
The DNA program that regulates the expression of ***endo16*** in sea urchin

- **THE FIRST GENE**



THE FIRST NETWORK

The View from the Genome



A Case Study

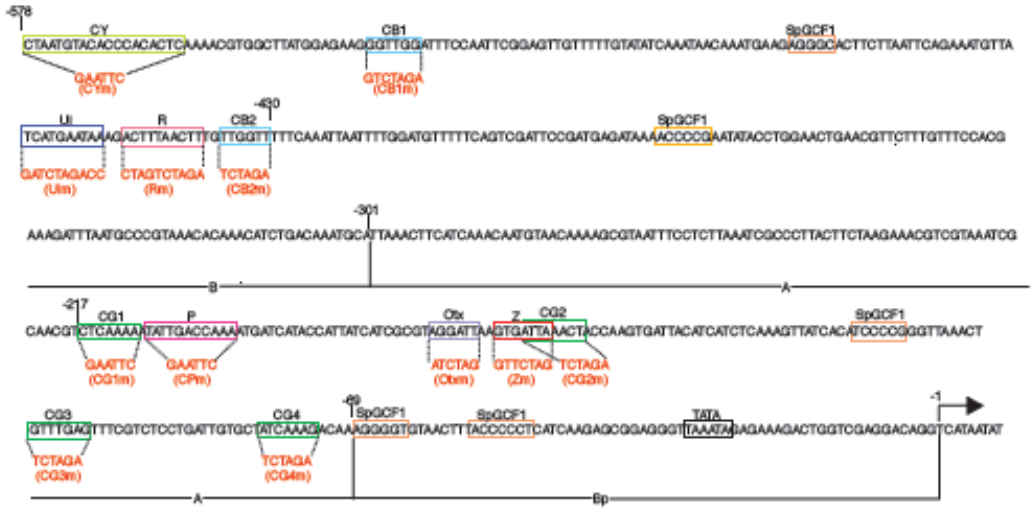


Figure 2: Quintessential diagram (from [25])

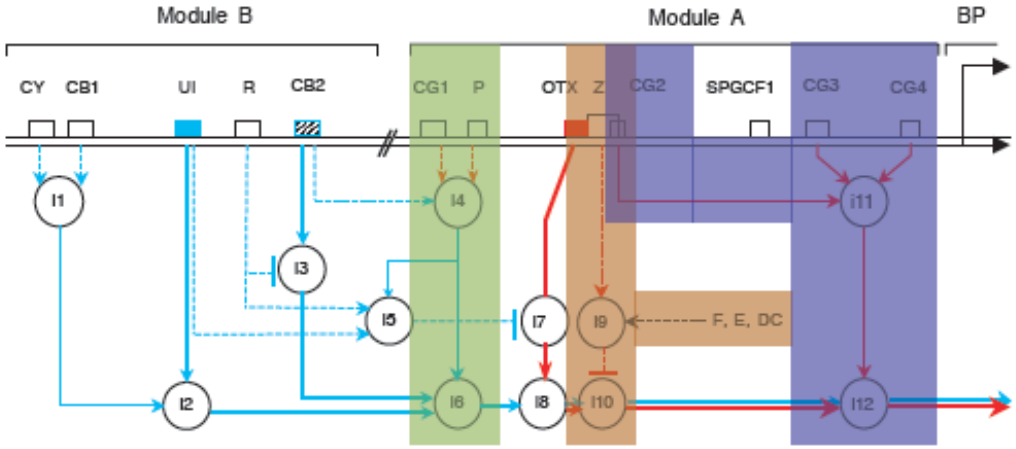
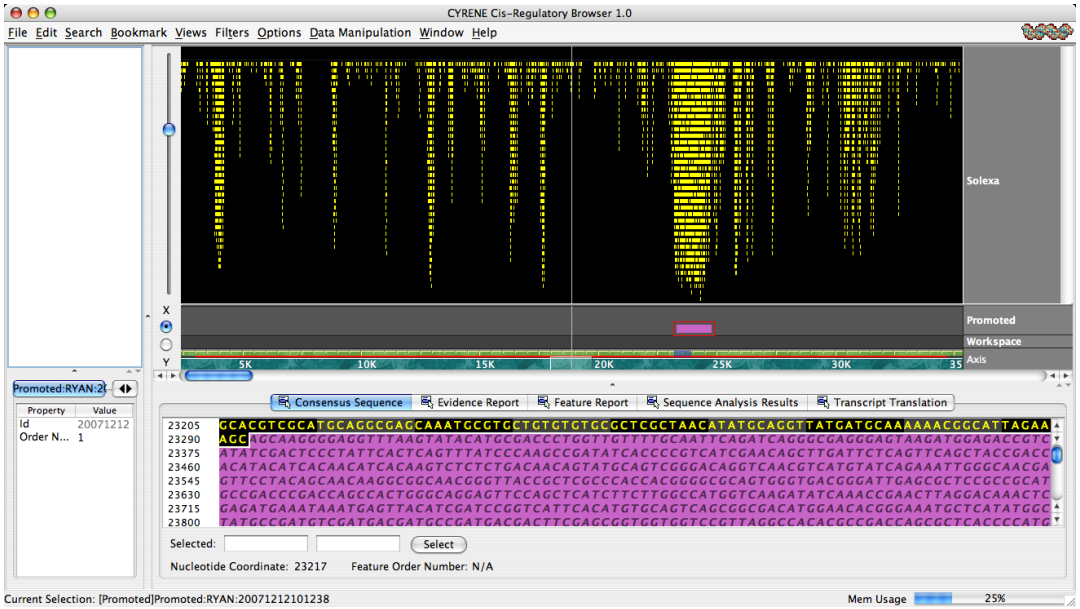


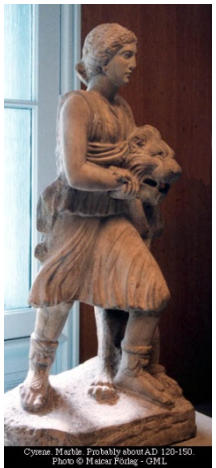
Figure 3: Computational logic model for Modules A and B of *endo16* (from [25])

Cyrene

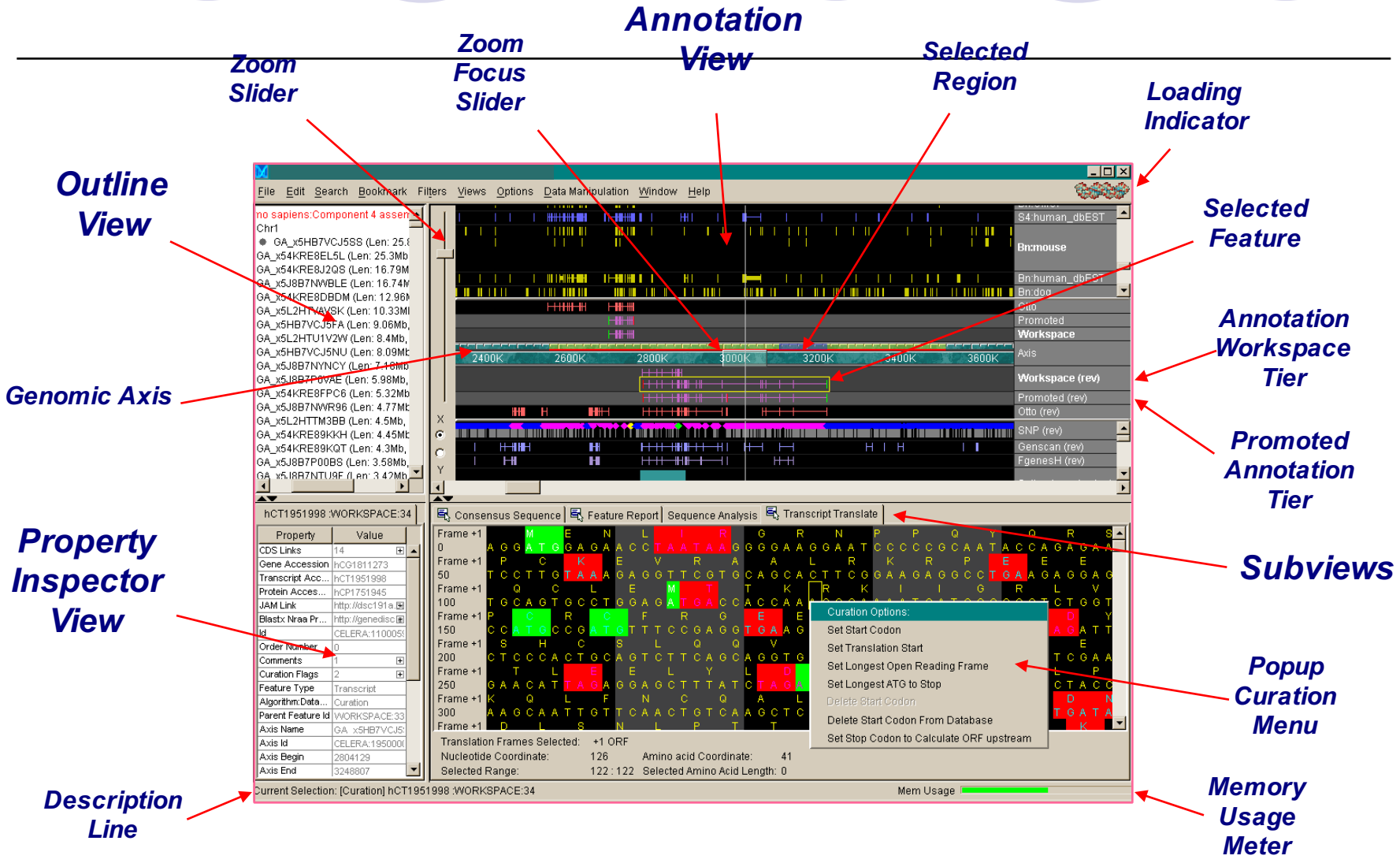
Ryan Tarpine

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.

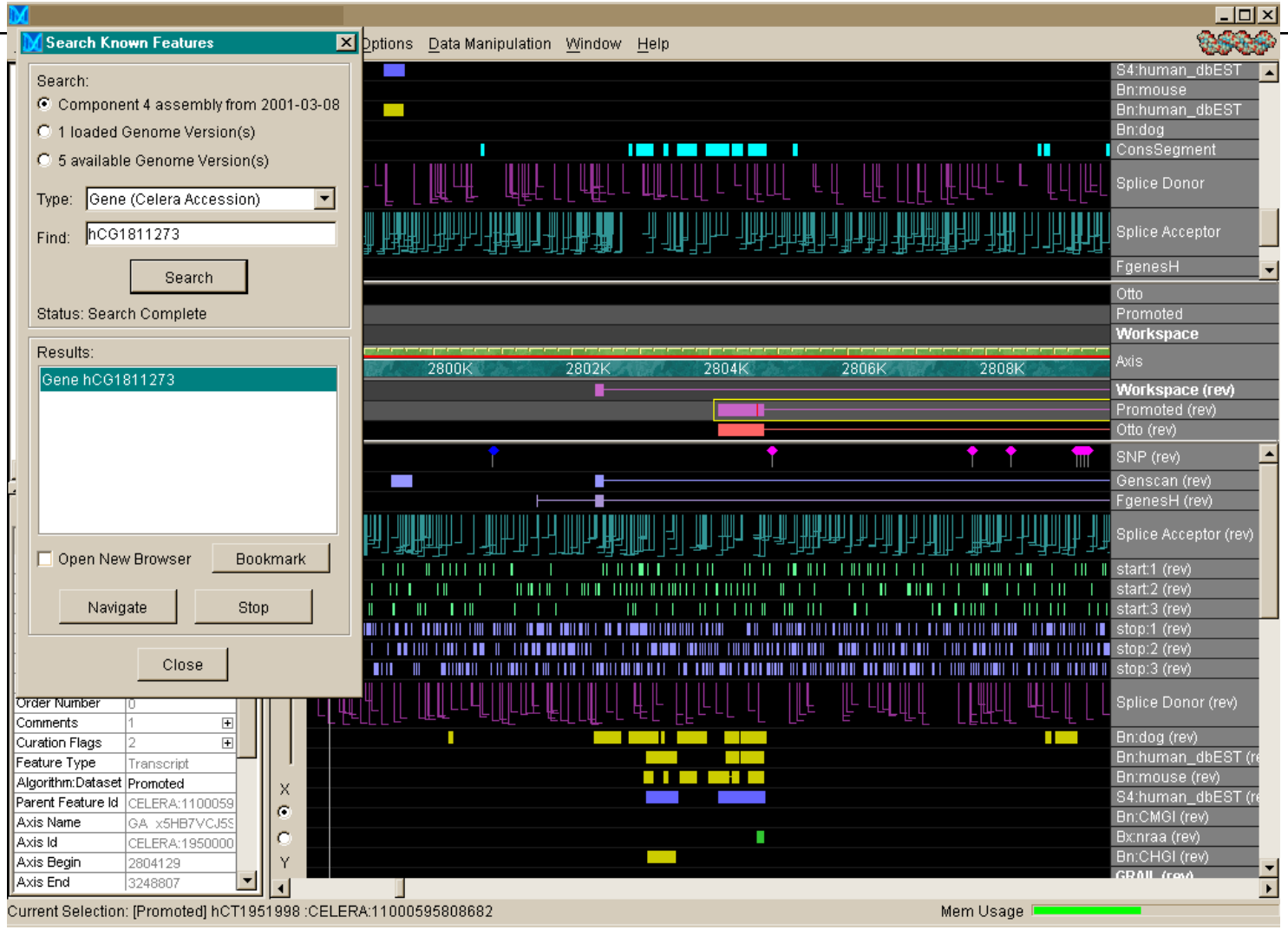




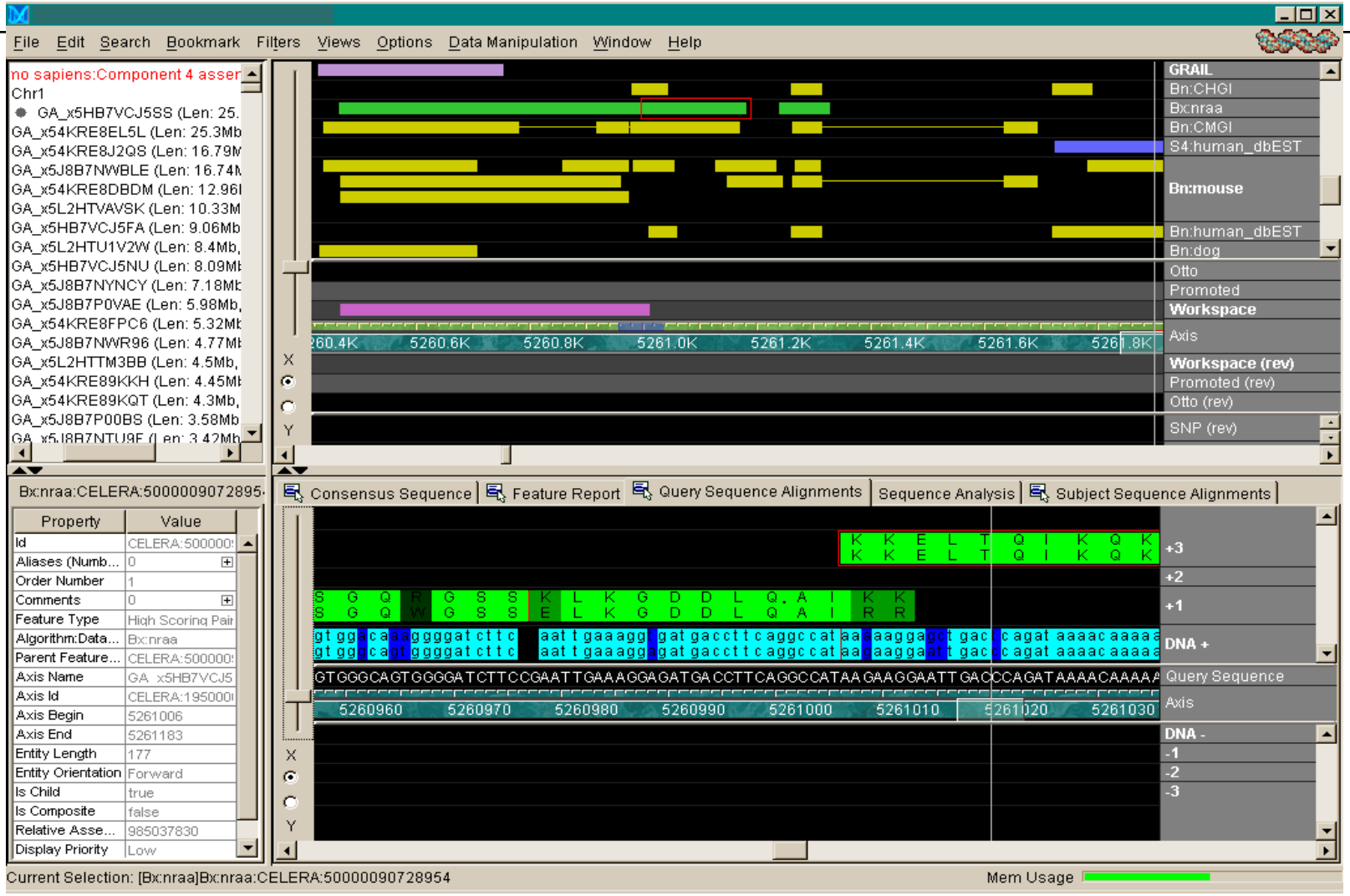
The cis-Browser

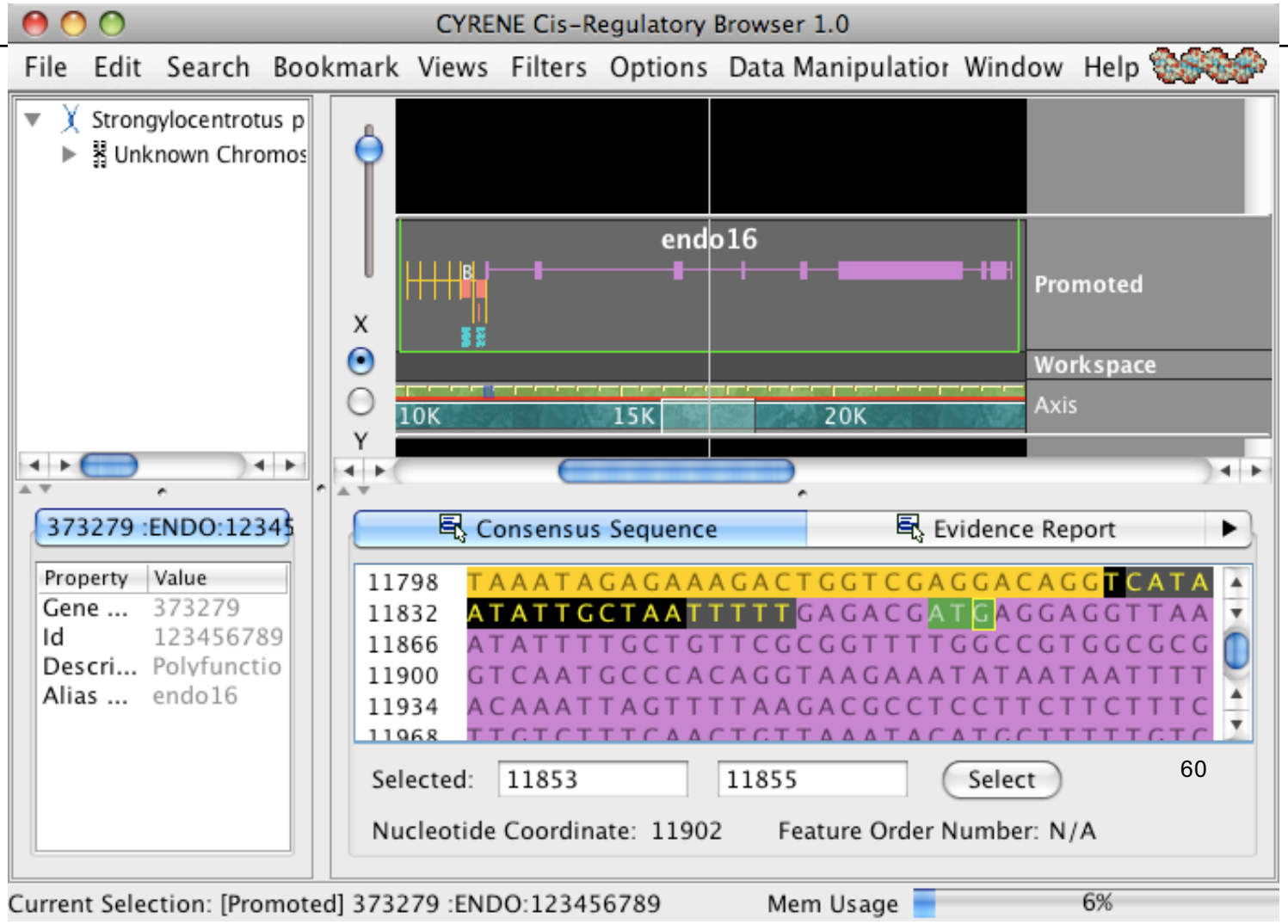


Transcript Curation

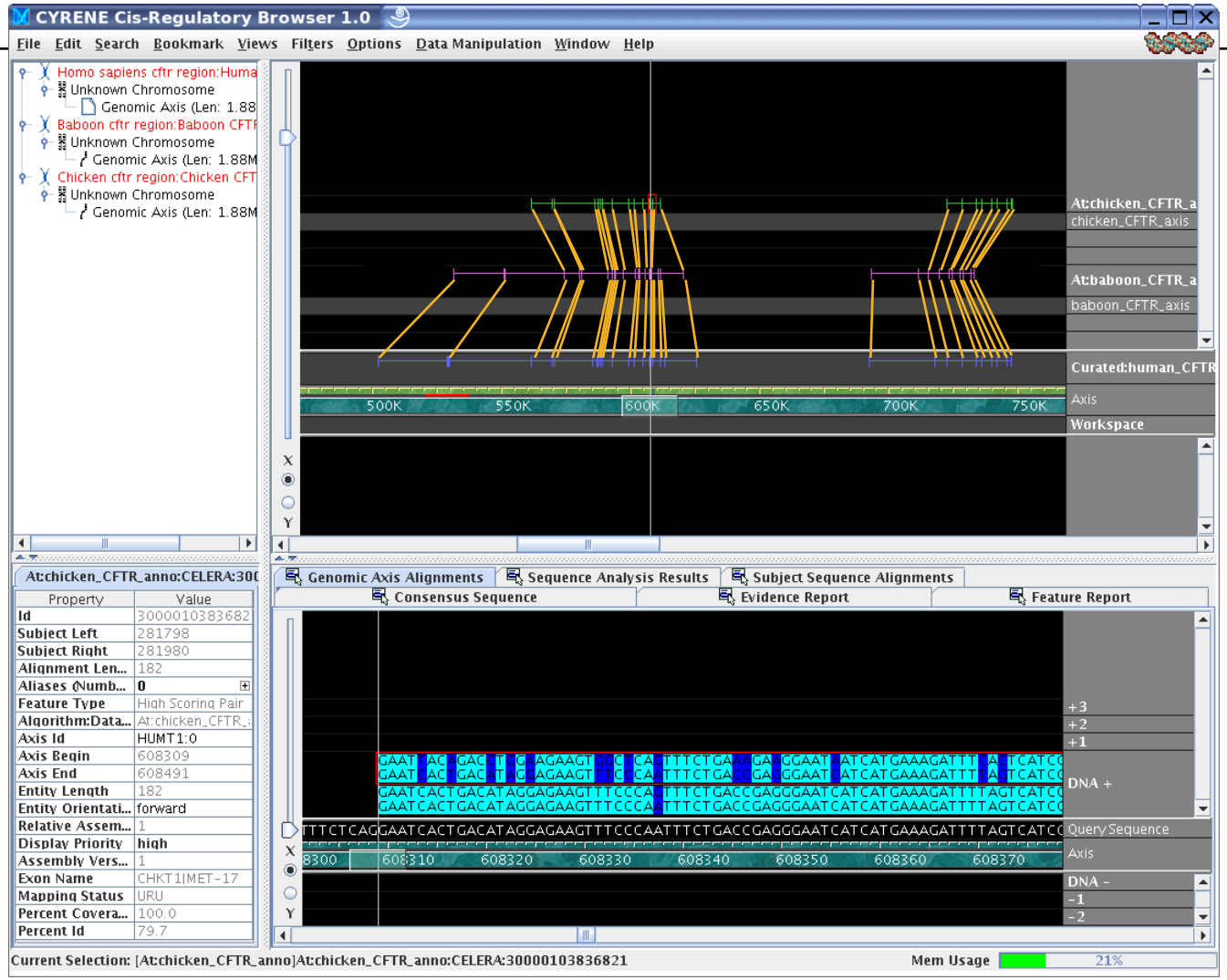


Sequence Comparison



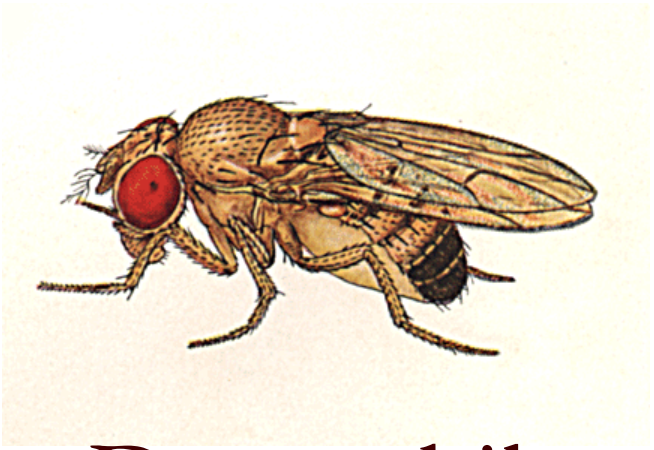

CYRENE Cis-Regulatory Browser 1.0
File Edit Search Bookmark Views Filters Options Data Manipulatior Window Help
Strongylocentrotus p
Unknown Chromos
endo16
Promoted
Workspace
Axis
X
Y
10K
15K
20K
373279 :ENDO:12345
Property Value
Gene ... 373279
Id 123456789
Descri... Polyfunctio
Alias ... endo16
Consensus 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

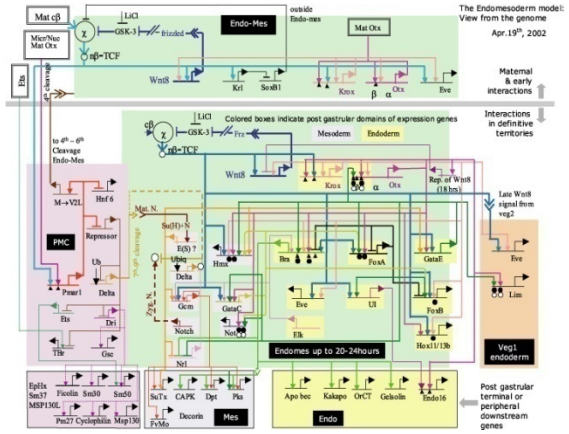


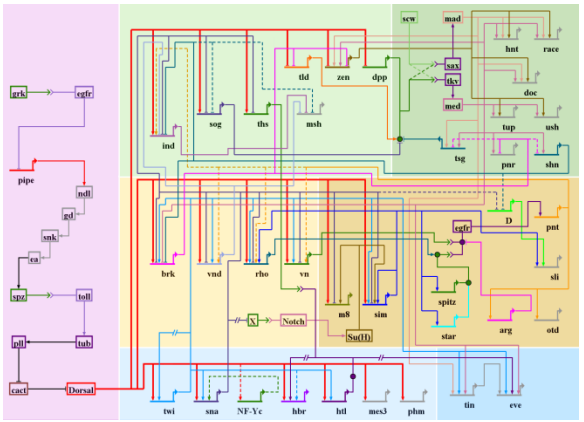
A Tale of Two Networks

Sea Urchin



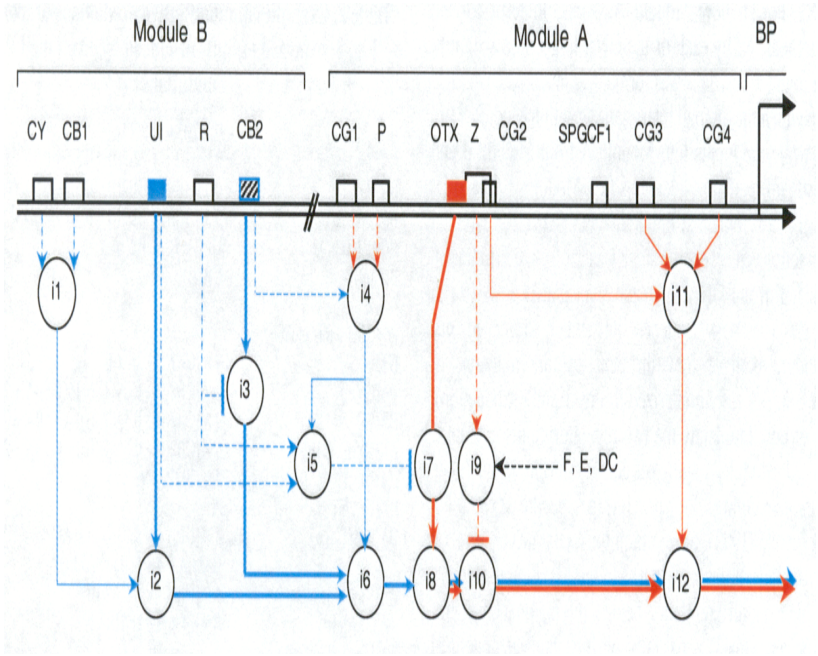
Drosophila





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

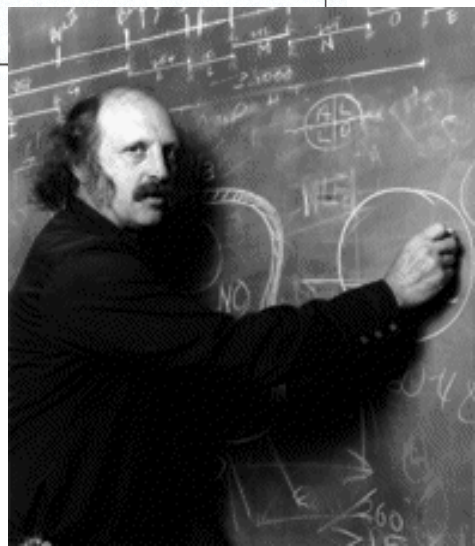
A Proposal for Nobel Prize



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."

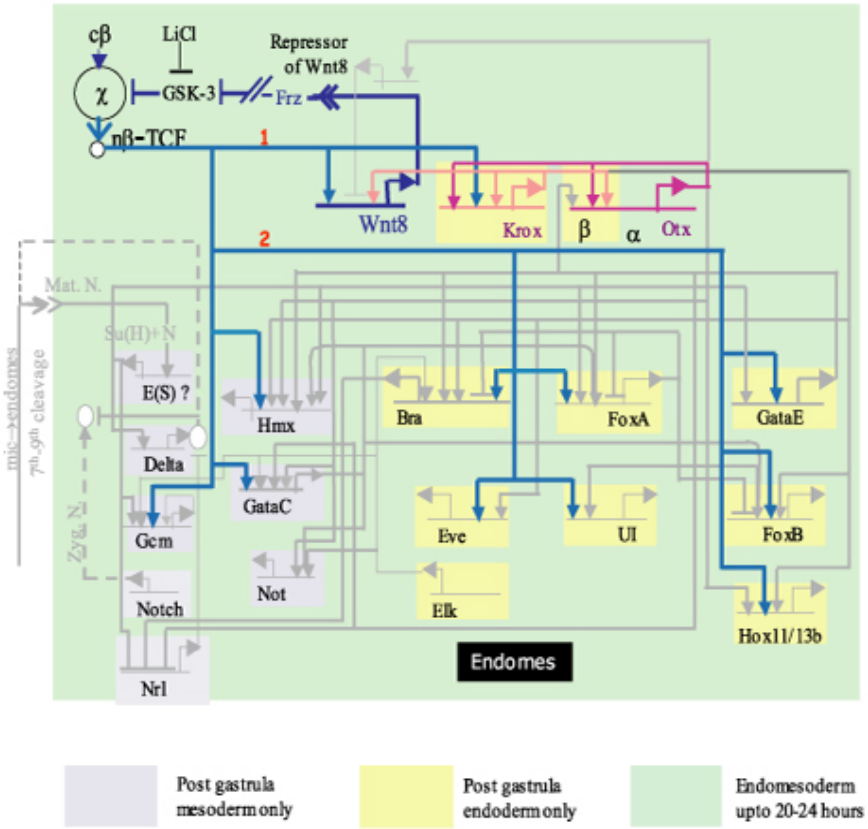
Eric H. Davidson



Genomic Regulatory Systems

The View from the Nucleus

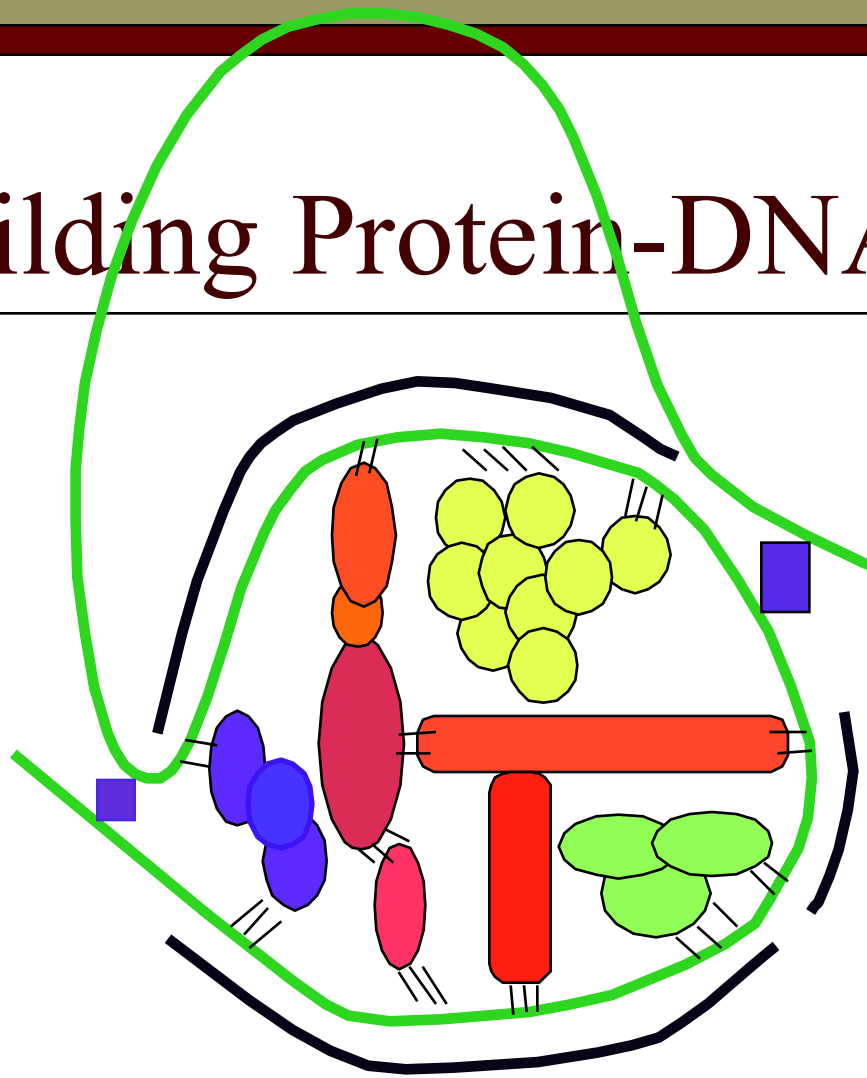
View from the nucleus: Endomesoderm nuclei to hatching blastula stage; the Wnt8/Tcf signalling loop and its genes.
Apr. 19th , 2002



Notes:

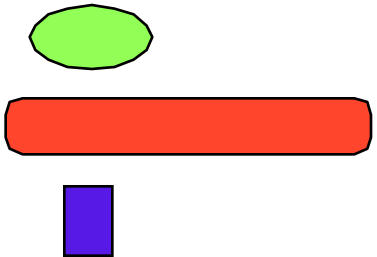
1. β-catenin/Tcf input now produced by a zygotic signaling loop driven by Wnt8 expression in endmesoderm cells.
2. β-catenin/Tcf input required for expression of many regulatory genes that become active in the veg_2 endomesodermal territory during early-mid blastula stage.

Building Protein-DNA Assemblies



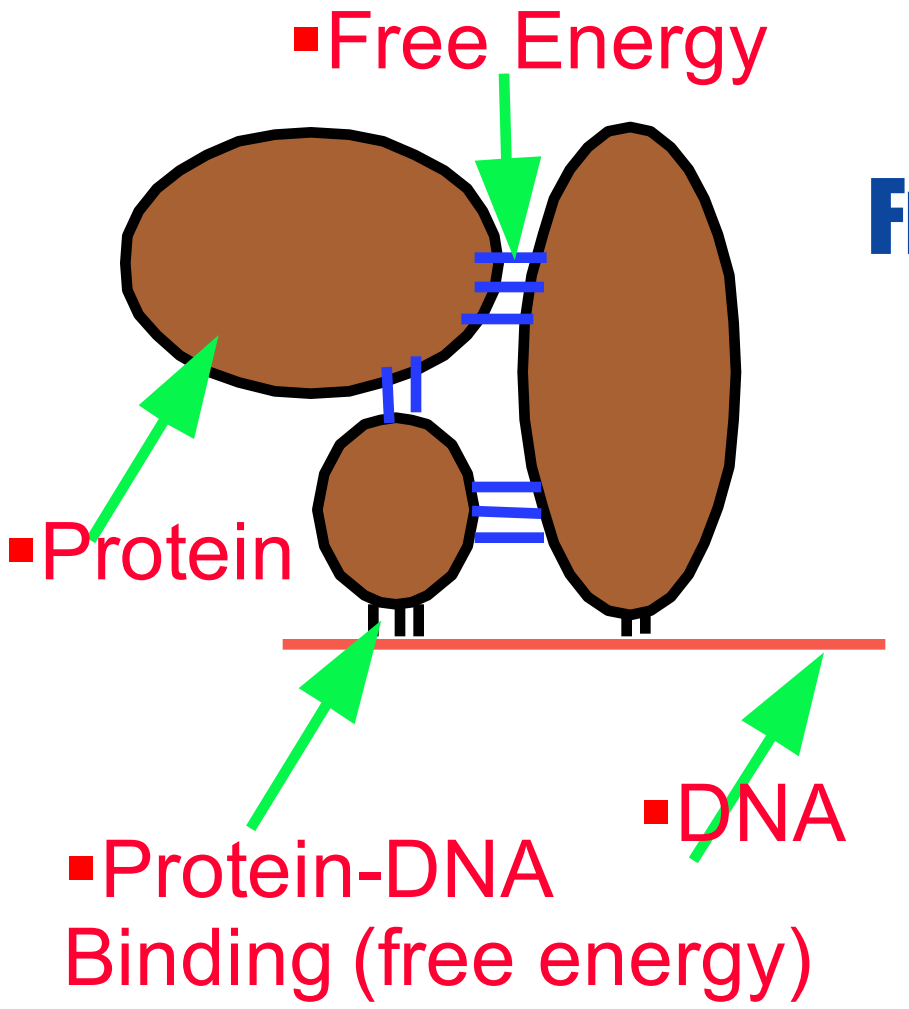
- DNA
- cismodule

- Cooperativity
- Linear-amp
- Gates
- Potentiality



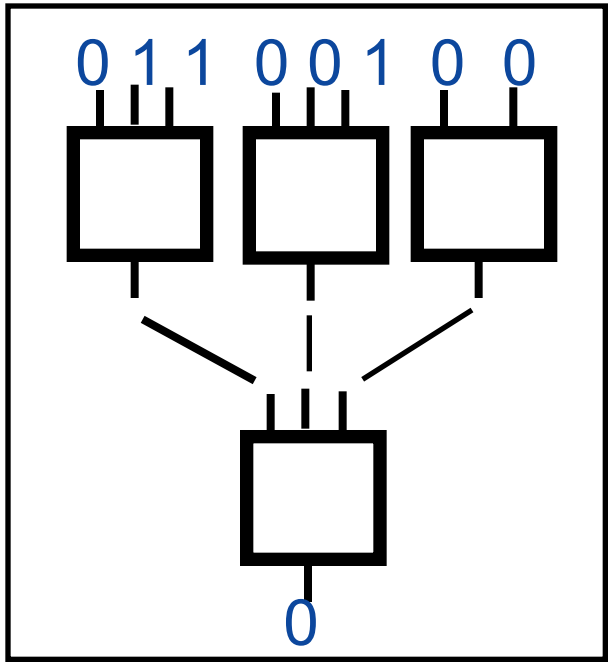
- Inter-cismodule linkage
- Insulation
- Communication

The Building Blocks

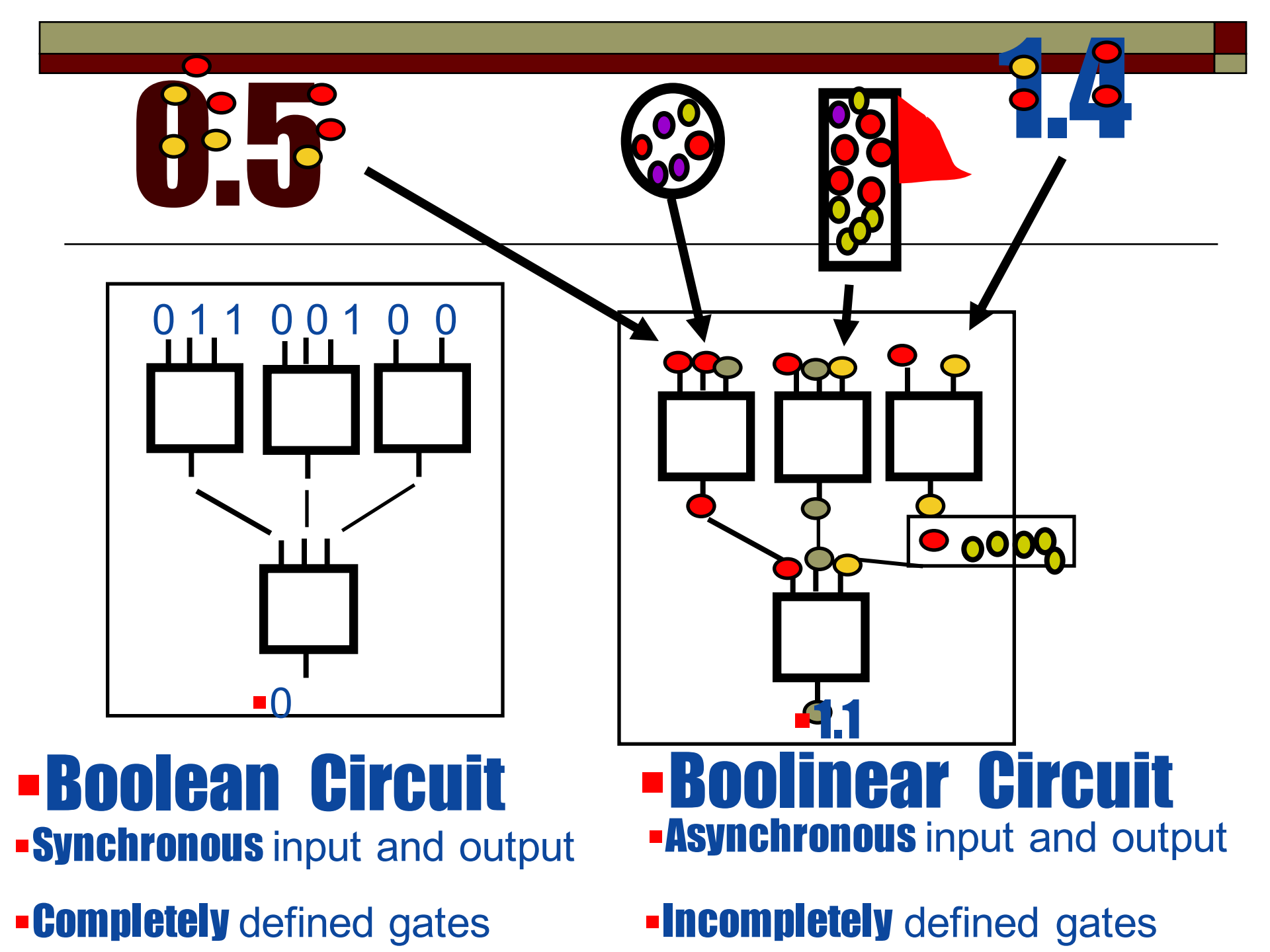


Free energy is the "GLUE"

Information Processing

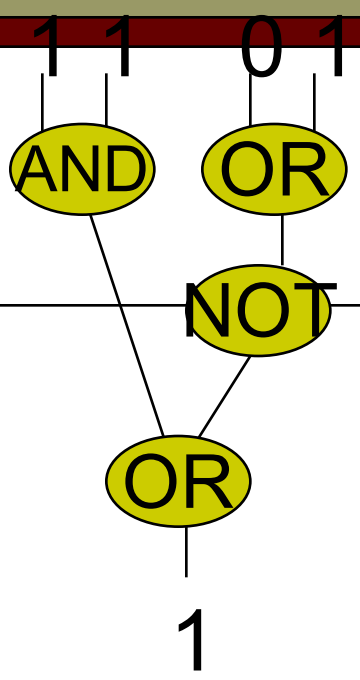


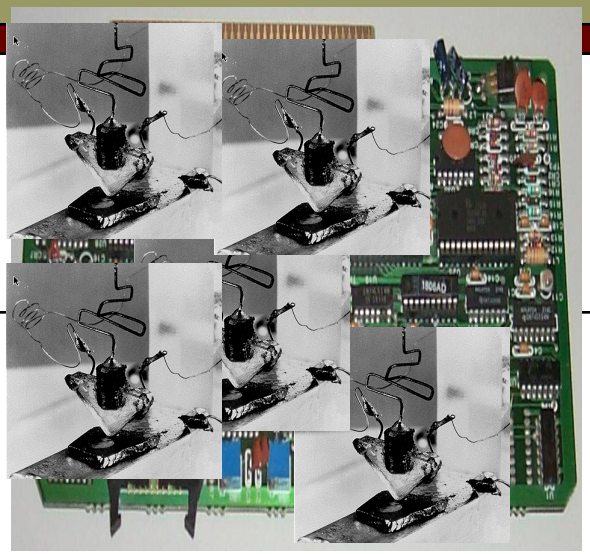
- **Boolean Circuit**
- **Synchronous** input and output
- **Completely** defined gates



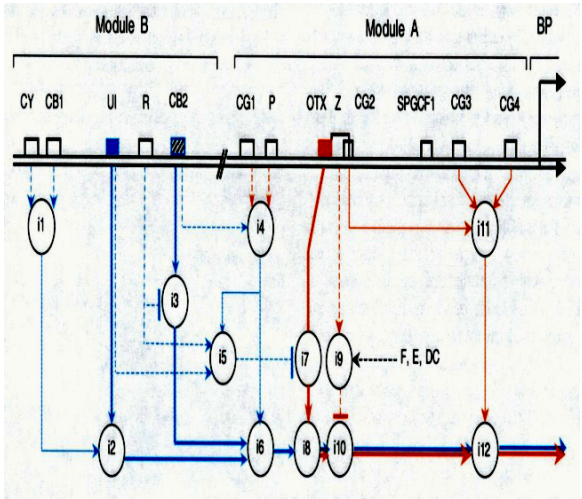
- **Boolean Circuit**
  - **Synchronous** input and output
  - **Completely** defined gates
- **Boolinear Circuit**
  - **Asynchronous** input and output
  - **Incompletely** defined gates

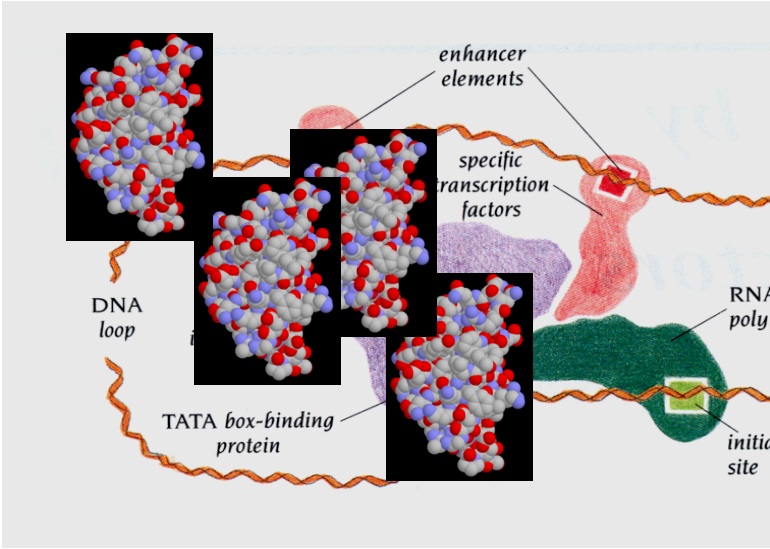
IF (x1 = 1
AND
x2= 1)
THEN
…..





GTAGGATTAAG
……
CATCCTAATTC
…….
GTATCTAGAAG
…….






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

The Gold-Bug – E.A. Poe

E. A. Poe "The father of cryptography"

Circumstances, and a certain biased of mind, have led me to take interest in such riddles, and it may well be doubted whether human ingenuity can construct an enigma of the kind which human ingenuity may not, by proper application, resolve.

What Language ?

“In the present case -- indeed in all cases of secret writing -- the first question regards the language of the cipher; for the principles of solution, so far, especially, as the more simple ciphers are concerned, depend upon, and are varied by, the genius of a particular idiom.

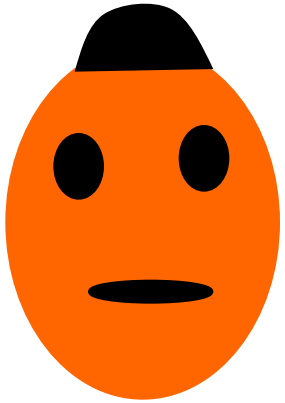
What Language ? …

… In general, there is no alternative but experiment (directed by probabilities) of every tongue known to him who attempts the solution, until the true one be attained.

… But for this consideration, I should have begun my attempts with the Spanish and French, as the tongues in which a secret of this kind would most naturally have been written by a pirate of the Spanish main. As it was, I assumed the cryptograph to be English.”

The Gold-bug – the story

Mr. William Legrand, left New Orleans and took residence on Sullivan's Island, near South Carolina.



His servant was Jupiter, an old negro. He calls Mr. Legrand "Massa Will."

One day, Massa Will found a bug, a scarabeus, which he believed is totally new.

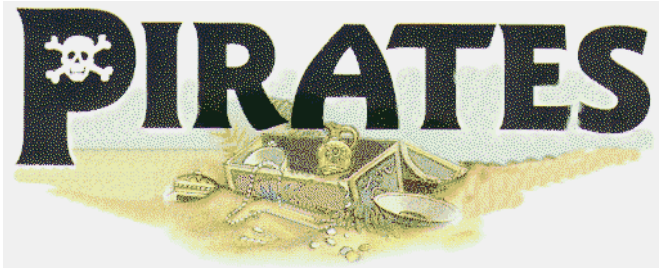
Jupiter describes the bug in his language:

"…de bug is a gole-bug, solid, ebery bit of him, inside and all, sep him wing – neber feel half so hebby a bug in my life."

The design on the bug's back resembled a death's-head …. And the story continues and they were searching for a big treasure hidden by a famous pirate Captain Kidd.

Captain Kidd's Code

**53||^305))6*;4826)4|.)4|);806*
;48^8%60))85;1|(;:|*8^83(88)5*^
;46(;88*96*?;8)*|(;485);5*^2:*|
(;4956*2(5*_4)8%8*;4069285);)
6^8)4||;1(|9;48081;8:8|1;48^85;4)
485^528806*81(|9;48;(88;4(|?34
;48)4|;161;:188;|?;**



Statistics

No division between words

Statistics of the character

8 there are 33.
; there are 26.
4 there are 19.
|) there are 16.
* there are 13.
5 there are 12.
6 there are 11.
^1 there are 8.
0 there are 6.
92 there are 5
:3 there are 4.
? there are 3.
% there are 2.
_ there are 1.

We found our first letter!

In English the letter which most frequently occurs is e.

Afterwards, the succession is:

a o i d h n r s t u y c f g l m w b k p q x z

About English text

"Now, in English, the letter which most frequently occurs is e.

Afterwards, the succession runs thus: a o i d h n r s t u y c f g l m w b k p q x z."

Captain Kidd's Code: 8 is "e"

53||^305))6*;4826)4|.)4|);806*
;48^8%60))85;1|(;:|*8^83(88)5*^
;46(;88*96*?;8)*|(;485);5*^2:*|
(;4956*2(5*_4)8%8*;4069285);)
6^8)4||;1(|9;48081;8:8|1;48^85;4)
485^52888 06*81(|9;48;(88;4(|?34
;48)4|;161;:188;|?;

8 is e

Progress…

As our predominant character is 8, we will commence by assuming it as the e of the natural alphabet. To verify the supposition, let us observe if the 8 be seen often in couples --for e is doubled with great frequency in English --in such words, for example, as 'meet,' 'fleet,' 'speed, 'seen,' 'been, ‘ 'agree,' &c.

Progress…

"Let us assume 8, then, as e. Now, of all words in the language, 'the' is the most usual; let us see, therefore, whether they are not repetitions of any three characters in the same order of collocation, the last of them being 8.

Progress…

On inspection, we find no less than seven such arrangements, the characters being ;48.

Progress ….

We may, therefore, assume that the semicolon represents t, that 4 represents h, and that 8 represents e --the last being now well confirmed. Thus a great step has been taken.

Captain Kidd's Code:

is **t** and **4** is **h**

53||^305))6*;4826)4|.)4|);806*
;48^8%60))85;1|(;:|*8^83(88)5*^
;46(;88*96*?;8)*|(;485);5*^2:*|
(;4956*2(5*_4)8%8*;4069285);)
6^8)4||;1(|9;48081;8:8|1;48^85;4)
485^52888806*81(|9;48;(88;4(|?34
;48)4|;161;:188;|?;

;48

must be "the" most frequent word

Progress …..

"But, having established a single word, we are enabled to establish a vastly important point; that is to say, several commencements and terminations of other words.
Let us refer, for example, to the last instance but one, in which the combination ;48 occurs --not far from the end of the cipher. We know that the semicolon immediately ensuing is the commencement of a word, and, of the six characters succeeding this 'the,' we are cognizant of no less than five. Let us set these characters down, thus, by the letters we know them to represent, leaving a space for the unknown--teeth.

Progress…

"Here we are enabled, at once, to discard the 'th,' as forming no portion of the word commencing with the first t; since, by experiment of the entire alphabet for a letter adapted to the vacancy we perceive that no word can be formed of which this th can be a part. We are thus narrowed into

t ee,

and, going through the alphabet, if necessary, as before, we arrive at the word 'tree,' as the sole possible reading. We thus gain another letter, r, represented by (, with the words 'the tree' in juxtaposition.

Progress…

"Looking beyond these words, for a short distance, we again see the combination ;48, and employ it by way of termination to what immediately precedes. We have thus this arrangement:

the tree ;4(+?34 the,

or substituting the natural letters, where known, it reads thus:

the tree thr+?3h the.

Progress…

"Now, if, in place of the unknown characters, we leave blank spaces, or substitute dots, we read thus:

the tree thr...h the,

when the word 'through' makes itself evident at once. But this discovery gives us three new letters, o, u and g, represented by + ? and 3.

Progress…

"Looking now, narrowly, through the cipher for combinations of known characters, we find, not very far from the beginning, this arrangement,

83(88, or egree,

which, plainly, is the conclusion of the word 'degree,' and gives us another letter, d, represented by !.

"Four letters beyond the word 'degree,' we perceive the combination

;46(;88*.

Progress…

"Translating the known characters, and representing the unknown by dots, as before, we read thus:

th.rtee.

an arrangement immediately suggestive of the word 'thirteen,' and again furnishing us with two new characters, i and n, represented by 6 and *.

"Referring, now, to the beginning of the cryptograph, we find the combination,

53++!.

Progress…

"Translating, as before, we obtain

.good,

which assures us that the first letter is A, and that the first two words are 'A good.'

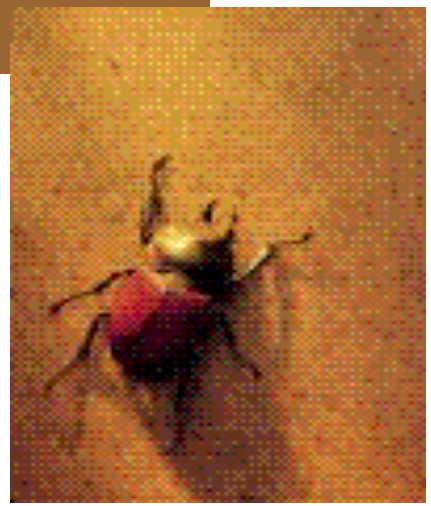
The decoding key

5	represents	a
!	"	d
8	"	e
3	"	g
4	"	h
6	"	i
*	"	n
+	"	o
(	"	r
;	"	t

The mystery text revealed

It now only remains to give you the full translation of the characters upon the parchment, as unriddled. Here it is:

"A good glass in the bishop's hostel in the devil's seat twenty-one degrees and thirteen minutes northeast and by north main branch seventh limb east side shoot from the left eye of the death's-head a bee line from the tree through the shot fifty feet out."

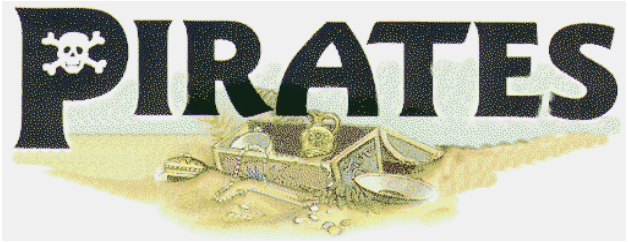


The Solution

“A good glass in the bishop’s hostel in the devil’s seat forty-one degrees and thirteen minutes northeast and by north main branch seventh limb east side shoot from the left eye of the death’s head a bee-line from the tree through the shot fifty feet out.”

“’ A good glass in the bishop's hostel in the devil's --twenty-one degrees and thirteen minutes --northeast and by north --main branch seventh limb east side --shoot from the left eye of the death's-head --a bee-line from the tree through the shot fifty feet out.’ ”

We estimated the entire contents of the chest, that night, at a million and a half of dollars; and, upon the subsequent disposal of the trinkets and jewels (a few being retained for our own use), it was found that we had greatly undervalued the treasure.



Jupiter's language

"Dey aint no tin in him, Massa Will, I keep a tellin on you," here interrupted Jupiter; "de bug is a goole bug, solid, ebery bit of him, inside and all, sep him wing --neber feel half so hebby a bug in my life."

Jupiter’s language

"Why, to speak de troof, massa, him not so berry well as mought be.“

Dar! dat's it! --him neber plain of notin --but him berry sick for all dat.“

"No, dat he ain't! --he ain't find nowhar --dat's just whar de shoe pinch --my mind is got to be berry hebby bout poor Massa Will."

Alignment

Massa	Massa-
Master	Master
Mass-a	Mas-sa
Master	Master

The End