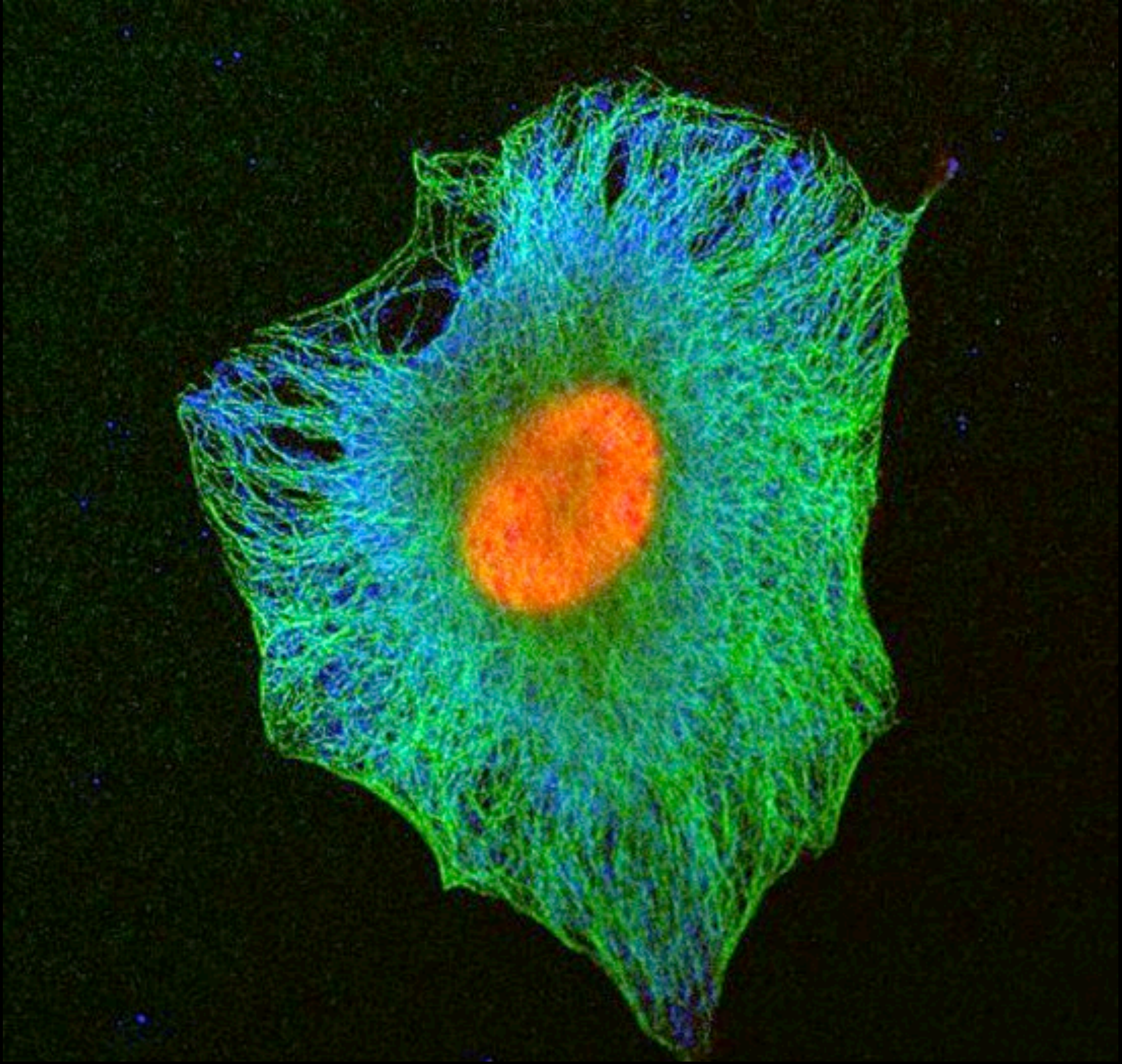
Introduction to Bioinformatics

Jérôme Waldispühl

School of Computer Science, McGill university

(includes slides from Mathieu Blanchette)

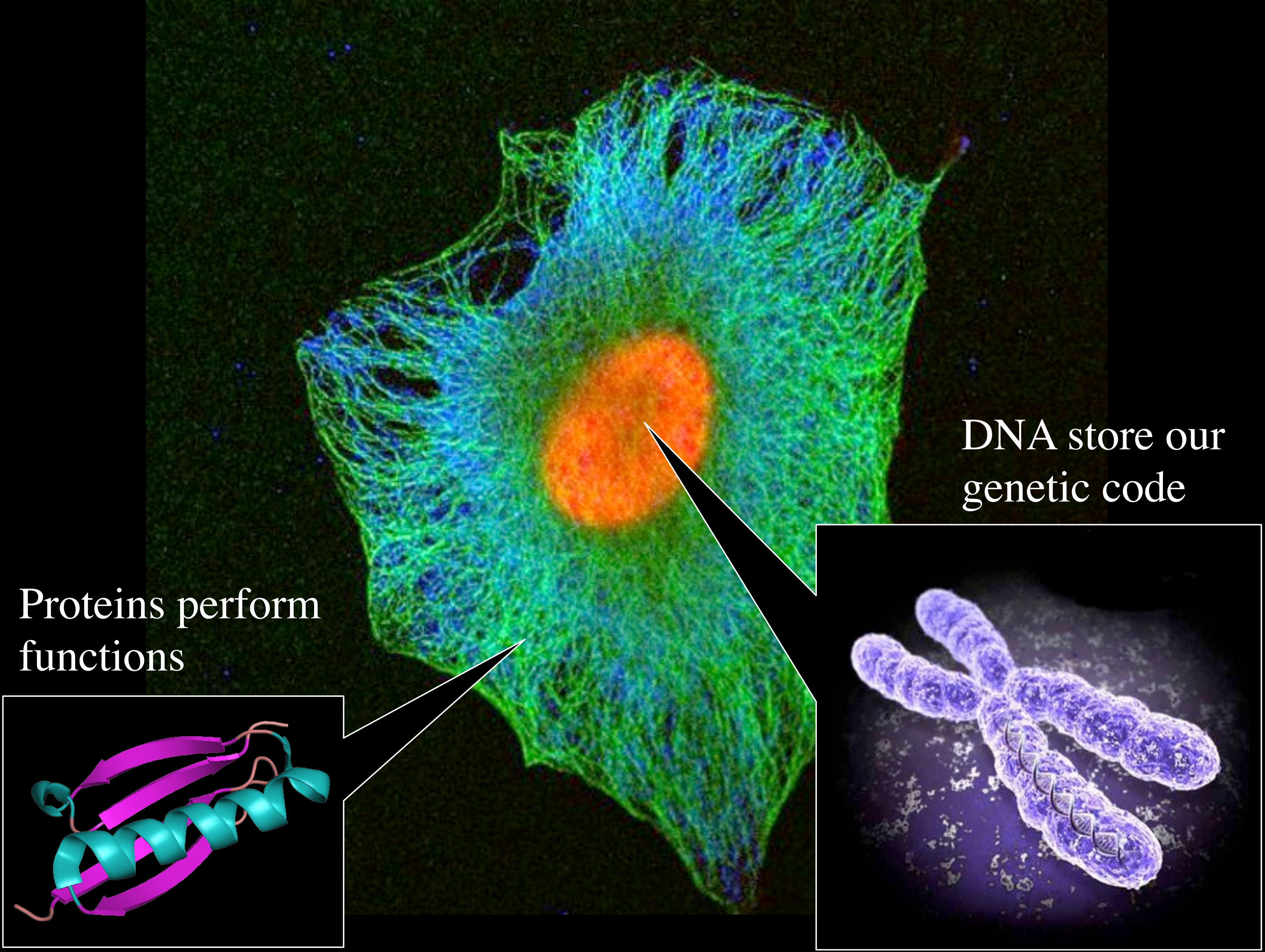


How do cells work?

Definition

Bioinformatics is an interdisciplinary field that develops and improves on methods for storing, retrieving, organizing and analyzing biological data. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge.

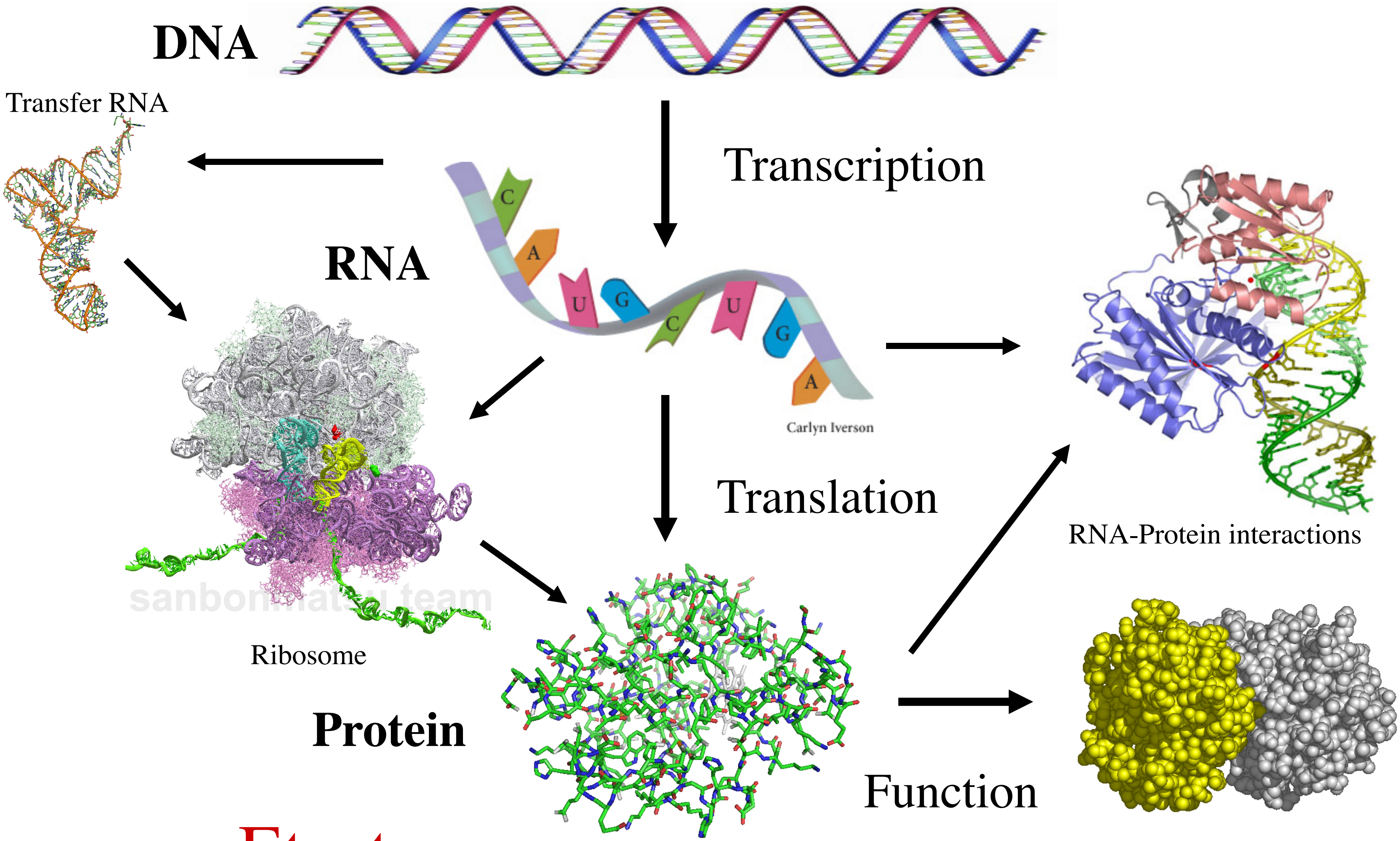
(from Wikipedia)



DNA store our genetic code

Proteins perform functions

Central dogma of biology



DNA

Transcription

Transfer RNA

RNA

Carlyn Iverson

RNA-Protein interactions

Translation

sanbonmatsu team

Ribosome

Protein

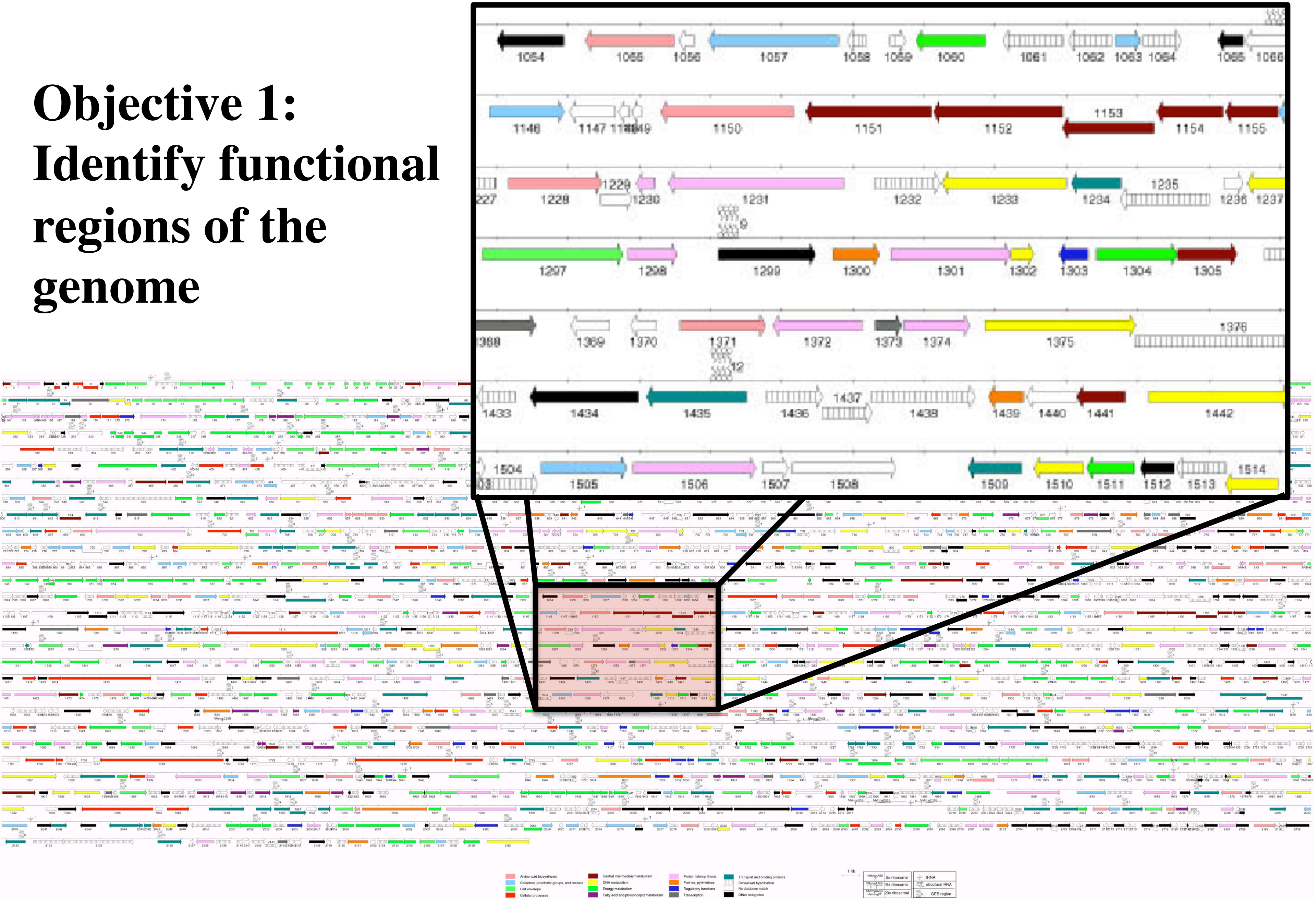
Function

Protein-protein interactions

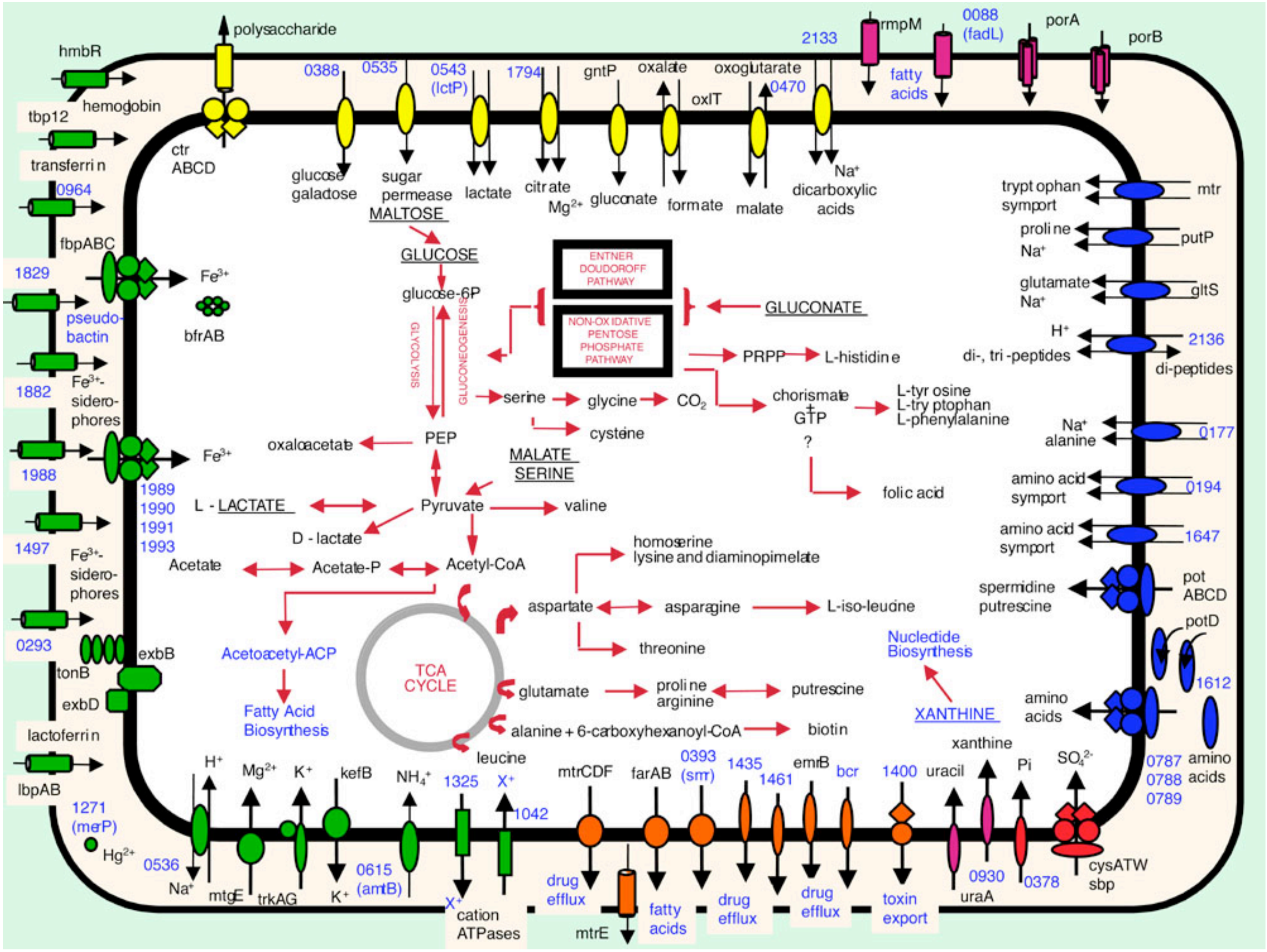
(Images not to scale)

Etcetera…

Objective 1: Identify functional regions of the genome

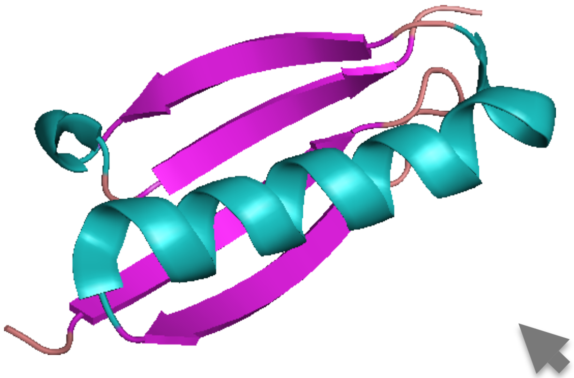


Objective 2: Cell Modeling

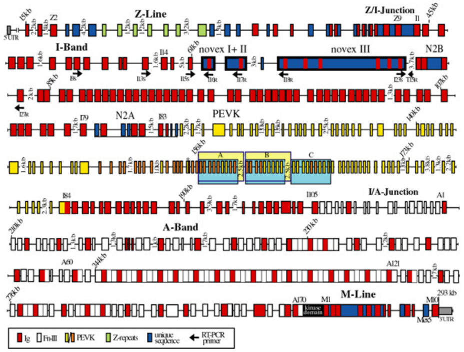


The 3 domains of bioinformatics

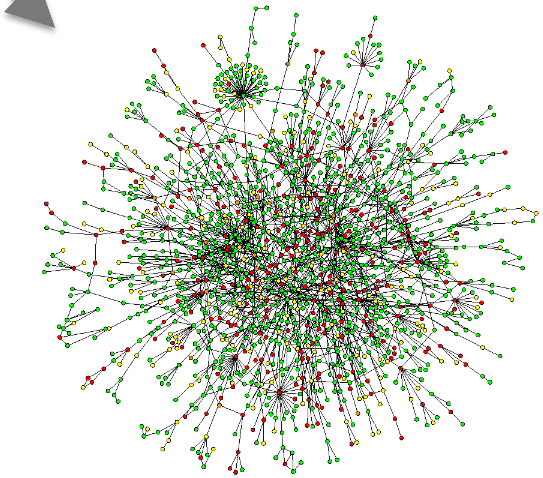
Structural Bioinformatics



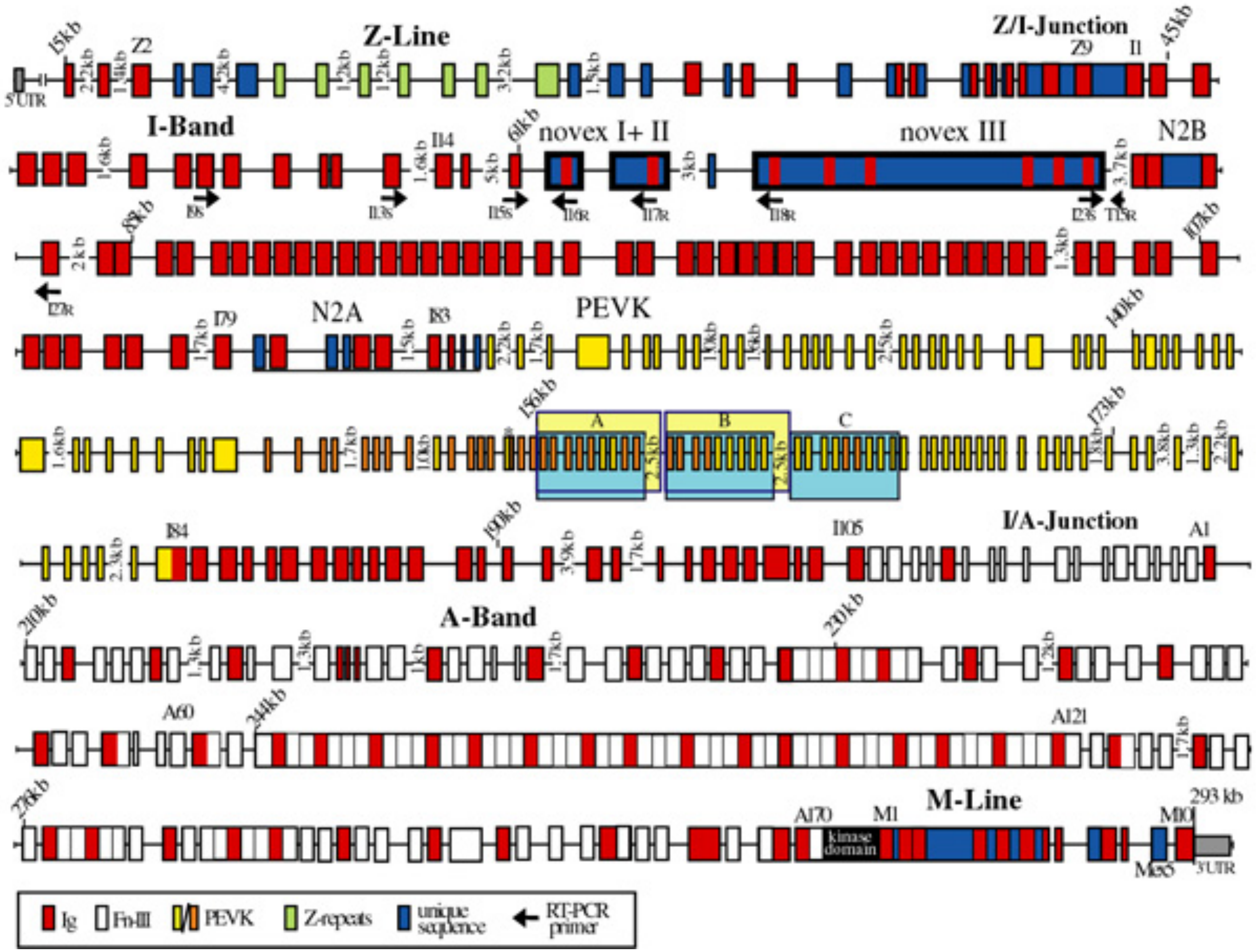
Genomics



System Biology

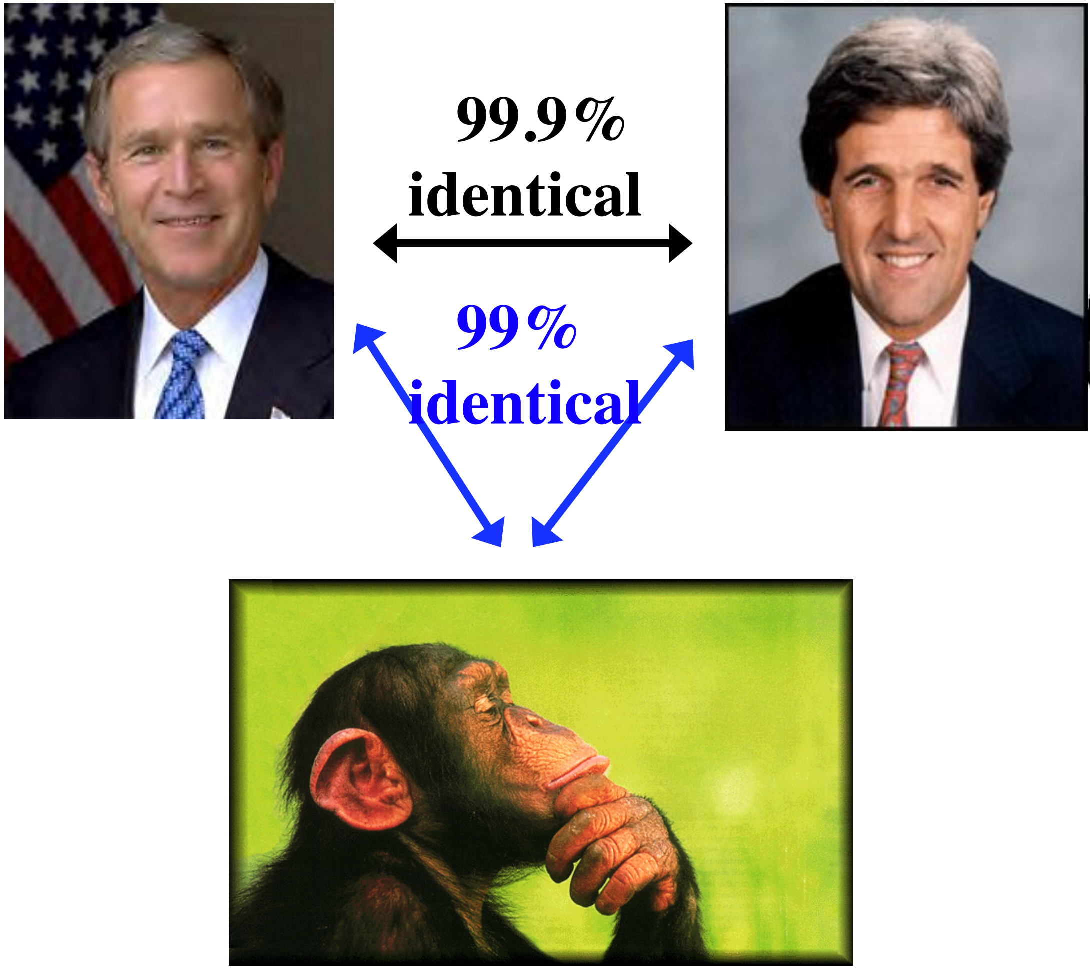


Genomic


Z-Line
Z/I-Junction
15kb
Z2
Z9
I1
45kb
5'UTR
I-Band
novex I+ II
novex III
N2B
I14
61kb
I8
I13
I15
I16R
I17R
I18R
I23
T15R
88kb
107kb
I27R
N2A
I79
I83
PEVK
140kb
156kb
A
B
C
173kb
I84
190kb
I105
I/A-Junction
A1
210kb
A-Band
230kb
A60
244kb
A121
276kb
A170
M1
M-Line
kinase domain
293 kb
M10
Mex5
3'UTR
Ig
Fn-III
PEVK
Z-repeats
unique sequence
RT-PCR primer

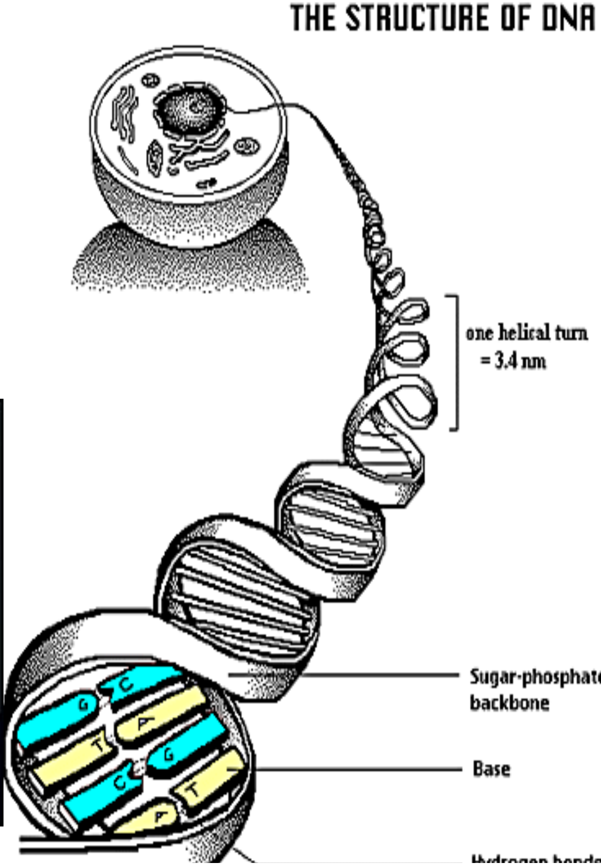
Genomes

- Human genome: $\{A,C,G,T\}^{3 \times 10^9}$
- Each of your 10^{14} cells has two copies

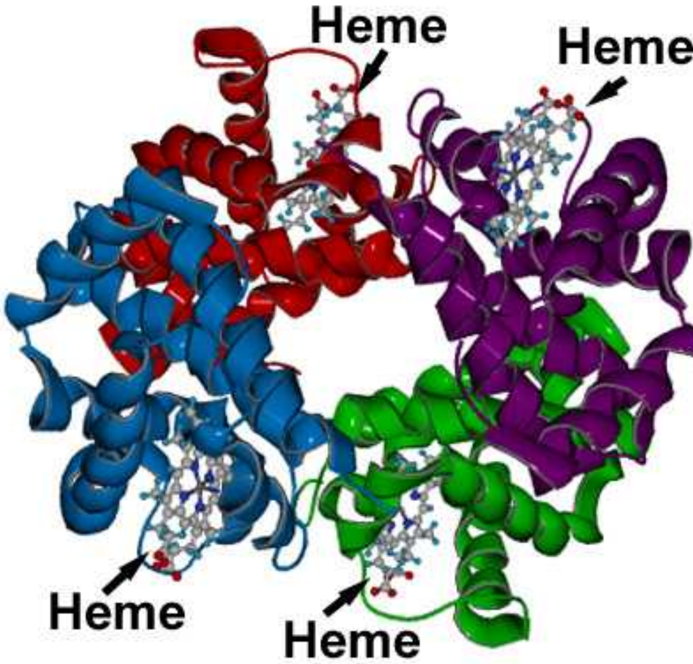


99.9% identical

99% identical

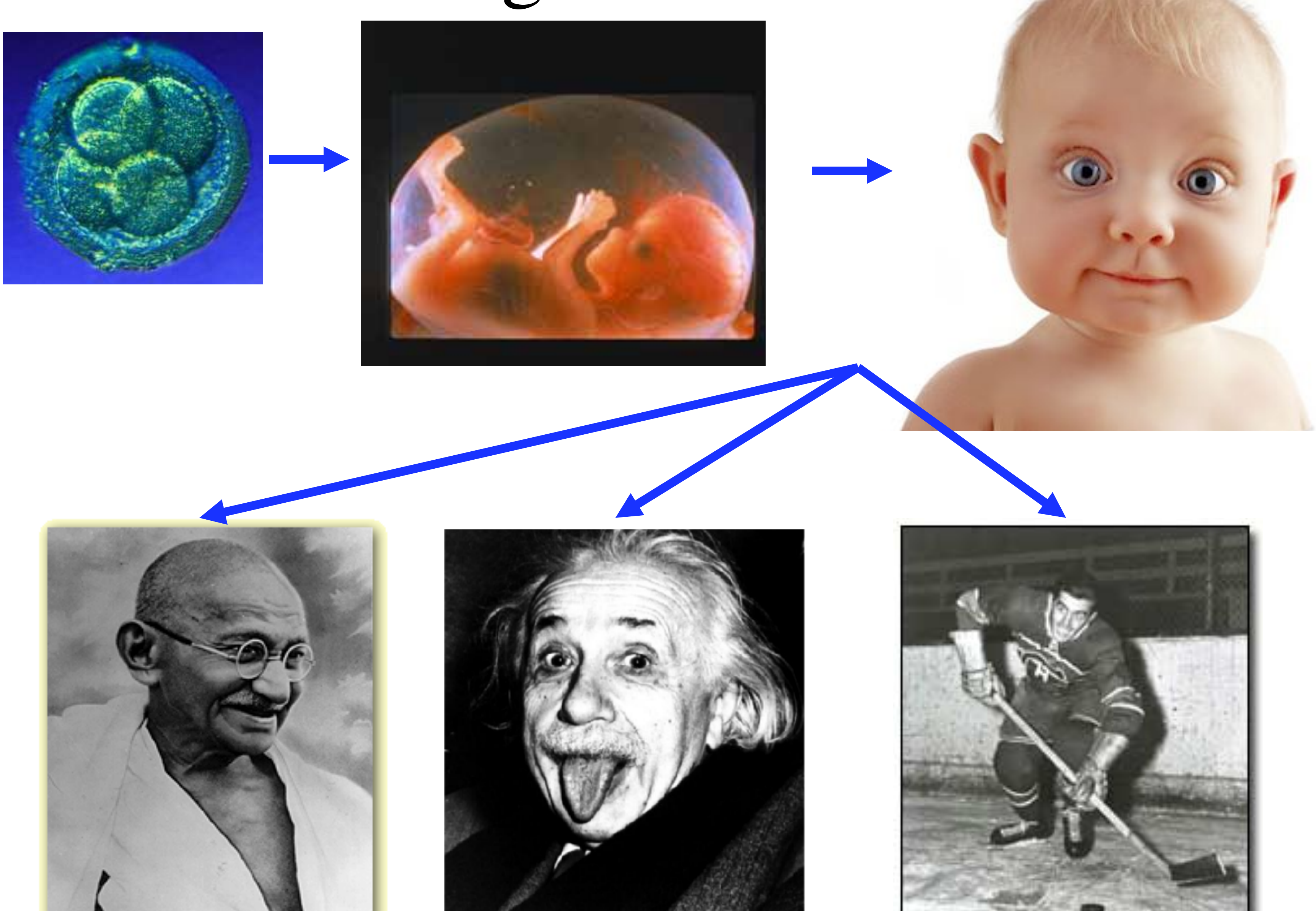


Roles of the genome



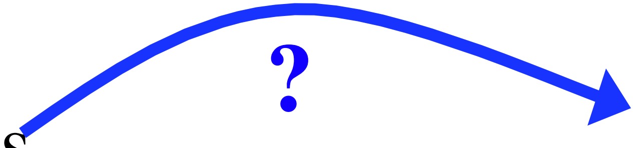
- Genome is a blue print for a cell
- Describes *how* to build proteins
  - 25,000 genes --> 25,000 proteins (+variations)
  - Each protein has its biochemical function
- Describes *when* to build each protein
  - Under which situations should a gene be expressed?
- Proteins allow:
  - Cell administration and maintenance
  - Reaction to stimuli
  - Protocols for communication between cells

Roles of the genome: development



Content of the genome - Genes

- Gene: region of DNA that encodes one protein

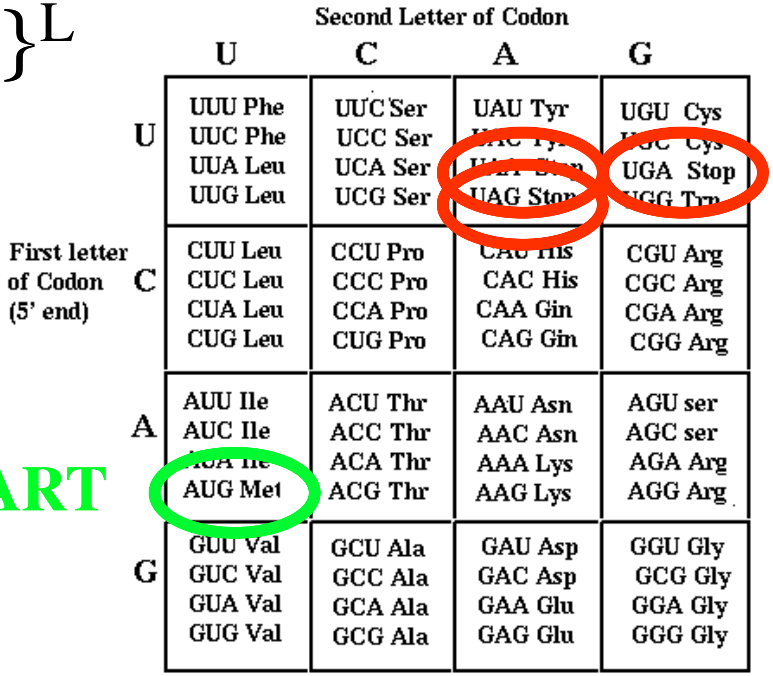


DNA sequences

$\{A,C,G,T\}^L$

Protein sequence

$\{A,C,D,E,F,G,H,I,K,L,$
$M,N,P,Q,R,S,T,V,W,Y\}^K$



Second Letter of Codon

First letter of Codon (5' end)

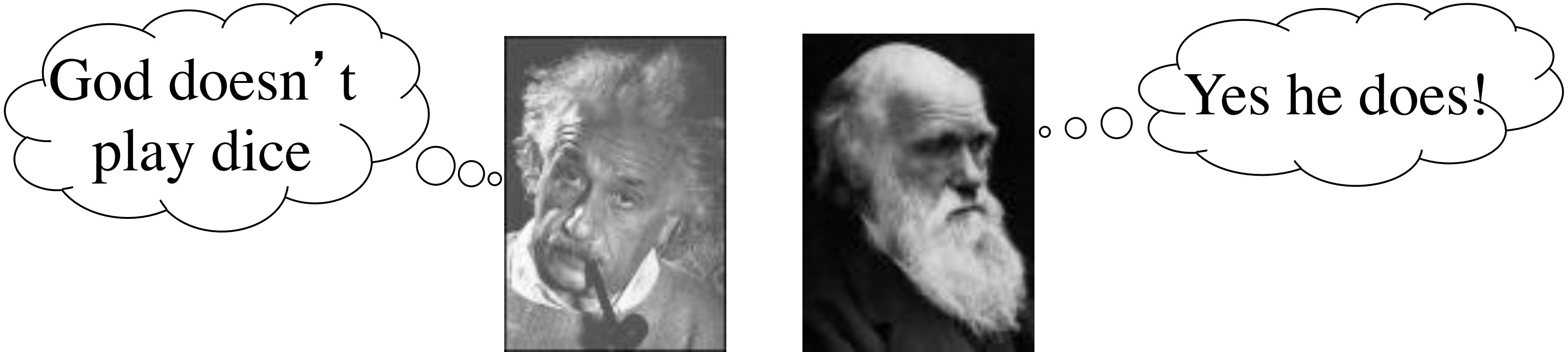
	U	C	A	G
U	UUU Phe UUC Phe UUA Leu UUG Leu	UUC Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CUG Pro	CAU His CAC His CAA Gin CAG Gin	CGU Arg CGC Arg CGA Arg CGG Arg
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU ser AGC ser AGA Arg AGG Arg
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GCG Gly GGA Gly GGG Gly

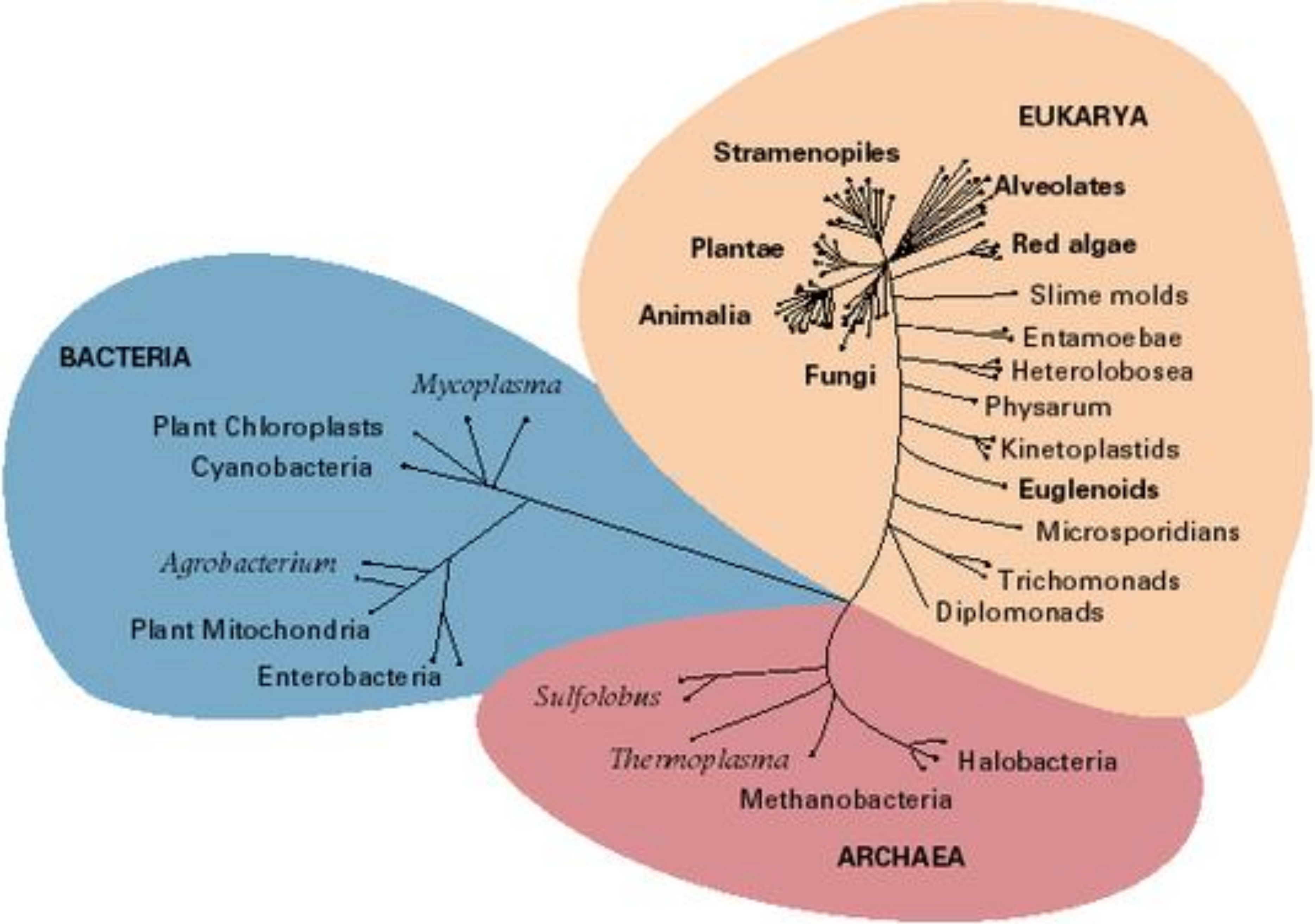
STOP

START

The Programmer - Evolution

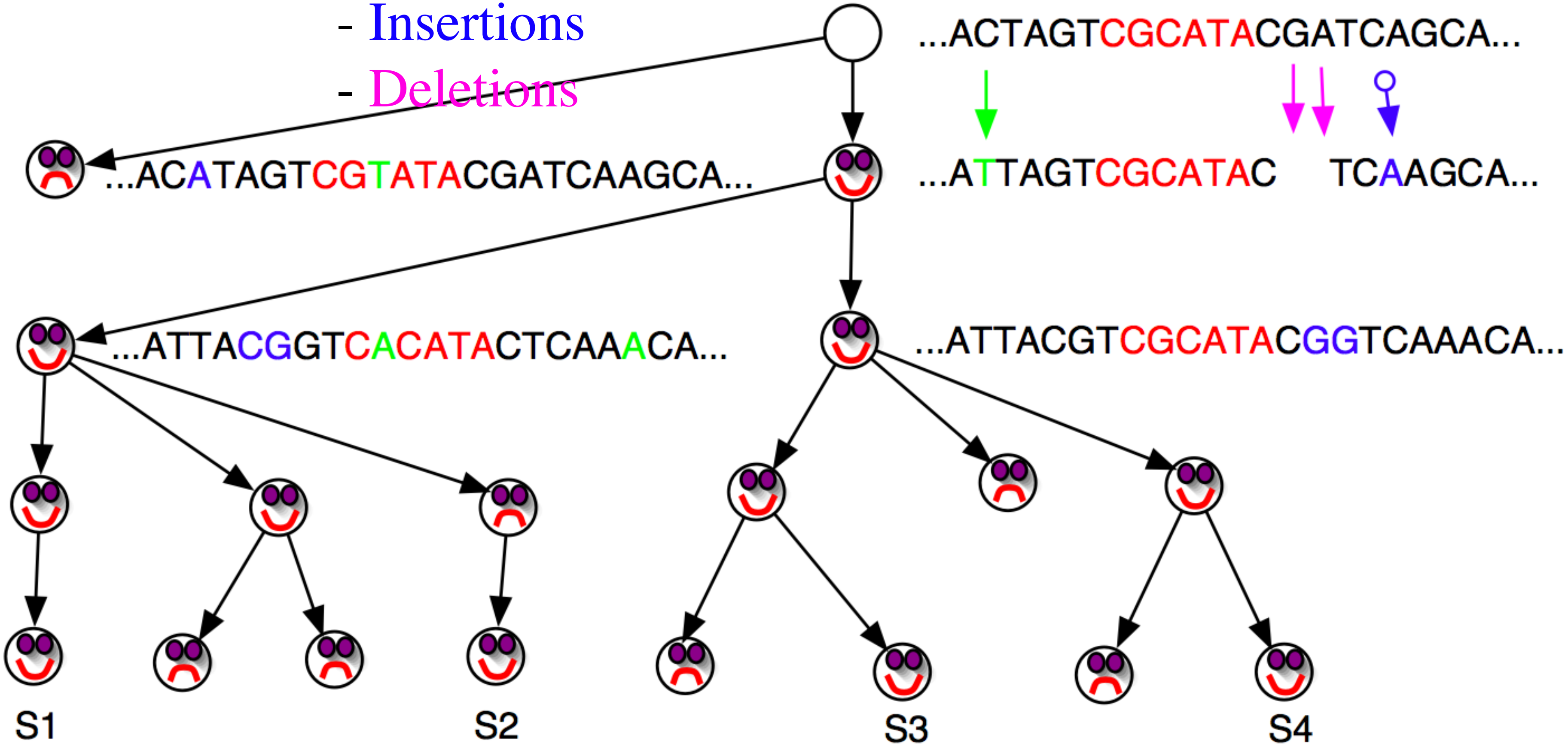
- Design principles:
  - Random modifications (variation)
  - Survival of the fittest (natural selection)
- 3 Billion years of evolution
- Today’s species are the current solution of the fitness optimization problem




EUKARYA
Stramenopiles
Alveolates
Plantae
Red algae
Slime molds
Animalia
Entamoebae
Fungi
Heterolobosea
Physarum
Kinetoplastids
Euglenoids
Microsporidians
Trichomonads
Diplomonads
BACTERIA
Mycoplasma
Plant Chloroplasts
Cyanobacteria
Agrobacterium
Plant Mitochondria
Enterobacteria
Sulfolobus
Thermoplasma
Halobacteria
Methanobacteria
ARCHAEA

Central dogma of comparative genomics

Mutations:
- Substitutions
- Insertions
- Deletions



...ACTAGTCGCATACGATCAGCA...

...ATTAGTCGCATAC TCAAGCA...

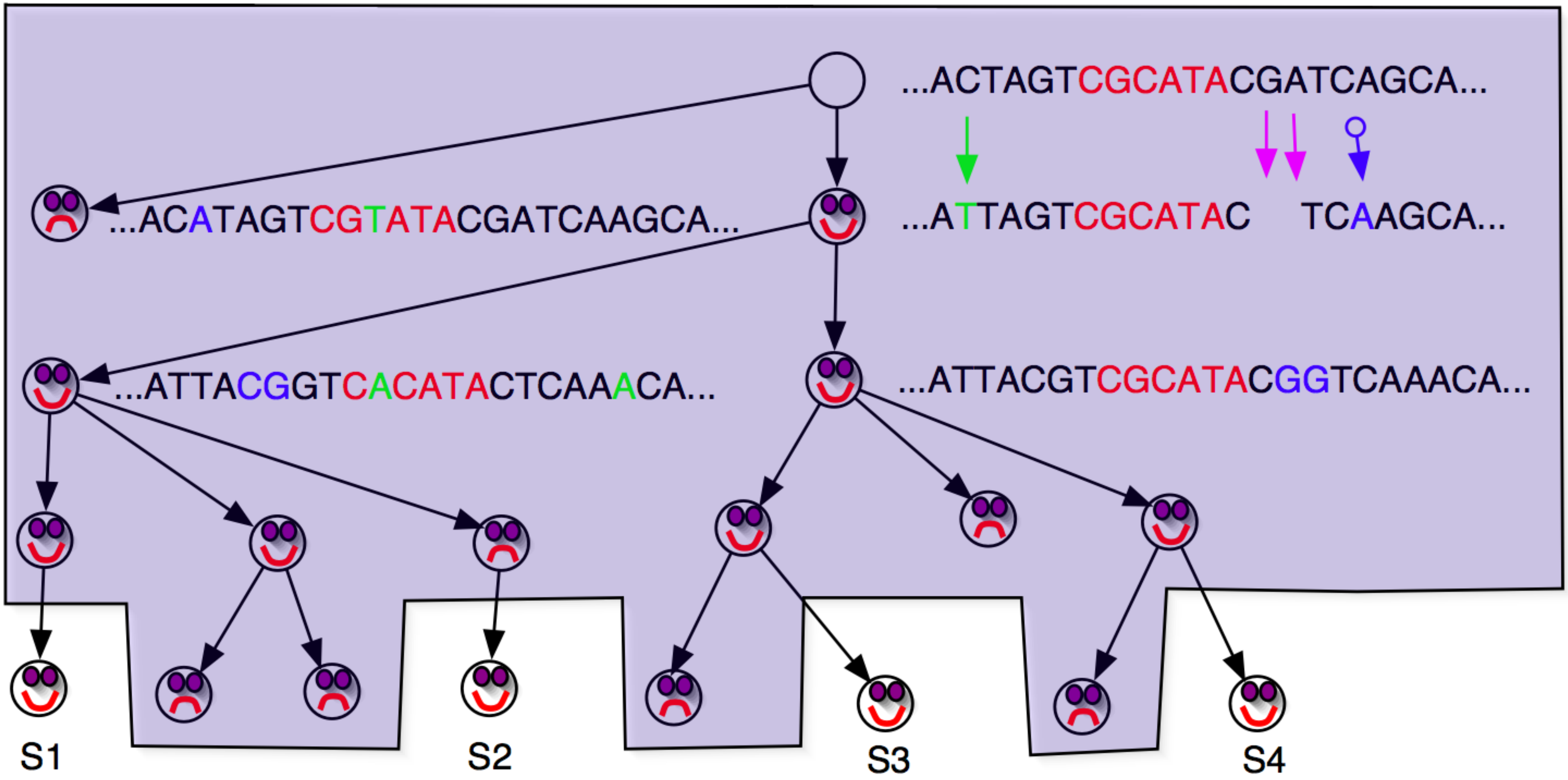
...ACATAGTCGTATACGATCAAGCA...

...ATTACGGTCACATACTCAAACA...

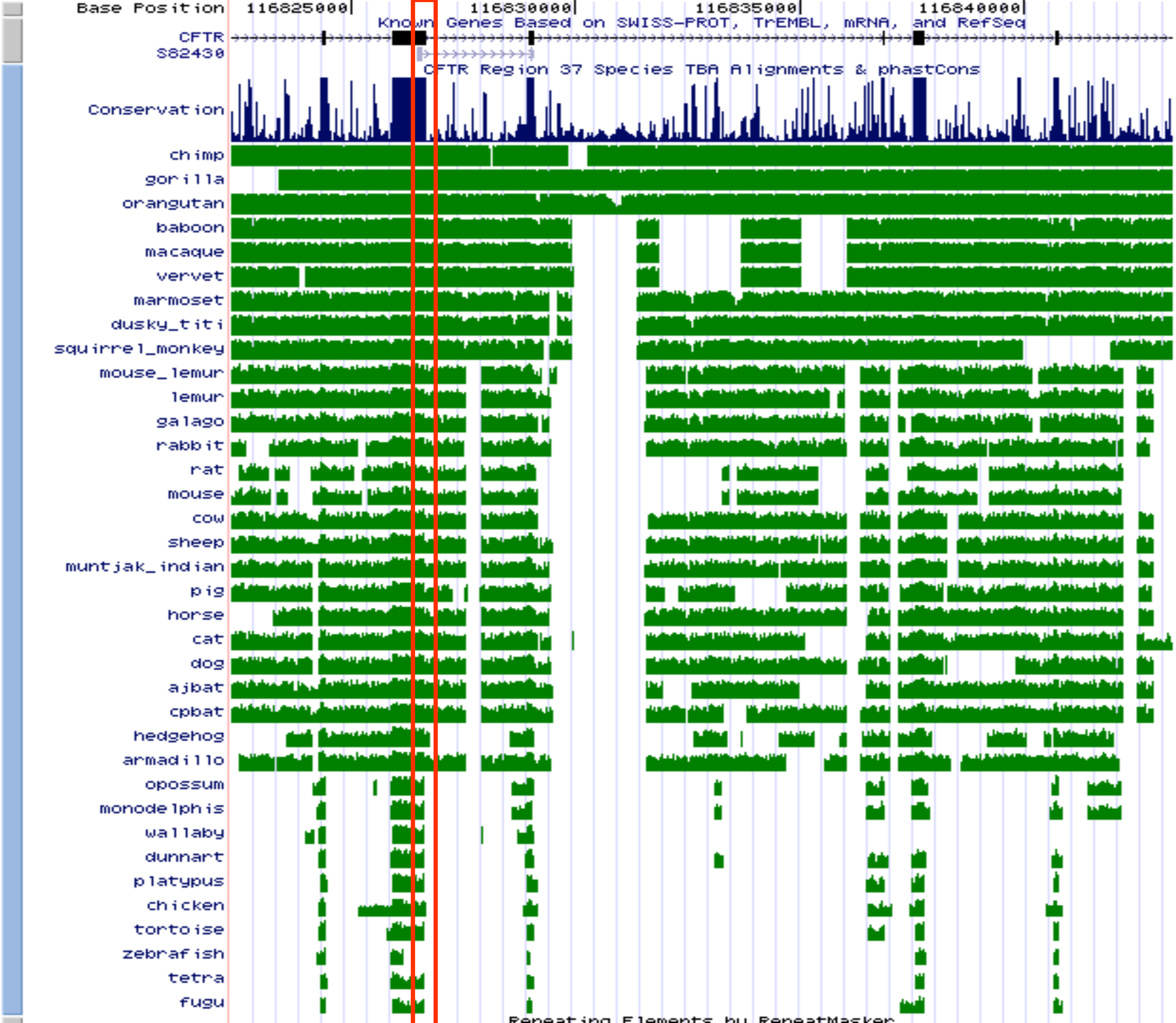
...ATTACGTCGCATACGGTCAAACA...

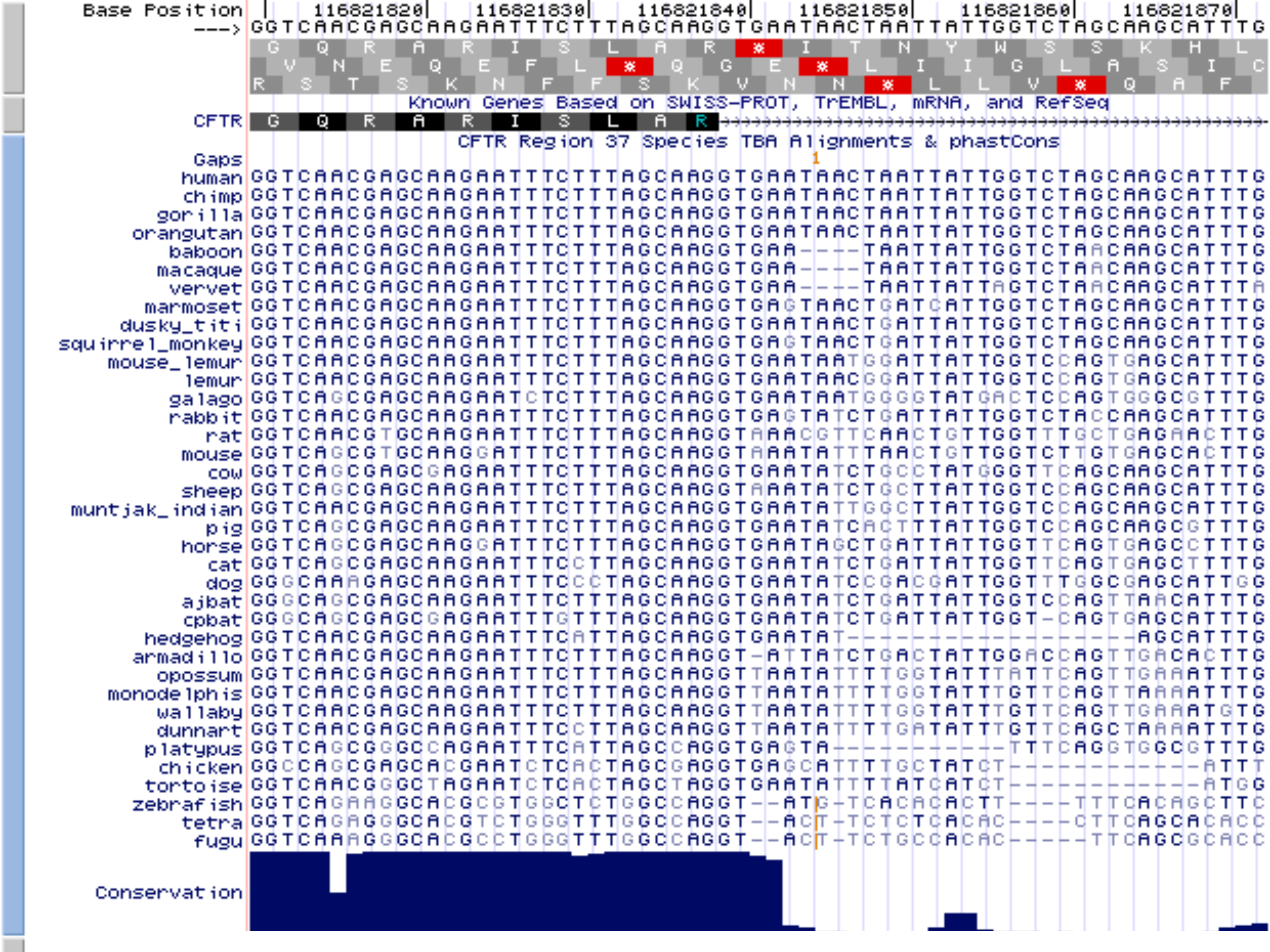
S1	...GTTACGGTCACATACTGAAACA...
S2	...GTTATGGTCACATACTGAAACTGA...
S3	...ATTACTCGCATACGGTCTAACA
S4	...ATTTACTCGCATACGGTCTAGCACT

Unavailable!


...ACTAGTCGCATACGATCAGCA...
...ACATAGTCGTATACGATCAAGCA...
...ATTAGTCGCATAC TCAAGCA...
...ATTACGGTCACATACTCAAACA...
...ATTACGTCGCATACGGTCAAACA...
S1
S2
S3
S4

```
S1:   GTTACGGTCACATACTGAAACA
S2:   GTTATGGTCACATACTGAAACTGA
S3:   ATTACTCGCATACGGTCTAACA
S4:   ATTTACTCGCATACGGTCTAGCAC
```


Base Position
116825000
116830000
116835000
116840000
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
CFTR
S82430
CFTR Region 37 Species TBA Alignments & phastCons
Conservation
chimp
gorilla
orangutan
baboon
macaque
vervet
marmoset
dusky_titi
squirrel_monkey
mouse_lemur
lemur
galago
rabbit
rat
mouse
cow
sheep
muntjak_indian
pig
horse
cat
dog
ajbat
cpbat
hedgehog
armadillo
opossum
monodelphis
wallaby
dunnart
platypus
chicken
tortoise
zebrafish
tetra
fugu
Repeating Elements by RepeatMasker



```
Base Position  |    116821820|    116821830|    116821840|    116821850|    116821860|    116821870|
        --->   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
                G  Q  R  A  R  I  S  L  A  R  *  I  T  N  Y  W  S  S  K  H  L
                 V  N  E  Q  E  F  L  *  Q  G  E  *  L  I  I  G  L  A  S  I  C
               R  S  T  S  K  N  F  F  S  K  V  N  N  *  L  L  V  *  Q  A  F
                     Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
        CFTR    G  Q  R  A  R  I  S  L  A  R
                     CFTR Region 37 Species TBA Alignments & phastCons
        Gaps                                        1
       human   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
       chimp   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
     gorilla   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
   orangutan   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTAATTATTGGTCTAGCAAGCATTTG
      baboon   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAA----TAATTATTGGTCTAACAAGCATTTG
     macaque   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAA----TAATTATTGGTCTAACAAGCATTTG
      vervet   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAA----TAATTATTAGTCTAACAAGCATTTA
    marmoset   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAGTAACTGATCATTGGTCTAGCAAGCATTTG
  dusky_titi   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACTGATTATTGGTCTAGCAAGCATTTG
squirrel_monkey GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAGTAACTGATTATTGGTCTAGCAAGCATTTG
 mouse_lemur   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAATGGATTATTGGTCCAGTGAGCATTTG
       lemur   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATAACGGATTATTGGTCCAGTGAGCATTTG
      galago   GGTCAGCGAGCAAGAATCTCTTTAGCAAGGTGAATAATGGGGTATGACTCCAGTGGGCGTTTG
      rabbit   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAGTATCTGATTATTGGTCTACCAAGCATTTG
         rat   GGTCAACGTGCAAGAATTTCTTTAGCAAGGTAAACGTTCAACTGTTGGTTTGCTGAGAACTTG
       mouse   GGTCAGCGTGCAAGGATTTCTTTAGCAAGGTAAATATTTAACTGTTGGTCTTGTGAGCACTTG
         cow   GGTCAGCGAGCGAGAATTTCTTTAGCAAGGTGAATATCTGCCTATGGGTTCAGCAAGCATTTG
       sheep   GGTCAGCGAGCAAGAATTTCTTTAGCAAGGTAAATATCTGCTTATTGGTCCAGCAAGCATTTG
muntjak_indian GGTCAACGAGCAAGAATTTCTTTAGCAAGGTGAATATTGGCTTATTGGTCCAGCAAGCATTTG
         pig   GGTCAGCGAGCAAGAATTTCTTTAGCAAGGTGAATATCACTTTATTGGTCCAGCAAGCGTTTG
       horse   GGTCAGCGAGCAAGGATTTCTTTAGCAAGGTGAATAGCTGATTATTGGTTCAGTGAGCCTTTG
         cat   GGTCAGCGAGCAAGAATTTCCTTAGCAAGGTGAATATCTGATTATTGGTTCAGTGAGCTTTTG
         dog   GGGCAAAGAGCAAGAATTTCCCTAGCAAGGTGAATATCCGACGATTGGTTTGGCGAGCATTGG
       ajbat   GGGCAGCGAGCAAGAATTTCTTTAGCAAGGTGAATATCTGATTATTGGTCCAGTTAACATTTG
       cpbat   GGGCAGCGAGCGAGAATTTGTTTAGCAAGGTGAATATCTGATTATTGGT-CAGTGAGCATTTG
    hedgehog   GGTCAACGAGCAAGAATTTCATTAGCAAGGTGAATAT------------------AGCATTTG
   armadillo   GGTCAACGAGCAAGAATTTCTTTAGCAAGGT-ATTATCTGACTATTGGACCAGTTGACACTTG
     opossum   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTTAATATTTTGGTATTTATTCAGTTGAAATTTG
 monodelphis   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTTAATATTTTGGTATTTGTTCAGTTAAAATTTG
     wallaby   GGTCAACGAGCAAGAATTTCTTTAGCAAGGTTAATATTTTGGTATTTGTTCAGTTGAAATGTG
     dunnart   GGTCAACGAGCAAGAATTTCCTTAGCAAGGTTAATATTTTGATATTTGTTCAGCTAAAATTTG
    platypus   GGTCAGCGGGCCAGAATTTCATTAGCCAGGTGAGTA-----------TTTCAGGTGGCGTTTG
     chicken   GGCCAGCGAGCACGAATCTCACTAGCGAGGTGAGCATTTTGCTATCT------------ATTT
    tortoise   GGTCAACGGGCTAGAATCTCACTAGCTAGGTGAATATTTTATCATCT------------ATGG
   zebrafish   GGTCAGAAGGCACGCGTGGCTCTGGCCAGGT--ATG-TCACACACTT----TTTCACAGCTTC
       tetra   GGTCAGAGGGCACGTCTGGGTTTGGCCAGGT--ACT-TCTCTCACAC----CTTCAGCACACC
        fugu   GGTCAAAGGGCACGCCTGGGTTTGGCCAGGT--ACT-TCTGCCACAC-----TTCAGCGCACC

Conservation
```

Algorithms to compare and align genomes

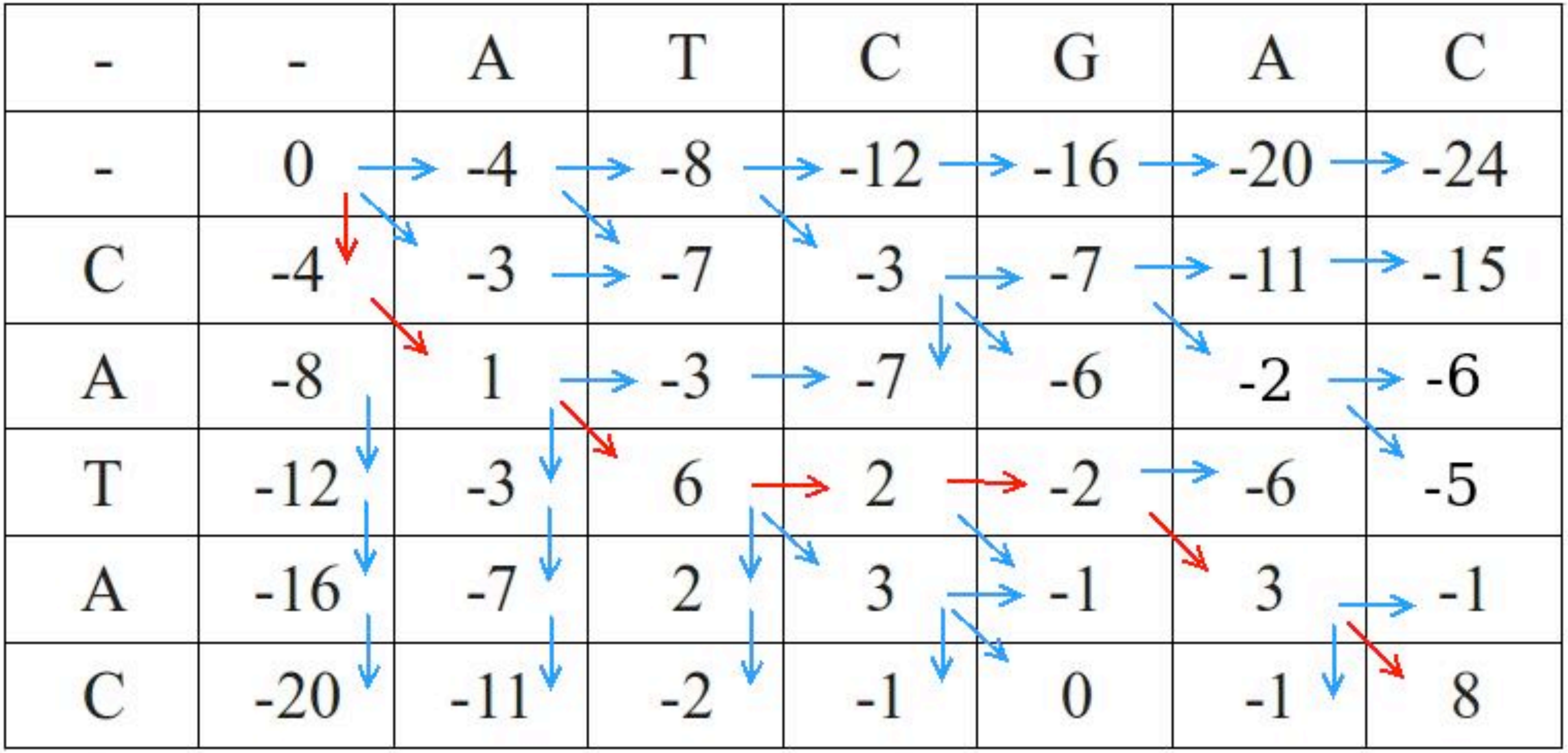
Input

```
ATCGAC
CATAC
```

→

Output

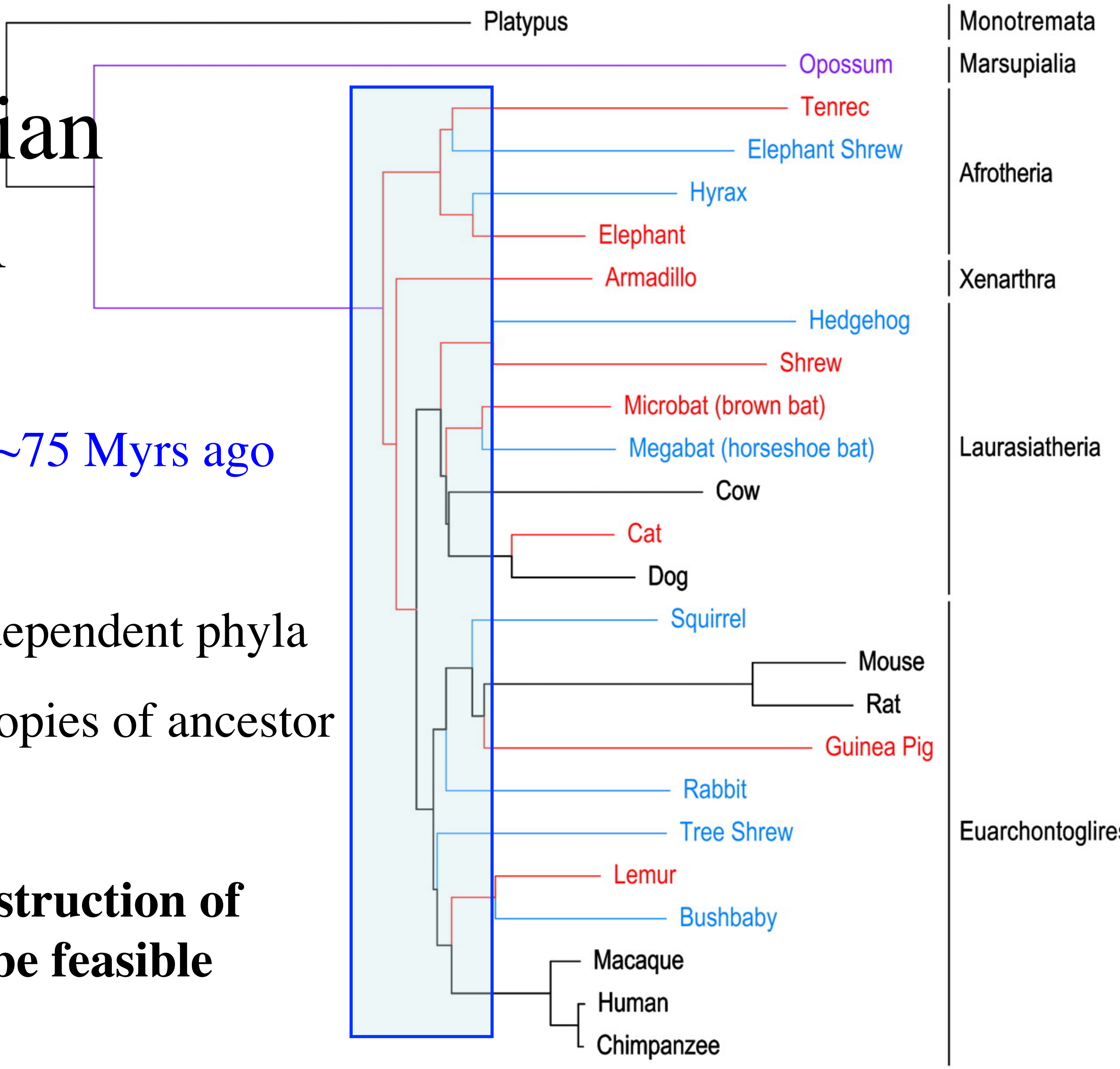
```
-ATCGAC
CAT--AC
```



-	-	A	T	C	G	A	C
-	0	-4	-8	-12	-16	-20	-24
C	-4	-3	-7	-3	-7	-11	-15
A	-8	1	-3	-7	-6	-2	-6
T	-12	-3	6	2	-2	-6	-5
A	-16	-7	2	3	-1	3	-1
C	-20	-11	-2	-1	0	-1	8

(Needleman-Wunsch, 1970)

Mammalian evolution



- Rapid radiation ~75 Myrs ago

↓

- Many nearly independent phyla
- Many "noisy" copies of ancestor

↓

- **Accurate reconstruction of ancestors may be feasible**

Margulies et al., PNAS 2005

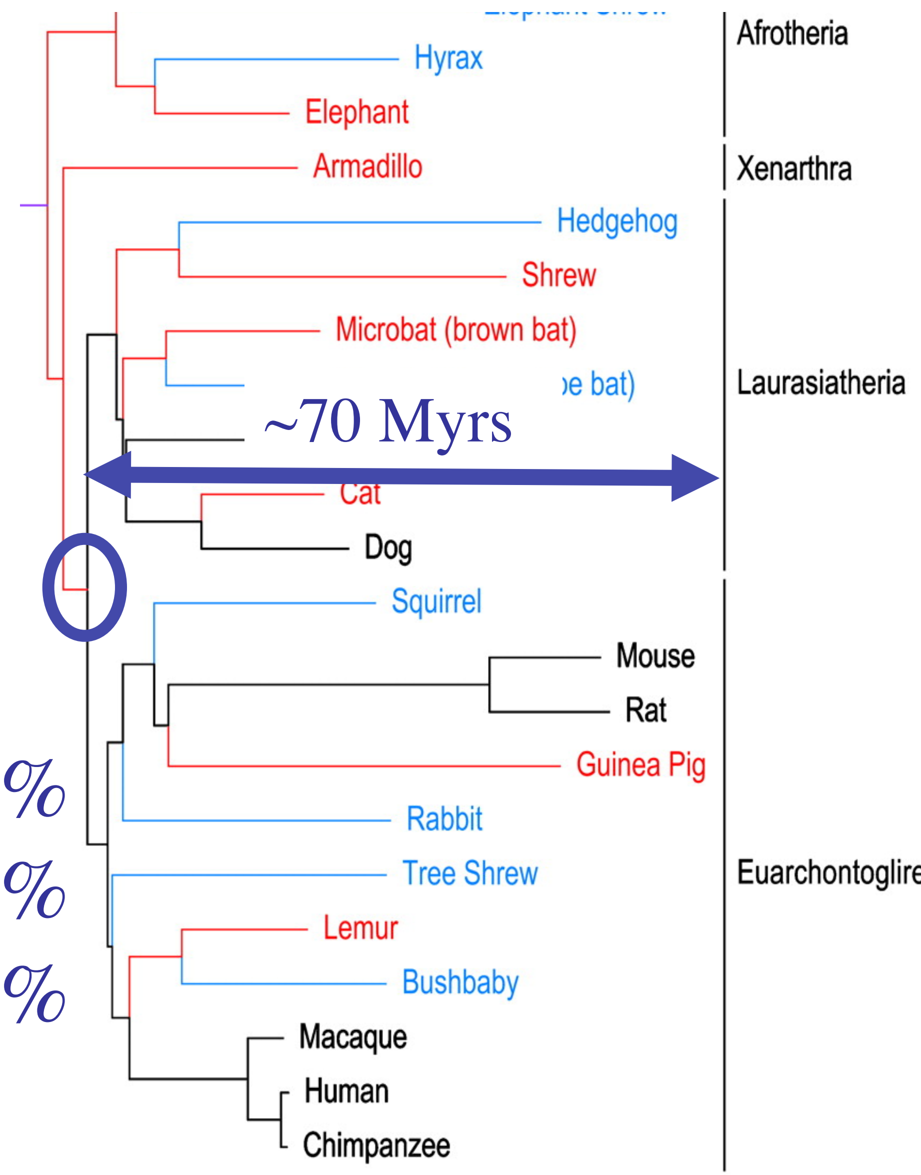
Ancestral mammalian genome reconstruction

[Miller, Haussler, Blanchette]

Base-by-base reconstruction of complete ancestral genomes

- Including coding, non-coding, repetitive regions

Boreoeutherian ancestor



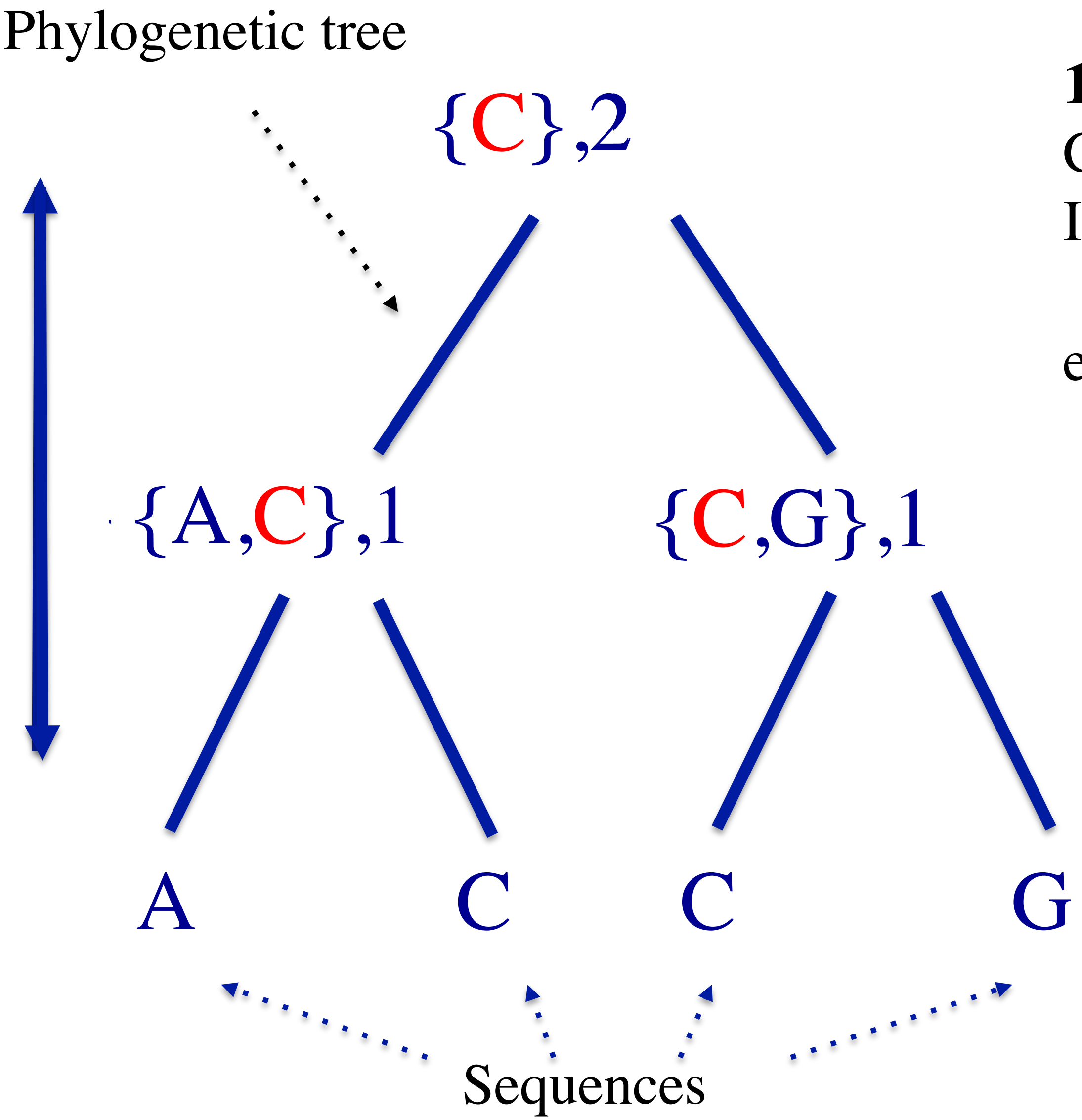
Expected reconstruction accuracy(*):

- From ideal choice of extant mammals 99%
- From soon-to-be available genomes: 96%
- With currently available sequences: (full or 2X coverage) 90%

(*) For >90% of euchromatic genome

Tree from Margulies et al., PNAS 200

The Parsimony Score and Fitch's Algorithm



1. **Bottom-up phase**

```
Compare leaves
If intersection==Ø :
    assign union & score++
else :
    assign intersection
```

2. **Top-down phase**

Compute ancestors

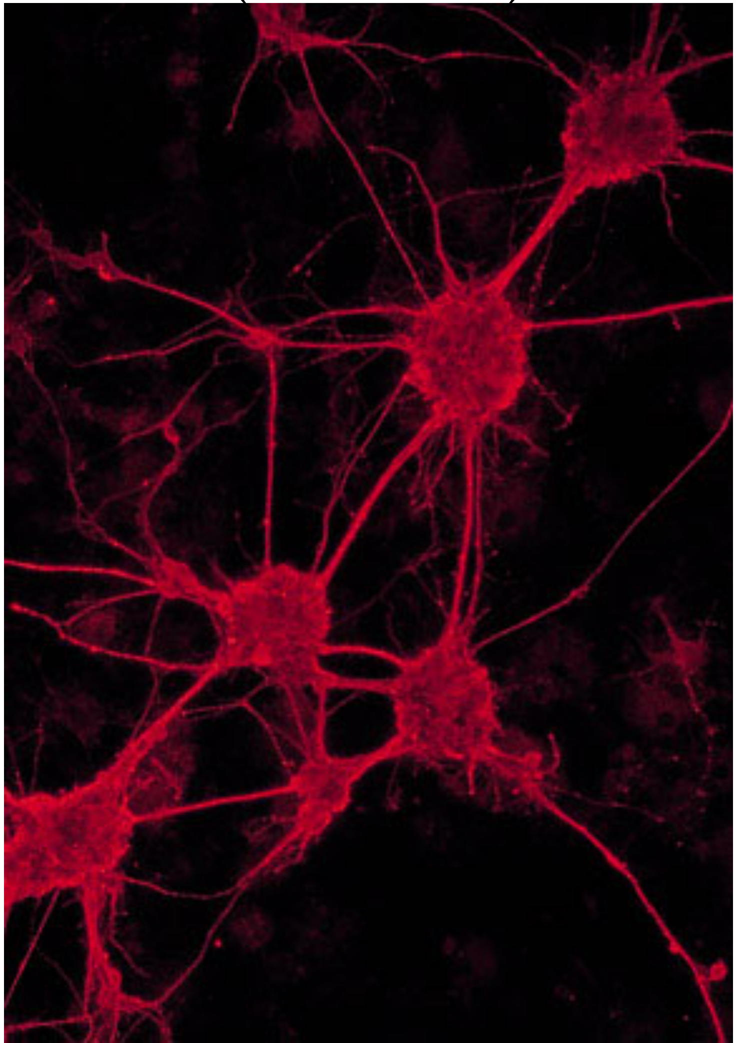
Why should we care?

- See genome evolution *happening*, rather than just see its outcome
  - Assign directionality to events (indels, substitutions)
  - Reconstructing the timing of events

- Boreoeutherian ancestor = Ye good olde mammalian stuff
  - BorEut is an archetypical mammalian genome
  - None of the species-specific quirks
  - Human is ~4 times close to BorEut than to mouse

One program, many functions

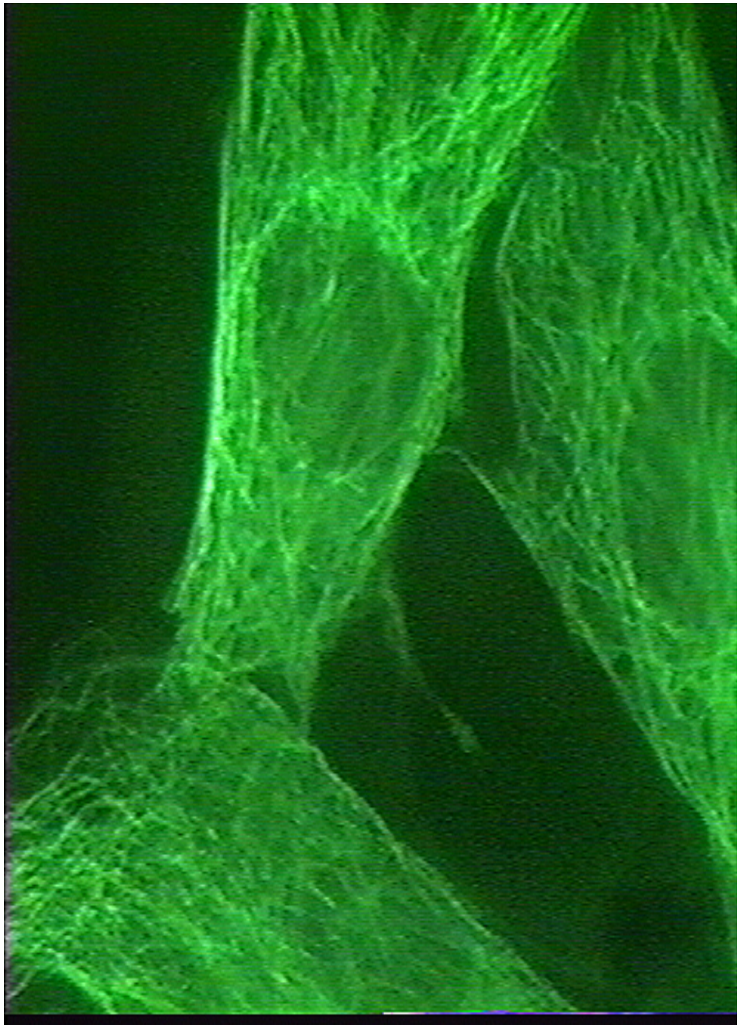
Neuron (nerve cell)



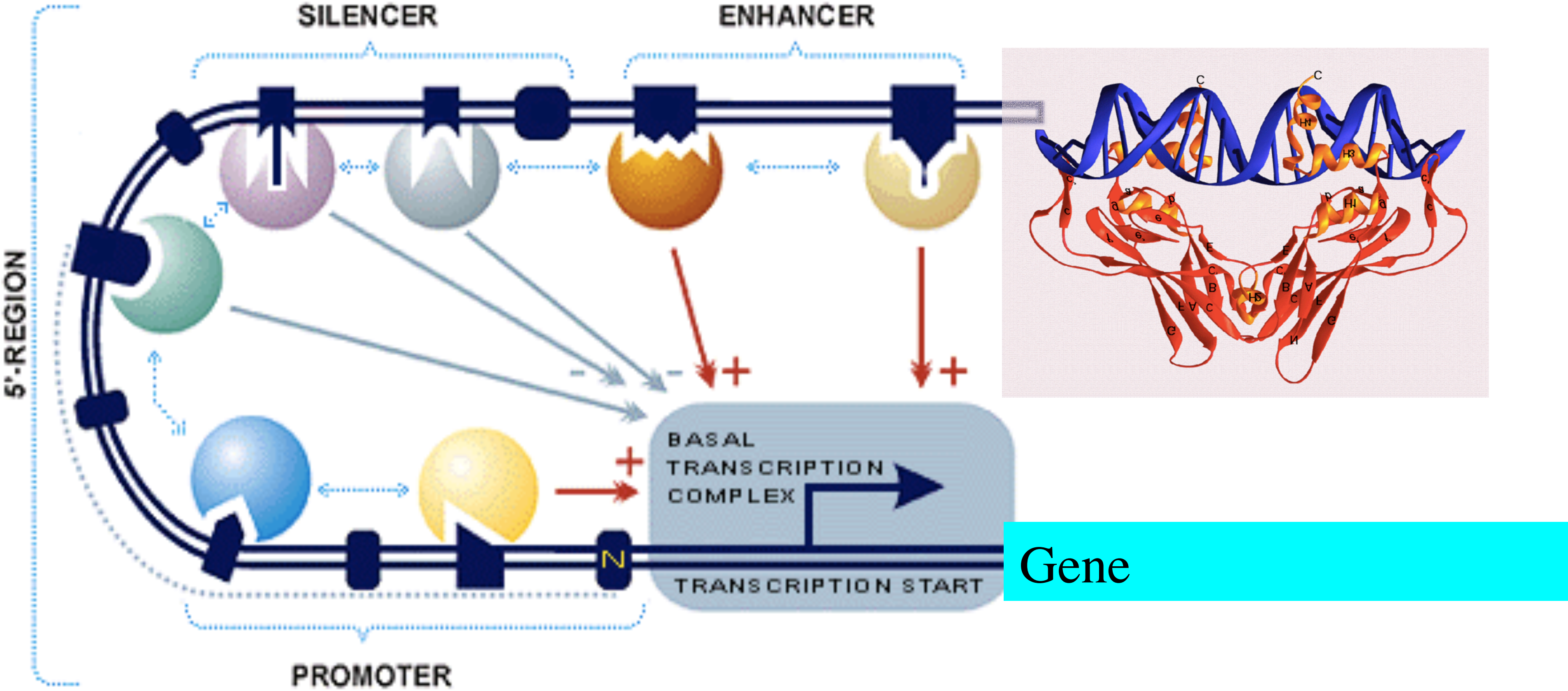
Same genome!

↔

Muscle cell



Regulation of gene expression



Transcription factor binding sites:

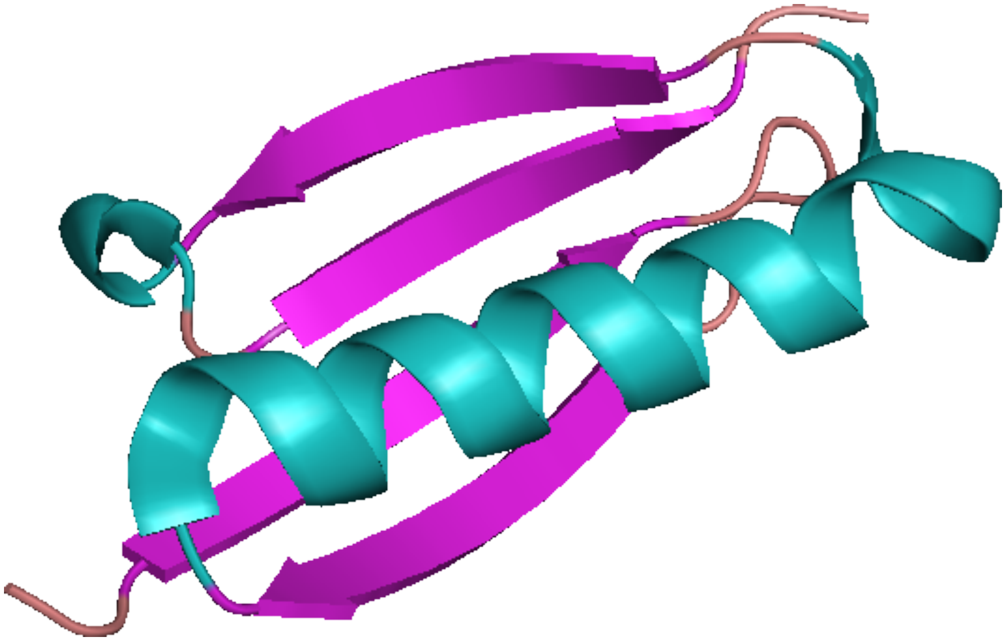
☹ Short: 6 to 20 nucleotides

☹ No specific signature; each TF has different binding site

☹ Can be up to 1 million nucleotides upstream of gene regulated

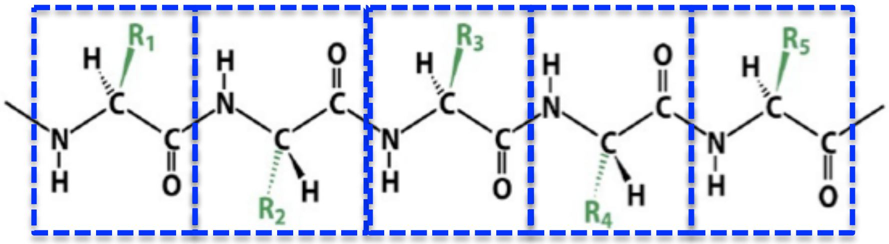
☺ Often clustered with other binding sites, forming modules

Structural Bioinformatics



Protein Structure

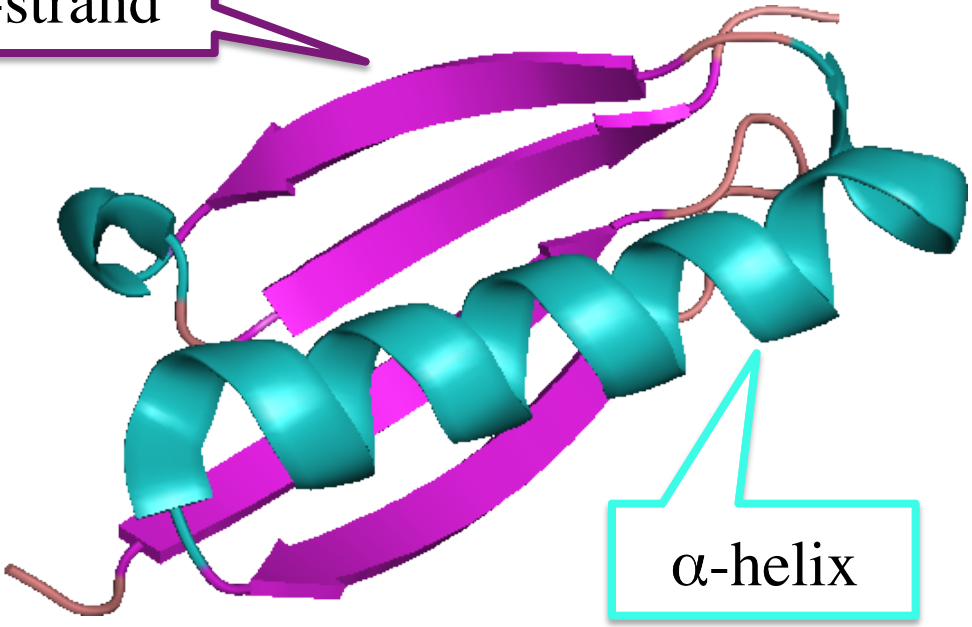
Sequence



A_1 A_2 A_3 A_4 A_5

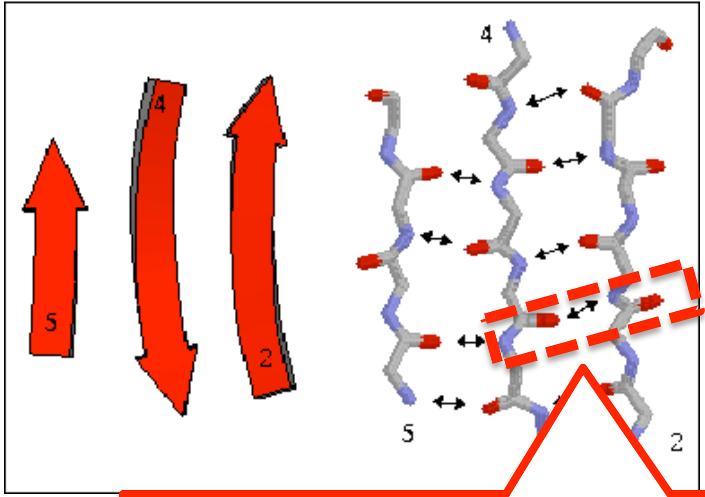
Structure

β-strand



α-helix

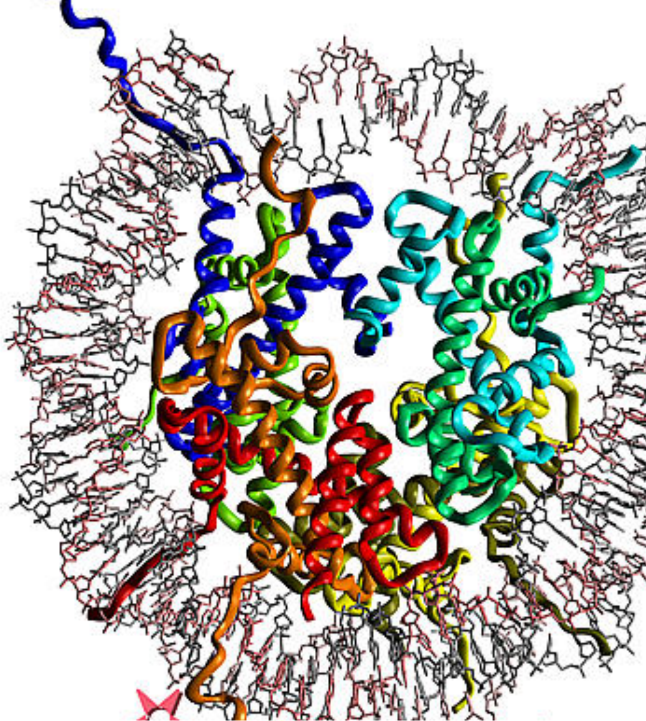
β-sheet



Long-range interactions stabilize β–sheet.

The hardware - Proteins

- Molecules only obey laws of physics and chemistry.
- Cell organization only relies on interactions between molecules
- Stochastic, dynamic, “chaotic” system
- High error rate in interactions
- Replicating, self-assembling, self-repairing system



Maybe software engineers have something to learn here...

Ab-initio vs knowledge-based methods

- Ab inito folding (simulation-based method)

  1998 Duan and Kollman

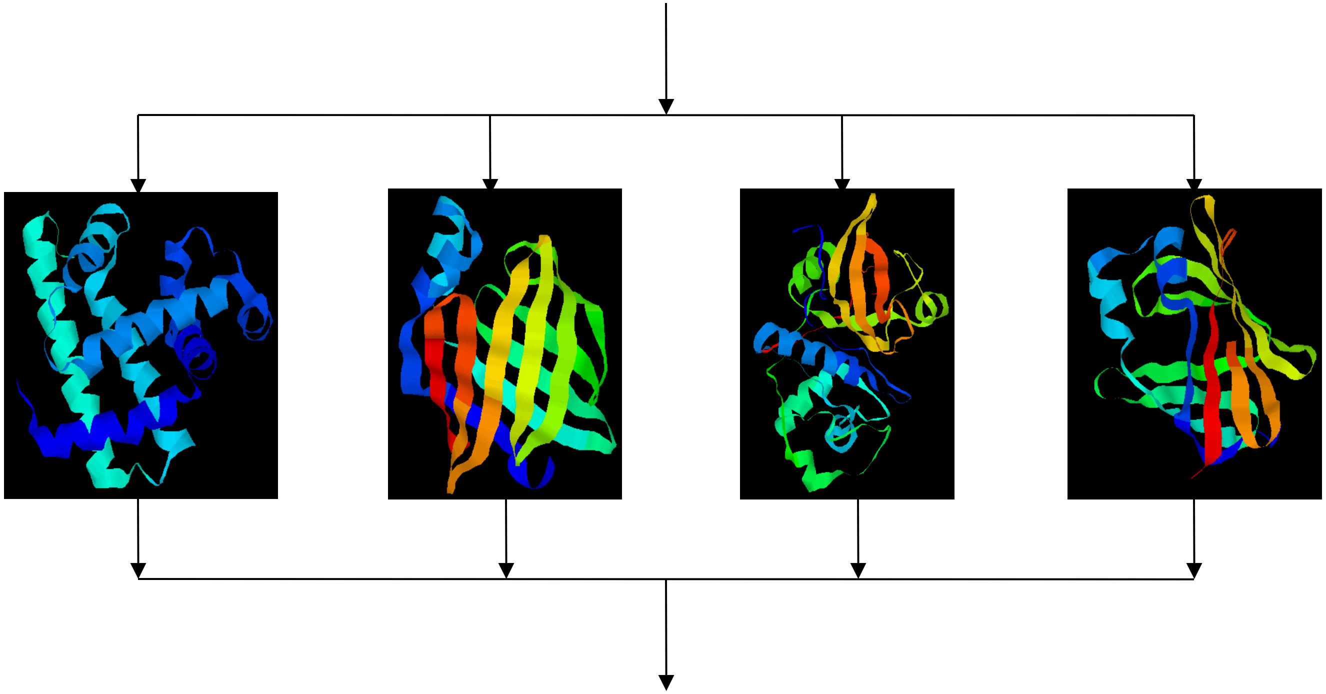
  36 residues, 1000 ns, 256 processors, 2 months

  Do not find native structure

- Template-based (or knowledge-based) methods
  - Homology modeling: sequence-sequence alignment, works if sequence identity > 25%
  - Protein threading: sequence-structure alignment, can go beyond the 25% limit

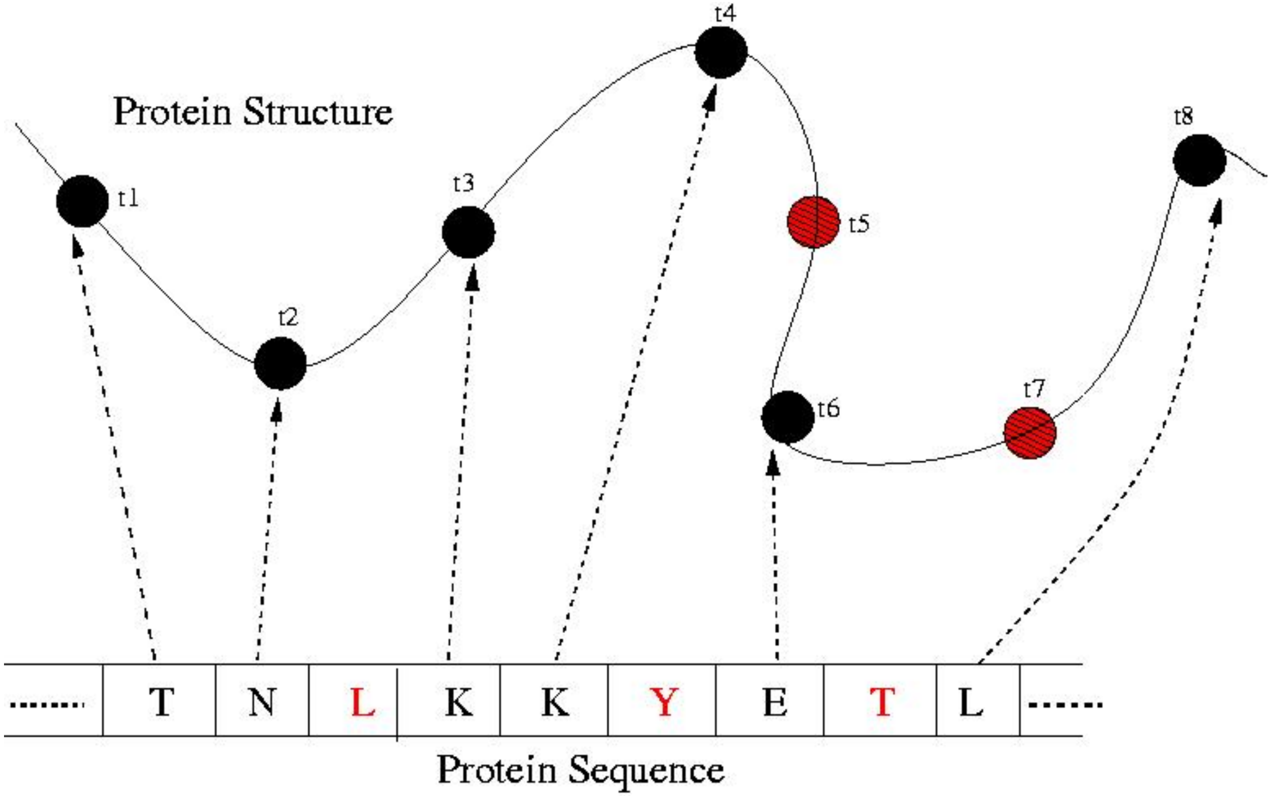
Protein Threading

Query Sequence: **DRVYIHPFADRVYIHPFA**



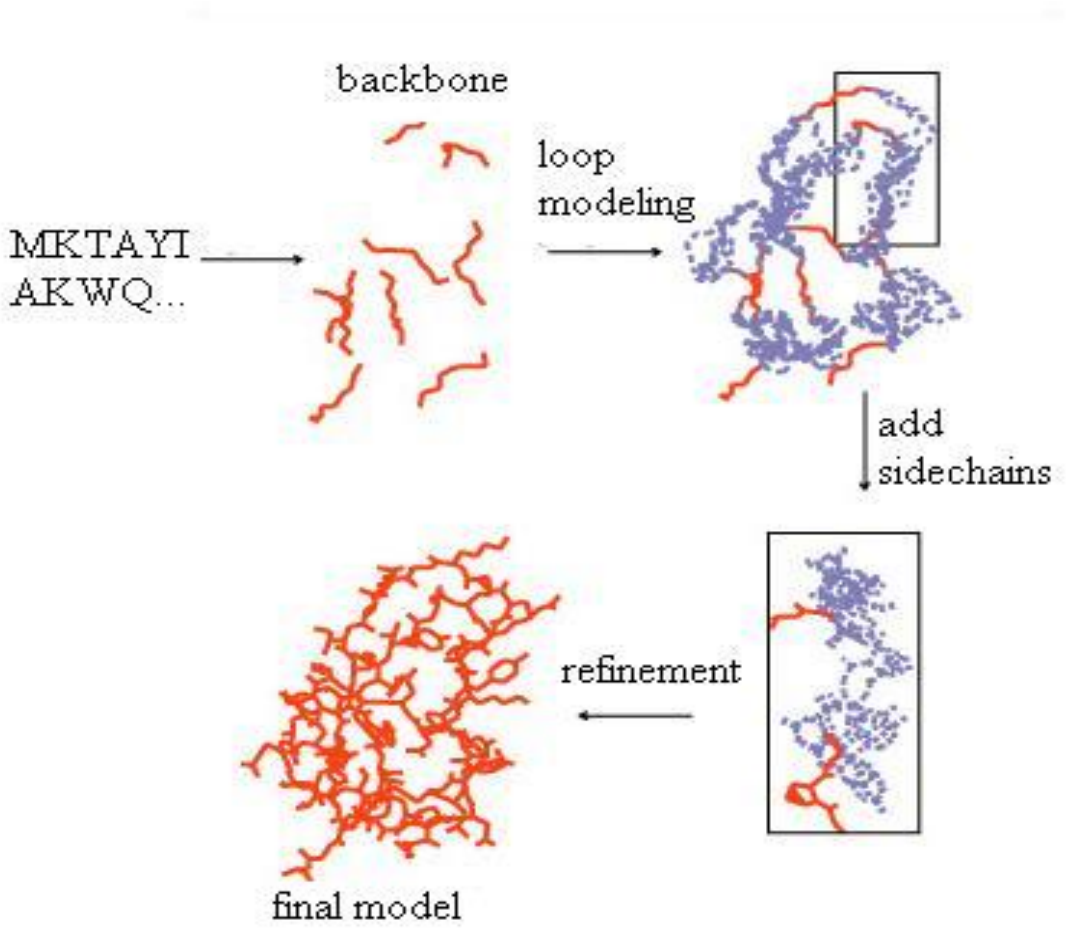
The Best Match

Threading Example



Positions or residues in red are gaps

Protein Structure Prediction

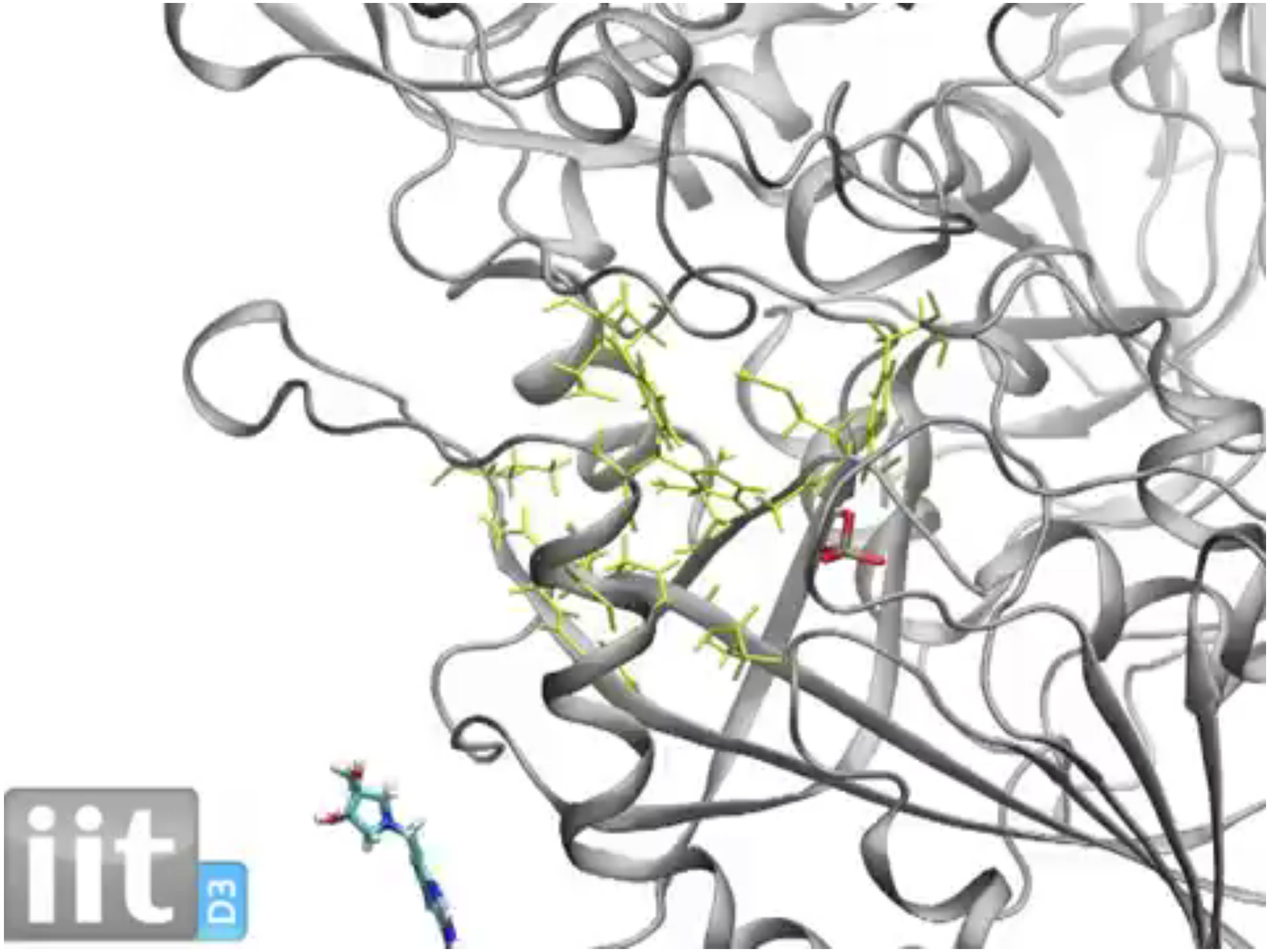


- Stage 1: Backbone Prediction
  - Ab initio folding
  - Homology modeling
  - Protein threading
- Stage 2: Loop Modeling
- Stage 3: Side-Chain Packing
- Stage 4: Structure Refinement

The picture is adapted from http://www.cs.ucdavis.edu/~koehl/ProModel/fillgap.html

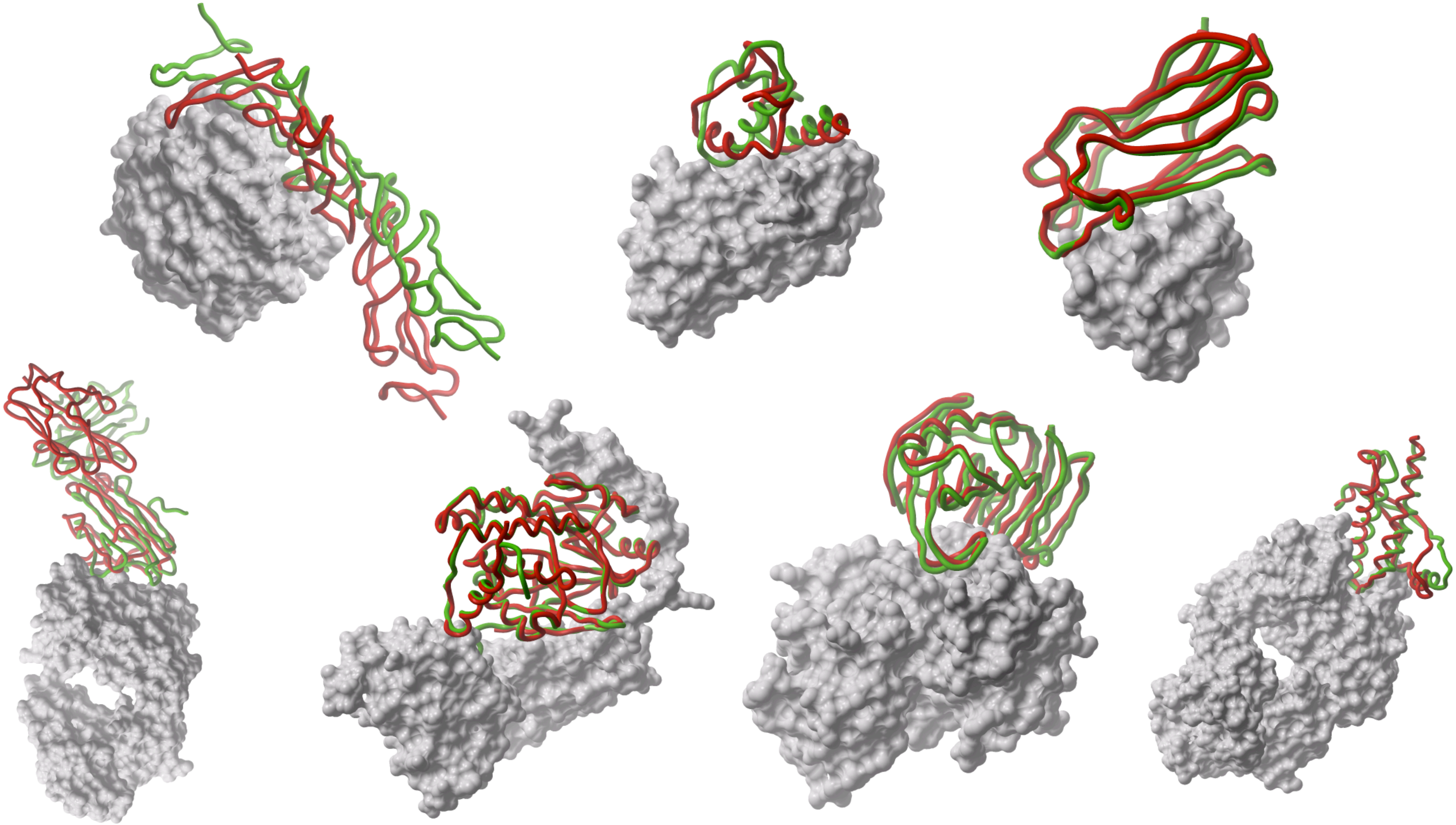
Why are structure useful?

Structure helps to understand function

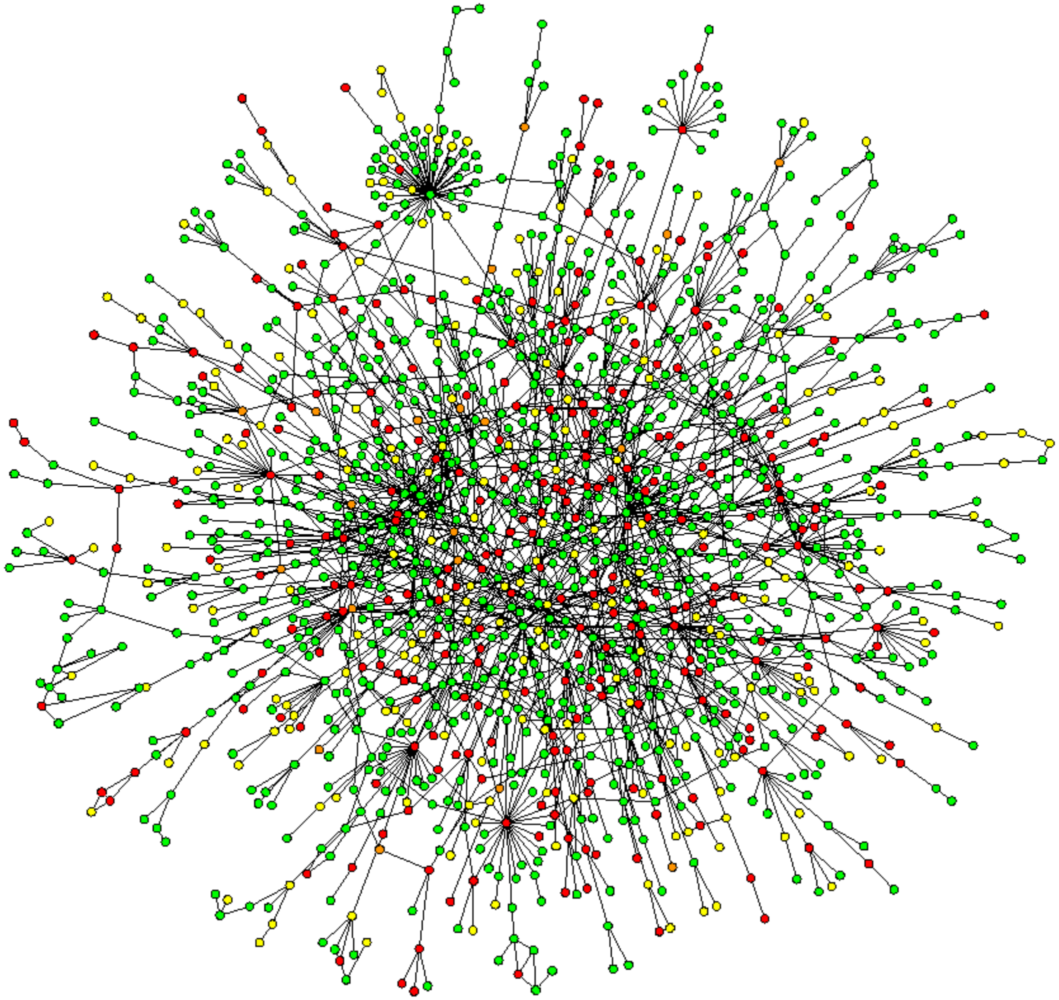


Simulation of a drug entering binging site of a protein

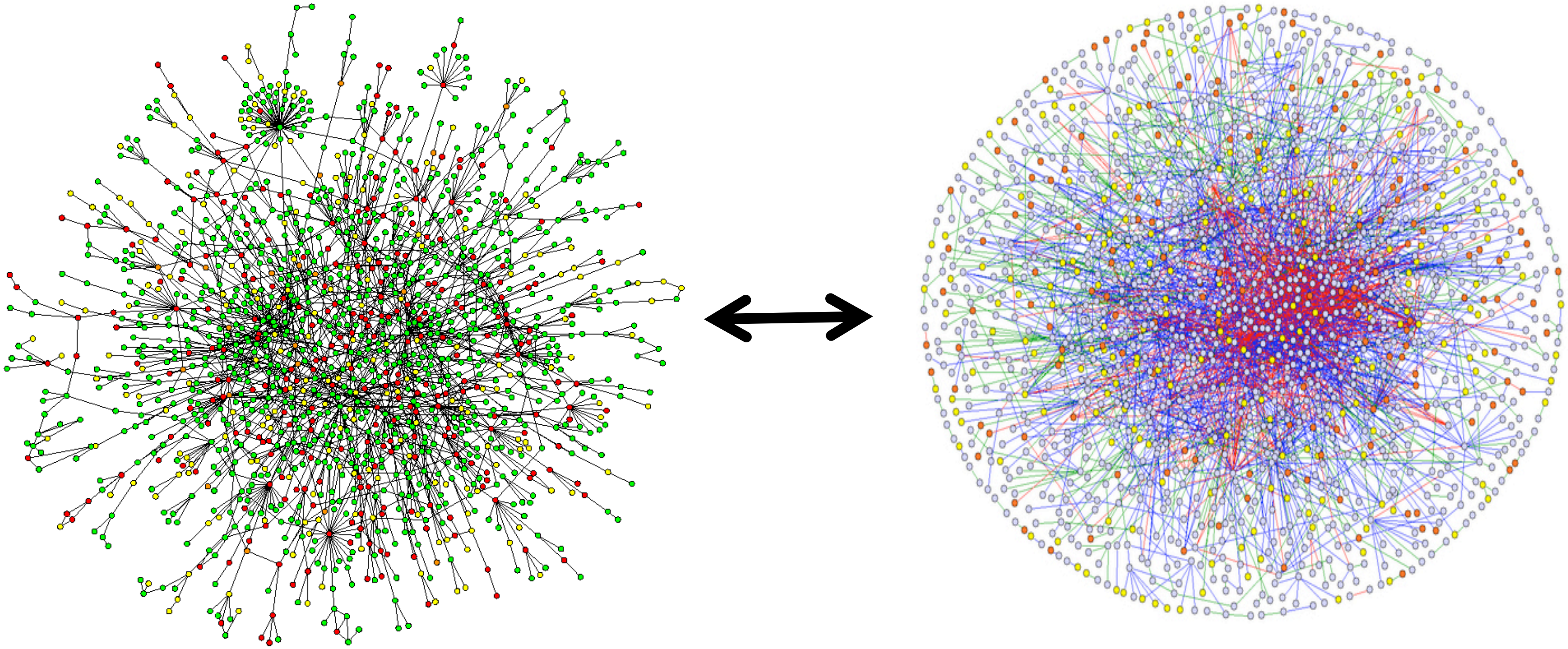
Why are structure useful?



System Biology



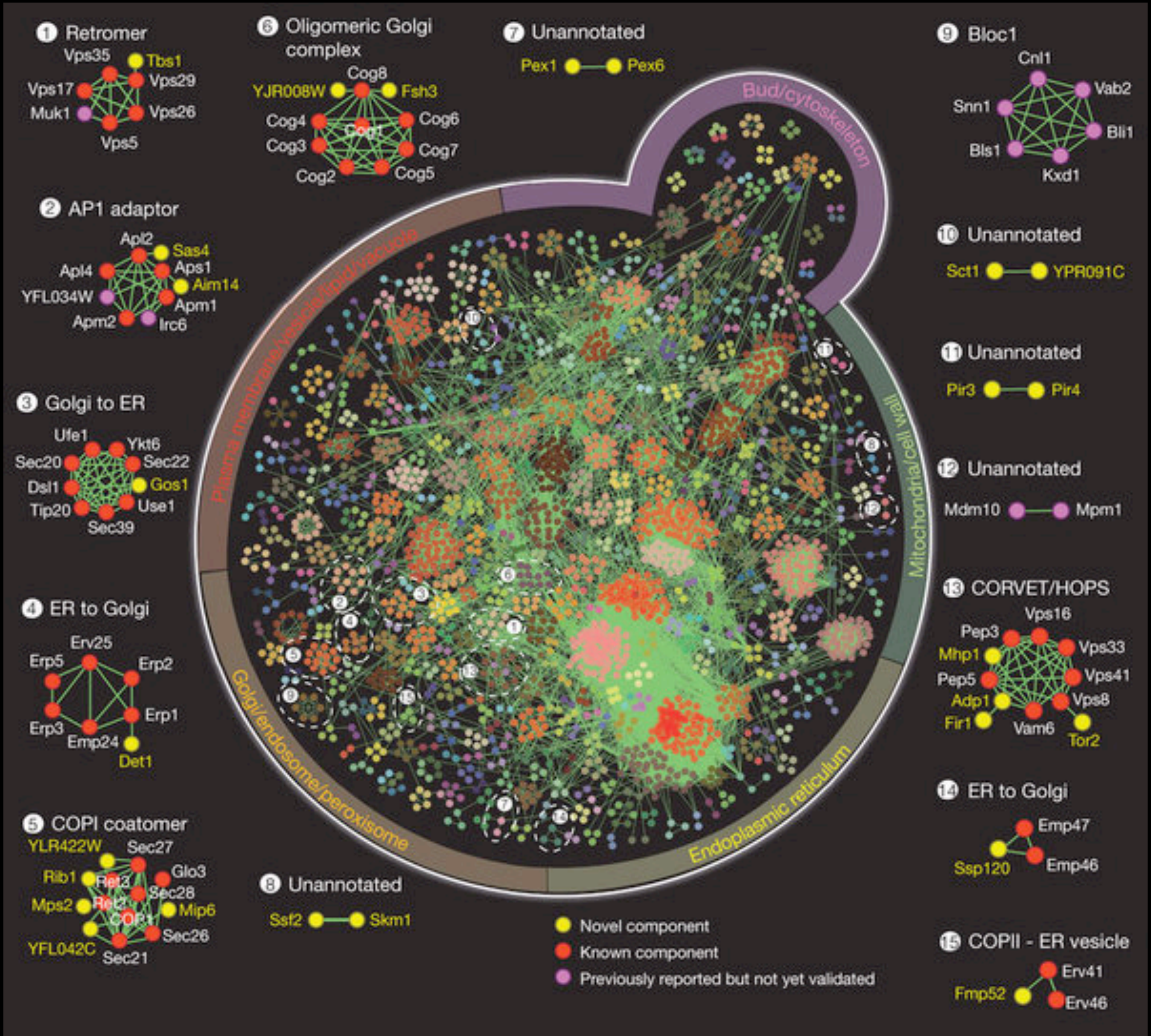
Interaction network alignment



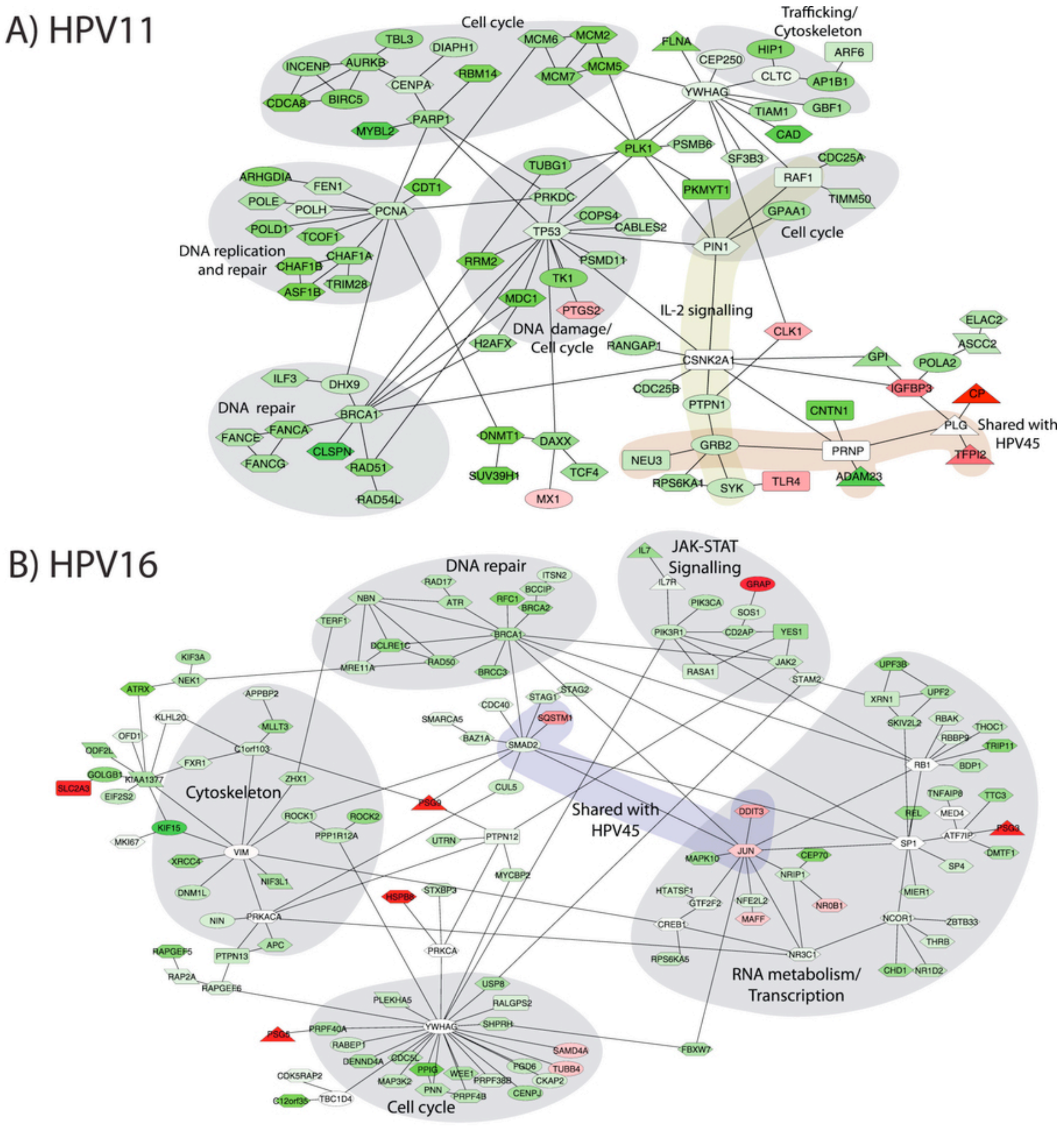
Yeast

Human

Discovery of functional motifs in PPI



How do interaction network change with disease?



Bioinformatics @ SOCS

McGill Centre for Bioinformatics is located in the Trottier Building

- Mathieu Blanchette
- Michael Hallett (Bellini Bld)
- Derek Ruths
- Jérôme Waldispühl