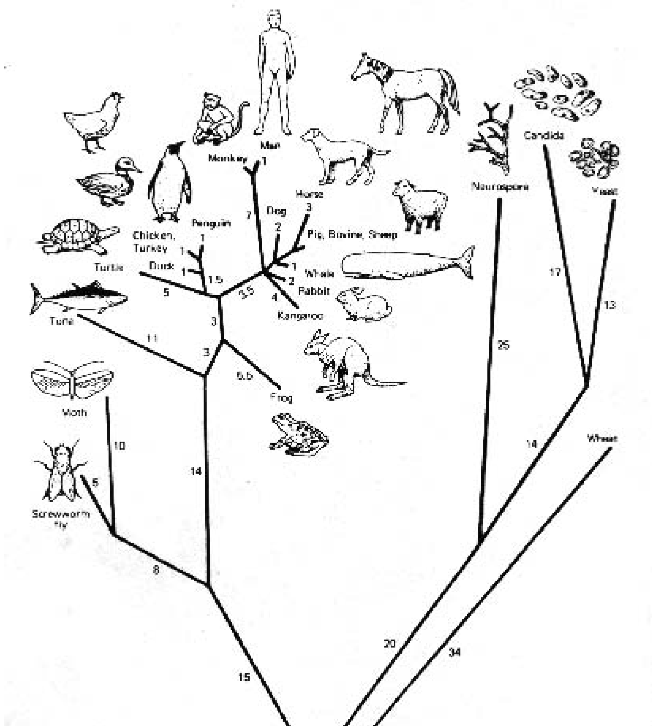
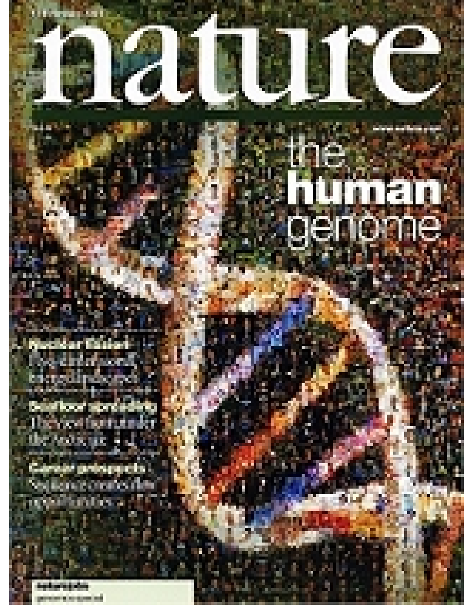
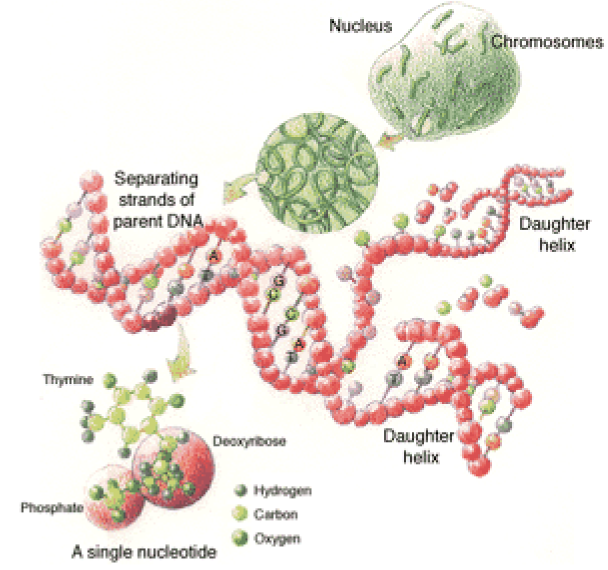






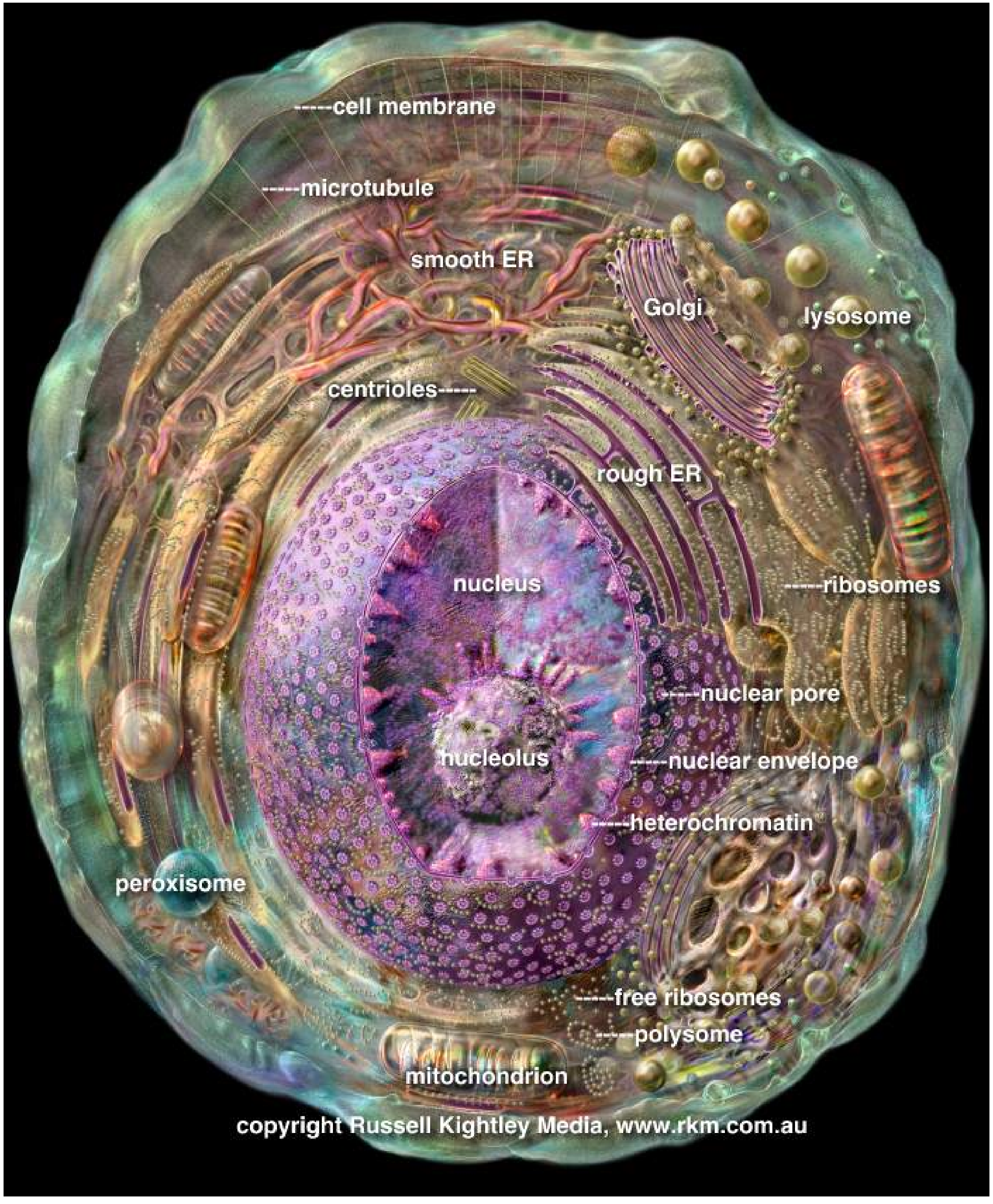
How can computers help cure cancer? (computational biology and bioinforamtics)

COS116

Instructor: Olga Troyanskaya

Molecular biology 101
or
“why bother?”

Cells are fundamental working units of all organisms



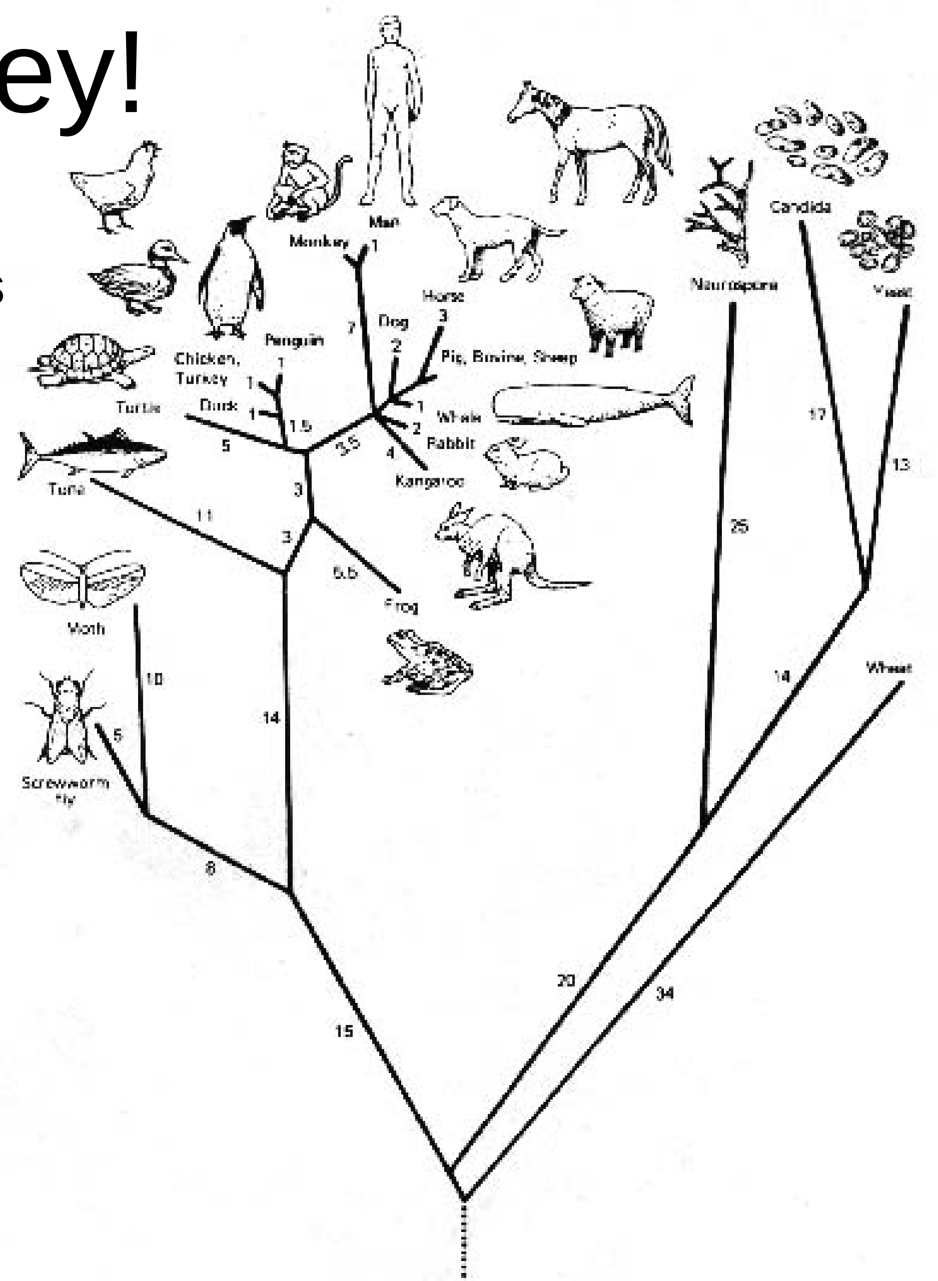
Evolution is key!

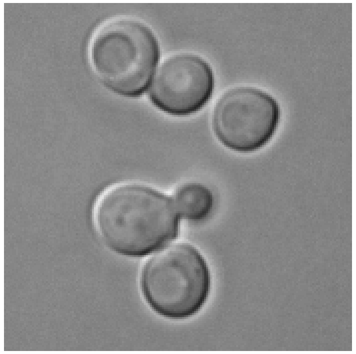
Common descent of organisms implies that they will share many “basic technologies.”

Development of new adaptations in response to environmental pressure can lead to “specialized technologies.”

More recent divergence implies more shared technologies between species.

All of biology is about two things: understanding shared or unshared features.





Yeast are unicellular organisms

Humans are multi-cellular organisms

Understanding **how a cell works** is critical to understanding how the organism functions

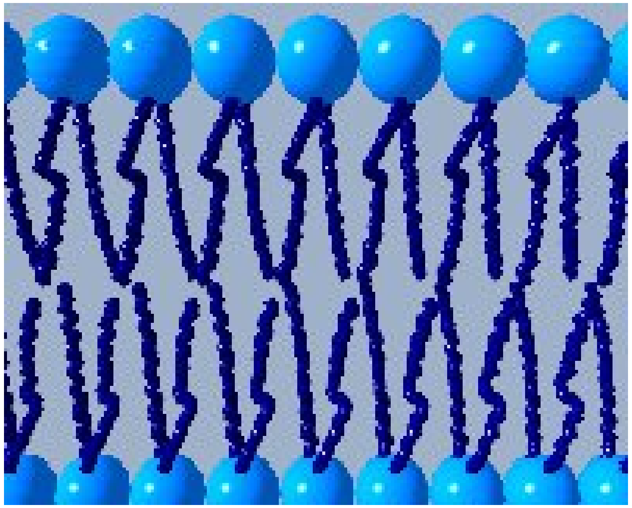
Because of evolutionary similarities – we can use yeast and other small organisms to study human biology and disease!

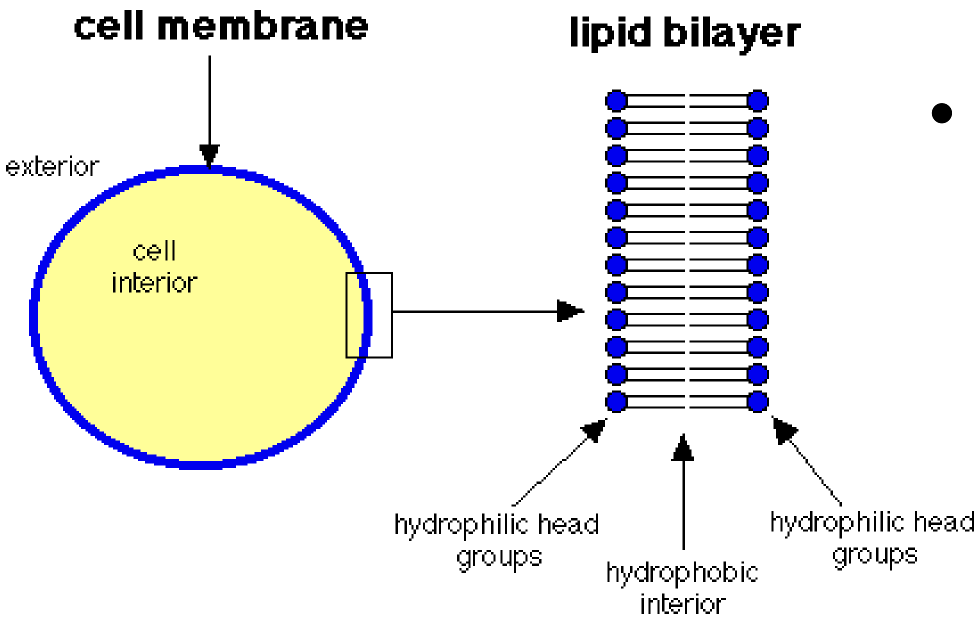
Biological macromolecules

What are the main players in molecular biology?

What is DNA, RNA, protein, lipid?

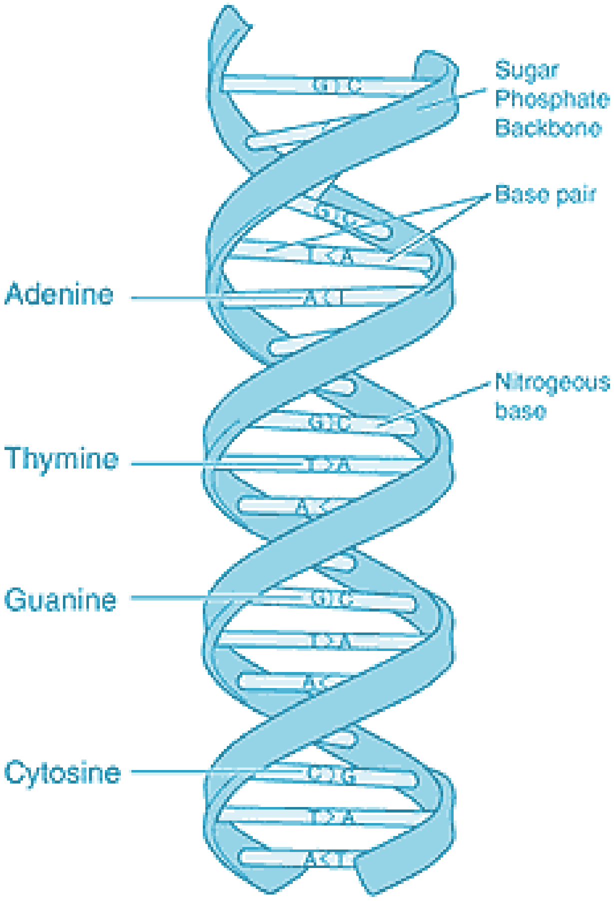
Lipids





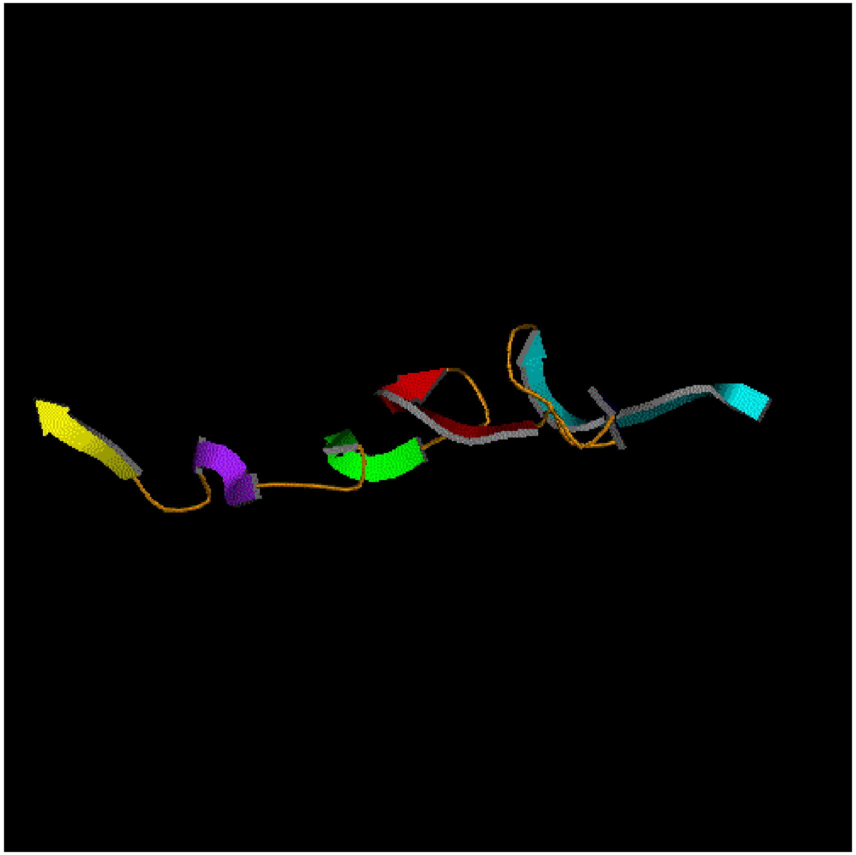
- Each lipid consists of a hydrophilic (water loving) and hydrophobic fragment
- Spontaneously form lipid bilayers => membranes

DNA



- Uses alphabet of 4 letters {ATCG}, called bases
- Encodes genetic information in triplet code
- Structure: a double helix

Proteins



Courtesy of the Zhou Laboratory, The State University of New York at Buffalo

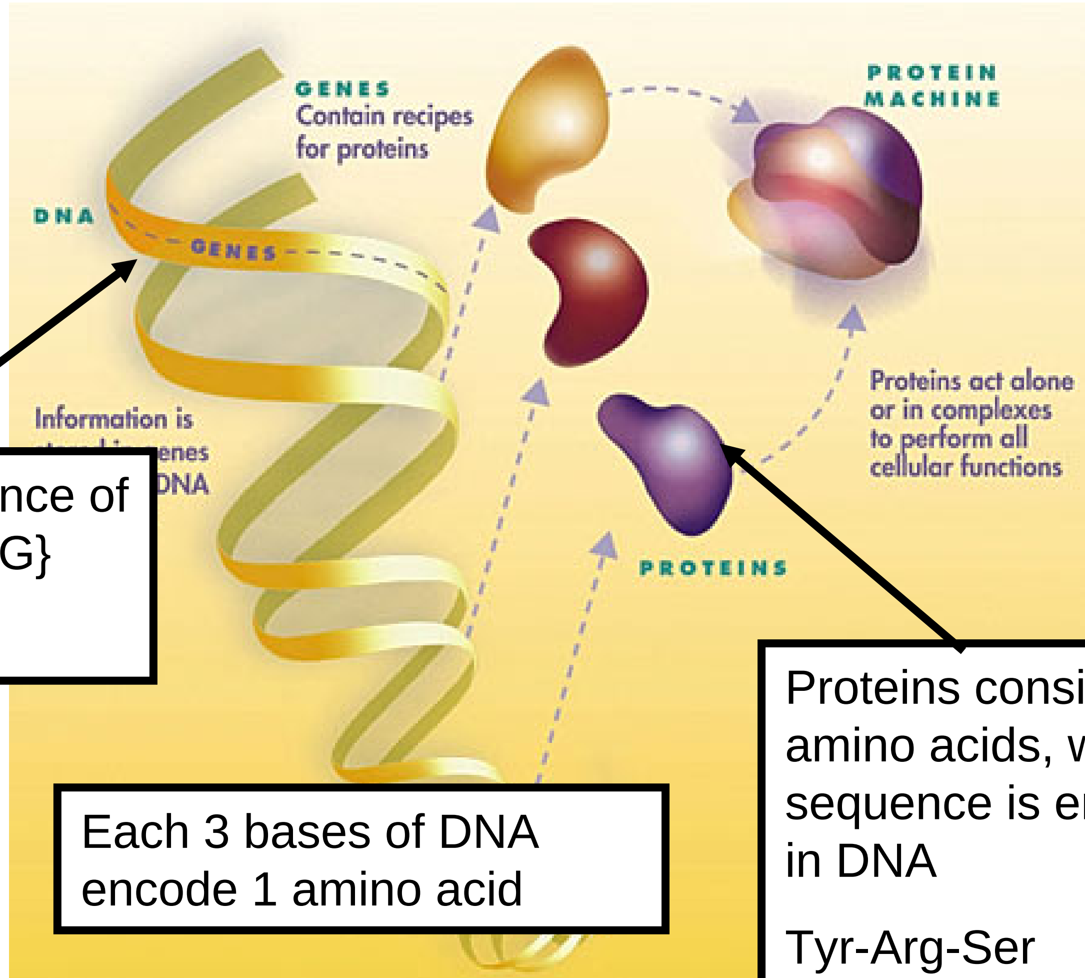
- A sequence of amino acids (alphabet of 20)
- Each amino acid encoded by 3 DNA bases
- Perform most of the actual work in the cell
- Fold into complex 3D structure

How does a cell function?
The “Central Dogma” of biology

How are proteins made?

What are translation & transcription?

How does a cell function?



DNA is a sequence of bases {A, T, C, G}

TAT-CGT-AGT

Each 3 bases of DNA encode 1 amino acid

Proteins consist of amino acids, whose sequence is encoded in DNA

Tyr-Arg-Ser

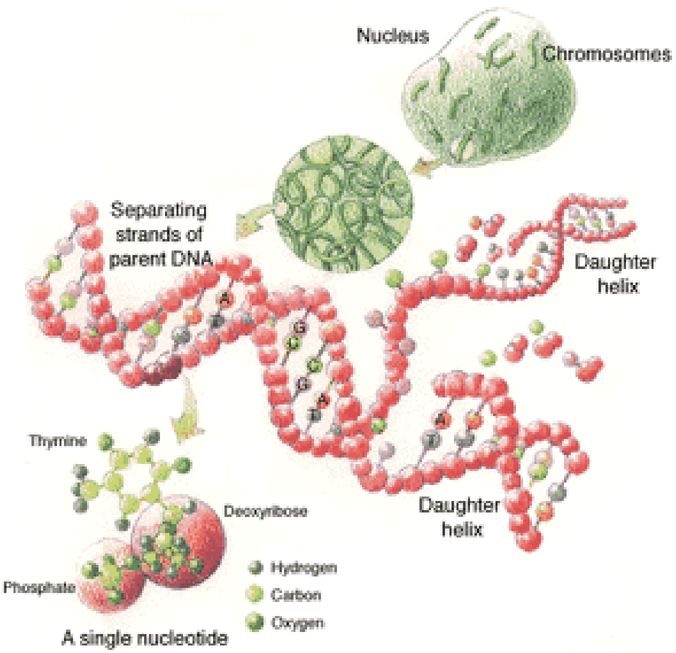
Courtesy U.S. Department of Energy Genomes to Life program

So what con computer science do to help study biology?

Case study 1:

sequencing the human genome

How is the genome sequenced?



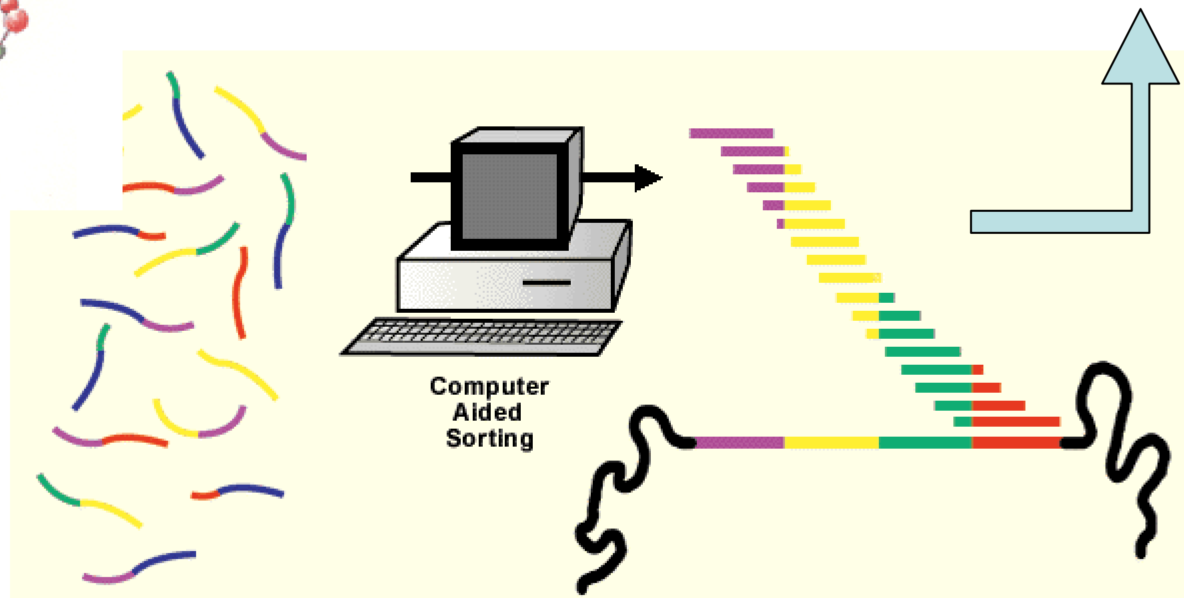


Fig 2: Short fragments of DNA sequence are ordered by overlapping data to recreate the whole genome sequence

We have a sequence – now what?

Where are the genes?

*start with “atg”

*go in triplets

*end with “act”

So how do we find them?

```
000001   gctgctggaaggggagctggccggtgggccatggccggctgcaggctctgggtttcgctgctgc
000065   tggcggcggcgttggcttgcttggccacggcactgtggccgtggccccagtacatccaaaccta
000129   ccaccggcgctacaccctgtaccccaacaacttccagttccggtaccatgtcagttcggccgcg
000193   caggcgggctgcgtcgtcctcgacgaggcctttcgacgctaccgtaacctgctcttcggttccg
000257   gctcttggccccgacccagcttctcaaataaacagcaaacgttggggaagaacattctggtggt
000321   ctccgtcgtcacagctgaatgtaatgaatttcctaatttggagtcggtagaaaattacaccccta
000385   accattaatgatgaccagtgtttactcgcctctgagactgtctggggcgctctccgaggtctgg
000449   agactttcagtcagcttgtttggaaatcagctgagggcacgttctttatcaacaagacaaagat
000513   taaagactttcctcgattccctcaccggggcgtactgctggatacatctcgccattacctgcca
000577   ttgtctagcatcctggatacactggatgtcatggcatacaataaattcaacgtgttccactggc
000641   acttggtggacgactcttccttcccatatgagagcttcactttcccagagctcaccagaaaggg
000705   gtccttcaaccctgtcactcacatctacacagcacaggatgtgaaggaggtcattgaatacgca
000769   aggcttcggggtatccgtgtgctggcagaatttgacactcctggccacactttgtcctgggggc
000833   caggtgcccctgggttattaacaccttgctactctgggtctcatctctctggcacatttggacc
000897   ggtgaaccccagtctcaacagcacctatgacttcatgagcacactcttcctggagatcagctca
000961   gtcttcccggacttttatctccacctgggagggggatgaagtcgacttcacctgctggaagtcca
001025   accccaacatccaggccttcatgaagaaaaagggctttactgacttcaagcagctggagtcctt
001089   ctacatccagacgctgctggacatcgtctctgattatgacaagggctatgtggtgtggcaggag
001153   gtatttgataataaagtgaaggttcggccagatacaatcatacaggtgtggcgggaagaaatgc
001217   cagtagagtacatgttggagatgcaagatatcaccagggctggcttccgggccctgctgtctgc
001281   tccctggtacctgaaccgtgtaaagtatggccctgactggaaggacatgtacaaagtggagccc
001345   ctggcgtttcatggtacgcctgaacagaaggctctggtcattggagggggaggcctgtatgtggg
001409   gagagtatgtggacagcaccaacctggtccccagactctggcccagagcgggtgccgtcgctga
001473   gagactgtggagcagtaacctgacaactaatatagactttgcctttaaacgtttgtcgcatttc
001537   cgttgtgagctggtgaggagaggaatccaggcccagcccatcagtgtaggctgctgtgagcagg
001601   agtttgagcagacttgagccaccagtgctgaacacccaggaggttgctgtcctttgagtcagct
001665   gcgctgagcacccaggagggtgctggccttaagagagcaggtcccggggcagggctaatctttc
001729   actgcctcccggccaggggagagcaccccttgcccgtgtgcccctgtgactacagagaaggagg
001793   ctggtgctggcactggtgttcaataaagatctatgtggcatttctc
```

P53

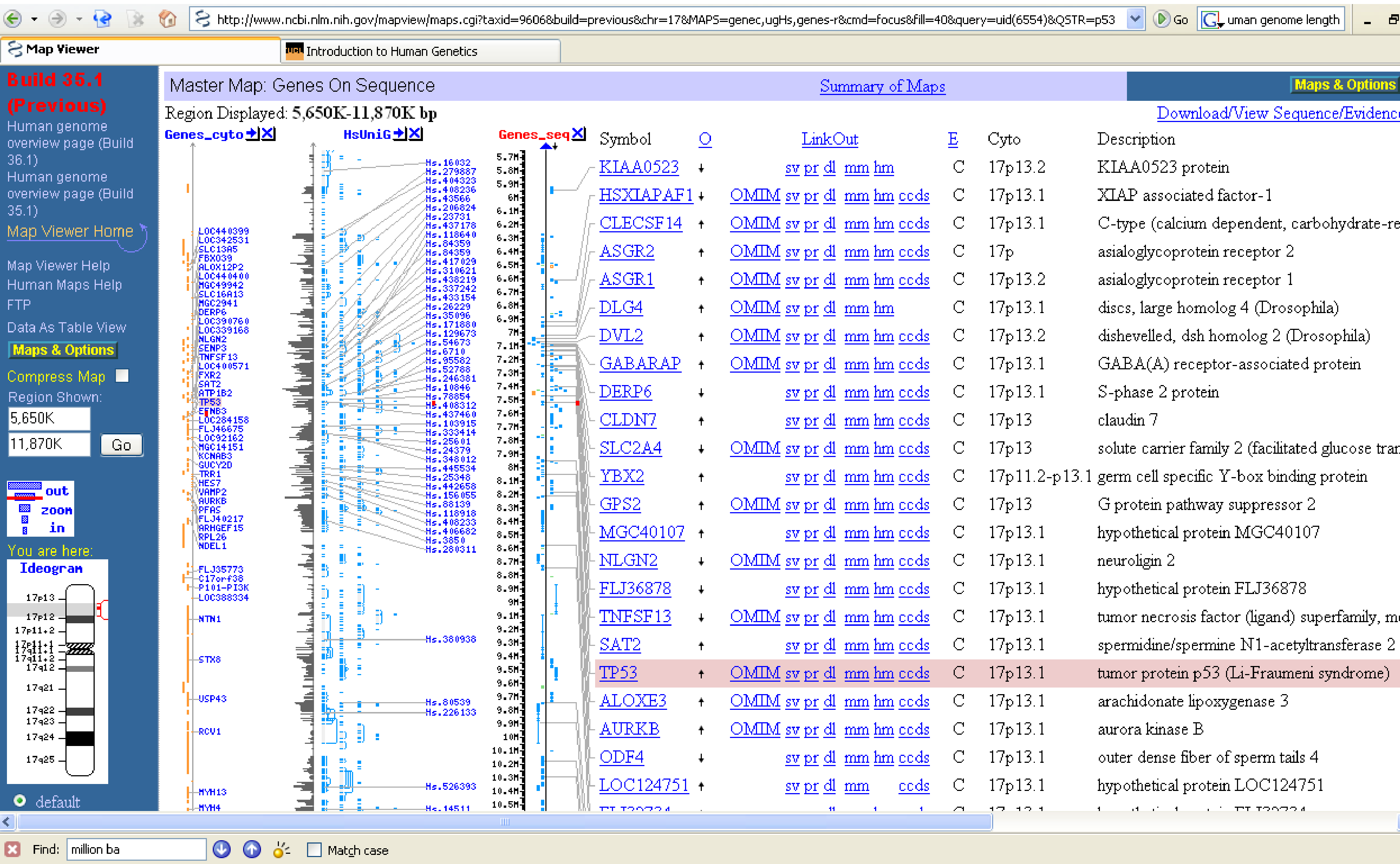
how would you find this?

```
   1 ctggcgcgcg cggccctgcg ggtgacaggc aggcgggaag gggcggggcc tcgggcgggg
  61 ccgccgtggg gaggagggcg gtgggagggg aggagtggag atggcggcgg cggcggctca
 121 gggggcgggg ggcggggagc cccgtagaac cgaggggtc ggcccggggg tcccggggga
 181 ggtggagatg gtgaaggggc agccgttcga cgtgggcccg cgctacacgc agttgcagta
 241 catcggcgag ggcgcgtacg gcatggtcag ctcggcctat gaccacgtgc gcaagactcg
 301 cgtggccatc aagaagatca gcccccttcga acatcagacc tactgccagc gcacgctccg
 361 ggagatccag atcctgctgc gcttccgcca tgagaatgtc atcggcatcc gagacattct
 421 gcgggcgtcc accctggaag ccatgagaga tgtctacatt gtgcaggacc tgatggagac
 481 tgacctgtac aagttgctga aaagccagca gctgagcaat gaccatatct gctacttcct
 541 ctaccagatc ctgcggggcc tcaagtacat ccactccgcc aacgtgctcc accgagatct
 601 aaagccctcc aacctgctca tcaacaccac ctgcgacctt aagatttgtg atttcggcct
 661 ggcccggatt gccgatcctg agcatgacca caccggcttc ctgacggagt atgtggctac
 721 gcgctggtac cgggccccag agatcatgct gaactccaag ggctatacca agtccatcga
 781 catctggtct gtgggctgca ttctggctga gatgctctct aaccggccca tcttccctgg
 841 caagcactac ctggatcagc tcaaccacat tctgggcatc ctgggctccc catcccagga
 901 ggacctgaat tgtatcatca acatgaaggc ccgaaactac ctacagtctc tgccctccaa
 961 gaccaaggtg gcttgggcca agcttttccc caagtcagac tccaaagccc ttgacctgct
1021 ggaccggatg ttaaccttta accccaataa acggatcaca gtggaggaag cgctggctca
1081 cccctacctg gagcagtact atgacccgac ggatgaggtg ggccagtccc cagcagcagt
1141 ggggctgggg gcaggggagc agggggcac gtaggcatcc cccatgccag gcctgagcct
1201 tgctgtctct accaccccag ccagtggccg aggagccctt caccttcgcc atggagctgg
1261 atgacctacc taaggagcgg ctgaaggagc tcatcttcca ggagacagca cgcttccagc
1321 ccggagtgct ggaggccccc tagcccagac agacatctct gcaccctggg gcctggacct
1381 gcctcctgcc tgcccctctc ccgccagact gttagaaaat ggacactgtg cccagcccgg
1441 accttggcag cccaggccgg ggtggagcat gggcctggcc acctctctcc tttgctgagg
1501 cctccagctt caggcaggcc aaggccttct cctccccacc cgccctcccc acggggcctc
1561 gggacctcag gtggccccag ttcaatctcc cgctgctgct gctgcgccct taccttcccc
1621 agcgtcccag tctctggcag ttctggaatg gaagggttct ggctgcccca acctgctgaa
1681 gggcagaggt ggagggtggg gggcgctgag tagggactca gggccatgcc tgcccccctc
1741 atctcattca aaccccaccc tagtttccct gaaggaacat tccttagtct caagggctag
1801 catccctgag gagccaggcc gggccgaatc ccctccctgt caaagctgtc acttcgcgtg
1861 ccctcgctgc ttctgtgtgt ggtgagcaga agtggagctg gggggcgtgg agagcccggc
1921 gccccctgcca cctccctgac ccgtctaata tataaatata gagatgtgtc tatggctgaa
1981 aaaaaaaaaa aaaaaaaaaa aaaaa
```

Where is this gene in the genome?

- Human genome is 3 billion bases long
- TP53 is on:
  - Chromosone 17
  - Small arm
  - Position 17.3
  - Around base 9.5 mil in chromosome 17 (Human chromosomes range in length from 51 million to 245 million base pairs)
- Could you find this by hand?

Need large databases for all the information!



So what con computer science do to help study biology?

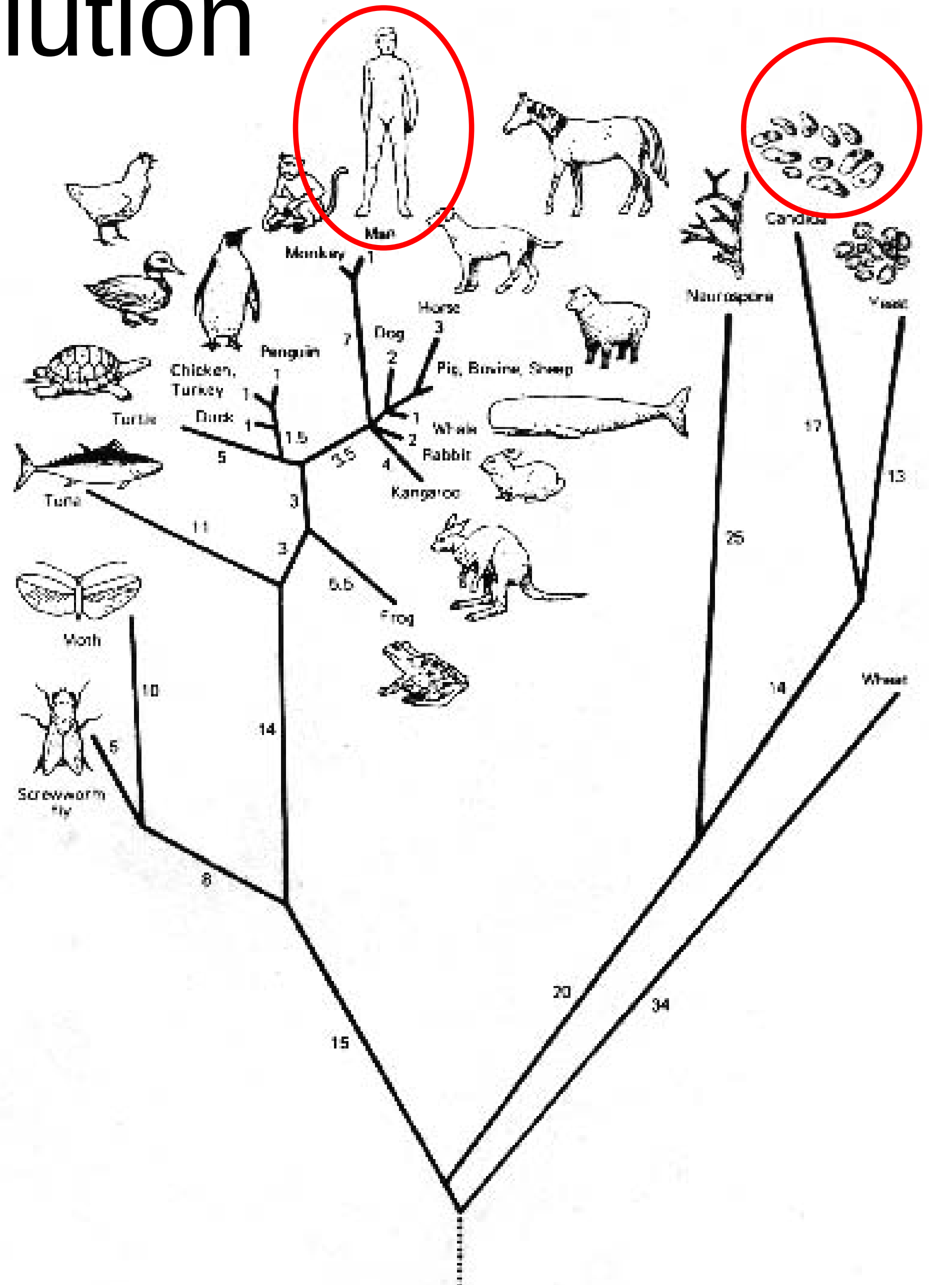
Case study 2:

so what do these genes do?

Remember – evolution can help!

Common descent of organisms implies that they will share many "basic technologies."

Thus, can use yeast to understand what proteins do in humans!



Human hereditary colon cancer gene was found by looking for a gene similar to MSH2 gene in yeast (these genes are 65% similar)!

(Yeast sequence on top).

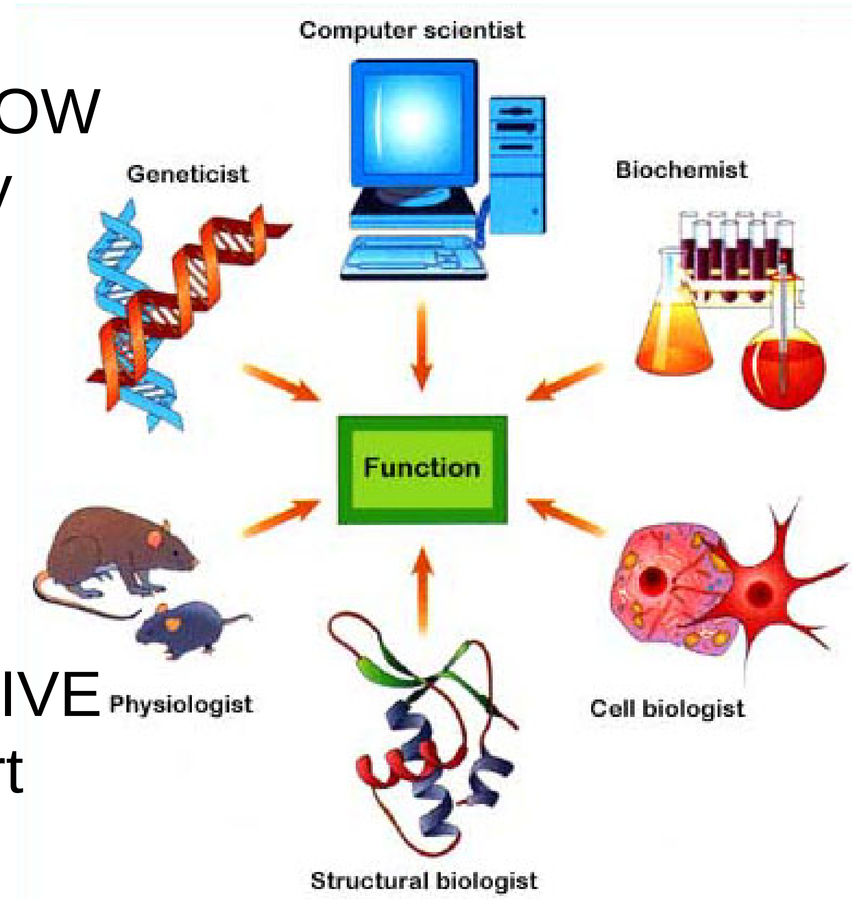
BLAST P-value: 3.8e-255

Percent Similarity (|+|): 64.7   Percent Identity (|): 43.0

```
  1 MSSTRPELKFSDVSEERNFYKKYTGLPKKPLKTIRLVDKGDYYTVIGSDA  50
    |.  ..|    | ..| .| | | |||.|| .||||.|||||||. |.||
  1 MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYTAHGEDA  50

 51 IFVADSVYHTQSVLKNCQLDPVTAKNFHEPTKYVTVSLQVLATLLKLCLL 100
    ||.| .||.||||||   |||..|||||.    |.||   |..|||  ||
 51 LLAAREVFKTQGVIKY..MGPAGAKNLQS....VVLSKMNFESFVKDLLL  94

101 DLGYKVEIY..........DKGWKLIKSASPGNIEQVNELMNMNIDSSII 140
      .||||||          |.|| |  .||||||.|.||||  | | |
 95 VRQYRVEVYKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNNDMSAS 144

141 IASLKVQWNSQDGNCIIGVAFIDTTAYKVGMLDIVDNEVYSNLESFLIQL 190
    || | |. .. ||.  |||||||... |||| ||.||| |||||.|||||
145 IGVVGVKMSAVDGQRQVGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQI 194

191 GVKECLVQDLTSNSNSNAEMQKVINVIDRCGCVVTLLKNSEFSEKDVELD 240
    |.|||||.|    ||..|||.|| .||||.|.|||  |..|||.|||  |
195 GPKECVLPG....GETAGDMGKLRQIIQRGGILITERKKADFSTKDIYQD 240

241 LTKLL....GDDL.ALSLPQKYSKLSMGACNALIGYLQLLSEQDQVGKYE 285
    |.|||    |||| .  |||. ..|.||. .||| ||||||||...|.||
241 LNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIKFLELLSDDSNFGQFE 290
```

So what's next? Genomics is a collaborative discipline

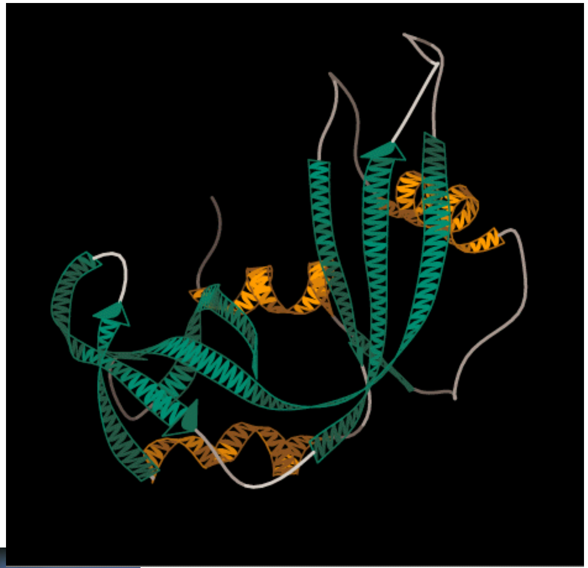
- To study WHAT proteins DO, HOW they INTERACT, and HOW they are REGULATED, need data beyond genomic sequence



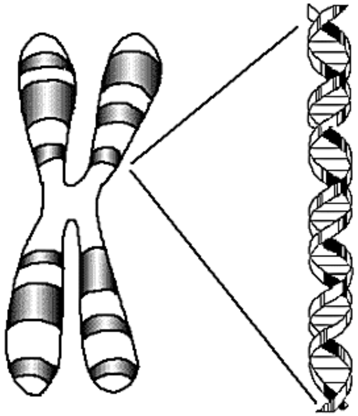
- Genomics/Bioinformatics is fundamentally a COLLABORATIVE and MULTIDISCIPLINARY effort

Now we are trying to understand how small differences in DNA lead to large differences in phenotypes

Proteins



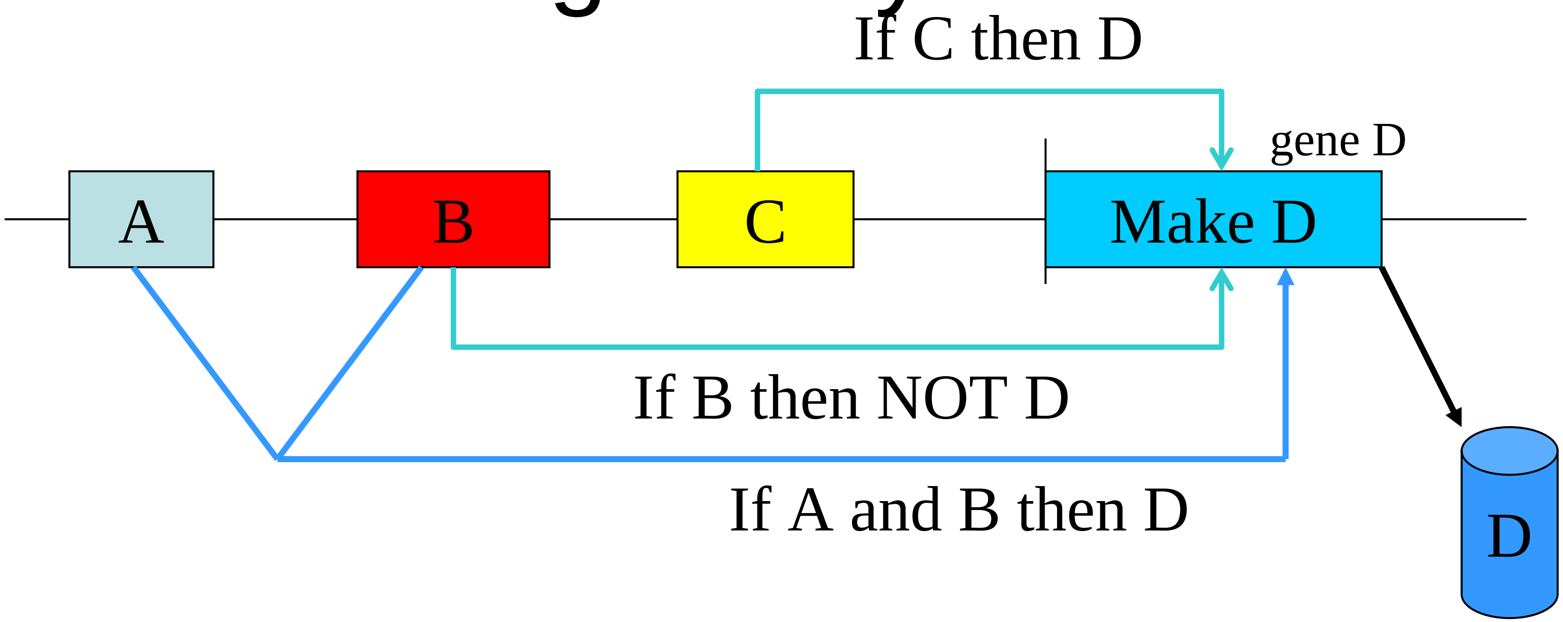
Gene Expression



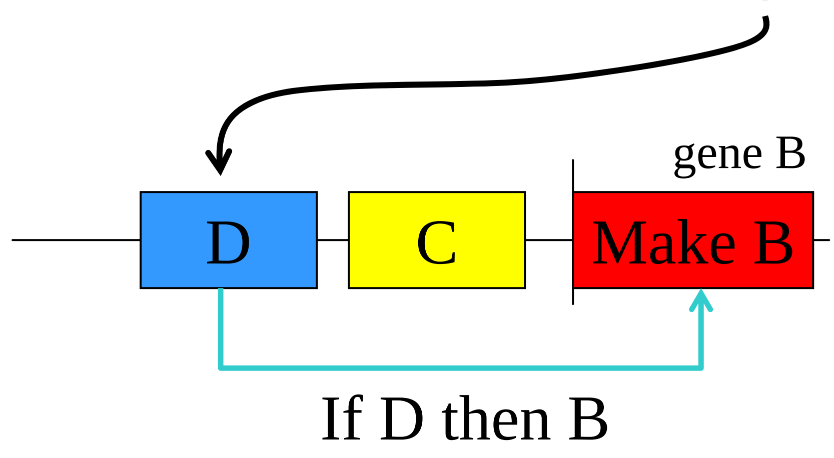
DNA

People

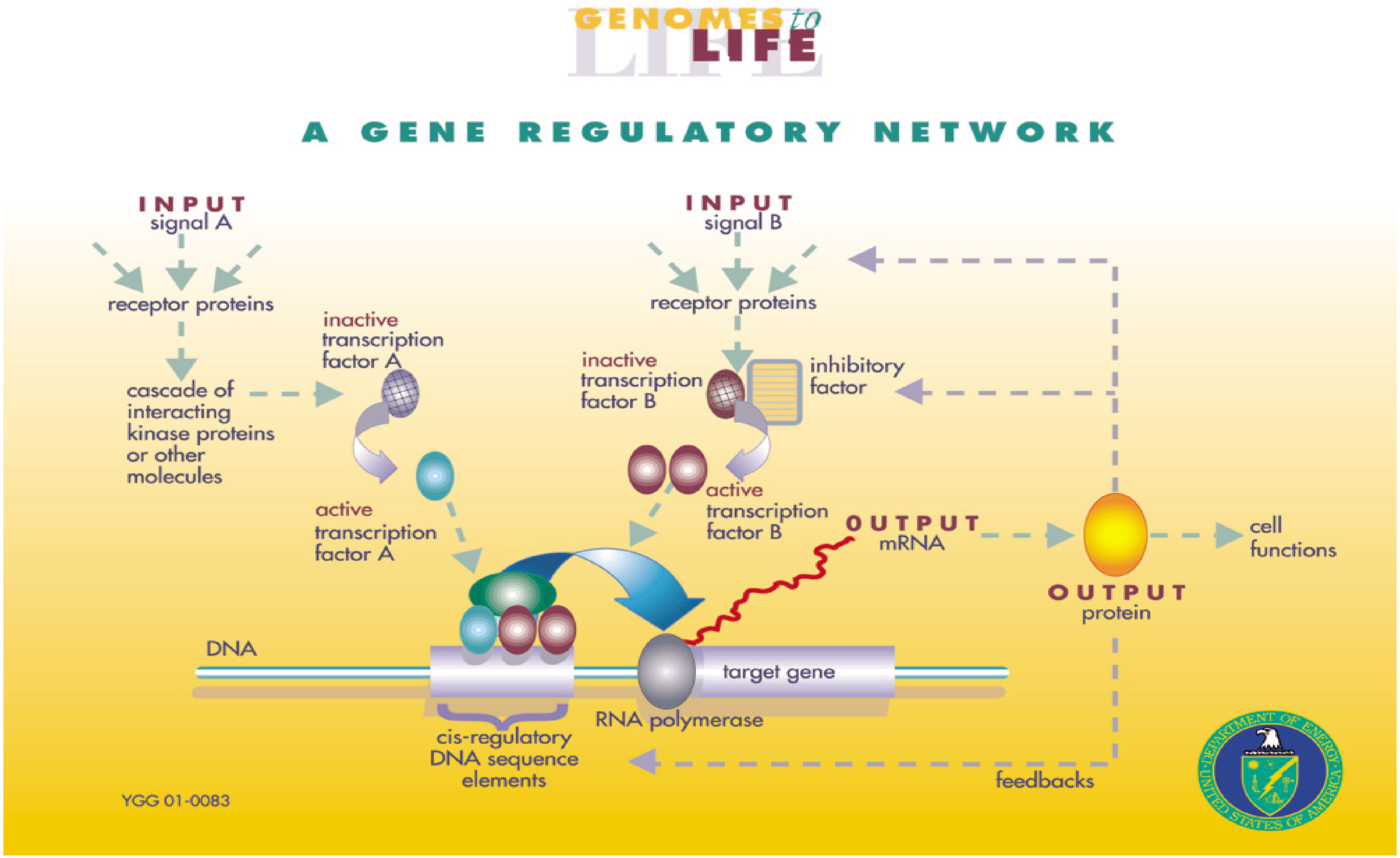
The answer is probably in regulation "Gene Regulatory Circuits"



- Genes =? wires
- Motifs =? gates



Regulatory Networks



Any questions?

What is genomics all about?

The “omes” in biology.

Why bioinformatics?

What is “systems biology”?

The "omes"

- Genome – organism's complete set of DNA
  - Relatively stable through an organism's lifetime
  - Size: from 600,000 to several billion bases
  - Gene is a basic unit of heredity (only 2% of the human genome)
- Proteome – organism's complete set of proteins
  - Dynamic – changes minute to minute
  - Proteins actually perform most cellular functions, they are encoded by genes (not a 1-to-1 relationship)
  - Protein function and structure form molecular basis for disease

Beyond the “omes” – systems biology

- Understanding the function and regulation of cellular machinery, as well as cell-to-cell communication on the molecular level
- Why? Because most important biological problems are fundamentally systems-level problems
  - Systems-level understanding of disease (e.g. cancer)
  - Molecular medicine
  - Gene therapy

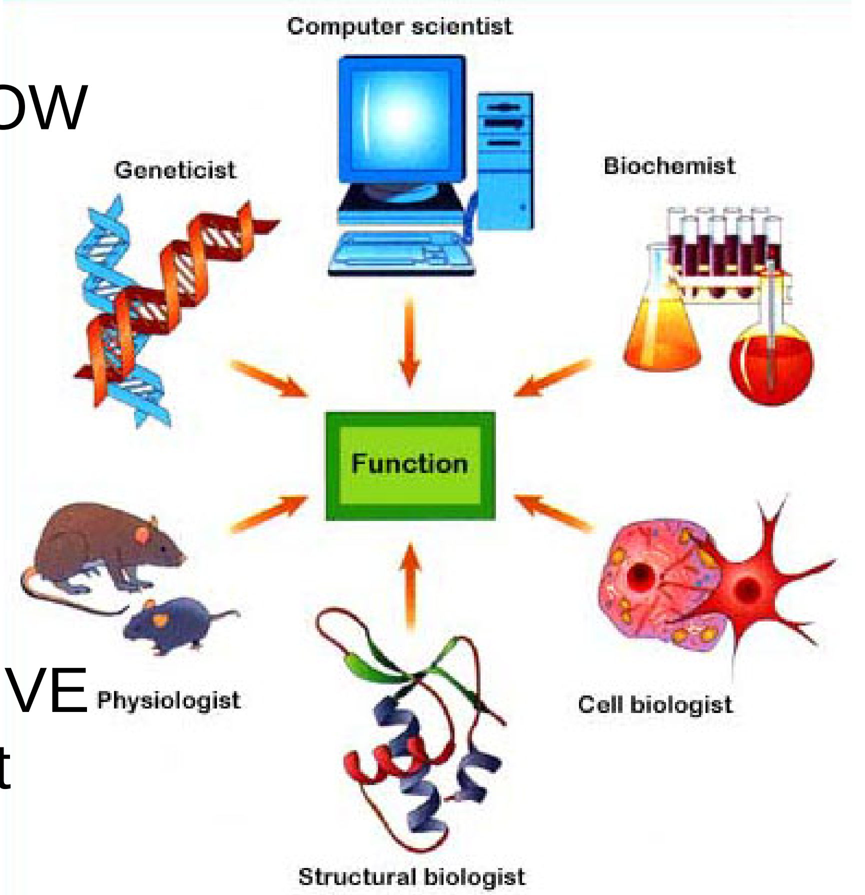
Systems-level challenges

- **Gene function annotation – what does a gene do**
  - ~30,000 genes in the human genome => systems-level approaches necessary
  - A modern human microarray experiment produces ~500,000 data points => computational analysis & visualization necessary
  - Many high-throughput functional technologies => computational methods necessary to integrate the data
- **Biological networks – how do proteins interact**
  - Large amounts of high-throughput data => computation necessary to store and analyze it
  - Data has variable specificity => computational approaches necessary to separate reliable conclusions from random coincidences
- **Comparative genomics – comparing data between organisms**
  - Need to map concepts across organisms on a large scale => practically impossible to do by hand
  - High amount of variable quality data => computational methods needed for integration, visualization, and analysis
  - Data often distributed in databases across the globe, with variable schemas etc => data storage and consolidation methods needed

Function

- To study WHAT proteins DO, HOW they INTERACT, and HOW they are REGULATED, need data beyond genomic sequence

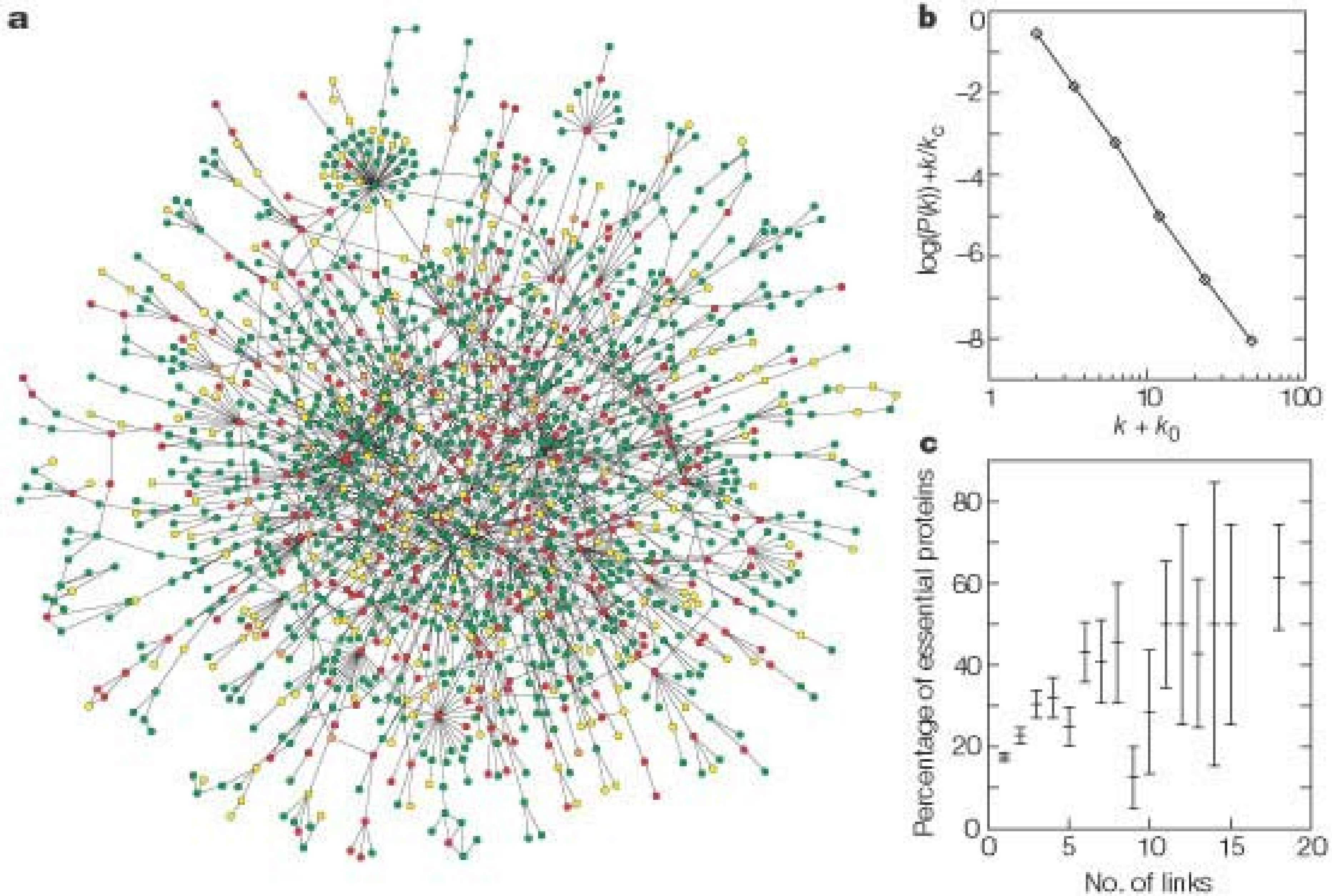
- Genomics/Bioinformatics is fundamentally a COLLABORATIVE and MULTIDISCIPLINARY effort



Biological networks

- Interaction maps (no directions)
- Pathway models (dynamic or static)
- Metabolic networks
- Genetic regulatory networks

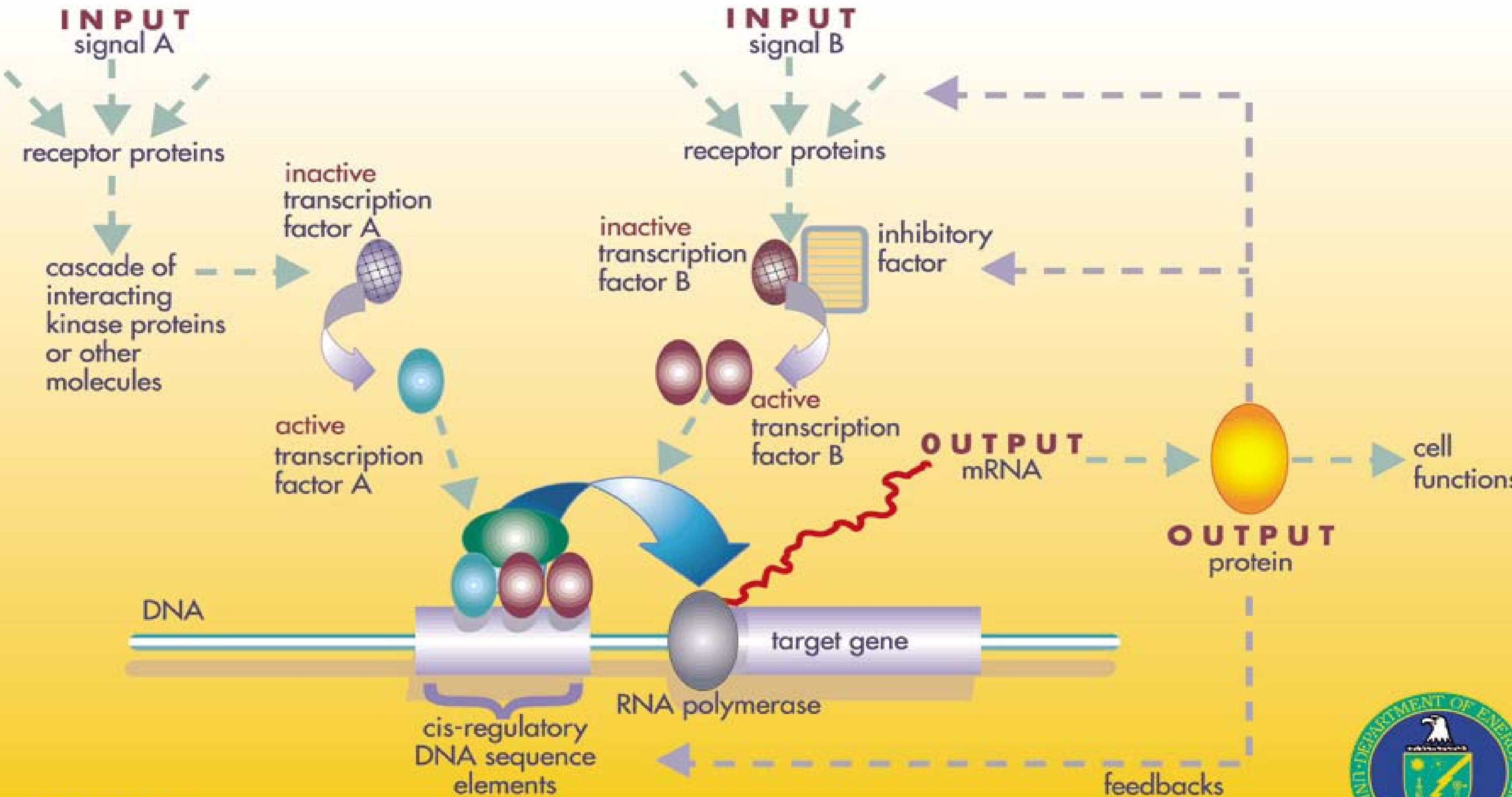
Yeast interaction network



Hawoong Jeong et al. Oltvai Centrality and lethality of protein networks. Nature 411, 41-42 (2001)

GENOMES to LIFE

A GENE REGULATORY NETWORK



YGG 01-0083

Gene expression microarrays – one type of high-throughput functional data

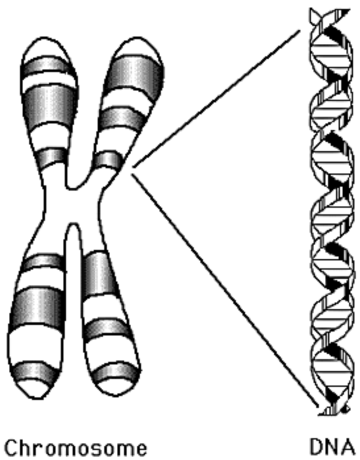
Why microarray analysis: the questions

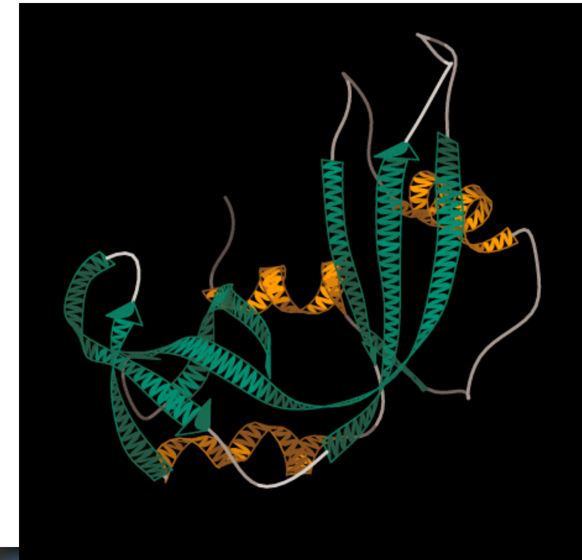
- Large-scale study of biological processes
- What is going on in the cell at a certain point in time?
- On the large-scale genetic level, what accounts for differences between phenotypes?
- Sequence important, but genes have effect through expression

Why study gene expression

Proteins

Gene Expression





DNA

People

Computational biology/bioinformatics

What does it study?

Where do we get the data?

Computational Molecular Biology

- In order to gather insight into the ways in which genes and gene products (proteins) function, we:
- 1. Analyze DNA and protein sequences, searching for clues about structure, function, and control.

  SEQUENCE ANALYSIS
- 2. Analyze biological structures, searching for clues about sequence, function and control.

  STRUCTURE ANALYSIS
- 3. Understand how cellular components function in living systems.

  FUNCTION ANALYSIS

What are functions of genes?

- Signal transduction: sensing a physical signal and turning into a chemical signal
- Structural support: creating the shape and pliability of a cell or set of cells
- Enzymatic catalysis: accelerating chemical transformations otherwise too slow.
- Transport: getting things into and out of separated compartments

What are the functions of genes?

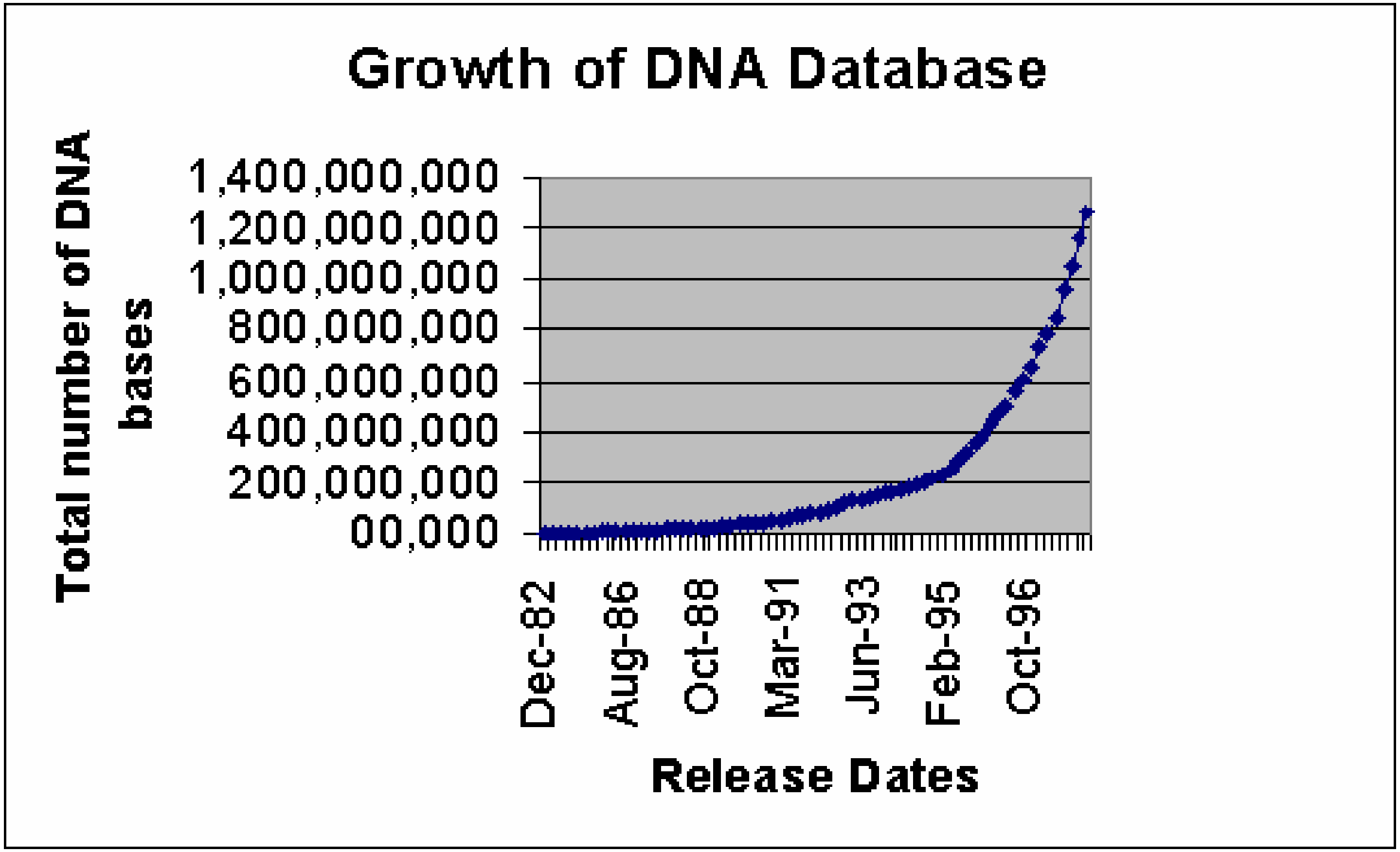
- Movement: contracting in order to pull things together or push things apart.

- Transcription control: deciding when other genes should be turned ON/OFF

- Trafficking: affecting where different elements end up inside the cell

Biology and Medicine are fundamentally information sciences.

http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html



Complete Genomes Known (900 currently available publically)

- Aquifex aeolicus
- Archaeoglobus fulgidus
- Bacillus subtilis
- Borrelia burgdorferi
- Chlamydia trachomatis
- Escherichia coli
- Haemophilus influenzae
-
- Methanobacterium thermoautotrophicum
- Caulobacter crescentus
- Helicobacter pylori
- Methanococcus jannaschii
- Mycobacterium tuberculosis
- Mycoplasma genitalium
- Mycoplasma pneumoniae
- Pyrococus horikoshii
- Treponema pallidum
- Saccharomyces cerevisiae
- Drosophila melanogaster
- Arabidopsis thaliana
- Caenorhabditis elegans
- Homo sapiens

http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/org.html

Computer Science & Genomics

- Computer science a discipline of itself
- BUT: it's also a tool applied to study of other disciplines
  - Bioinformatics: cs applied to biology & biochemistry
  - Neuroscience
  - Security and policy
  - Economics
  - Physics

Computational Molecular Biology (bioinformatics)

In order to gather insight into the ways in which genes and gene products (proteins) function, we:

1. Analyze DNA and protein sequences, searching for clues about structure, function, and control.

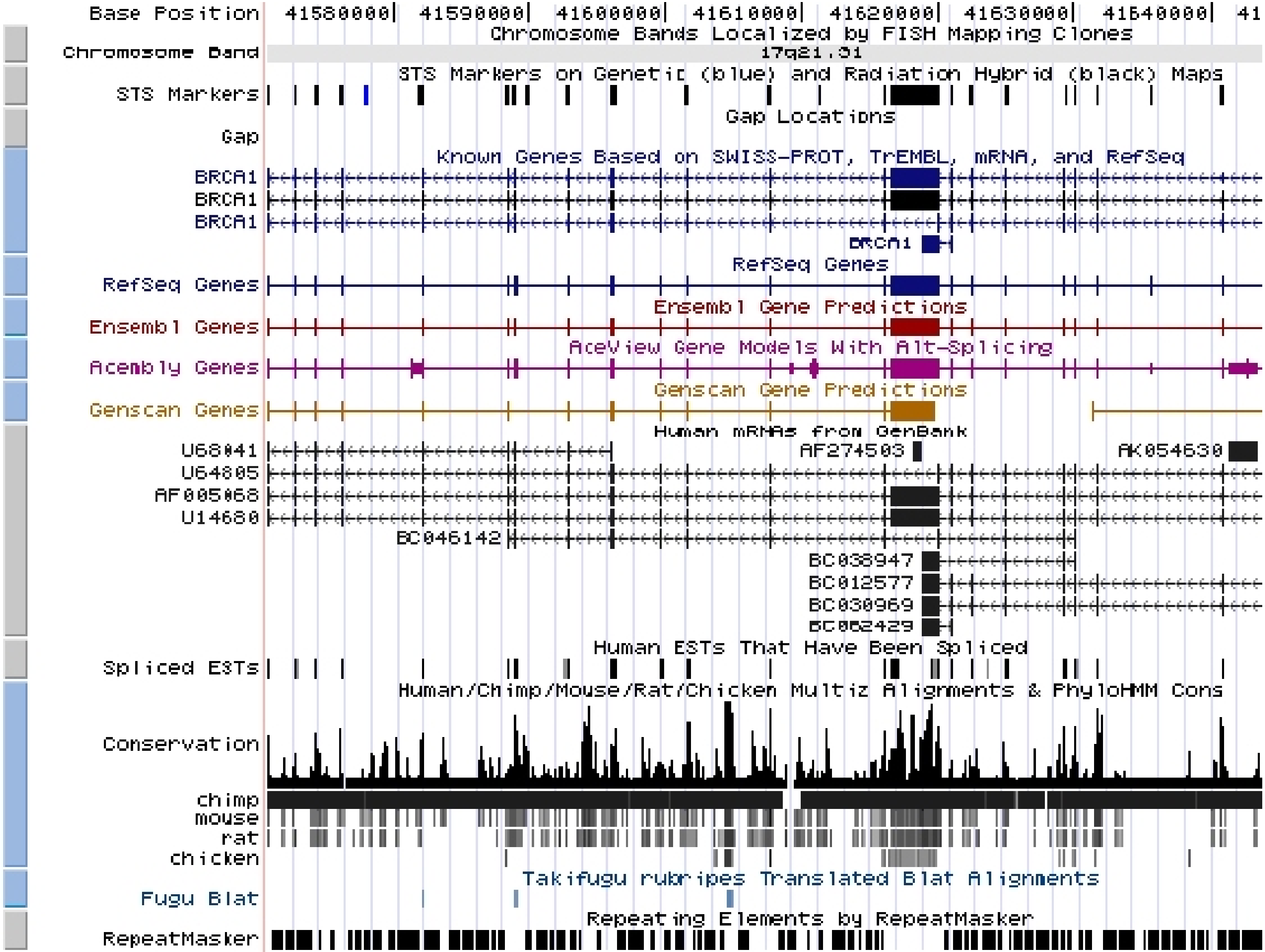
   SEQUENCE ANALYSIS

2. Analyze biological structures, searching for clues about sequence, function and control.

   STRUCTURE ANALYSIS

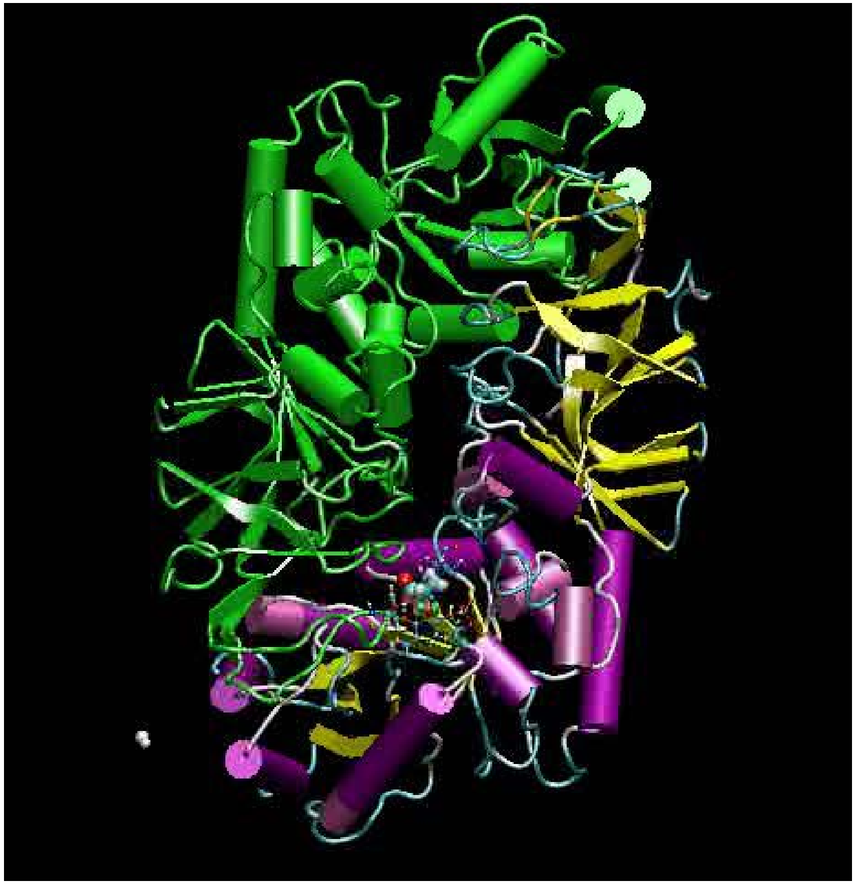
3. Understand how cellular components function in living systems.

   FUNCTION ANALYSIS

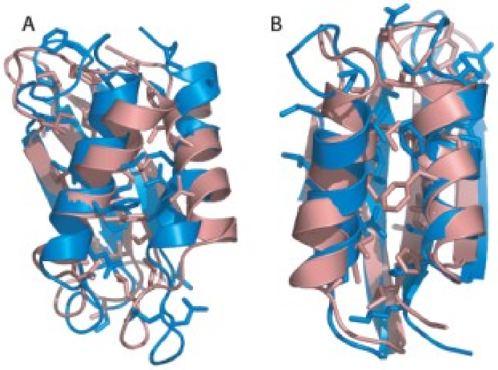


Base Position 41580000 41590000 41600000 41610000 41620000 41630000 41640000 41
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
17q21.31
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
BRCA1
BRCA1
BRCA1
BRCA1
RefSeq Genes
RefSeq Genes
Ensembl Gene Predictions
Ensembl Genes
AceView Gene Models With Alt-Splicing
Acembly Genes
Genscan Gene Predictions
Genscan Genes
Human mRNAs from GenBank
U68041
AF274503
AK054630
U64805
AF005068
U14680
BC046142
BC038947
BC012577
BC030969
BC062429
Human ESTs That Have Been Spliced
Spliced ESTs
Human/Chimp/Mouse/Rat/Chicken Multiz Alignments & PhyloHMM Cons
Conservation
chimp
mouse
rat
chicken
Takifugu rubripes Translated Blat Alignments
Fugu Blat
Repeating Elements by RepeatMasker
RepeatMasker

Structure analysis



Protein dynamics of secondary structure



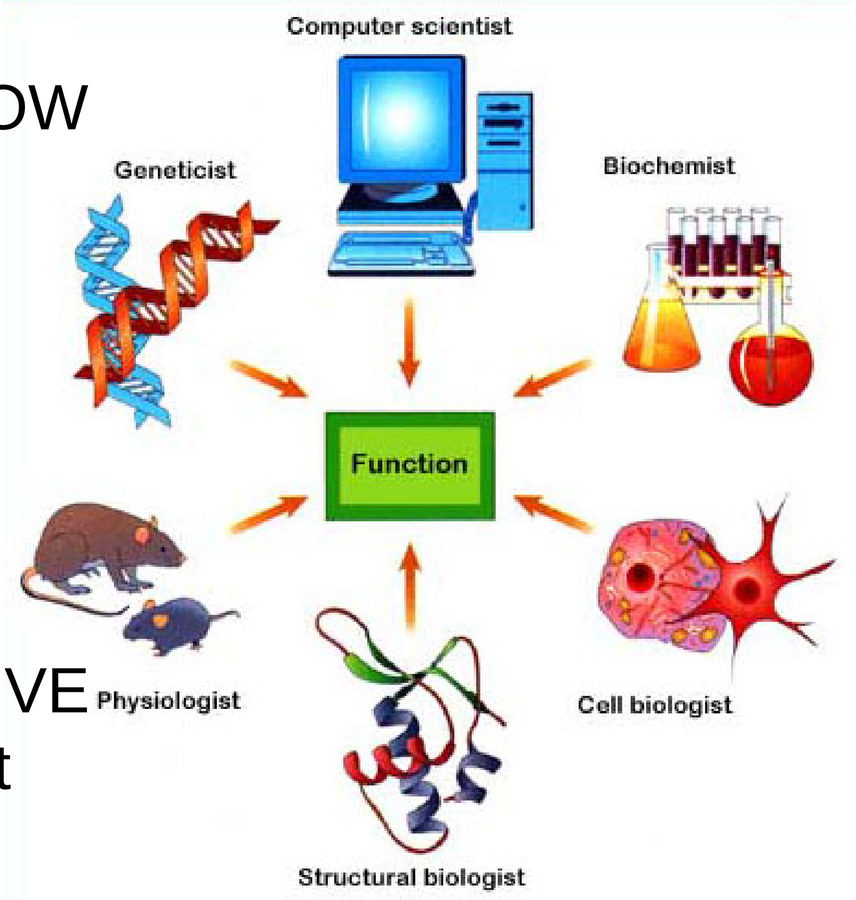
Protein structure prediction

- Protein structure prediction
- Docking
- Small molecule binding
- Molecular dynamics

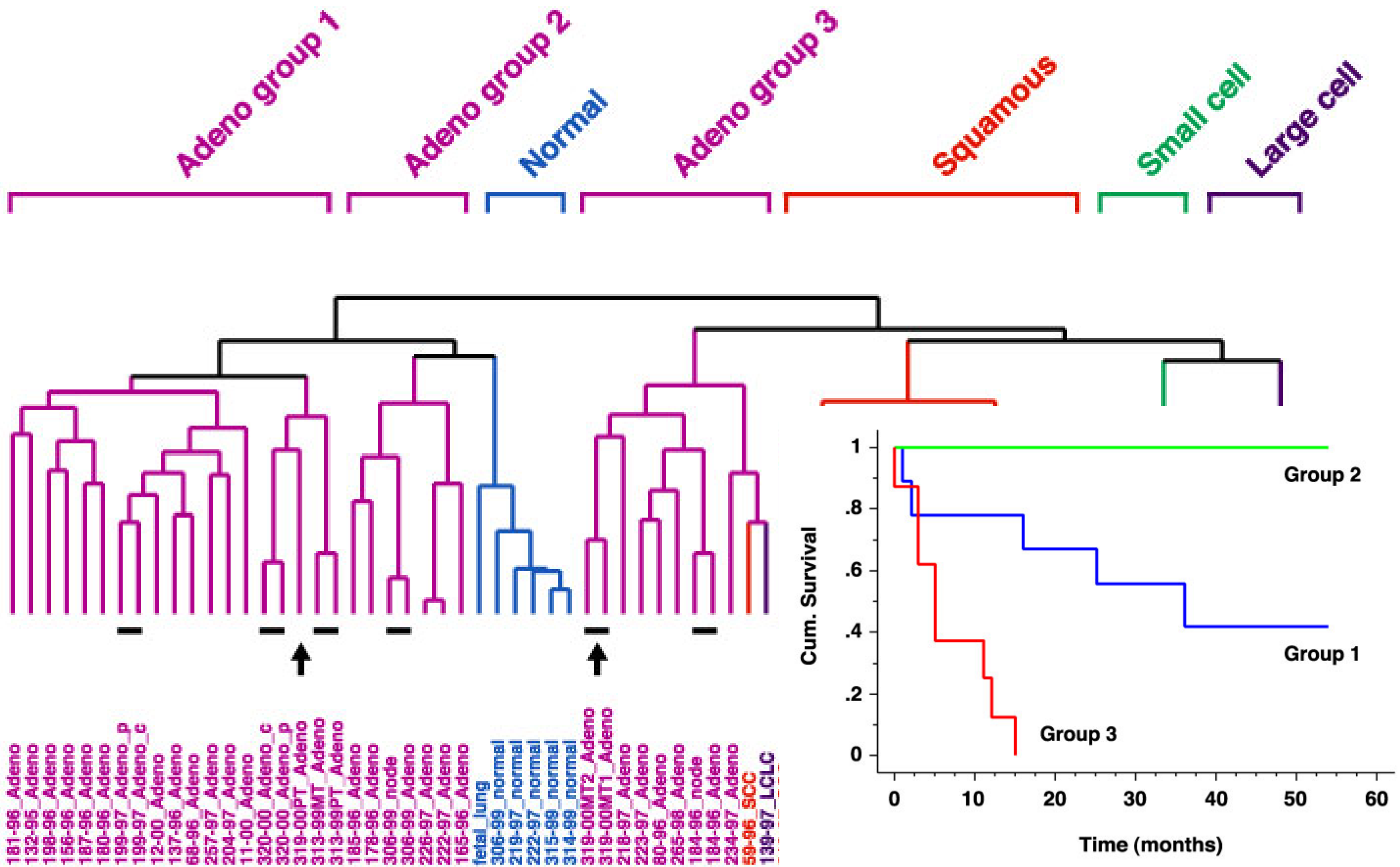
Function analysis

- To study WHAT proteins DO, HOW they INTERACT, and HOW they are REGULATED, need data beyond genomic sequence

- Genomics/Bioinformatics is fundamentally a COLLABORATIVE and MULTIDISCIPLINARY effort

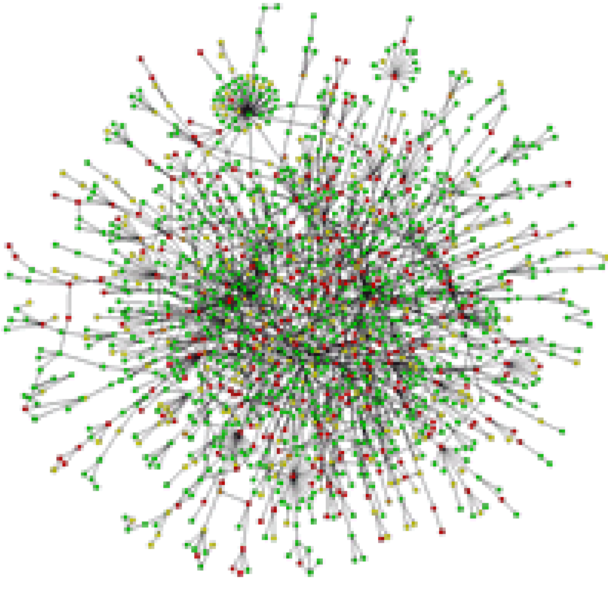


Microarray analysis

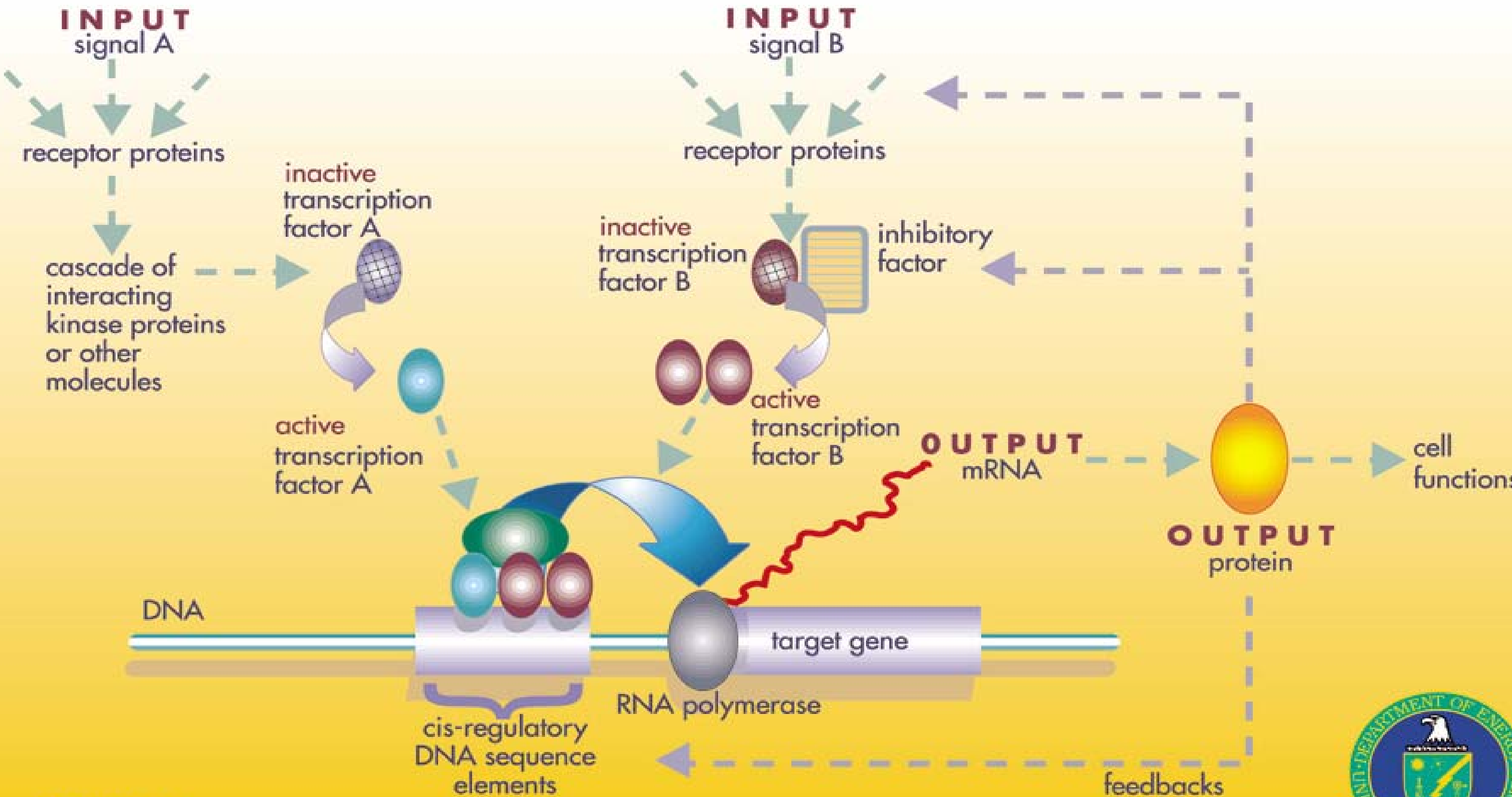


Biological networks

- Interaction maps (no directions)
- Pathway models (dynamic or static)
- Metabolic networks
- Genetic regulatory networks



A GENE REGULATORY NETWORK



YGG 01-0083