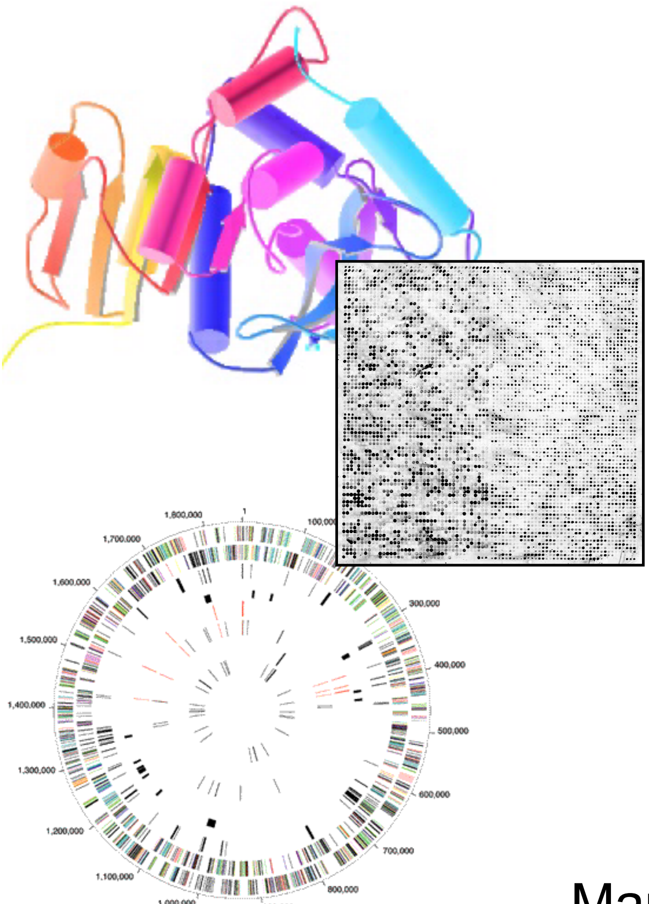
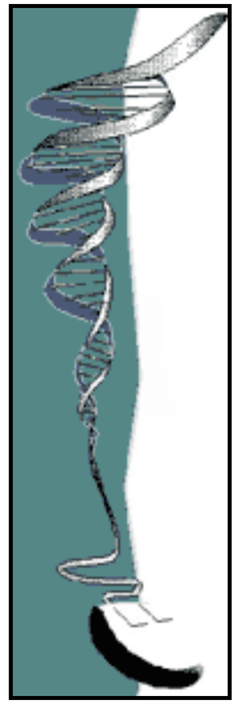
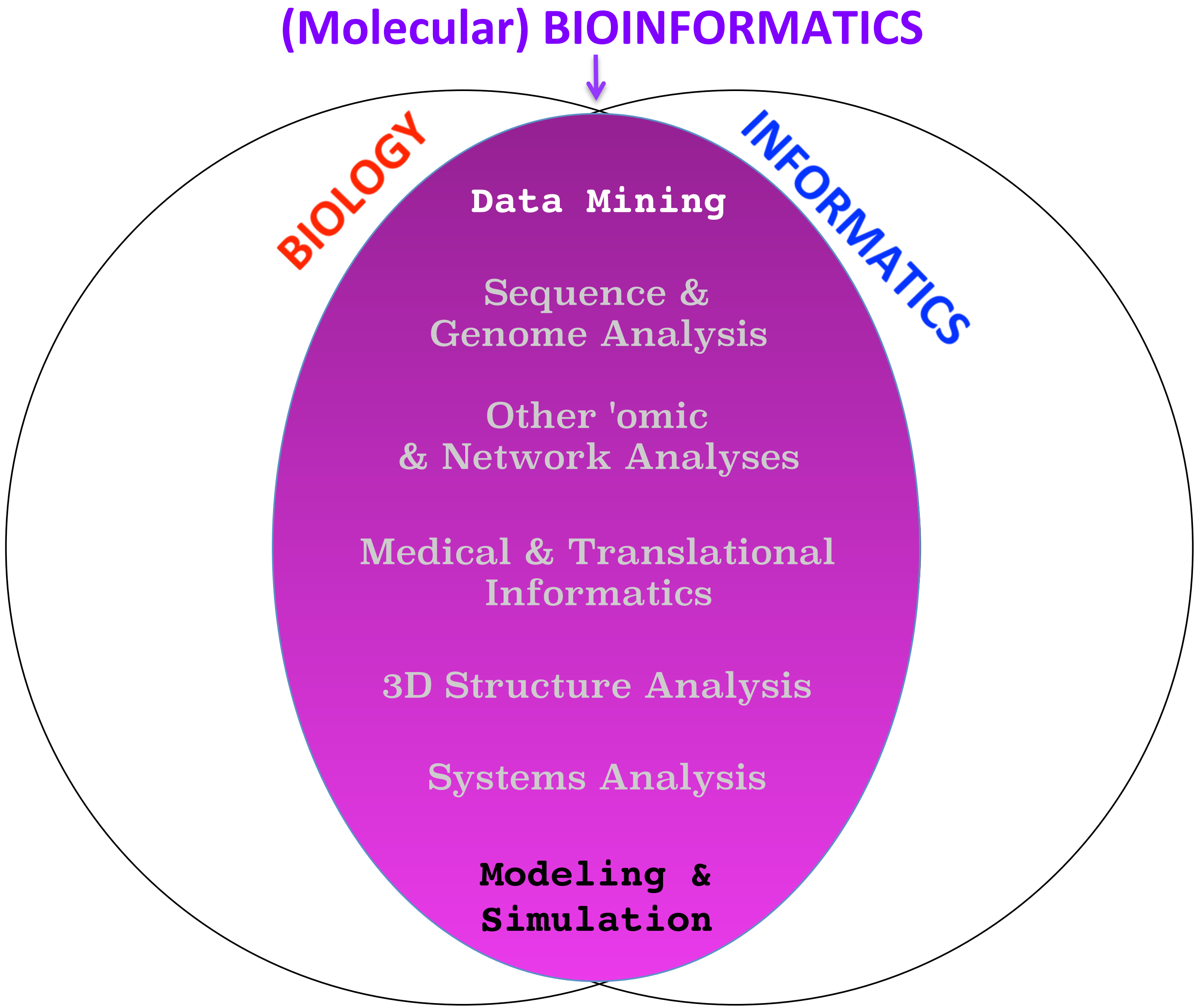
BIOINFORMATICS
Introduction



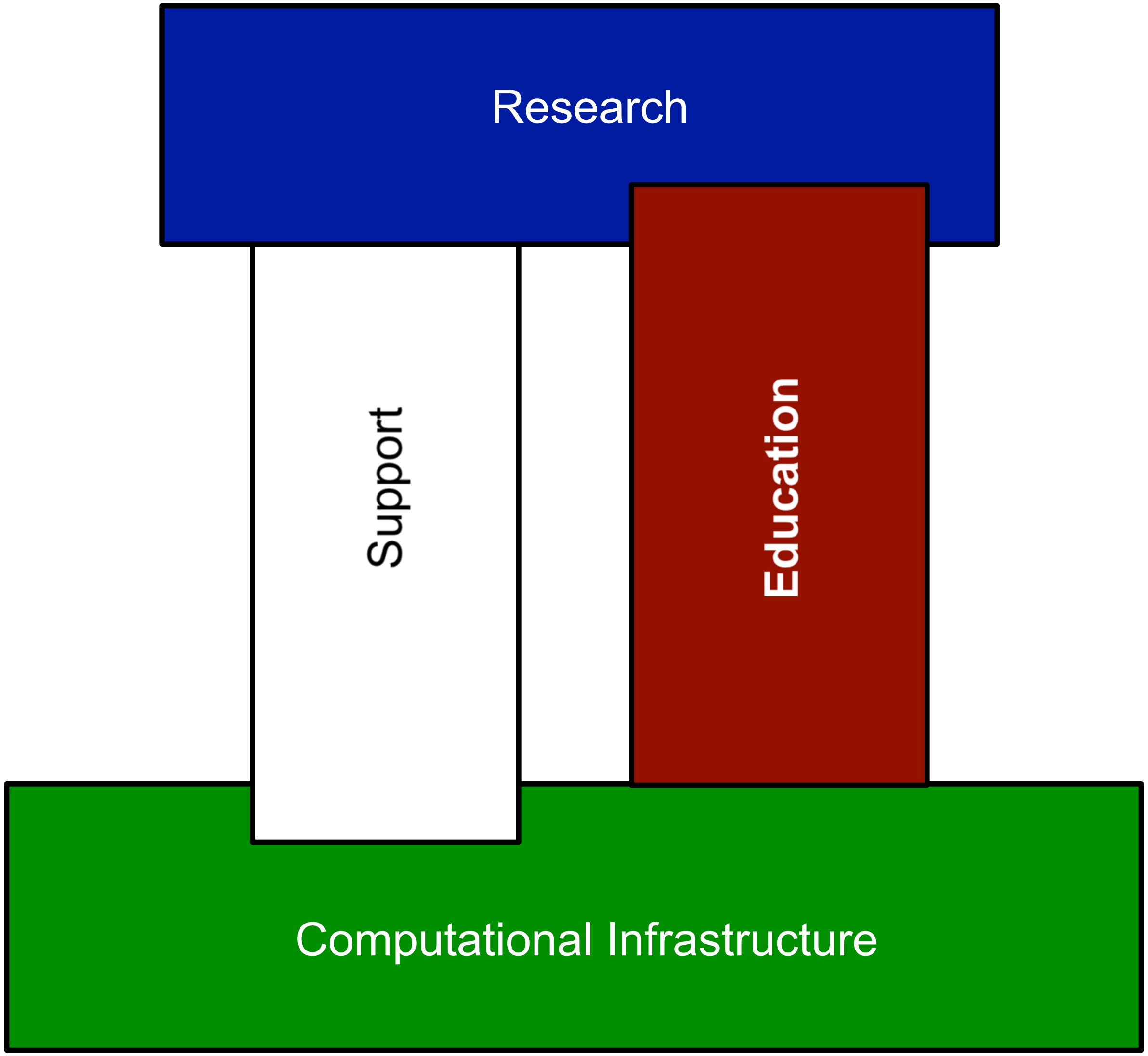


Mark Gerstein, Yale University
GersteinLab.org/courses/452
(last edit in spring ’16)



[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

Elements of Bioinformatics as a discipline



What is Bioinformatics?

- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, mine, model & understand the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

What **Information** to Organize?

- **Sequences** (DNA & Protein)
- 3D Structures
- Network & Pathway Connectivity
- Phylogenetic tree relationships
- Large-scale gene expression & functional genomics data
- Phenotypic data & medical records....

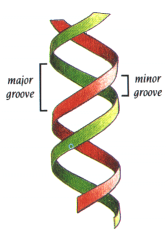
What is the Information?
Molecular Biology as an Information Science

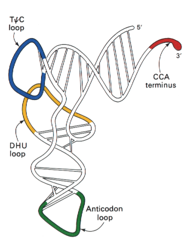
- Central Dogma
  of Molecular Biology

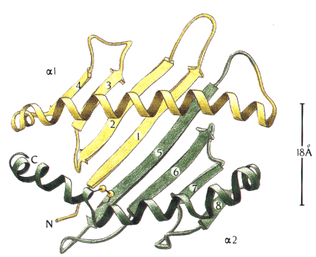
```
DNA
 -> RNA
  -> Protein
   -> Phenotype
    -> DNA
```

- Central Paradigm
  for Bioinformatics

```
Genomic Sequence Information
 -> mRNA (level)
  -> Protein Sequence
   -> Protein Structure
    -> Biological Function
     -> Organismal Phenotype
```







•Genetic material

•Information transfer (mRNA)
•Protein synthesis (tRNA/mRNA)
•Some catalytic activity

Molecular Biology Information - DNA

- Raw DNA Sequence
  - ◊ Coding or Not?
  - ◊ Parse into genes?
  - ◊ 4 bases: AGCT
  - ◊ ~1 K in a gene, ~2 M in genome
  - ◊ ~3 Gb Human

```
atggcaattaaaattggtatcaatggttttggtcgtatcggccgtatcgtattccgtgca
gcacaacaccgtgatgacattgaagttgtaggtattaacgacttaatcgacgttgaatac
atggcttatatgttgaaatatgattcaactcacggtcgtttcgacggcactgttgaagtg
aaagatggtaacttagtggttaatggtaaaactatccgtgtaactgcagaacgtgatcca
gcaaacttaaactggggtgcaatcggtgttgatatcgctgttgaagcgactggtttattc
ttaactgatgaaactgctcgtaaacatatcactgcaggcgcaaaaaaagttgtattaact
ggcccatctaaagatgcaacccctatgttcgttcgtggtgtaaacttcaacgcatacgca
ggtcaagatatcgtttctaacgcatcttgtacaacaaactgtttagctcctttagcacgt
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
gcaactcaaaaaactgtggatggtccatcagctaaagactggcgcggcggccgcggtgca
tcacaaaacatcattccatcttcaacaggtgcagcgaaagcagtaggtaaagtattacct
gcattaaacggtaaattaactggtatggctttccgtgttccaacgccaaacgtatctgtt
gttgatttaacagttaatcttgaaaaaccagcttcttatgatgcaatcaaacaagcaatc
aaagatgcagcggaaggtaaaacgttcaatggcgaattaaaaggcgtattaggttacact
gaagatgctgttgtttctactgacttcaacggttgtgctttaacttctgtatttgatgca
gacgctggtatcgcattaactgattctttcgttaaattggtatc . . .

. . .   caaaaatagggttaatatgaatctcgatctccattttgttcatcgtattcaa
caacaagccaaaactcgtacaaatatgaccgcacttcgctataaagaacacggcttgtgg
cgagatatctcttggaaaaactttcaagagcaactcaatcaactttctcgagcattgctt
gctcacaatattgacgtacaagataaaatcgccatttttgcccataatatggaacgttgg
gttgttcatgaaactttcggtatcaaagatggtttaatgaccactgttcacgcaacgact
acaatcgttgacattgcgaccttacaaattcgagcaatcacagtgcctatttacgcaacc
aatacagcccagcaagcagaatttatcctaaatcacgccgatgtaaaaattctcttcgtc
ggcgatcaagagcaatacgatcaaacattggaaattgctcatcattgtccaaaattacaa
aaaattgtagcaatgaaatccaccattcaattacaacaagatcctctttcttgcacttgg
```

Molecular Biology Information: Protein Sequence

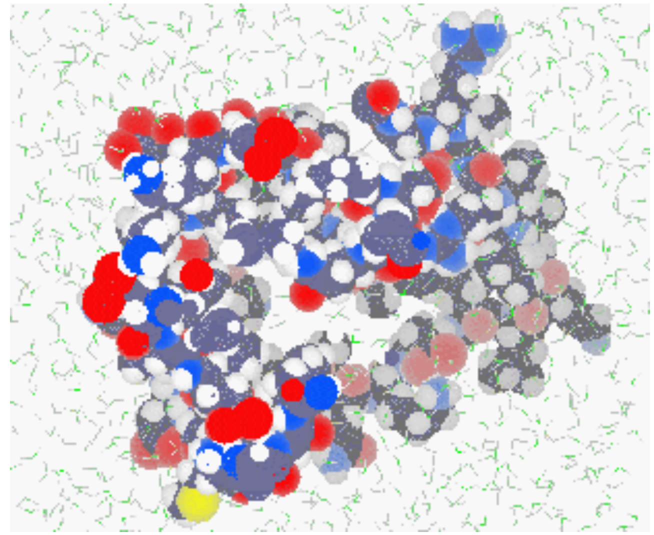
- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- >12 M known protein sequences (uniprot, http://www.ebi.ac.uk/uniprot/TrEMBLstats/, 2011)

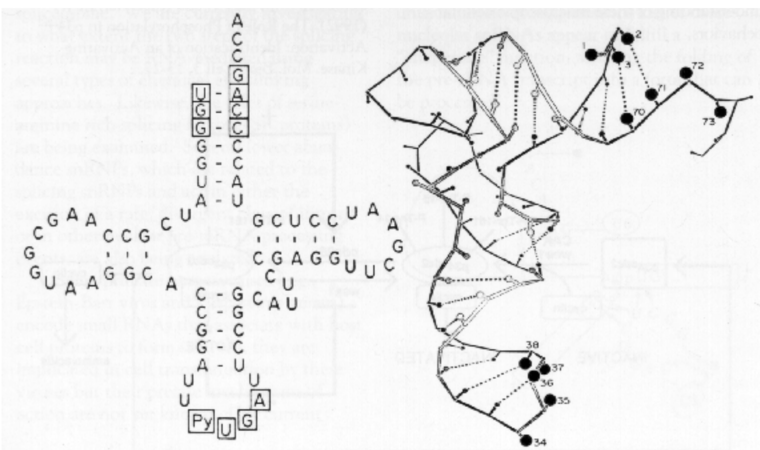
```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
d3dfr__ TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF

d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP------EIMVIGGGRVYEQFLPKA
d3dfr__ ---PKRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV

d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE----- IMVIGGGRVYEQFLPKA
d3dfr__ -P--KRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV
```

Molecular Biology Information: Macromolecular Structure

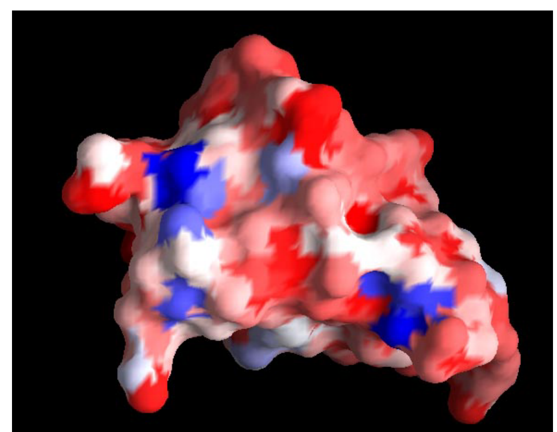
- DNA/RNA/Protein
  - ◊ Almost all protein

(RNA Adapted From D Soll Web Page,
Right Hand Top Protein from M Levitt web page)





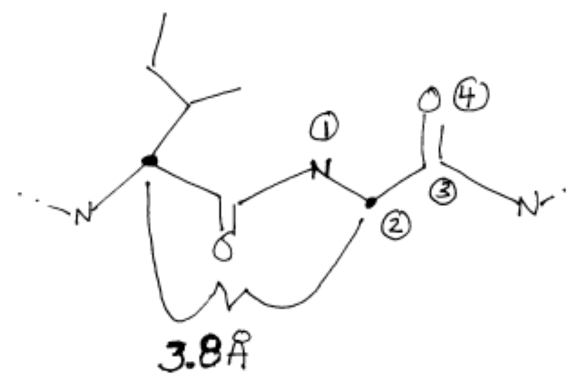
'Identity elements' in *Escherichia coli* glutamine tRNA.



Molecular Biology Information: Protein Structure Details

- Statistics on Number of XYZ triplets
  - ◊ 200 residues/domain -> 200 CA atoms, separated by 3.8 A
  - ◊ Avg. Residue is Leu: 4 backbone atoms + 4 sidechain atoms, 150 cubic A
    - => ~1500 xyz triplets (=8x200) per protein domain
  - ◊ >110K Domains, ~1200 folds (scop 1.75)

```
ATOM      1  C   ACE     0       9.401  30.166  60.595  1.00 49.88      1GKY  67
ATOM      2  O   ACE     0      10.432  30.832  60.722  1.00 50.35      1GKY  68
ATOM      3  CH3 ACE     0       8.876  29.767  59.226  1.00 50.04      1GKY  69
ATOM      4  N   SER     1       8.753  29.755  61.685  1.00 49.13      1GKY  70
ATOM      5  CA  SER     1       9.242  30.200  62.974  1.00 46.62      1GKY  71
ATOM      6  C   SER     1      10.453  29.500  63.579  1.00 41.99      1GKY  72
ATOM      7  O   SER     1      10.593  29.607  64.814  1.00 43.24      1GKY  73
ATOM      8  CB  SER     1       8.052  30.189  63.974  1.00 53.00      1GKY  74
ATOM      9  OG  SER     1       7.294  31.409  63.930  1.00 57.79      1GKY  75
ATOM     10  N   ARG     2      11.360  28.819  62.827  1.00 36.48      1GKY  76
ATOM     11  CA  ARG     2      12.548  28.316  63.532  1.00 30.20      1GKY  77
ATOM     12  C   ARG     2      13.502  29.501  63.500  1.00 25.54      1GKY  78
. . .
ATOM   1444  CB  LYS   186      13.836  22.263  57.567  1.00 55.06      1GKY1510
ATOM   1445  CG  LYS   186      12.422  22.452  58.180  1.00 53.45      1GKY1511
ATOM   1446  CD  LYS   186      11.531  21.198  58.185  1.00 49.88      1GKY1512
ATOM   1447  CE  LYS   186      11.452  20.402  56.860  1.00 48.15      1GKY1513
ATOM   1448  NZ  LYS   186      10.735  21.104  55.811  1.00 48.41      1GKY1514
ATOM   1449  OXT LYS   186      16.887  23.841  56.647  1.00 62.94      1GKY1515
TER    1450      LYS   186                                              1GKY1516
```



Molecular Biology Information: Whole Genomes

- The Revolution Driving Everything

  Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. & Venter, J. C. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae rd." *Science* 269: 496-512.

  (Picture adapted from TIGR website, http://www.tigr.org)

- Timeline

  1995, HI (bacteria): 1.6 Mb & 1600 genes done

  1997, yeast: 13 Mb & ~6000 genes for yeast

  1998, worm: ~100Mb with 19 K genes

  1999: >30 completed genomes!

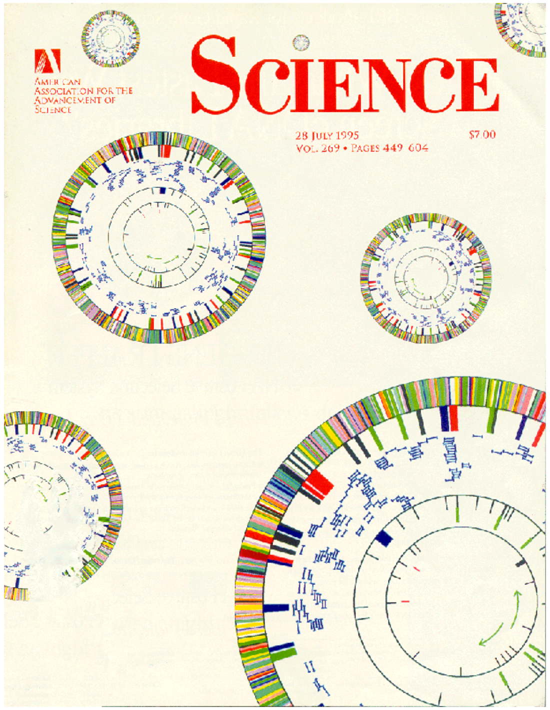
  2000, draft human

  2003, human: 3 Gb & 100 K genes...

  2010, 1000 human genomes!

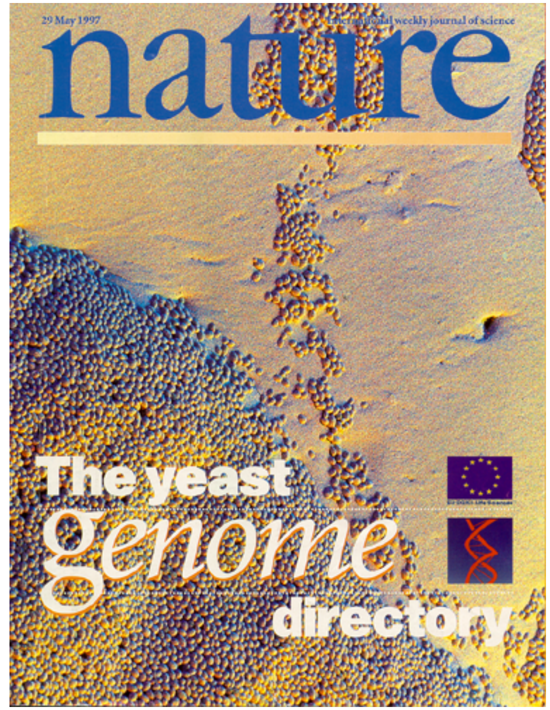
1995

Bacteria,
1.6 Mb,
~1600 genes
[*Science* **269**: 496]



1997

Eukaryote,
13 Mb,
~6K genes
[Nature **387**: 1]



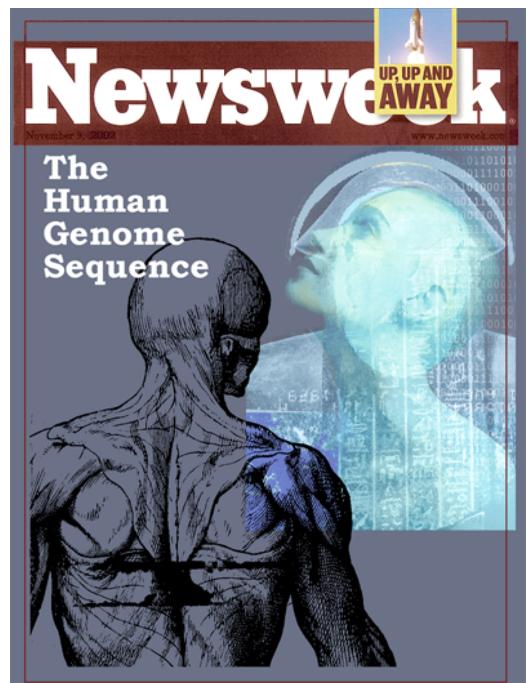
A Bioinformatics prediction that came true!

1998

Animal,
~100 Mb,
~20K genes
[*Science* **282**: 1945]

2000?

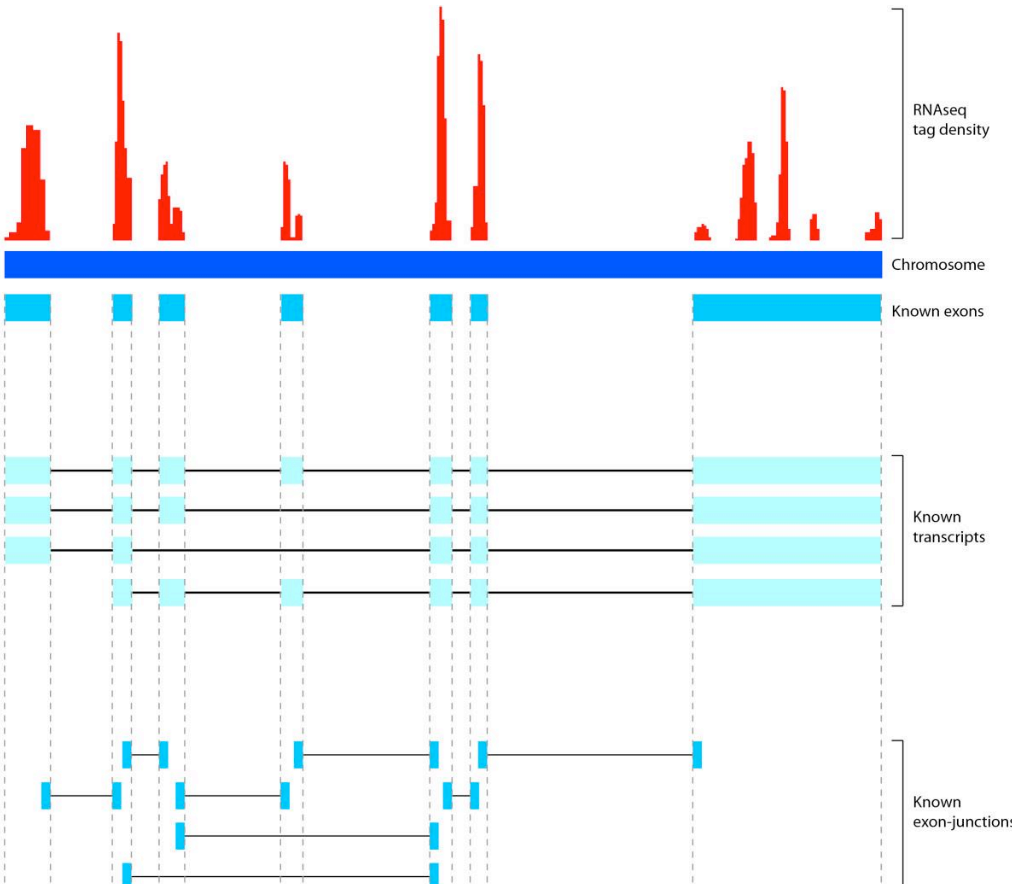
Human,
~3 Gb,
~20K genes



'98 spoof

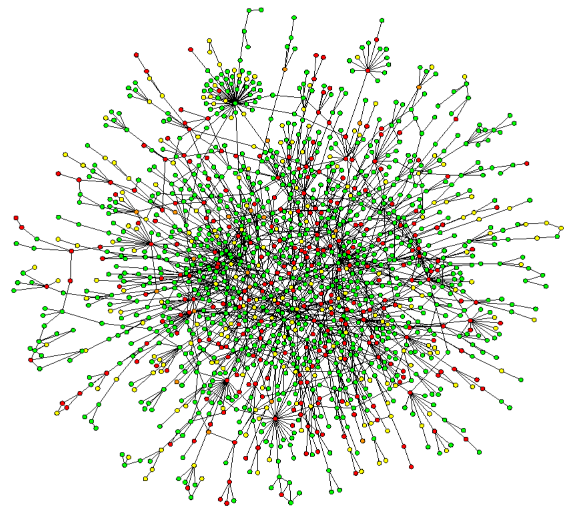
real thing, Apr '00

Gene Expression Data: On & Off

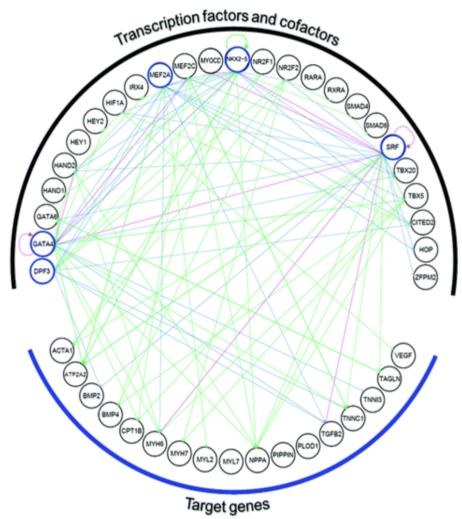


- Early experiments yeast
  - ◊ Complexity at 10 time points,
    6000 x 10 = 60K floats
- Then tiling array technology
  - ◊ 50 M data points to tile the human genome at ~50 bp res.
- Now Next-Gen Sequencing (RNAseq)
  - ◊ 10M+ reads on the human genome, counts
- Can only sequence genome once but can do an infinite variety of expression experiments

Molecular Networks: Connectivity



Protein-protein Interaction networks



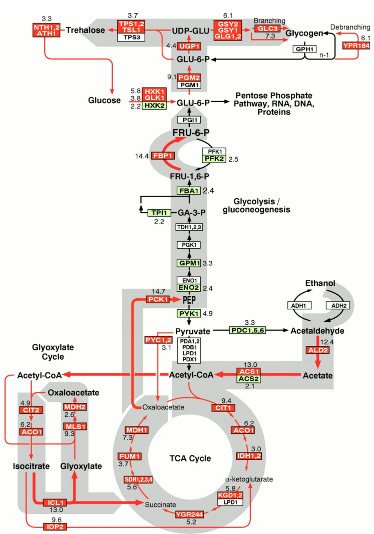
TF-target-gene Regulatory networks

Regulatory Networks

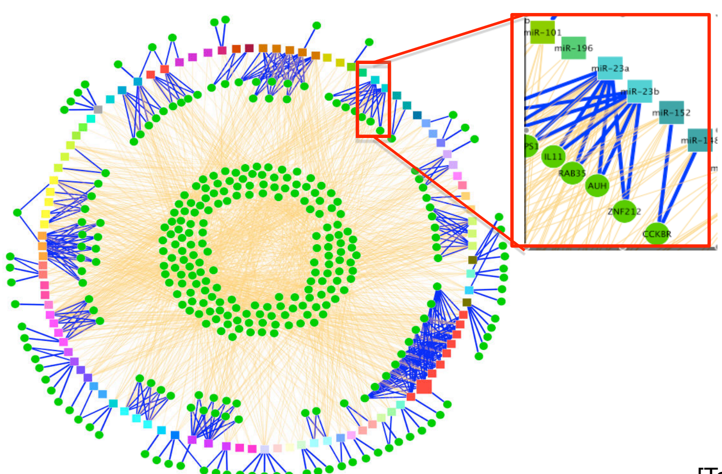
Get readouts of where proteins bind to DNA : Chip-chip then chip-seq

Protein Interaction Networks

For yeast: 6000 x 6000 / 2 ~ 18M possible interactions (maybe ~30K real)



Metabolic pathway networks



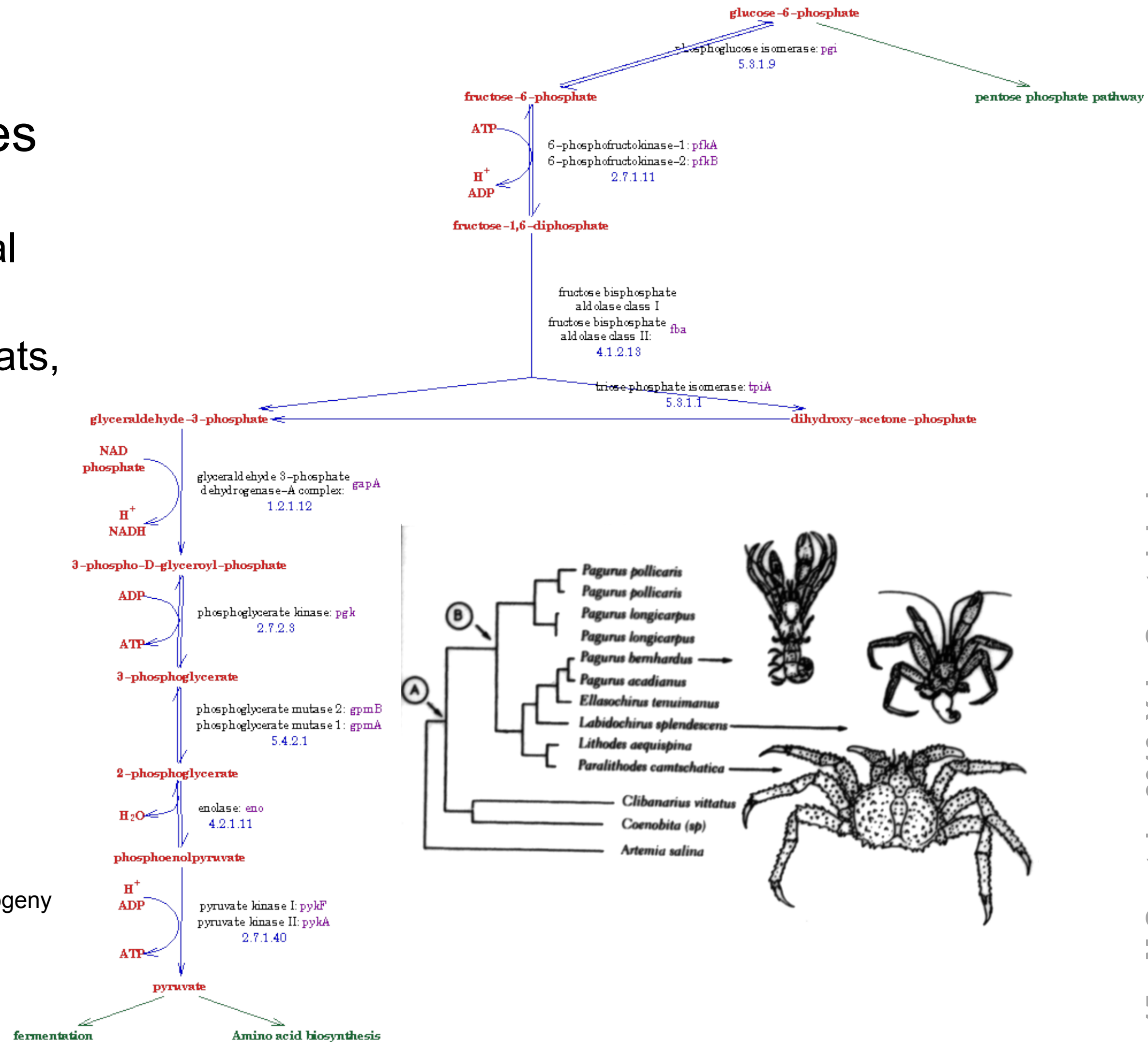
miRNA-target networks

[Toenjes, *et al*, *Mol. BioSyst.* (2008); Jeong *et al*, *Nature* (2001); [Horak, et al, Genes & Development, 16:3017-3033; DeRisi, Iyer, and Brown, Science, 278:680-686]

Molecular Biology Information: Other Integrative Data

- Information to understand genomes
  - ◊ Whole Organisms Phylogeny, traditional zoology
  - ◊ Environments, Habitats, ecology
  - ◊ Phenotype Experiments (large-scale KOs, transposons)
  - ◊ The Literature (MEDLINE)
- The Future....

(Pathway drawing from P Karp's EcoCyc, Phylogeny from S J Gould, Dinosaur in a Haystack)



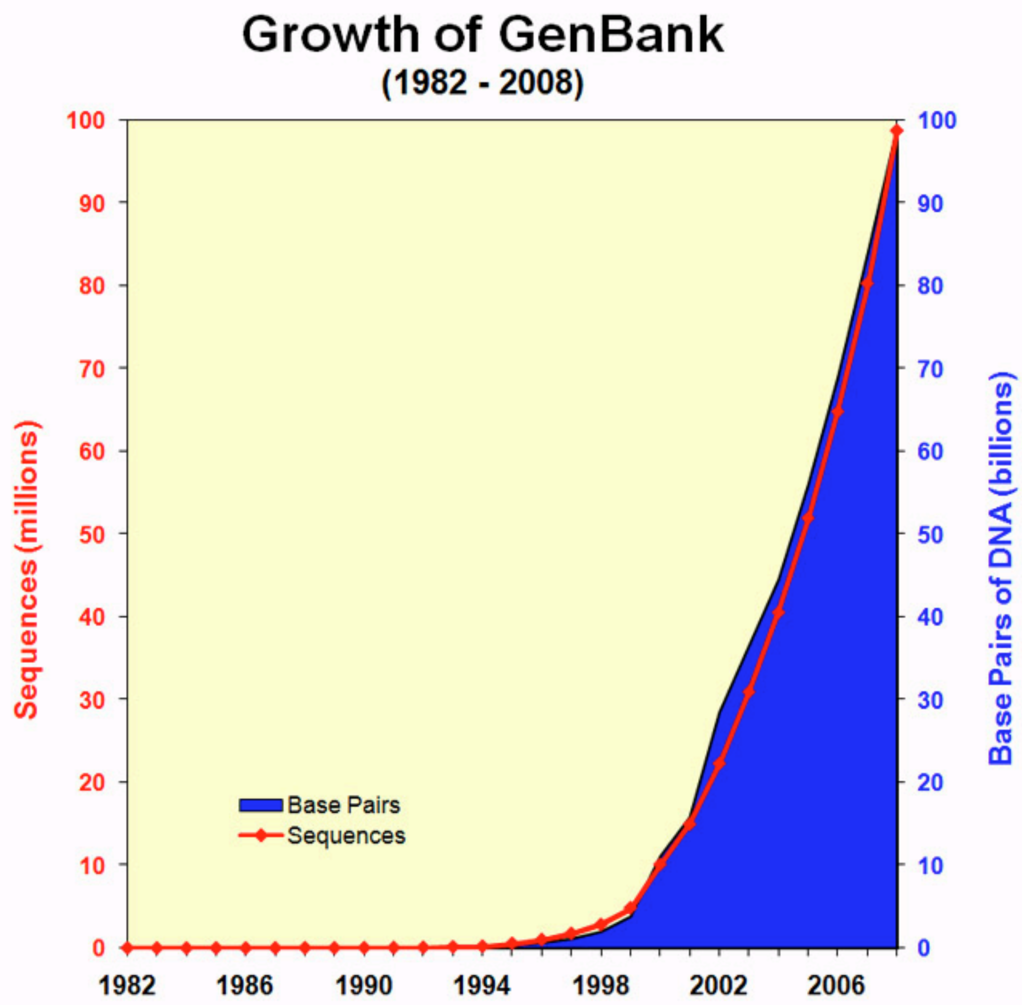
What is Bioinformatics?

- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, mine, model & understand the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

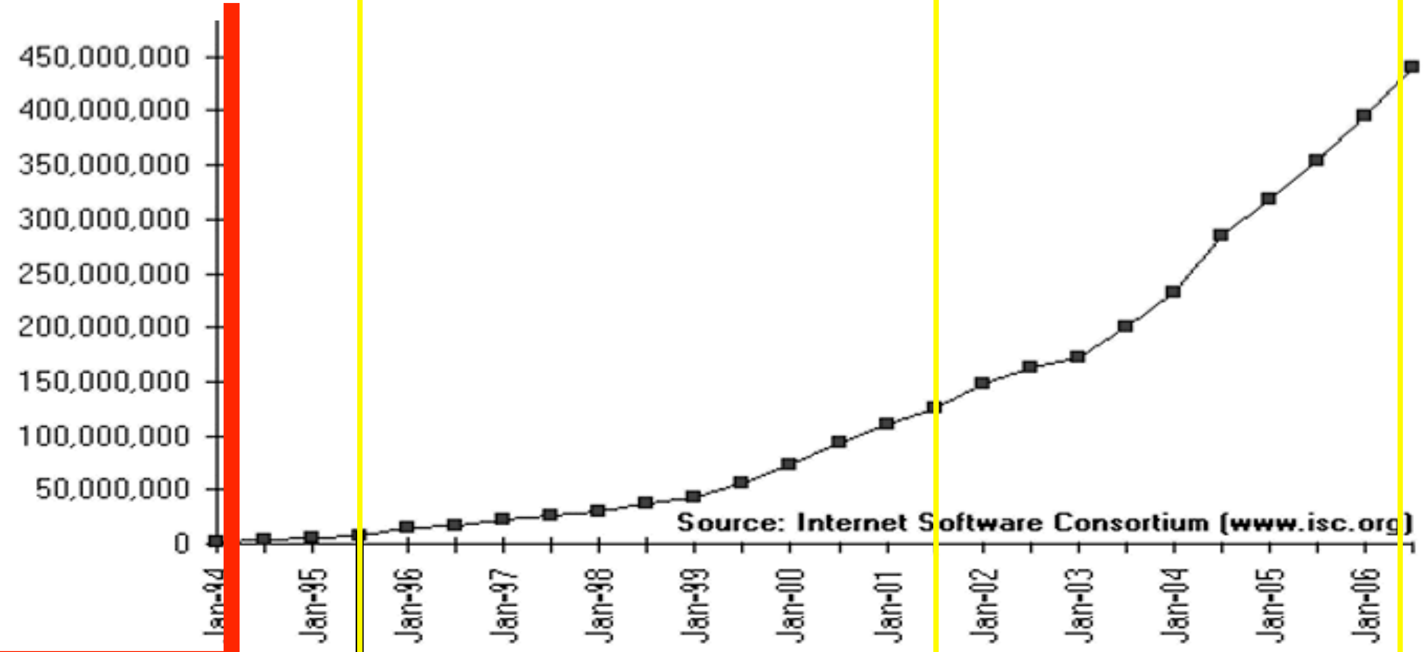
Large-scale Information: Explonential **Scaling** of Data Matched by Development of Computer Technology

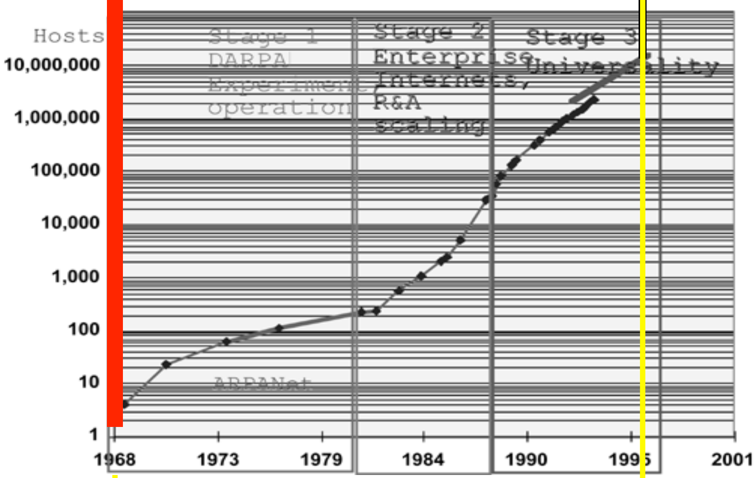
- CPU vs Disk & Net
  - ◊ As important as the increase in computer speed has been, the ability to store large amounts of information on computers is even more crucial
  - ◊ Comparison with **Moore's Law**
- A Driving Force in Bioinformatics



Internet Hosts

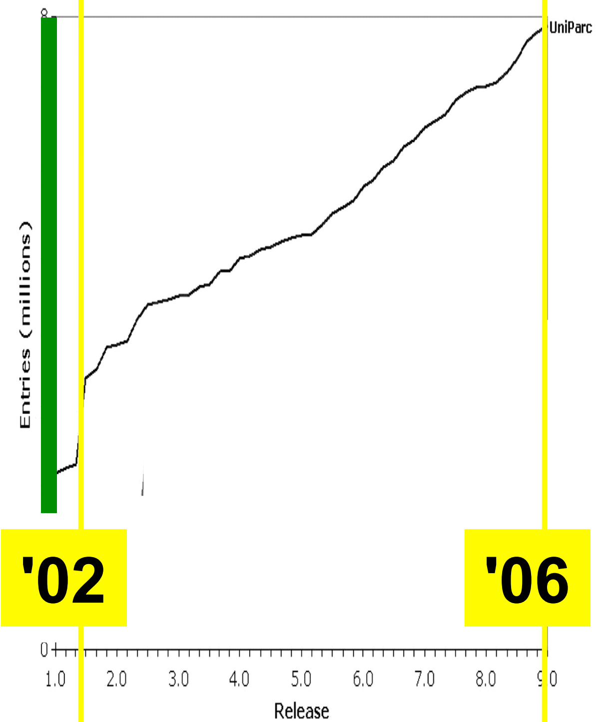
(adapted from D Brutlag, Stanford & http://navigators.com/stats.html)





Proteins

Suzek, B. E. et al. Bioinformatics 2007 23:1282-1288; doi: 10.1093/bioinformatics/btm098



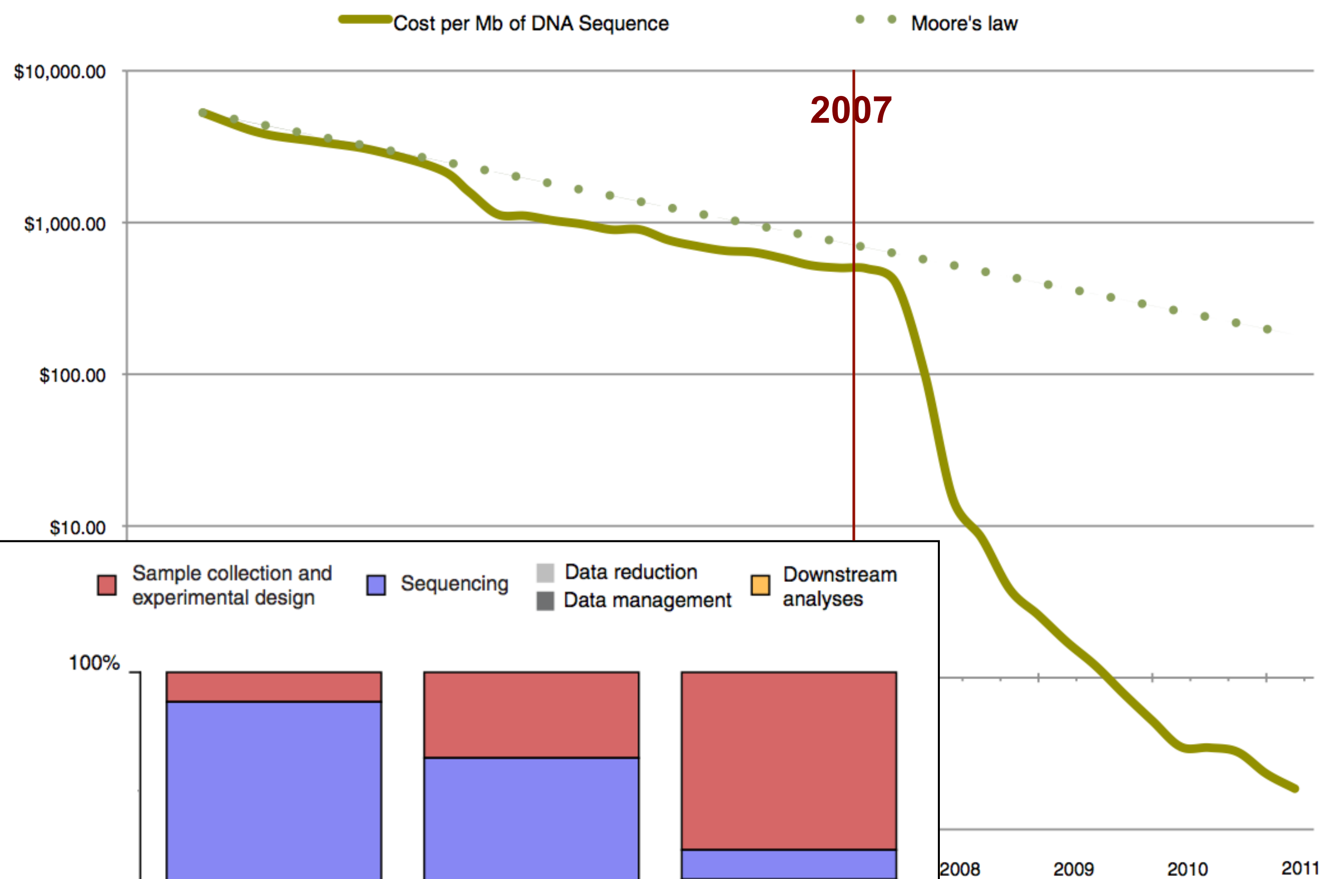
'68

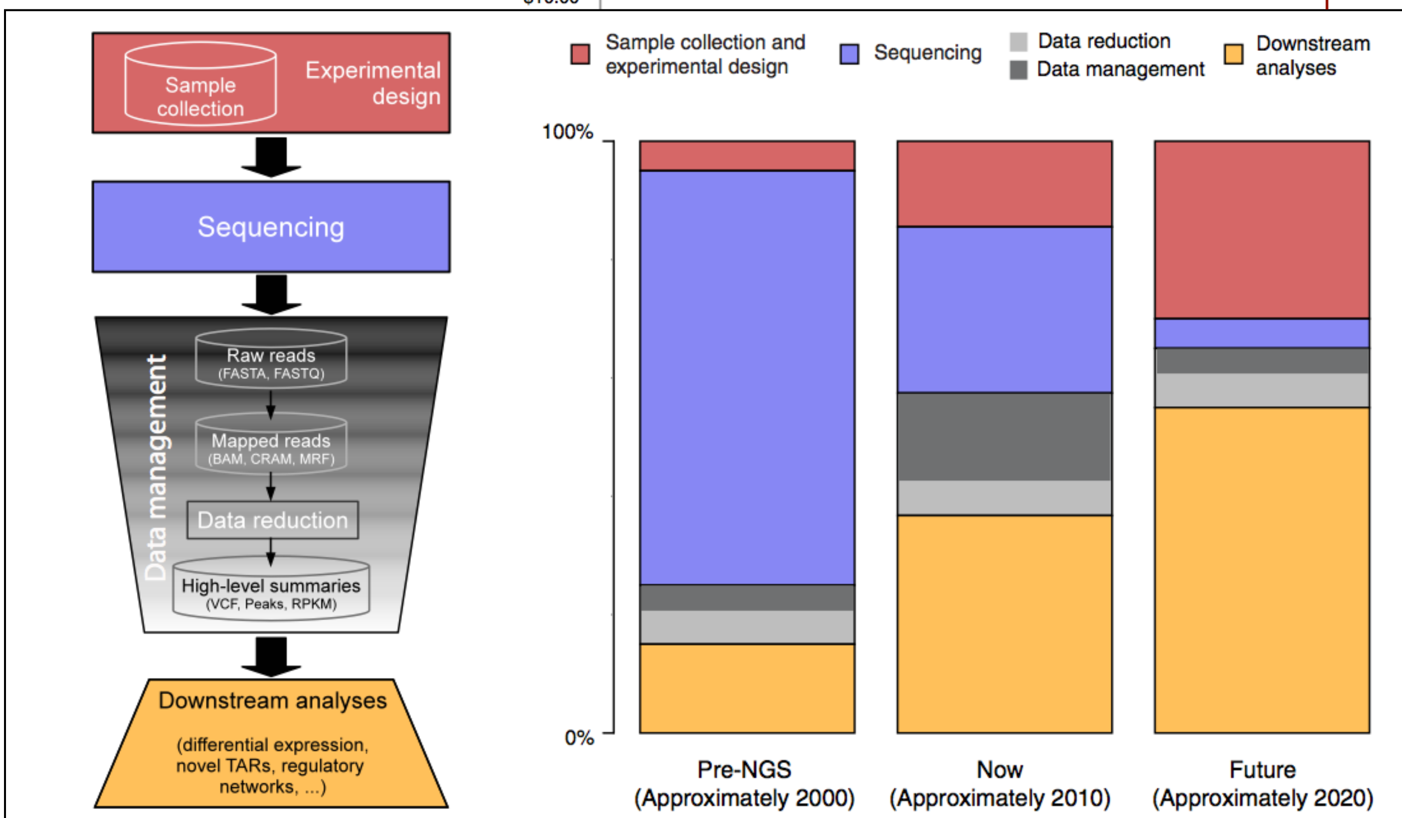
'95

'02

'06

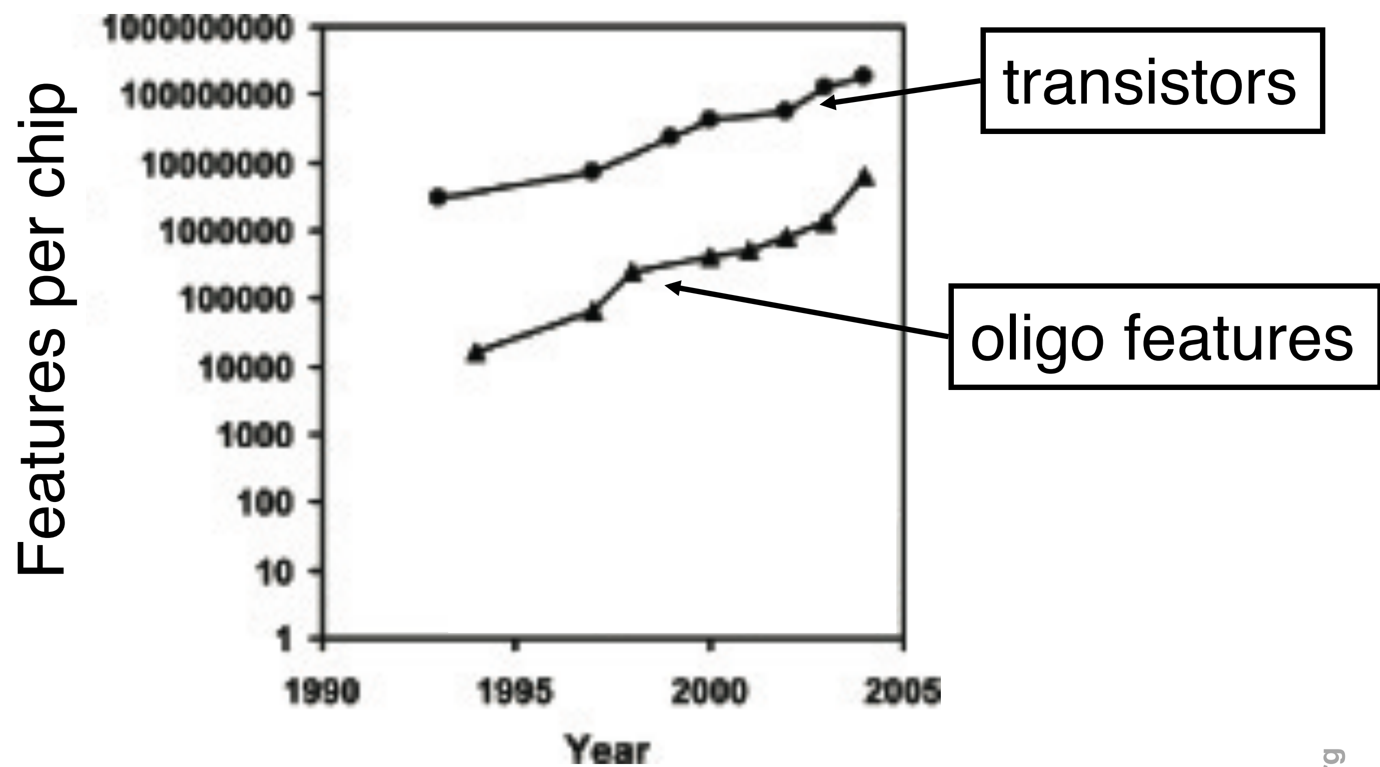
Sequencing Data Explosion: Going to $0/base





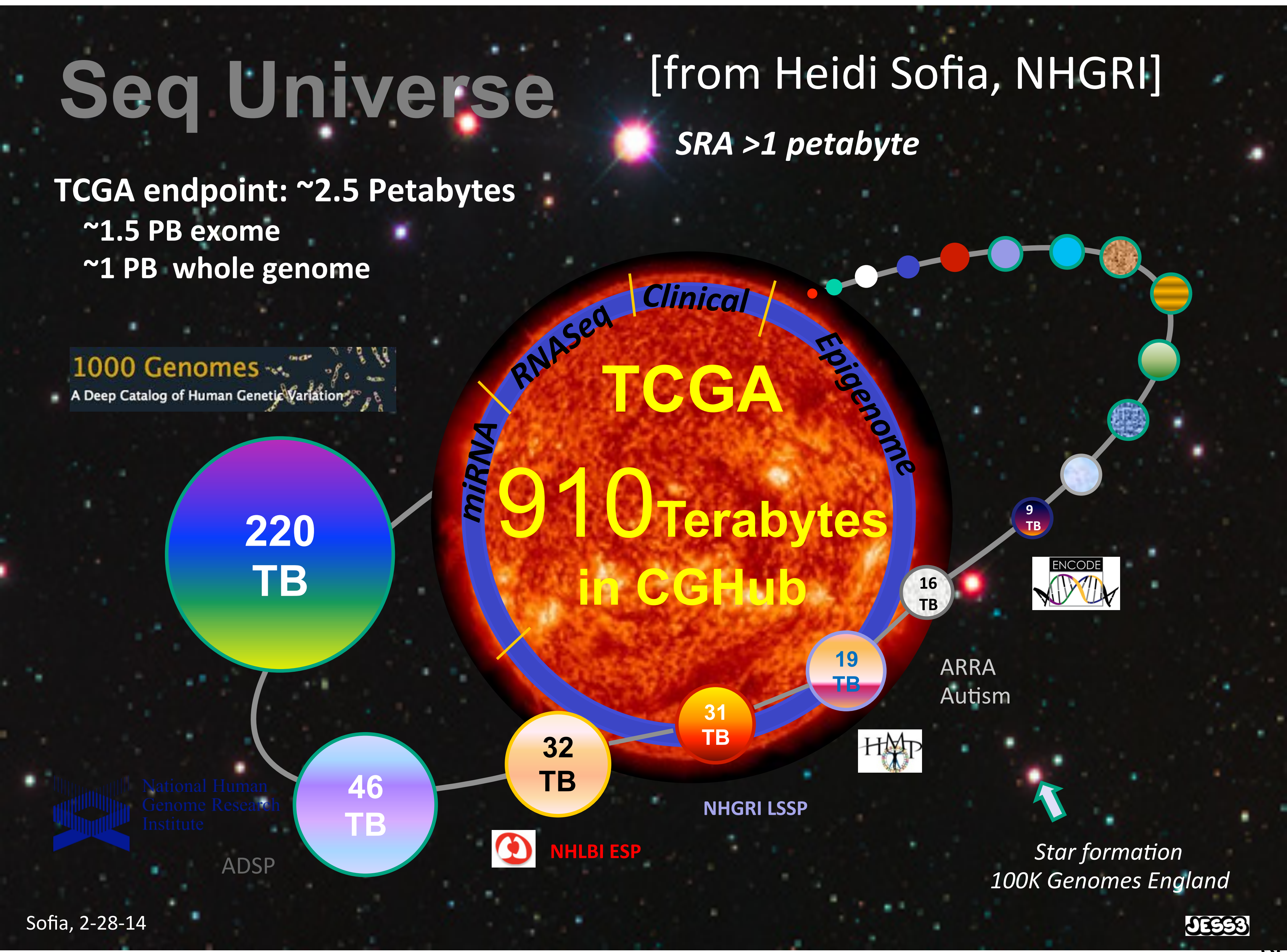
From ‘00 to ~’ 20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11) GenomeBiology]



Features per Slide

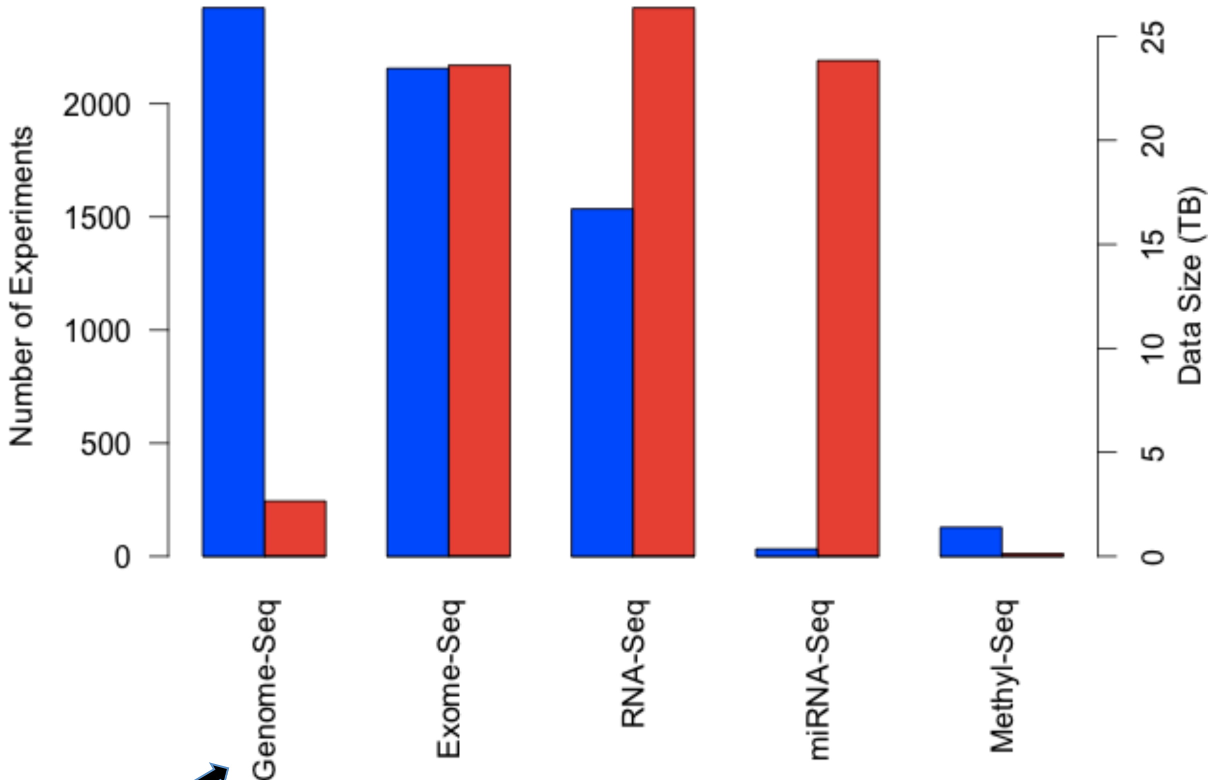
Chip Technology



Seq Universe
[from Heidi Sofia, NHGRI]
SRA >1 petabyte
TCGA endpoint: ~2.5 Petabytes
~1.5 PB exome
~1 PB whole genome
1000 Genomes
A Deep Catalog of Human Genetic Variation
Clinical
RNASeq
Epigenome
miRNA
TCGA
910 Terabytes in CGHub
220 TB
46 TB
32 TB
31 TB
19 TB
16 TB
9 TB
ENCODE
ARRA Autism
HMP
NHGRI LSSP
NHLBI ESP
National Human Genome Research Institute
ADSP
Star formation
100K Genomes England
Sofia, 2-28-14
JESS3

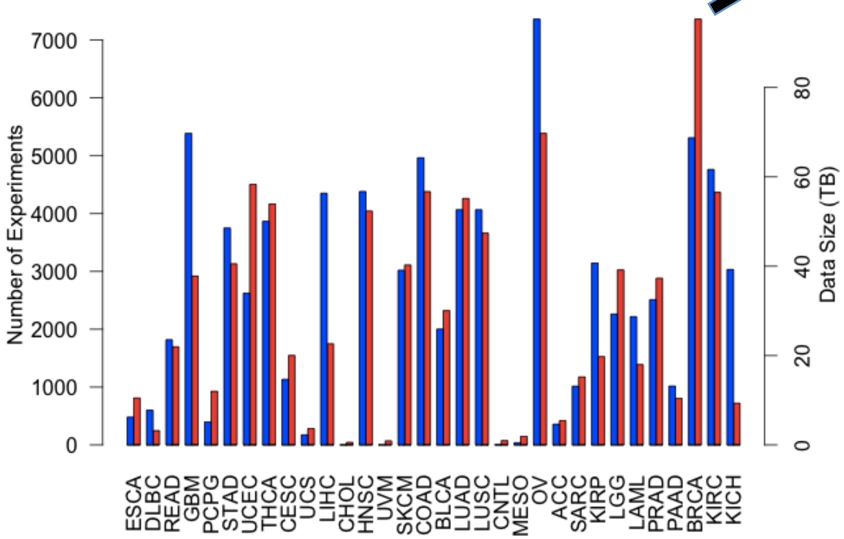
TCGA: What’s in a petabyte?

- >73,000 Expt
- 34 Cancer Types
- ~5,000 Patients



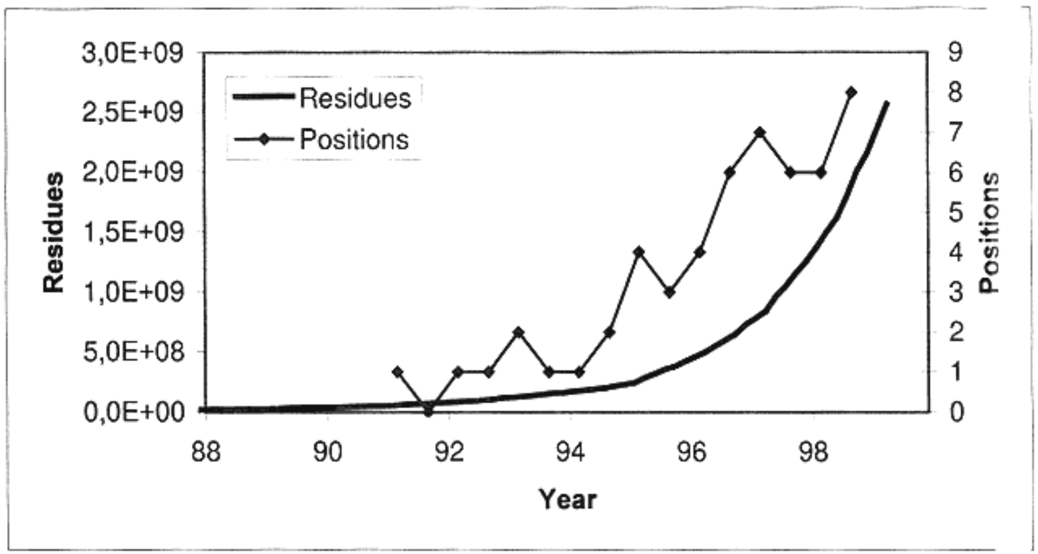
Experiments
Data Size (TB)

TCGA Cancer Types

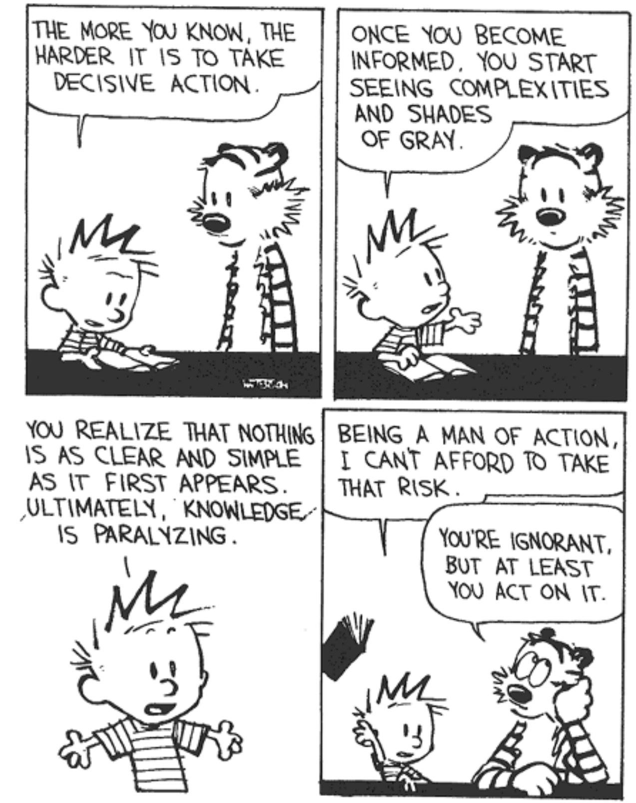


Breast Cancer Expt. Types

Jobs: Bioinformatics is born!



Growth in number of residues in Genbank, a central database for sequence data, compared to the request for people with competence in bioinformatics. The request for scientists is estimated from the number of relevant positions advertised in the first number of Nature in March and September of each year.



B. Watterson, "*There's treasure everywhere*", Andrews and McMeel, 1996.

(courtesy of Finn Drablos)

What is Bioinformatics?

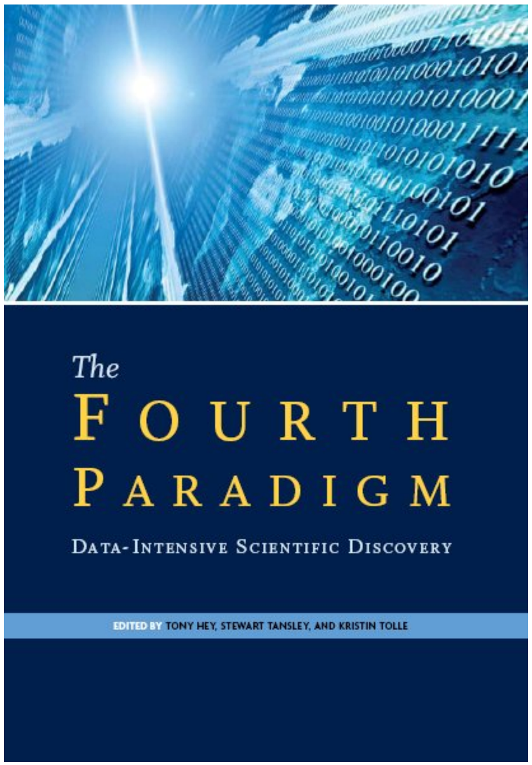
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, mine, model & understand the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

General Types of "Informatics" techniques in Computational Biology

- Databases
  - Building, Querying
  - Representing Complex data
- Data mining
  - Machine Learning techniques
  - Clustering & Tree construction
  - Rapid Text String Comparison & textmining
  - Detailed statistics of significance & association
- Network Analysis
  - Analysis of Topology (eg Hubs)
  - Predicting Connectivity
- Structure Analysis & Geometry
  - Graphics (Surfaces, Volumes)
  - Comparison & 3D Matching (Vision, recognition, docking)
- Physical Modeling
  - Newtonian Mechanics
  - Electrostatics
  - Numerical Algorithms
  - Simulation
  - Modeling Chemical Reactions & Cellular Processes

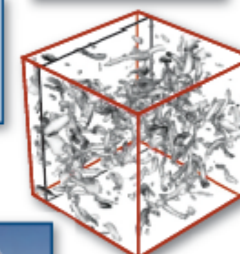
Jim Gray's 4th Paradigm



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - K\frac{c^2}{a^2}$$



- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

Jim Gray's 4th Paradigm

#3 - Simulation

Prediction based on physical principles (eg Exact Determination of Rocket Trajectory)

Emphasis on: Supercomputers

#4 - Data Mining

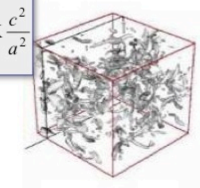
Classifying information & discovering unexpected relationships

Emphasis: networks, "federated" DBs

Science Paradigms

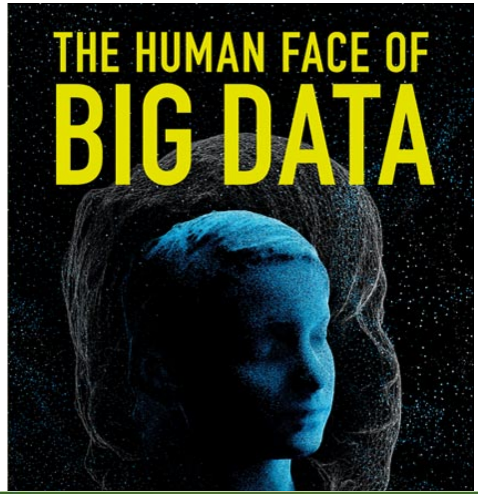
- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
  Or generated by simulator
  – Processed by software
  – Information/Knowledge stored in computer
  – Scientist analyzes database / files
  using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



Gray died in '07.
Book about his ideas came out in '09.....

[Slide from : http://research.microsoft.com/en-us/um/people/gray/talks/stanford%2520symbolic%2520systems%2520seminar.ppt]



Commercial World Data: Financial & Retail Data

Big Data:

a current buzz-word

Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

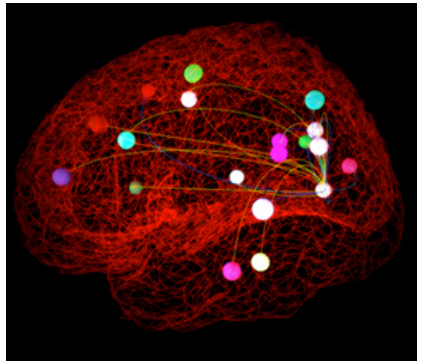
by Thomas H. Davenport and D.J. Patil

Artwork: **Tamar Cohen, *Andrew J Buboltz,*** 2011, silk screen on a page from a high

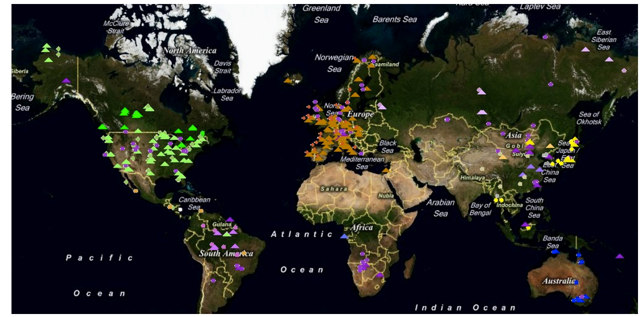
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne
up. The company had just under 8 million accounts, and the number was growing qu
friends and colleagues to join. But users weren't seeking out connections with the pe
rate executives had expected. Something was apparently missing in the social expe

[Oct. '12 issue]

Big data is transforming science



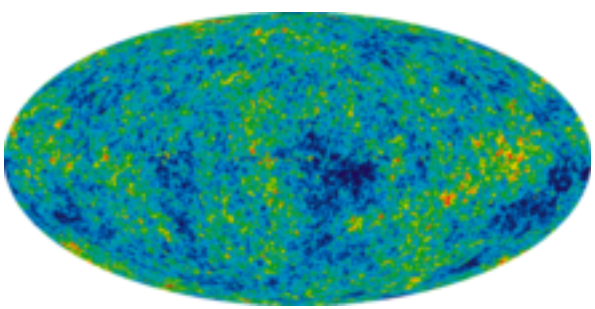
Neuroscience -
The Human Connectome Project



Ecology - Fluxnet

High energy physics -
Large Hadron Collider

Genomics
DNA sequencer



Astronomy -
Sloan Digital Sky survey

ISI Web of KNOWLEDGE
Transforming Research

Knowledge of knowledge
Meta-data of scientific documents

facebook YouTube bebo Technorati myspace.com USTREAM flickr del.icio.us WordPress

Computational social science
Online communities

What do people do with Big Data ?

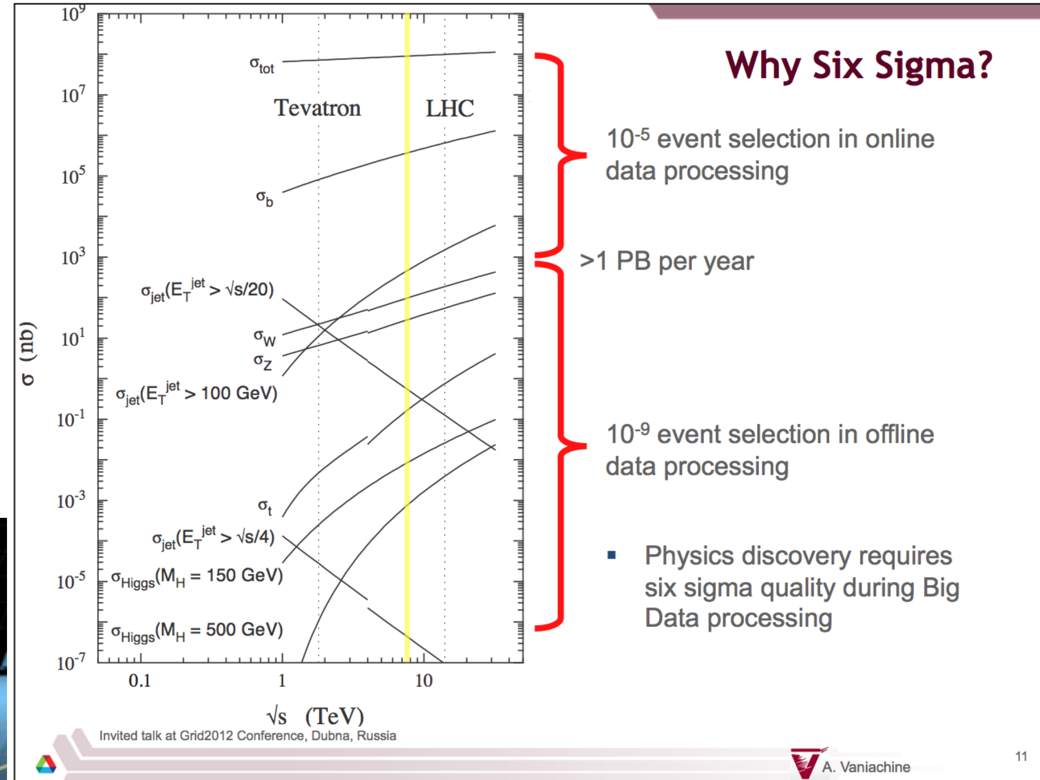
[*Nature* **489**: 208]

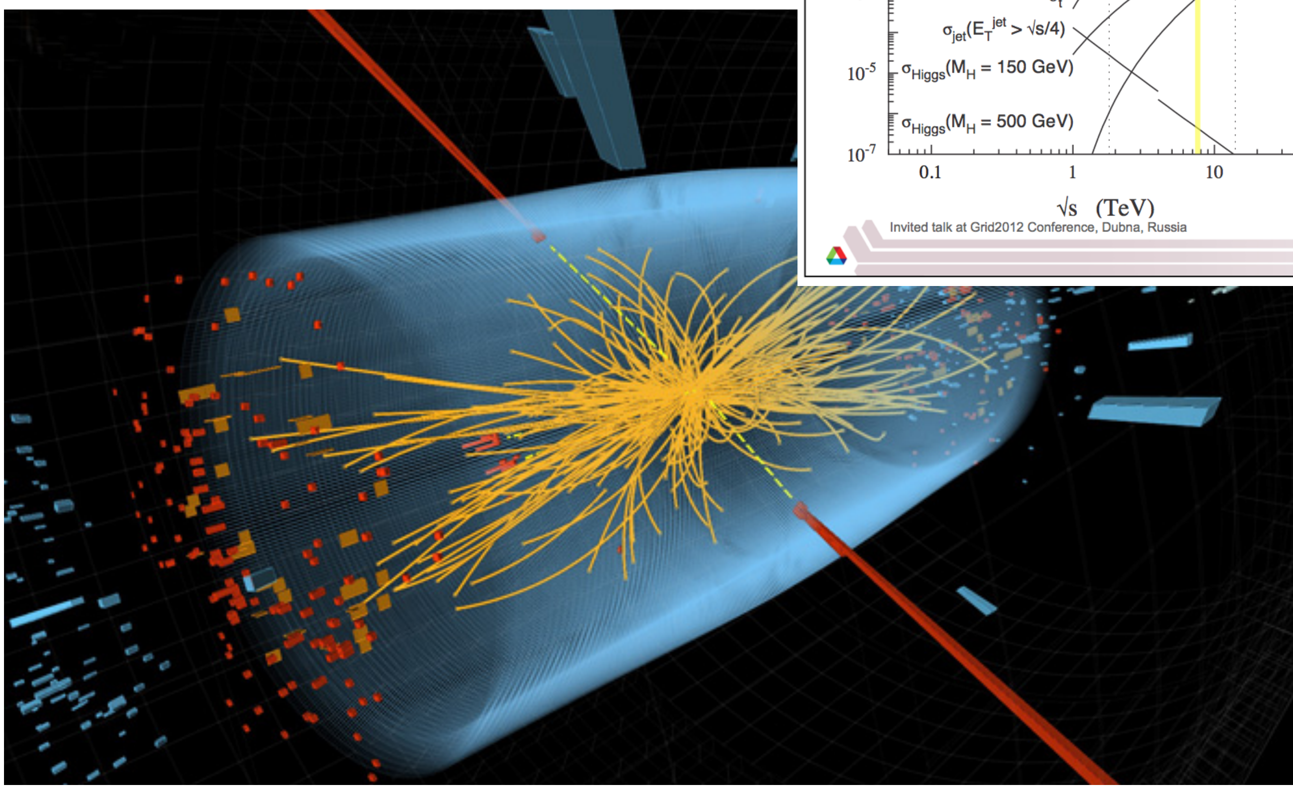
- Fundamental goal is general understanding & answering specific Qs:
  modeling & making predictions
- **Explicit Description of Data not Important --**
  Fast query, hiding underlying structure
  (e.g. Google **Search**)
- **Explicit Description of Data Important –**
  Organization highlighting underlying structure
  (e.g. Google **Maps**)

http://www.theatlantic.com/technology/archive/2012/01/to-know-but-not-understand-david-weinberger-on-science-and-big-data/250820/

http://images.iop.org/objects/phw/news/16/7/4/higgs-cms.jpg

Higgs Boson: Searching Through Many Events for a Few Needles



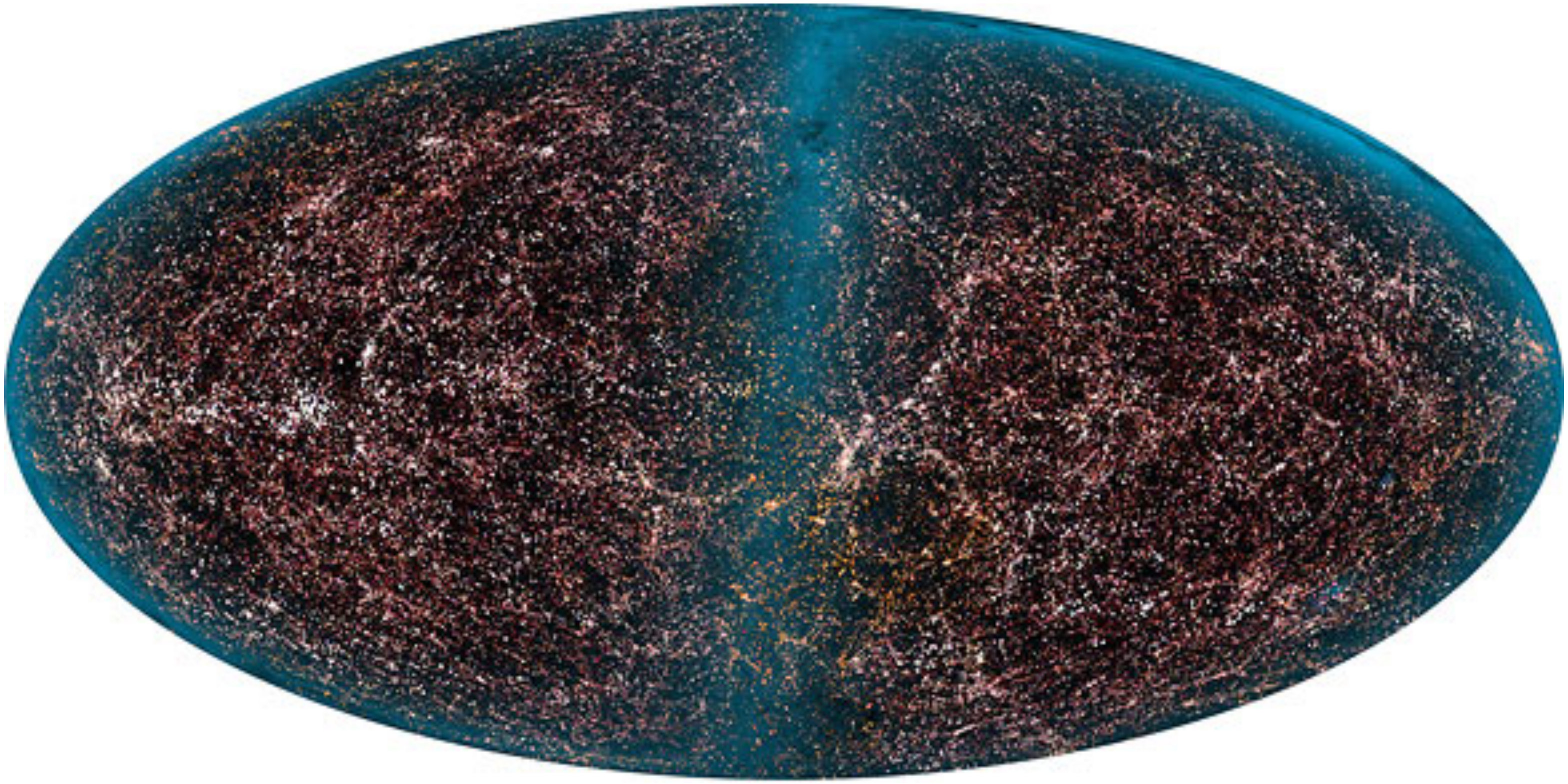


"Golden" Events

One H → 4 l / Billion

http://cdsweb.cern.ch/record/1462234/files/ATL-SOFT-SLIDE-2012-441.pdf

Making Intuitive Maps, Highlighting Large-scale Structure of Stars & the Earth

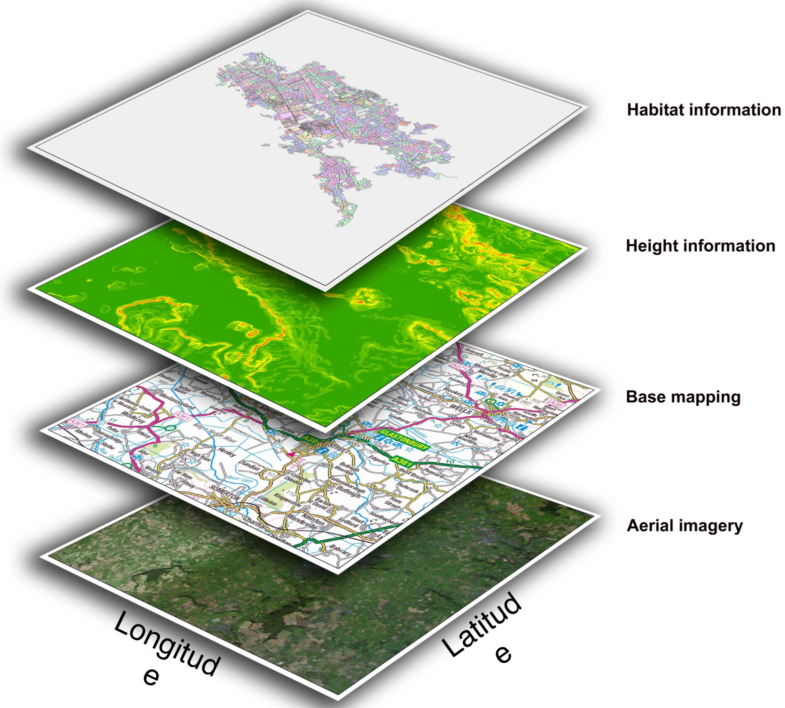


http://www.nrc-cnrc.gc.ca/eng/education/astronomy/tapping/2011/2011-03-02.html

[SDSS.org]

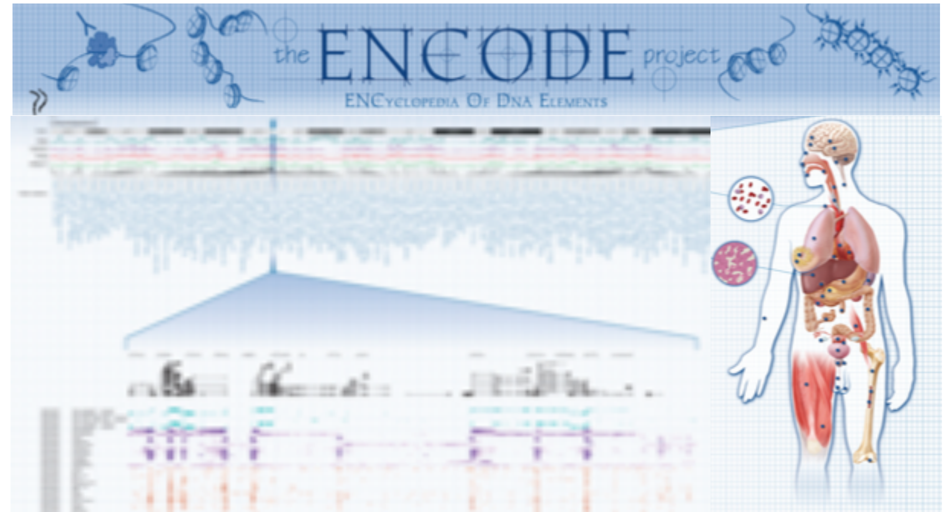
Human genome annotation — a non-intuitive map

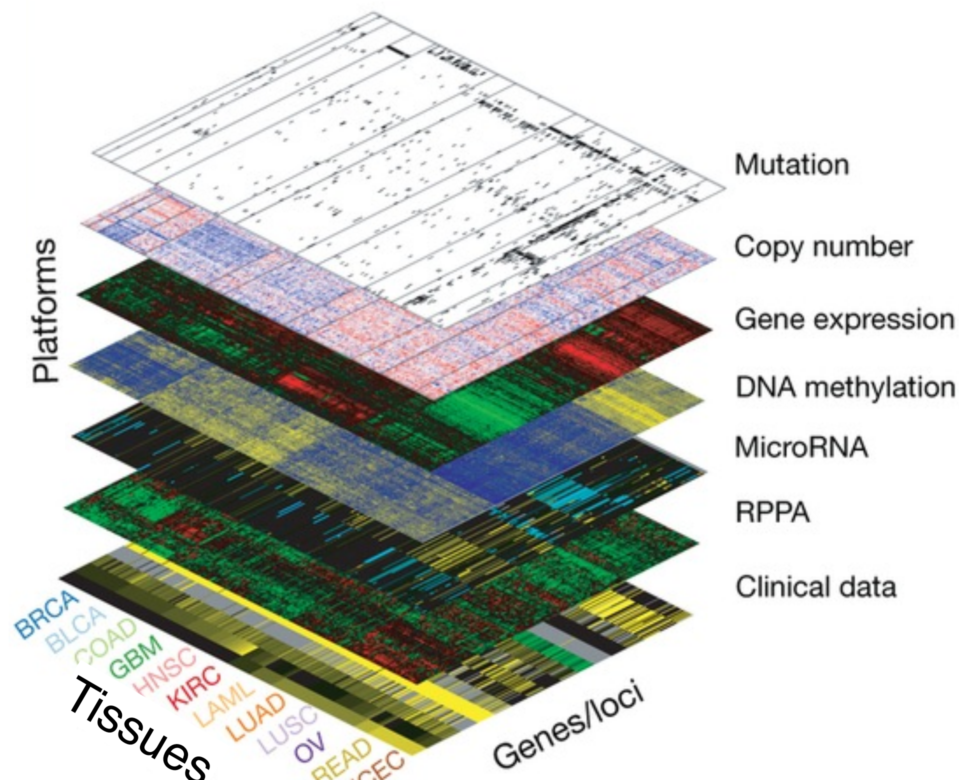
geographical information



- Large-scale organisation providing an overview of the genome
- Integration of heterogeneous data

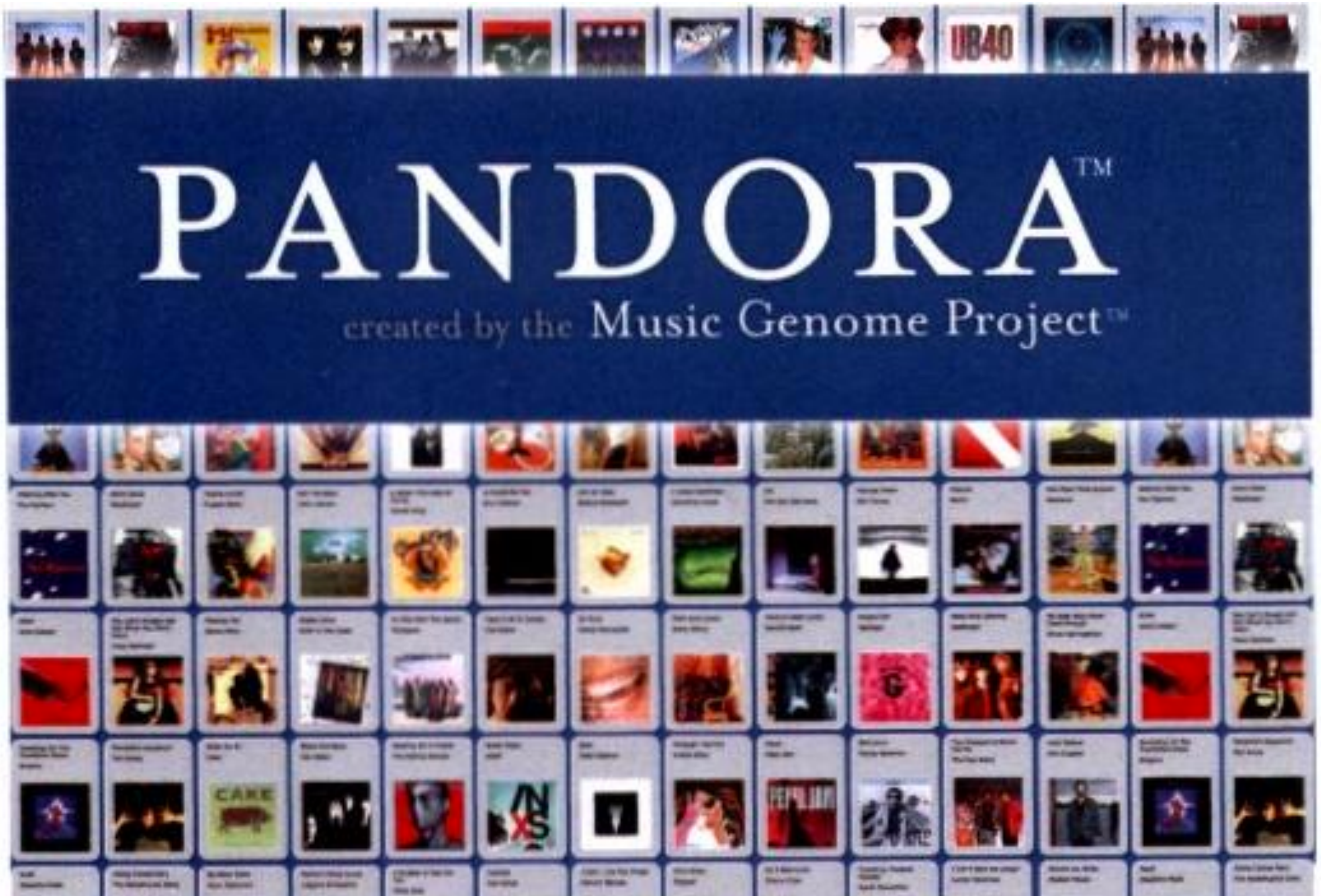
genomic information





Genomics: as an exemplar Data Science sub-discipline

- Developing ways of organizing & mining genomic information on a large scale
  - Very fundamental & early form of "Big Data"
- Perhaps we can learn from other disciplines &, in turn, teach them how to do this?

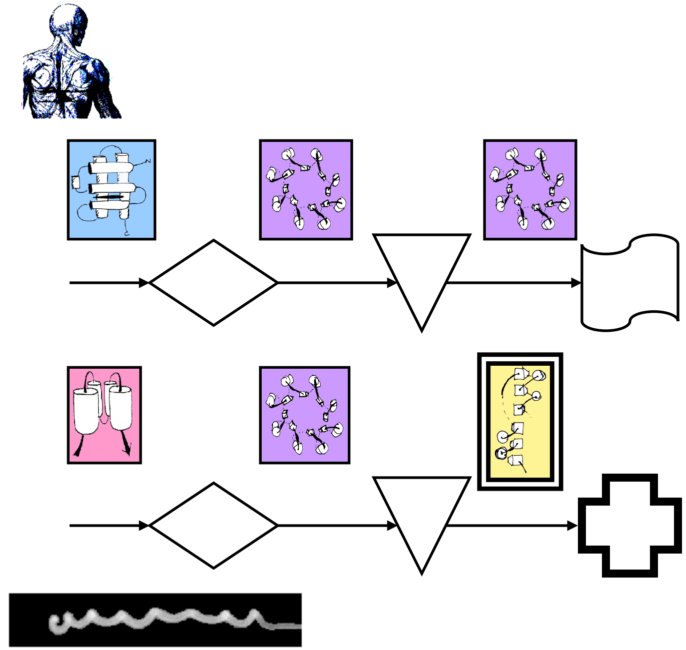


What is Bioinformatics?

- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, mine, model & understand the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

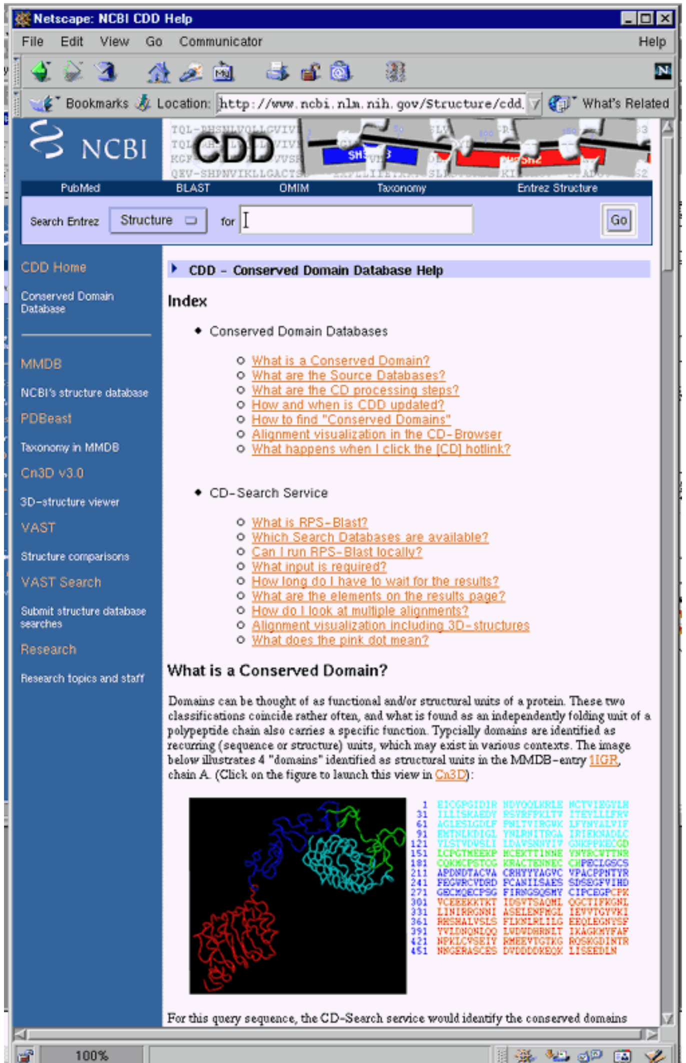
Organizing Molecular Biology Information: Redundancy and Multiplicity

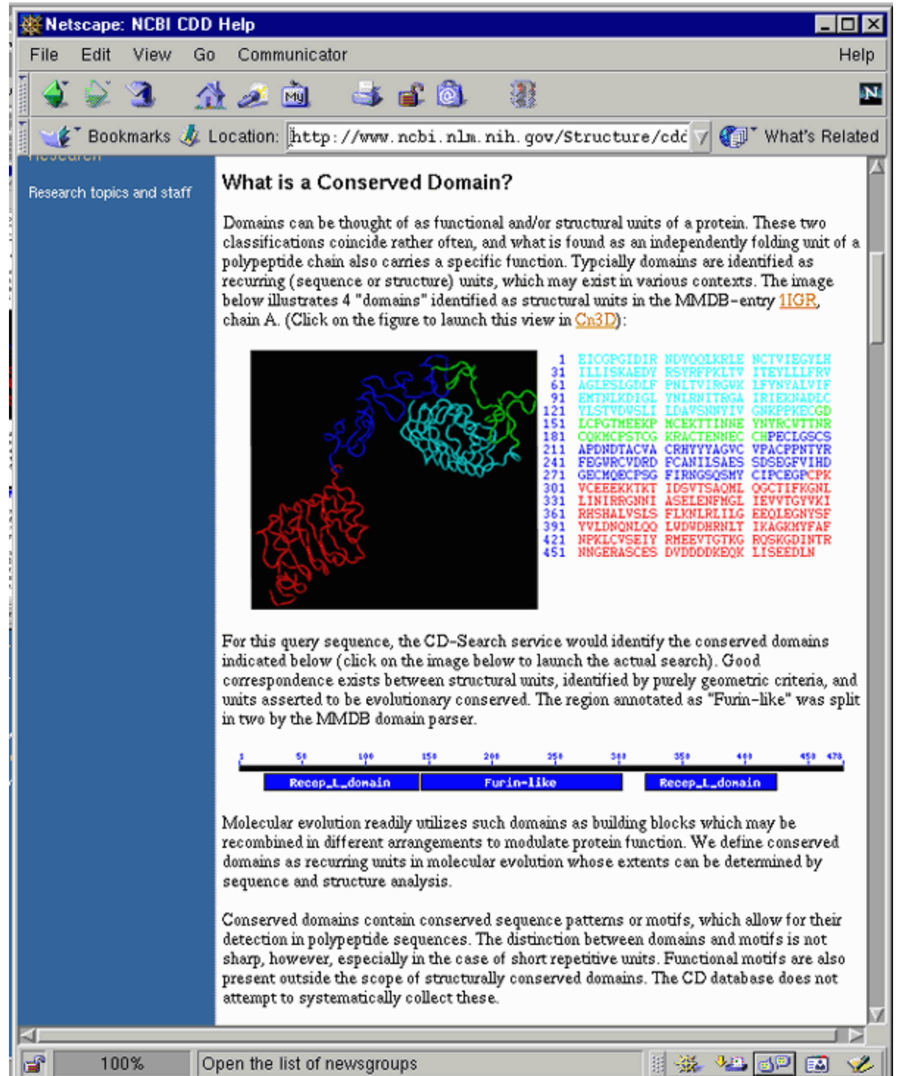


- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathway & Networks
- Genomic Sequence Redundancy due to the Genetic Code
- **How do we find the similarities?.....**

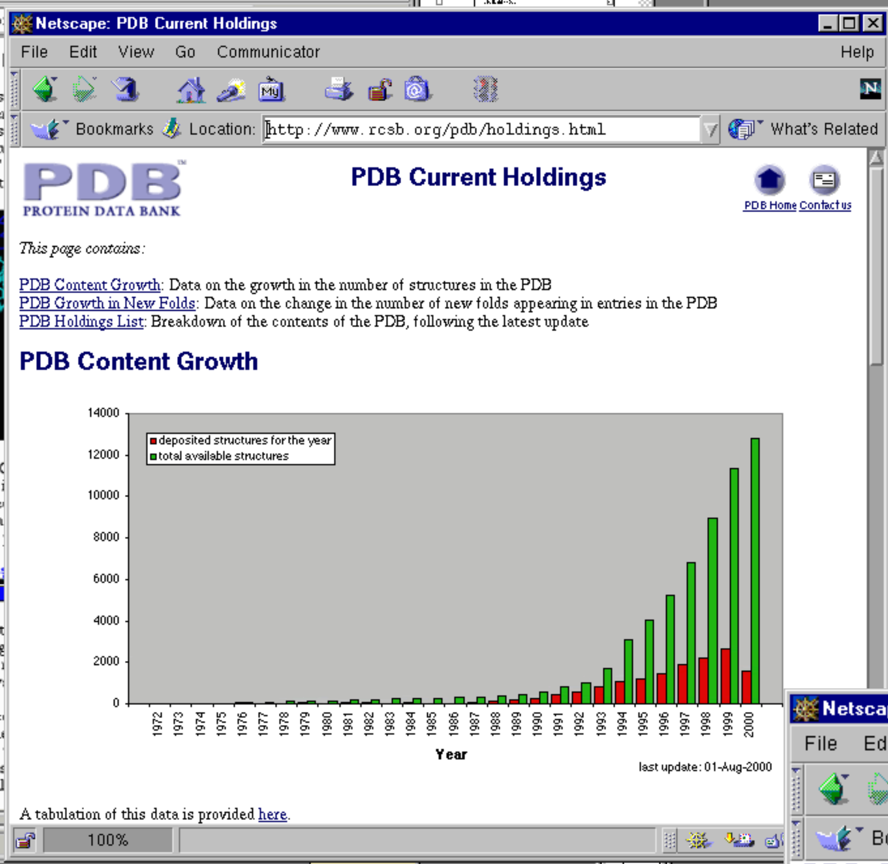
Integrative Genomics - genes ↔ structures ↔ **functions** ↔ **pathways** ↔ expression levels ↔ regulatory systems ↔ ….

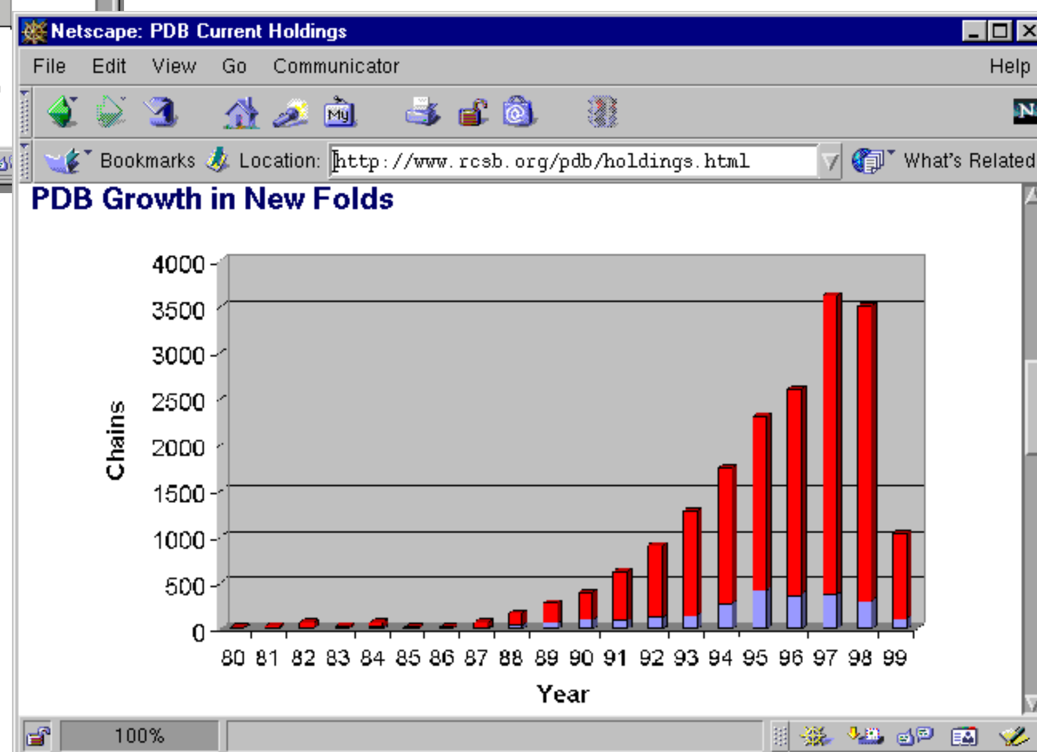
Molecular Parts = Conserved Domains, Folds, &c





Vast Growth in (Structural) Data... but number of Fundamentally New (Fold) Parts Not Increasing that Fast

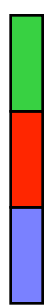




Total in Databank

New Submissions

New Folds

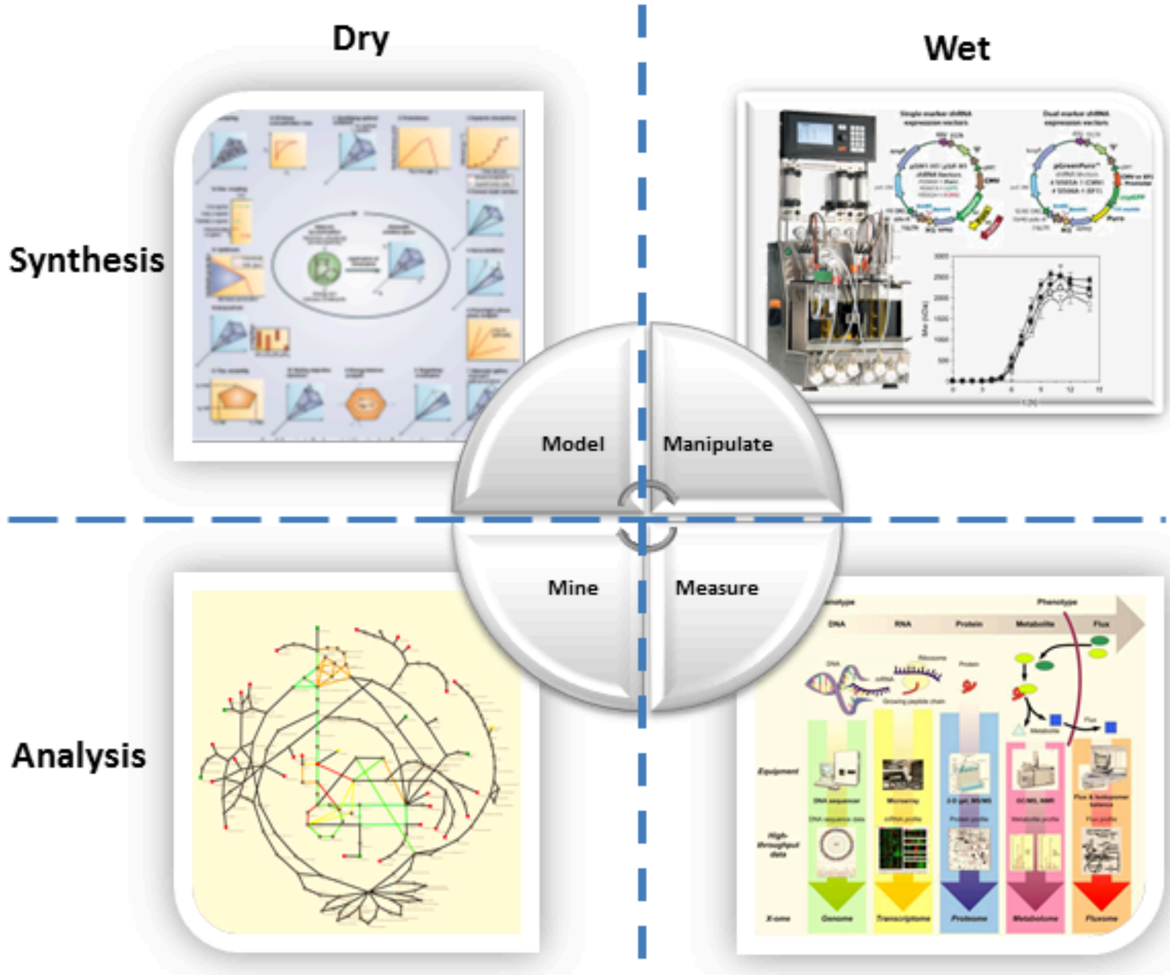


4Ms:

Measurement, Mining, Modeling & Manipulation

TREY IDEKER, L. RAIMOND WINSLOW & A. DOUGLAS LAUFFENBURGER ('06). "Bioengineering and Systems Biology," Annals of Biomedical Engineering DOI: 10.1007/s10439-005-9047-7

Image from http://web.aibn.uq.edu.au/cssb/ResearchProjects.html



What is Bioinformatics?

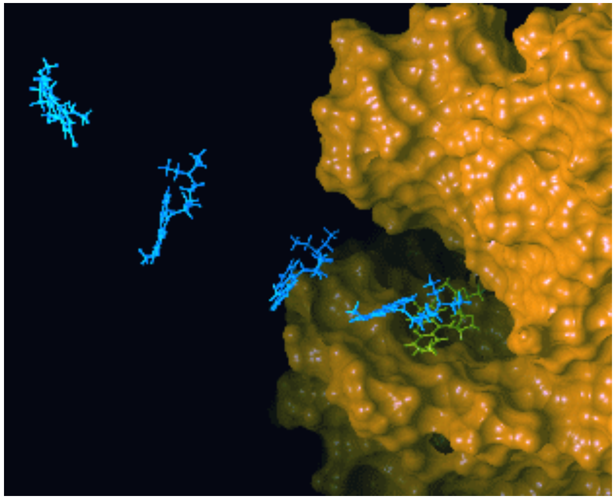
- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, mine, model & understand the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

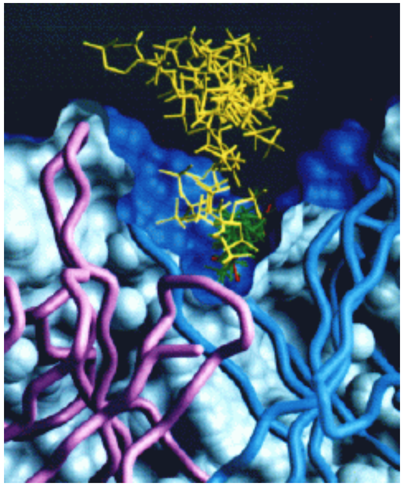
[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

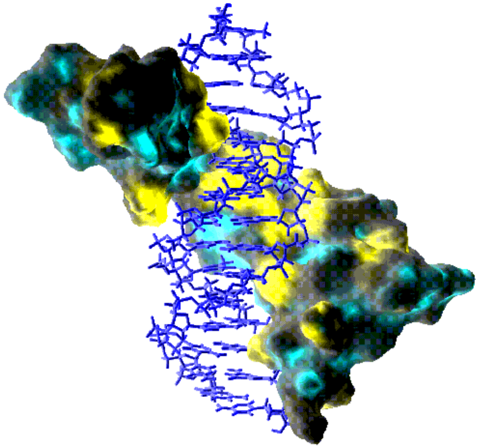
Major Application I: Designing Drugs

- Understanding How Structures Bind Other Molecules (Function)
- Designing Inhibitors
- Docking, Structure Modeling

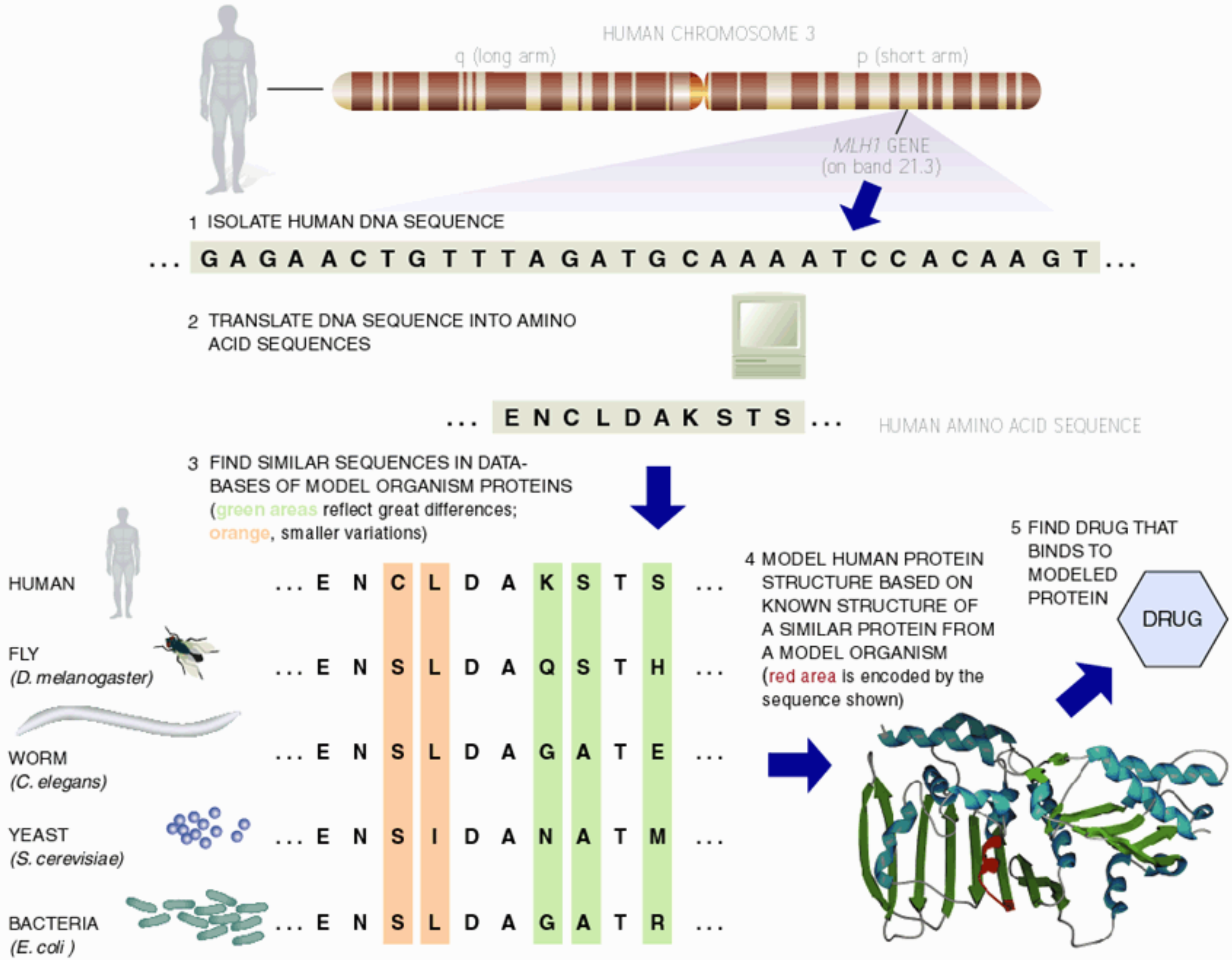
(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).





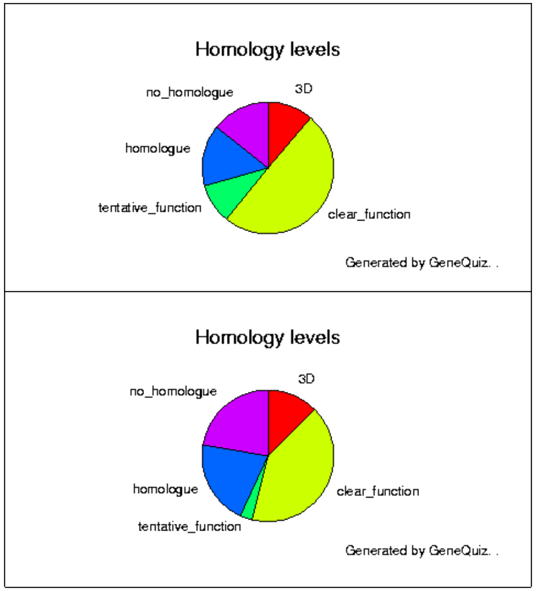


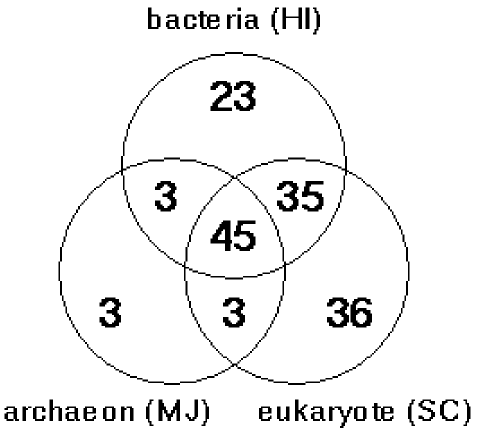
Major Application II: Finding Homologs



Major Application III: Overall Genome Characterization

- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics
- Using this for **picking drug targets**





(Clock figures, yeast v. Synechocystis, adapted from GeneQuiz Web Page, Sander Group, EBI)

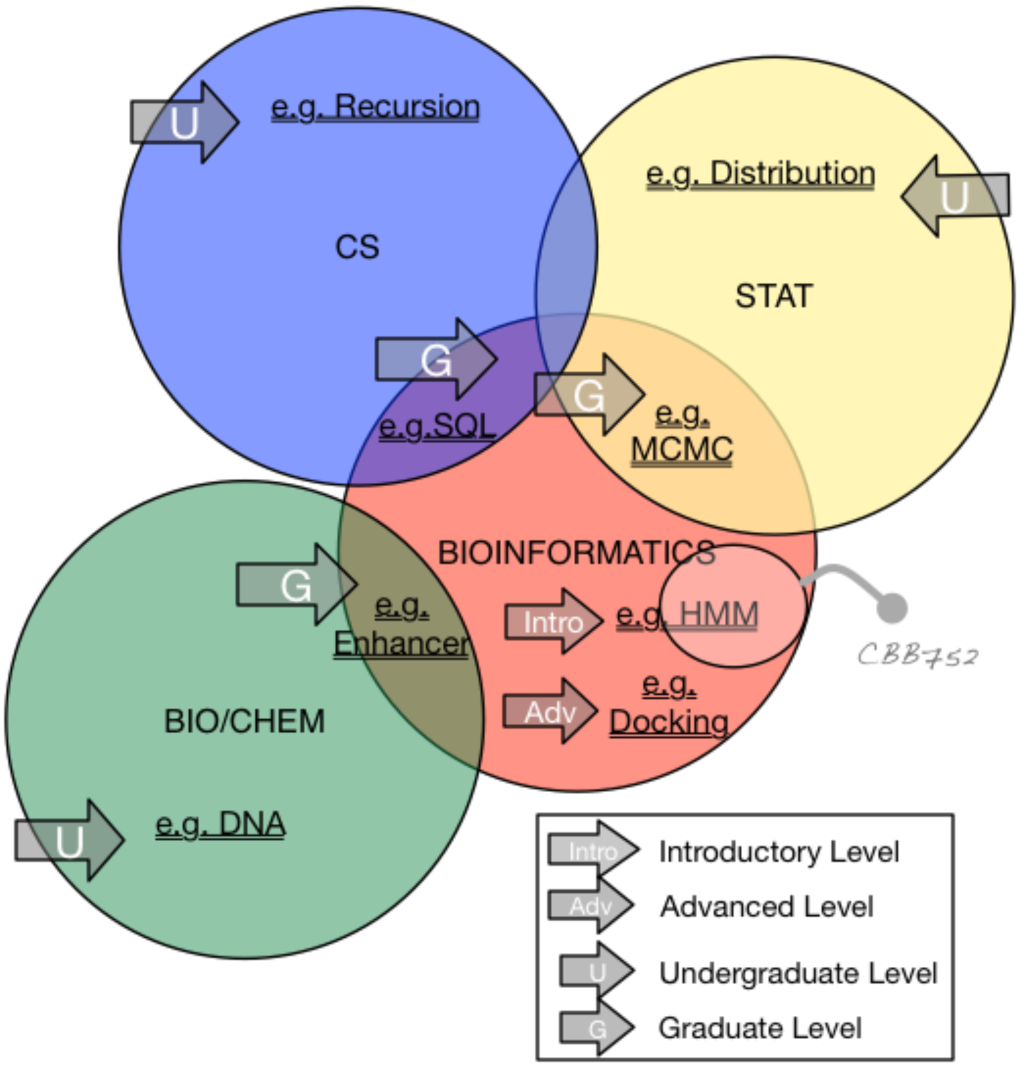
What is Bioinformatics?

- *(Molecular)* **Bio** - **`informatics`**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to **organize, mine, model & understand the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is a practical discipline with many **applications**.

[Luscombe et al. ('01). *Methods Inf Med* 40: 346]

Defining the Field

- Related terms
  - ◊ Bioinformatics & / or / vs Computational Biology
  - ◊ Biocomputing
  - ◊ Systems Biology
  - ◊ Qbio
- What are its boundaries
  - ◊ Determining the "Support Vectors”
- What are appropriate prerequisites



Are They or Aren't They Comp. Bio.? (#1)

- (Digital Libraries & Medical Record Analysis
  - ◊ Automated Bibliographic Search and Textual Comparison
  - ◊ Knowledge bases for biological literature
- (Motif Discovery Using Gibb's Sampling
- (Methods for Structure Determination
  - ◊ Computational Crystallography
    - Refinement
  - ◊ NMR Structure Determination
    - (Distance Geometry
- (Metabolic Pathway Simulation
- (The DNA Computer

Are They or Aren't They Comp. Bio.? (#1, `Answers`)

- **(YES?)** Digital Libraries & Medical Record Analysis
  - ◊ Automated Bibliographic Search and Textual Comparison
  - ◊ Knowledge bases for biological literature
- **(YES)** Motif Discovery Using Gibb's Sampling
- **(NO?)** Methods for Structure Determination
  - ◊ Computational Crystallography
    - Refinement
  - ◊ NMR Structure Determination
    - **(YES)** Distance Geometry
- **(YES)** Metabolic Pathway Simulation
- **(NO)** The DNA Computer

Are They or Aren't They Comp. Bio.? (#2)

- (Gene identification by sequence characteristics
  - ◊ Prediction of splice sites
- (DNA methods in forensics
- (Modeling of Populations of Organisms
  - ◊ Ecological Modeling (predator & prey)
- (Modeling the nervous system
  - ◊ Computational neuroscience
  - ◊ Understanding how brains think & using this to make a better computer
- (Molecular phenotype discovery – looking for gene expression signatures of cancer
  - ◊ What if it included non-molecular data such as age ?

Are They or Aren't They Comp. Bio.? (#2, `Answers`)

- **`(YES)`** Gene identification by sequence characteristics
  - ◊ Prediction of splice sites
- **`(YES)`** DNA methods in forensics
- **`(NO)`** Modeling of Populations of Organisms
  - ◊ Ecological Modeling (predator & prey)
- **`(NO?)`** Modeling the nervous system
  - ◊ Computational neuroscience
  - ◊ Understanding how brains think & using this to make a better computer
- **`(YES)`** Molecular phenotype discovery – looking for gene expression signatures of cancer
  - ◊ What if it included non-molecular data such as age ?

Are They or Aren't They Comp. Bio.? (#3)

- (RNA structure prediction
- (Radiological Image Processing
  - ◊ Computational Representations for Human Anatomy (visible human)
- (Artificial Life Simulations
  - ◊ Artificial Immunology / Computer Security
  - ◊ (Genetic Algorithms in molecular biology
- (Homology Modeling & Drug Docking
- (Char. drugs & other small molecules (QSAR)
- (Computerized Diagnosis based on Pedigrees
- (Processing of NextGen sequencing image files
- (Module finding in protein networks

Are They or Aren't They Comp. Bio.? (#3, Answers)

- **(YES)** RNA structure prediction
- **(NO)** Radiological Image Processing
  - ◊ Computational Representations for Human Anatomy (visible human)
- **(NO)** Artificial Life Simulations
  - ◊ Artificial Immunology / Computer Security
  - ◊ **(NO?)** Genetic Algorithms in molecular biology
- **(YES)** Homology Modeling & Drug Docking
- **(YES)** Char. drugs & other small molecules (QSAR)
- **(NO)** Computerized Diagnosis based on Pedigrees
- **(NO)** Processing of NextGen sequencing image files
- **(YES)** Module finding in protein networks

Class Web Page

http://GersteinLab.org/**courses/452**

Assignment #0 Page

https://goo.gl/**7MDP1I**