Modern Biology in Two Lectures (Part II)

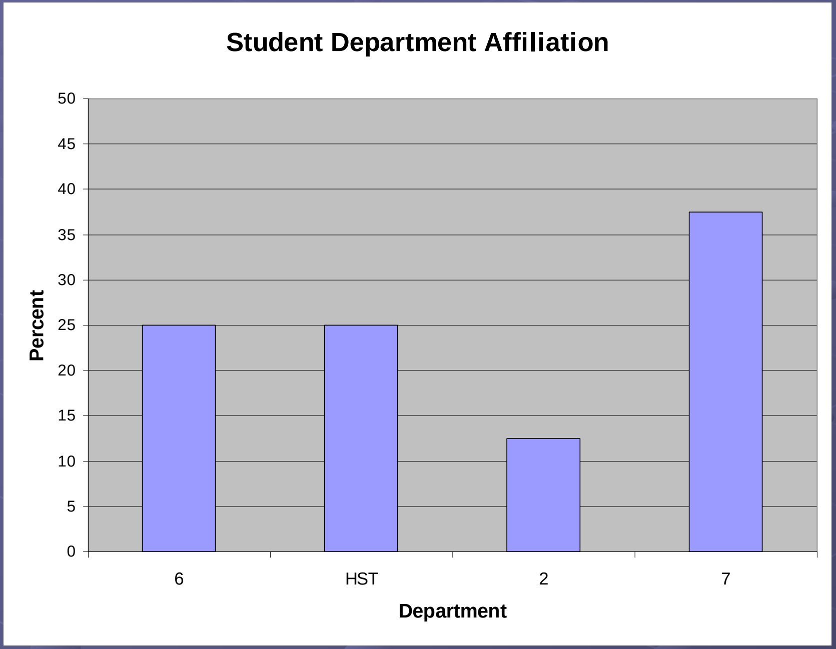
Gil Alterovitz

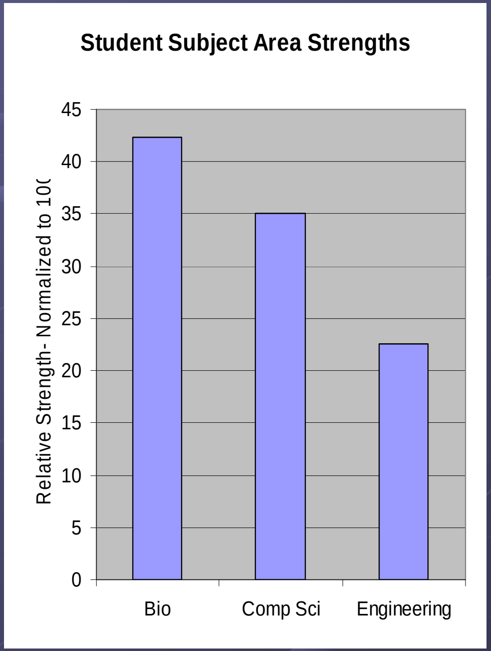
Harvard-MIT
Division of Health
Science & Technology

Course Administration

- Handouts
  - Open Courseware form- please turn in before leaving class
  - Matlab form- for free copy of Matlab for students in class for use in 6.092/HST.480 course. You can also use server Matlab or lab cluster.
  - Background sheet- complete and turn in by end of class so we can put you on course email list.
- Homework 1 (Due next Thurs.)
  - See assignments section in course site.

Background





Harvard-MIT
Division of Health
Science & Technology

Today

- Introduction, Part II- Gil Alterovitz
  - Review Part I
  - Splicing
  - Alternative Splicing
  - Post-Translational Modifications
- Sequence Analysis- Manolis Kellis

Genes to Proteins

Transcription → **Translation** →

DNA: “Lifetime Plan”

```
5′ATCTACAGATCAGCTACGACGCGACGAT
TTAGCAGCAGCGACGCGACAGCAGCTAGTG
ACGATAGCACATAGTTAGCACAGAGCAGAC
ACAGACAGCACAGCGACAGCGACGACG-3′
```

mRNA: “Task List”

```
5′AUCUACAGAUCAGCUACGACGCGACGAU
UUAGCAGCAGCGACGCGACAGCAGCUAGUG
ACGAUAGCACAUAGUUAGCACAGAGCAGAC
ACAGACAGCACAGCGACAGCGACGACG-3′
```

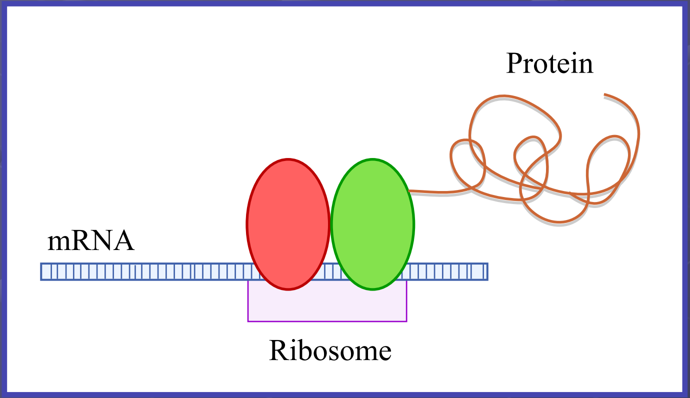


Figure by MIT OCW

Protein: Machines

```
MWTRFDSALPRSTPSTAKLVMPOILLLLEE
EDTYESAQYKTWLMVCSDETTTE
```

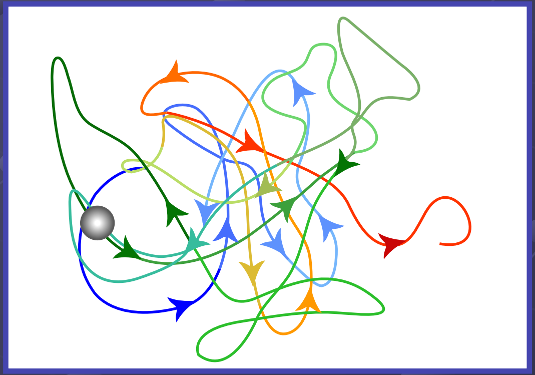


Figure by MIT OCW

DNA Sequencing

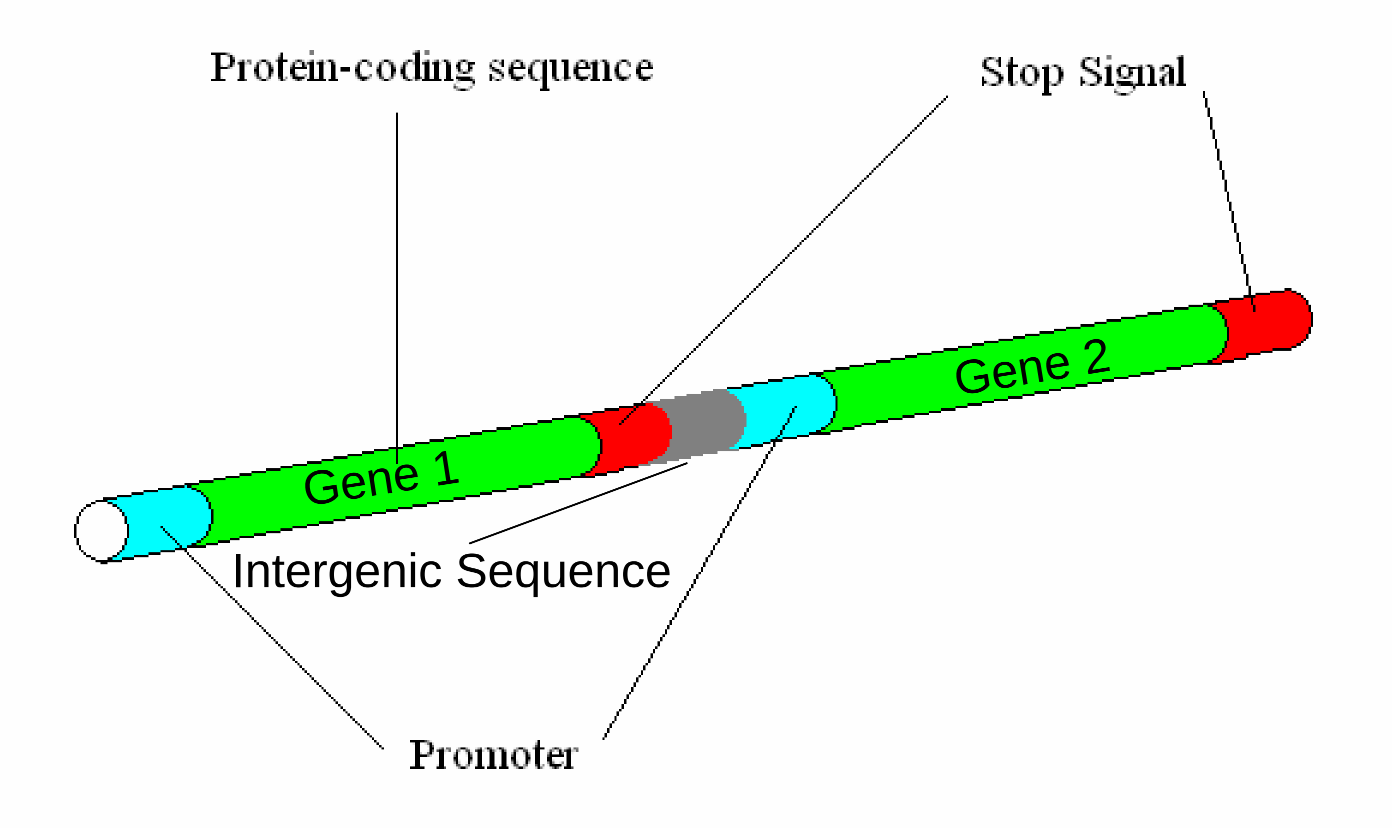
Relative Expression Levels

Identification
Post translation modification
Splicing variants
Relative expression levels

Source: HPCGG

Harvard-MIT Division of Health Science & Technology

Genes



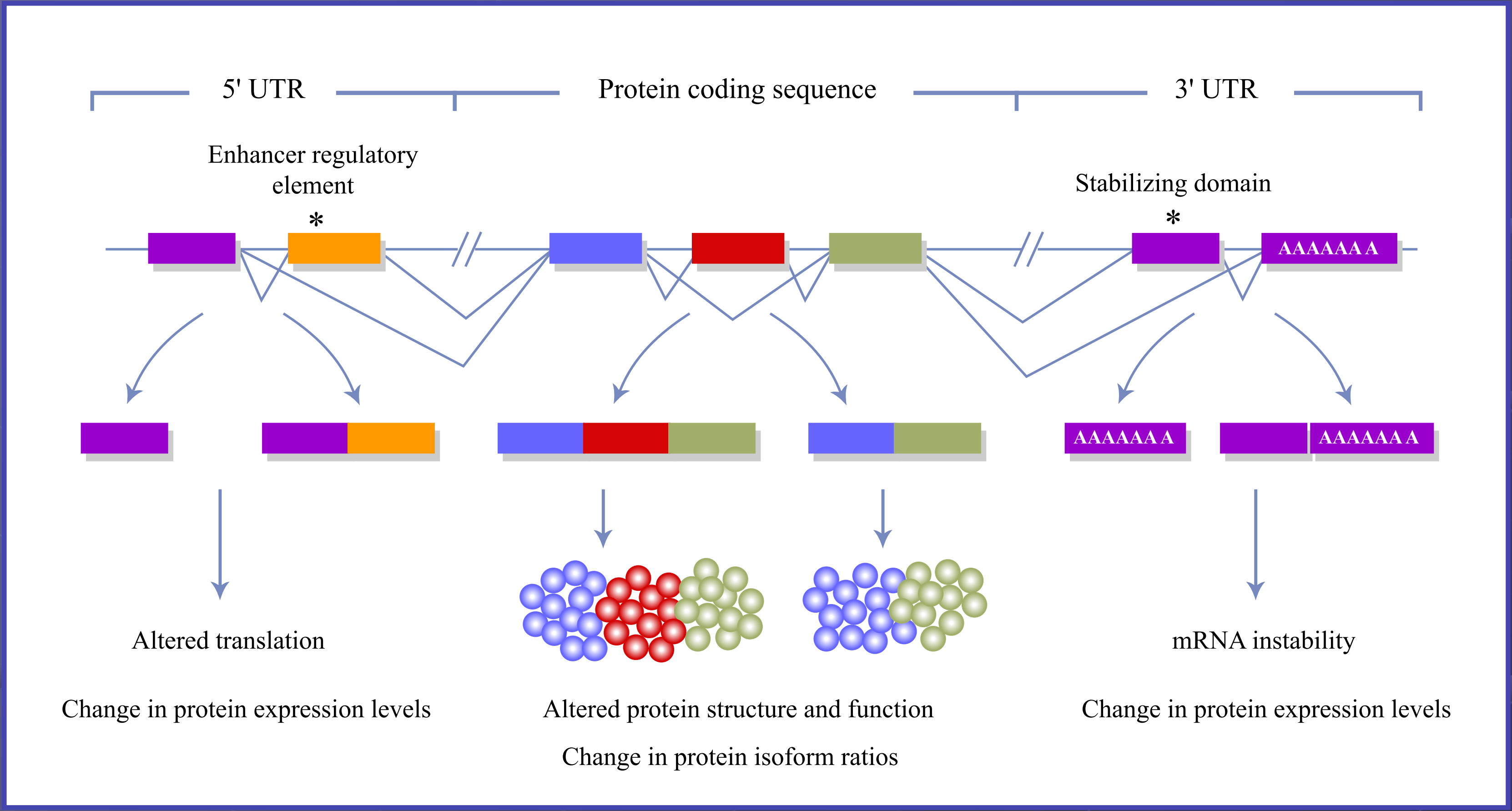
Communication analogy: start, message, stop.

Source: Ehsan Afkhami

Stereo Rack Analogy

Amplifier

Alternative Splicing



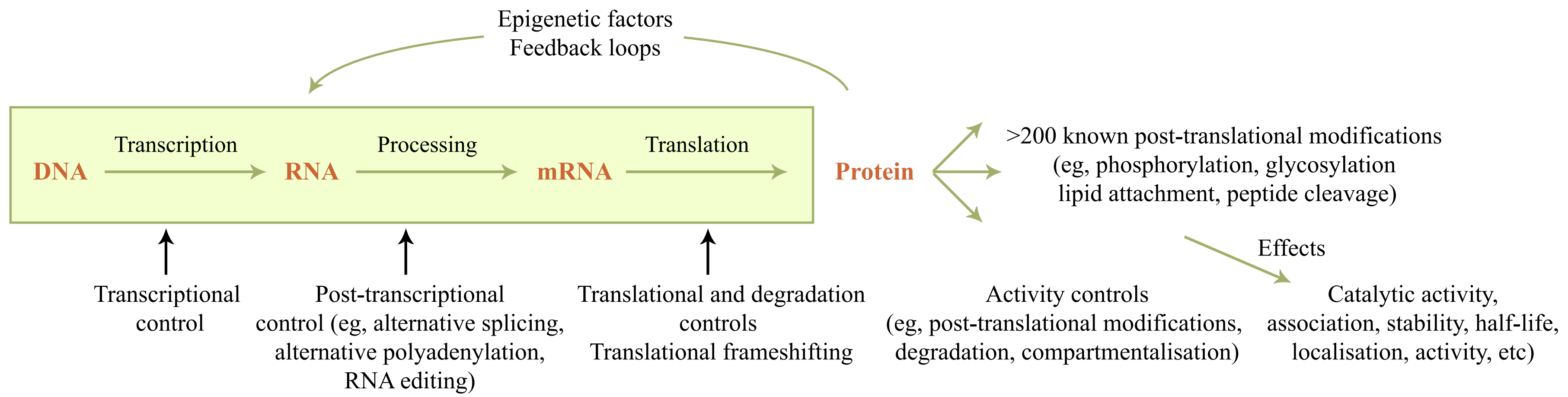
5' UTR
Protein coding sequence
3' UTR
Enhancer regulatory element
*
Stabilizing domain
*
AAAAAAA
AAAAAAA
AAAAAAA
Altered translation
Change in protein expression levels
Altered protein structure and function
Change in protein isoform ratios
mRNA instability
Change in protein expression levels

Figure by MIT OCW

Harvard-MIT
Division of Health
Science & Technology

Sequence Ordering

DNA	Coding Strand (Codons)	**5'** > > > - - - - - - T T C - - - - - - - > > > **3'**
	Template Strand (Anti-codons)	**3'** < < < - - - - - - - A A G - - - - - - - < < < **5'**
RNA	Message (Codons)	**5'** > > > - - - - - - U U C - - - - - - - > > > **3'**
Protein	Amino Acid	**Amino** > > > Phenylalanine > > > **Carboxy**



Epigenetic factors
Feedback loops

DNA Transcription RNA Processing mRNA Translation Protein

>200 known post-translational modifications
(eg, phosphorylation, glycosylation
lipid attachment, peptide cleavage)

Effects

Transcriptional
control

Post-transcriptional
control (eg, alternative splicing,
alternative polyadenylation,
RNA editing)

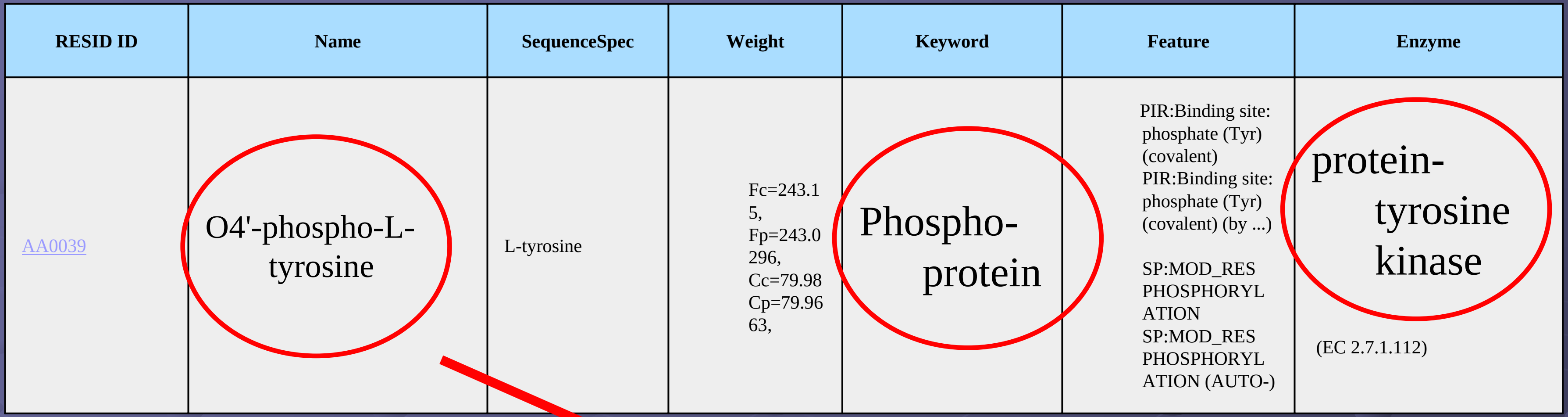
Translational and degradation
controls
Translational frameshifting

Activity controls
(eg, post-translational modifications,
degradation, compartmentalisation)

Catalytic activity,
association, stability, half-life,
localisation, activity, etc)

Figure by MIT OCW

Post-translational Modifications



RESID ID	Name	SequenceSpec	Weight	Keyword	Feature	Enzyme
AA0039	O4'-phospho-L-tyrosine	L-tyrosine	Fc=243.15, Fp=243.0296, Cc=79.98 Cp=79.9663,	Phosphoprotein	PIR:Binding site: phosphate (Tyr) (covalent) PIR:Binding site: phosphate (Tyr) (covalent) (by ...) SP:MOD_RES PHOSPHORYLATION SP:MOD_RES PHOSPHORYLATION (AUTO-)	protein-tyrosine kinase (EC 2.7.1.112)

339 modifications in RESID Database

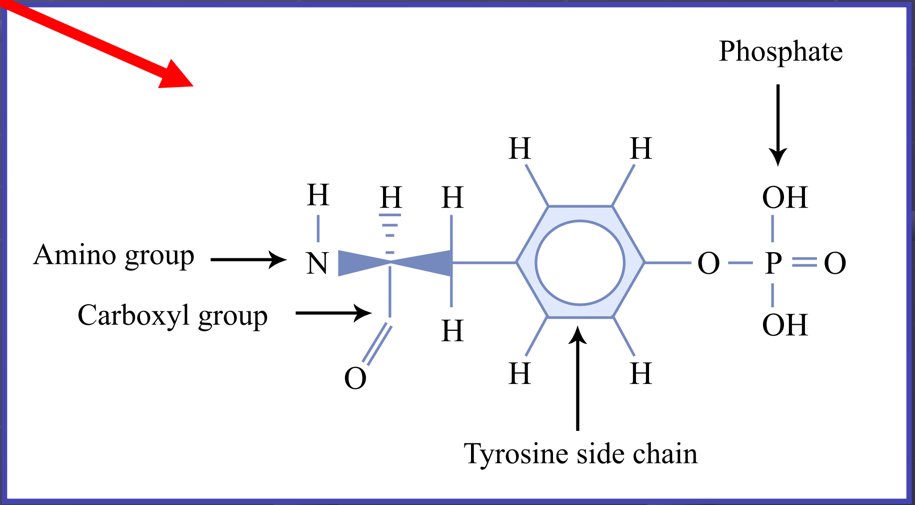


Figure by MIT OCW

Harvard-MIT
Division of Health
Science & Technology

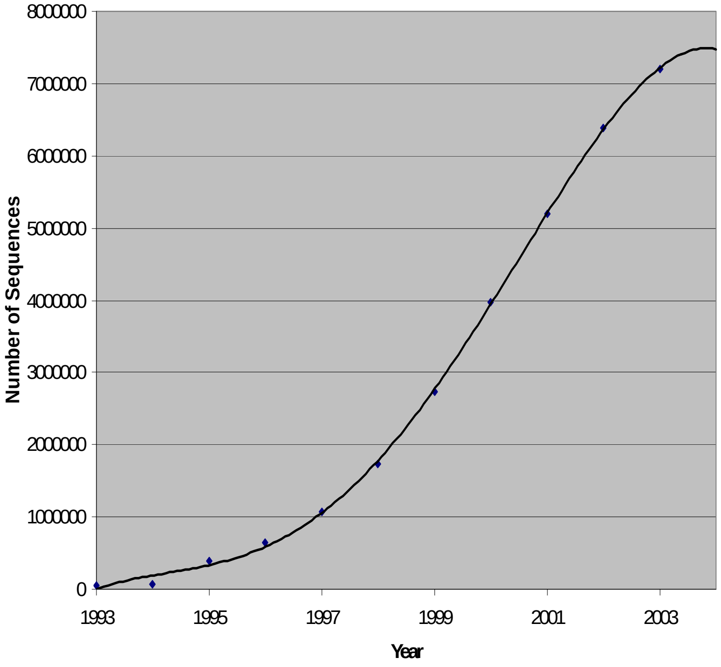
Bioinformatics: Trends, Tools, and Databases

What kind of problems need to be solved?

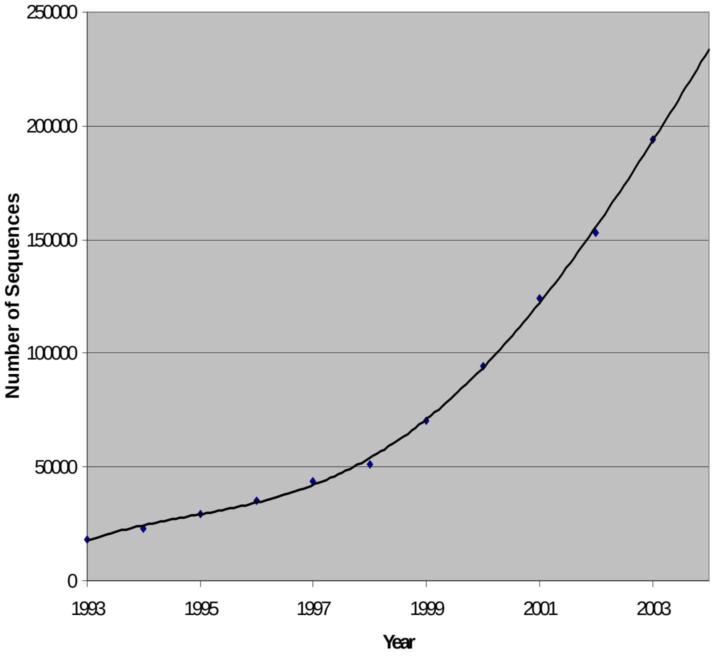
How have previous problems in the field been approached?

Databases Needed to Store Growing List of Sequence Data

Entrez Human Nucleotide Sequences



Entrez Human Protein Sequences



* Alterovitz, G., Afkhami, E. & Ramoni, M. in *Focus on Robotics and Intelligent Systems Research*, ed. Columbus, F. Nova Science Publishers, Inc., New York, 2005 (In press).

HST Harvard-MIT
Division of Health
Science & Technology

Paradigm Shifts in Bioinformatics

- **Sequencing** (1980’s to early 1990’s)
  - DNA/RNA/Protein Sequence Analysis/sequence storage
- **3-D Protein Structure Prediction** (Mid-1980’s-late 1990’s)
  - Databases of Protein structures
- **DNA/RNA Microarray Expression Experiments** (Mid-1990’s to 2000’s)
  - Databases of expression data
- **Protein interaction experiments** (Early 2000’s to Present)
  - Databases with pairwise interactions
- **Mass Spec proteomic pattern experiments** (Early 2000’s to Present)
  - Databases with mass spec, protein identifications, proteomic patterns
- Integration of multiple modalities (Ongoing)

HST

Harvard-MIT
Division of Health
Science & Technology

Human Genome Project

- ~ 99% of human genome has been sequenced (2004). Nature 431: 931-945.
- Error rate: ~1 event per 100,000 bases
- Number of protein-coding genes: 20,000-25,000
- Number of protein-coding genes in worm: ~18,000
- Genes comprise only about 2% of the human genome.
    - The rest consists of non-coding regions: functions may include providing chromosomal structural integrity and gene regulation.