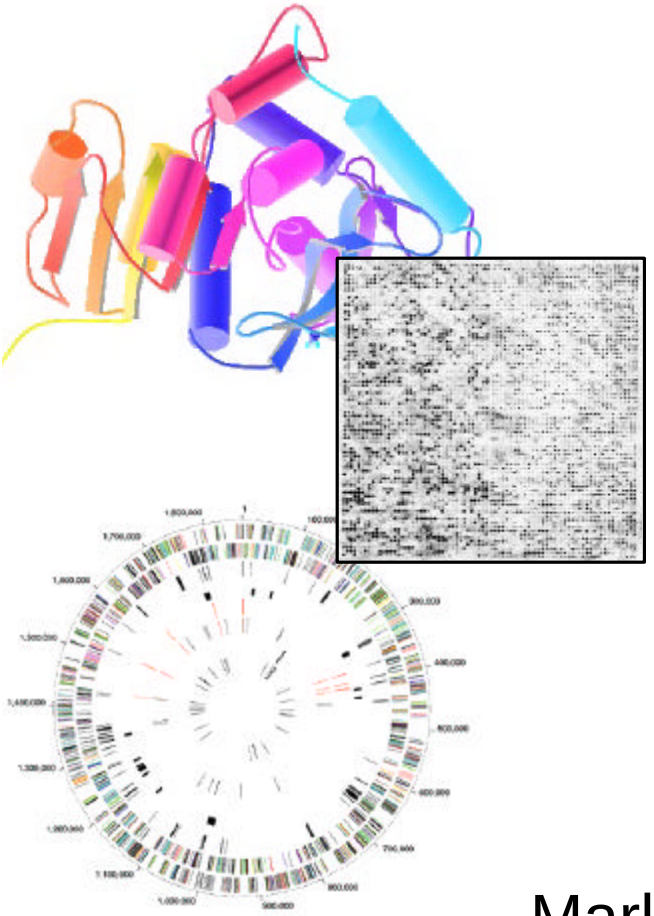
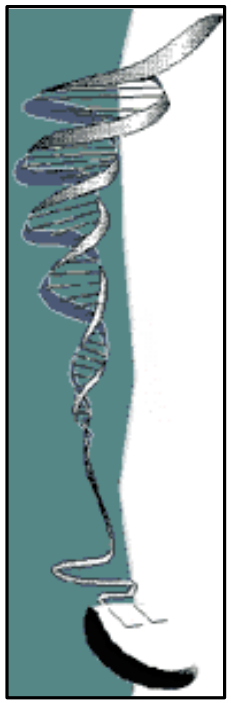
BIOINFORMATICS
Introduction





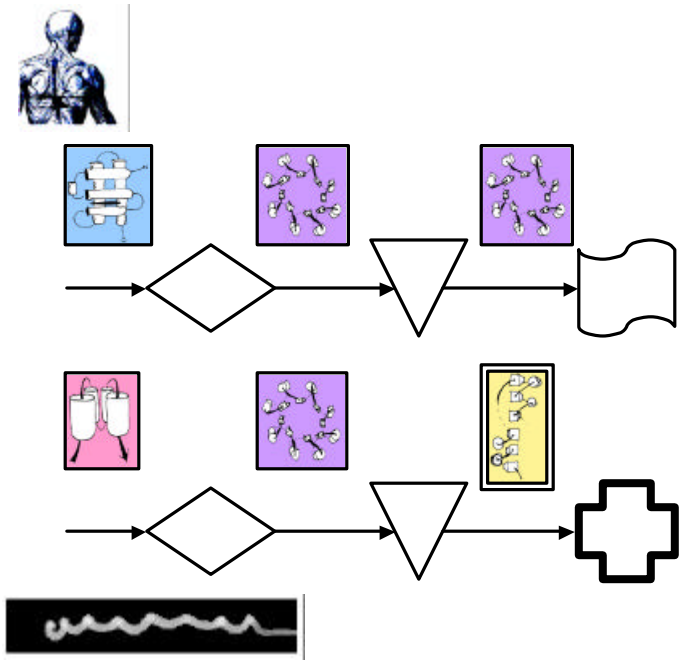
Mark Gerstein, Yale University

bioinfo.mbb.yale.edu/mbb452a

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 understand and **organize the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is "MIS" for Molecular Biology Information. It is a practical discipline with many **applications**.

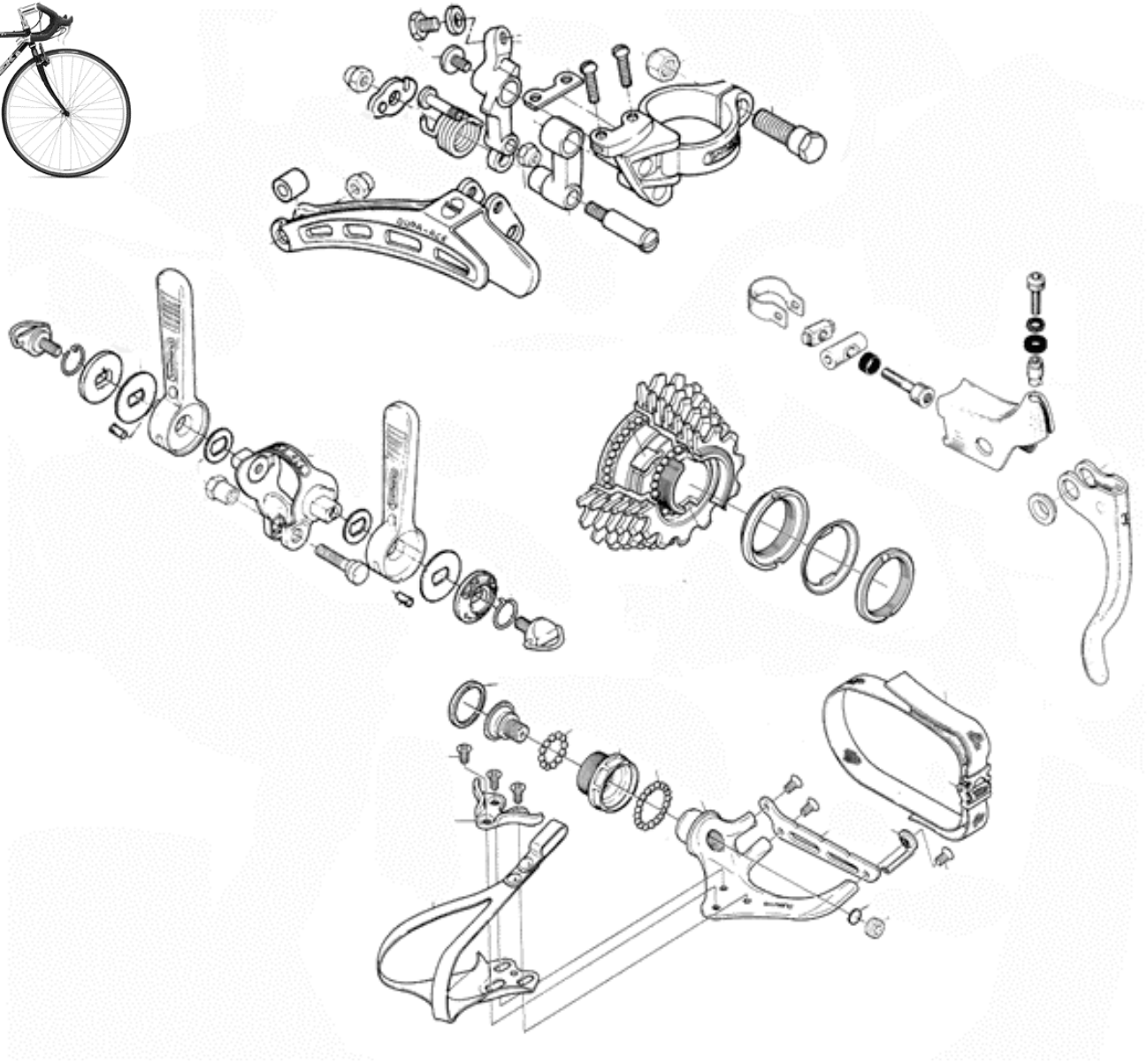
Organizing Molecular Biology Information: Redundancy and Multiplicity



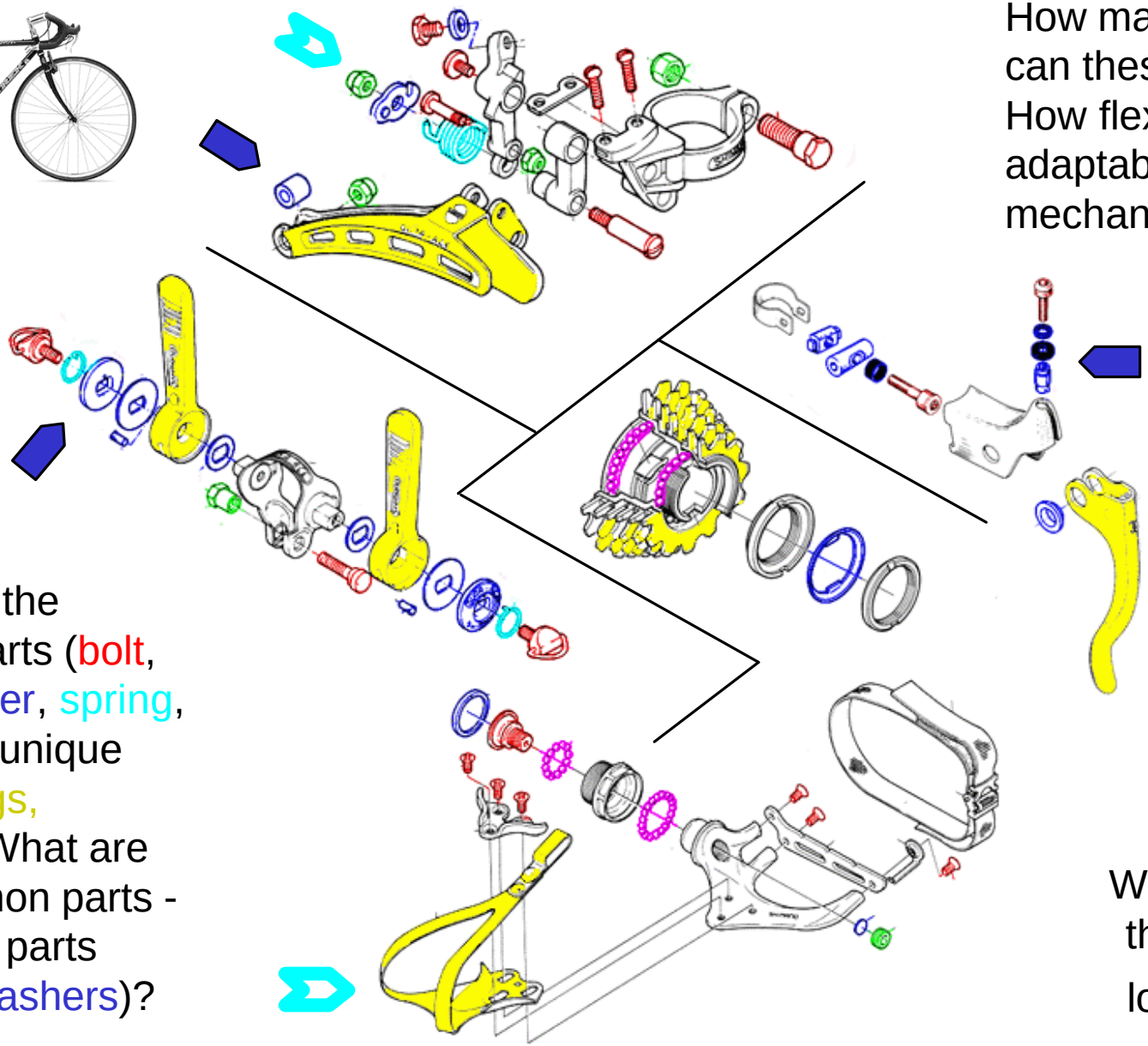
- Different Sequences Have the Same Structure
- Organism has many similar genes
- Single Gene May Have Multiple Functions
- Genes are grouped into Pathways
- Genomic Sequence Redundancy due to the Genetic Code
- **How do we find the similarities?.....**

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

A Parts List Approach to Bike Maintenance



A Parts List Approach to Bike Maintenance



How many roles can these play? How flexible and adaptable are they mechanically?

What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?

Where are the parts located?

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 understand and **organize the information associated** with these molecules, **on a large-scale.**
- Bioinformatics is "MIS" for Molecular Biology Information. It is a practical discipline with many **applications**.

General Types of "Informatics" techniques in Bioinformatics

- Databases
  - ◊ Building, Querying
  - ◊ Object DB
- Text String Comparison
  - ◊ Text Search
  - ◊ 1D Alignment
  - ◊ Significance Statistics
  - ◊ Alta Vista, grep
- Finding Patterns
  - ◊ AI / Machine Learning
  - ◊ Clustering
  - ◊ Datamining
- Geometry
  - ◊ Robotics
  - ◊ Graphics (Surfaces, Volumes)
  - ◊ Comparison and 3D Matching (Visision, recognition)
- Physical Simulation
  - ◊ Newtonian Mechanics
  - ◊ Electrostatics
  - ◊ Numerical Algorithms
  - ◊ Simulation

New Paradigm for Scientific Computing

- Because of increase in data and improvement in computers, new calculations become possible
- But Bioinformatics has a new style of calculation...
  - ◊ Two Paradigms
- Physics
  - ◊ Prediction based on physical principles
  - ◊ Exact Determination of Rocket Trajectory
  - ◊ Supercomputer, CPU
- Biology
  - ◊ Classifying information and discovering unexpected relationships
  - ◊ globin ~ colicin~ plastocyanin~ repressor
  - ◊ networks, “federated” database

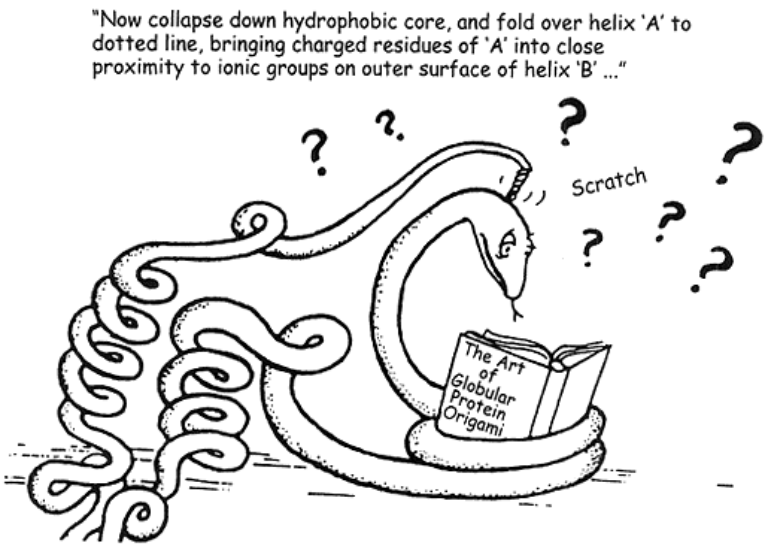
Bioinformatics Topics -- Genome Sequence

- Finding Genes in Genomic DNA
  - ◊ introns
  - ◊ exons
  - ◊ promotors
- Characterizing Repeats in Genomic DNA
  - ◊ Statistics
  - ◊ Patterns
- Duplications in the Genome

Bioinformatics Topics -- Protein Sequence

- Sequence Alignment
  - ◊ non-exact string matching, gaps
  - ◊ How to align two strings optimally via Dynamic Programming
  - ◊ Local vs Global Alignment
  - ◊ Suboptimal Alignment
  - ◊ Hashing to increase speed (BLAST, FASTA)
  - ◊ Amino acid substitution scoring matrices
- Multiple Alignment and Consensus Patterns
  - ◊ How to align more than one sequence and then fuse the result in a consensus representation
  - ◊ Transitive Comparisons
  - ◊ HMMs, Profiles
  - ◊ Motifs
- Scoring schemes and Matching statistics
  - ◊ How to tell if a given alignment or match is statistically significant
  - ◊ A P-value (or an e-value)?
  - ◊ Score Distributions (extreme val. dist.)
  - ◊ Low Complexity Sequences

Bioinformatics Topics -- Sequence / Structure



Reproduced in U. Tollemar, "*Protein Engineering i USA*", Sveriges Tekniska Attachéer, 1988

- Secondary Structure "Prediction"
  - ◊ via Propensities
  - ◊ Neural Networks, Genetic Alg.
  - ◊ Simple Statistics
  - ◊ TM-helix finding
  - ◊ Assessing Secondary Structure Prediction
- Tertiary Structure Prediction
  - ◊ Fold Recognition
  - ◊ Threading
  - ◊ Ab initio
- Function Prediction
  - ◊ Active site identification
- Relation of Sequence Similarity to Structural Similarity

Topics -- Structures

- Basic Protein Geometry and Least-Squares Fitting
  - ◊ Distances, Angles, Axes, Rotations
    - Calculating a helix axis in 3D via fitting a line
  - ◊ LSQ fit of 2 structures
  - ◊ Molecular Graphics
- Calculation of Volume and Surface
  - ◊ How to represent a plane
  - ◊ How to represent a solid
  - ◊ How to calculate an area
  - ◊ Docking and Drug Design as Surface Matching
  - ◊ Packing Measurement
- Structural Alignment
  - ◊ Aligning sequences on the basis of 3D structure.
  - ◊ DP does not converge, unlike sequences, what to do?
  - ◊ Other Approaches: Distance Matrices, Hashing
  - ◊ Fold Library

Topics -- Databases

- Relational Database Concepts
  - ◊ Keys, Foreign Keys
  - ◊ SQL, OODBMS, views, forms, transactions, reports, indexes
  - ◊ Joining Tables, Normalization
    - Natural Join as "where" selection on cross product
    - Array Referencing (perl/dbm)
  - ◊ Forms and Reports
  - ◊ Cross-tabulation
- Protein Units?
  - ◊ What are the units of biological information?
    - sequence, structure
    - motifs, modules, domains
  - ◊ How classified: folds, motions, pathways, functions?
- Clustering and Trees
  - ◊ Basic clustering
    - UPGMA
    - single-linkage
    - multiple linkage
  - ◊ Other Methods
    - Parsimony, Maximum likelihood
  - ◊ Evolutionary implications
- The Bias Problem
  - ◊ sequence weighting
  - ◊ sampling

Topics -- Genomics

- Expression Analysis
  - ◊ Time Courses clustering
  - ◊ Measuring differences
  - ◊ Identifying Regulatory Regions
- Large scale cross referencing of information
- Function Classification and Orthologs
- The Genomic vs. Single-molecule Perspective

- Genome Comparisons
  - ◊ Ortholog Families, pathways
  - ◊ Large-scale censuses
  - ◊ Frequent Words Analysis
  - ◊ Genome Annotation
  - ◊ Trees from Genomes
  - ◊ Identification of interacting proteins
- Structural Genomics
  - ◊ Folds in Genomes, shared & common folds
  - ◊ Bulk Structure Prediction
- Genome Trees
-

Topics -- Simulation

- Molecular Simulation
  - ◊ Geometry -> Energy -> Forces
  - ◊ Basic interactions, potential energy functions
  - ◊ Electrostatics
  - ◊ VDW Forces
  - ◊ Bonds as Springs
  - ◊ How structure changes over time?
    - How to measure the change in a vector (gradient)
  - ◊ Molecular Dynamics & MC
  - ◊ Energy Minimization
- Parameter Sets
- Number Density
- Poisson-Boltzman Equation
- Lattice Models and Simplification

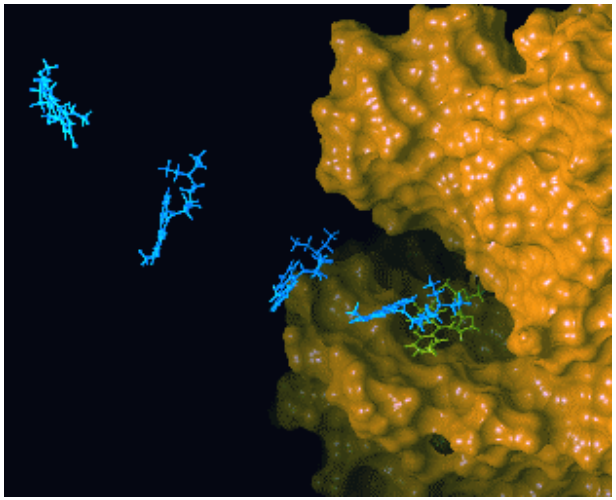
What is Bioinformatics?

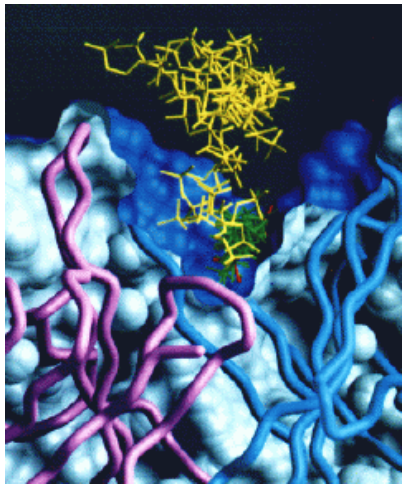
- *(Molecular)* **Bio** - **informatics**
- One idea for a definition?
  Bioinformatics is conceptualizing **biology in terms of molecules** (in the sense of physical-chemistry) and then applying **"informatics" techniques** (derived from disciplines such as applied math, CS, and statistics) to understand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

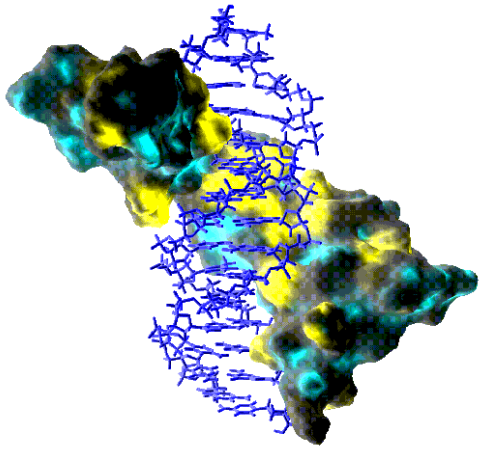
Major Application I: Designing Drugs

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

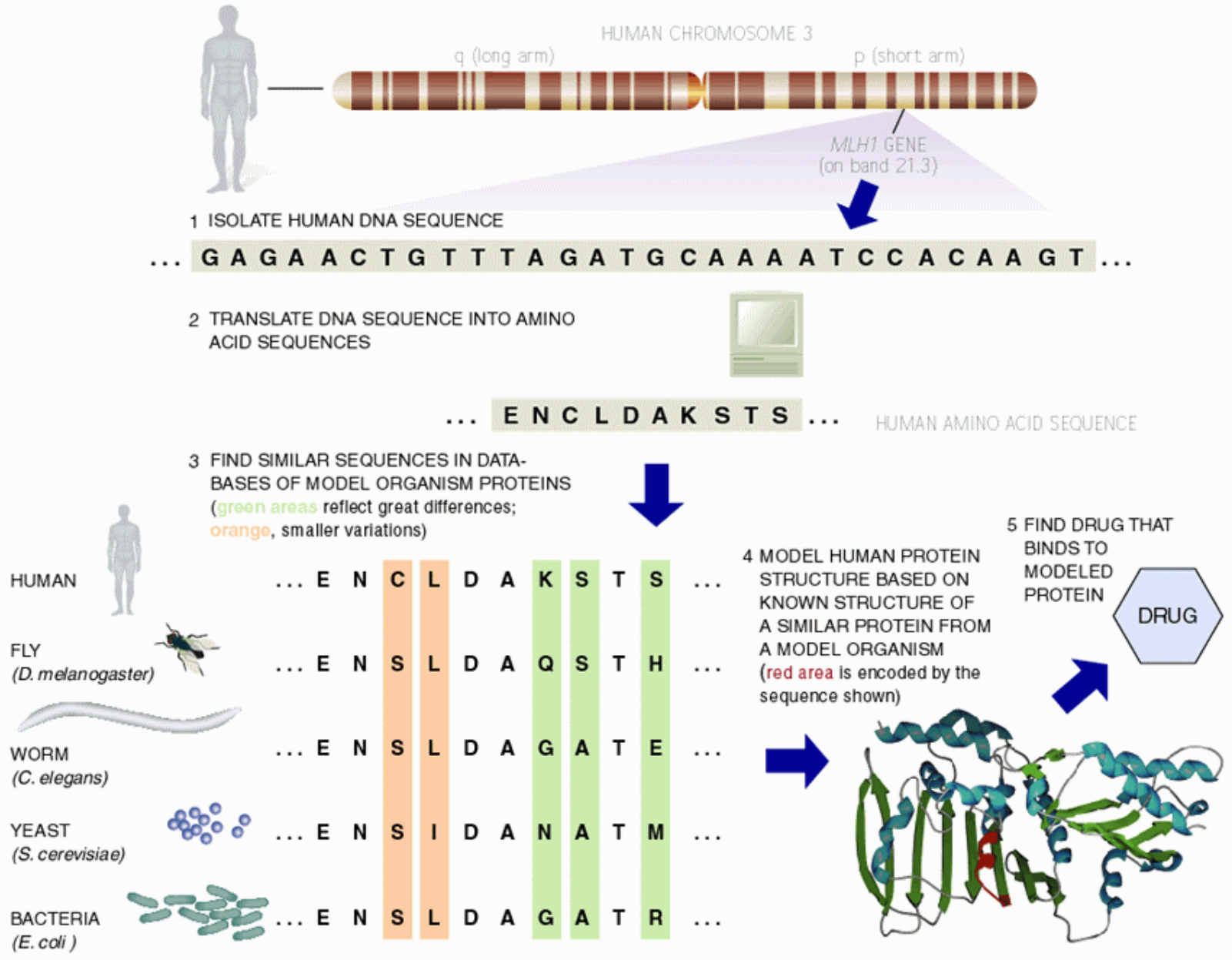
(From left to right, figures adapted from Olsen Group Docking Page at Scripps, Dyson NMR Group Web page at Scripps, and from Computational Chemistry Page at Cornell Theory Center).







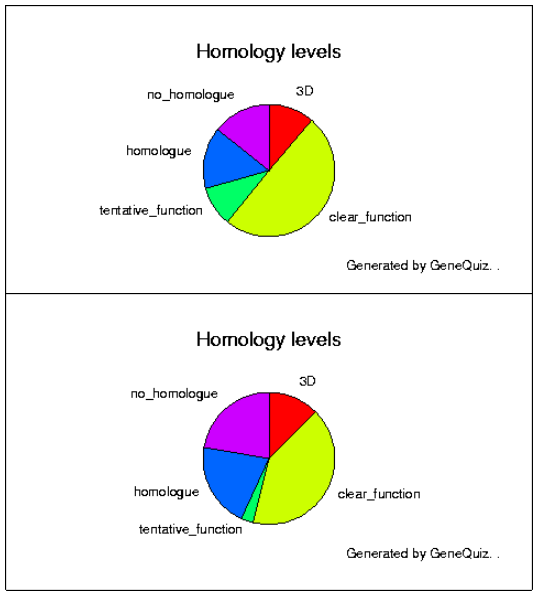
Major Application II: Finding Homologs

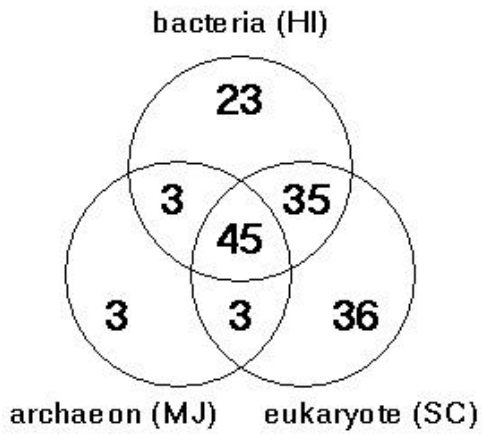


Major Application III: Overall Genome Characterization

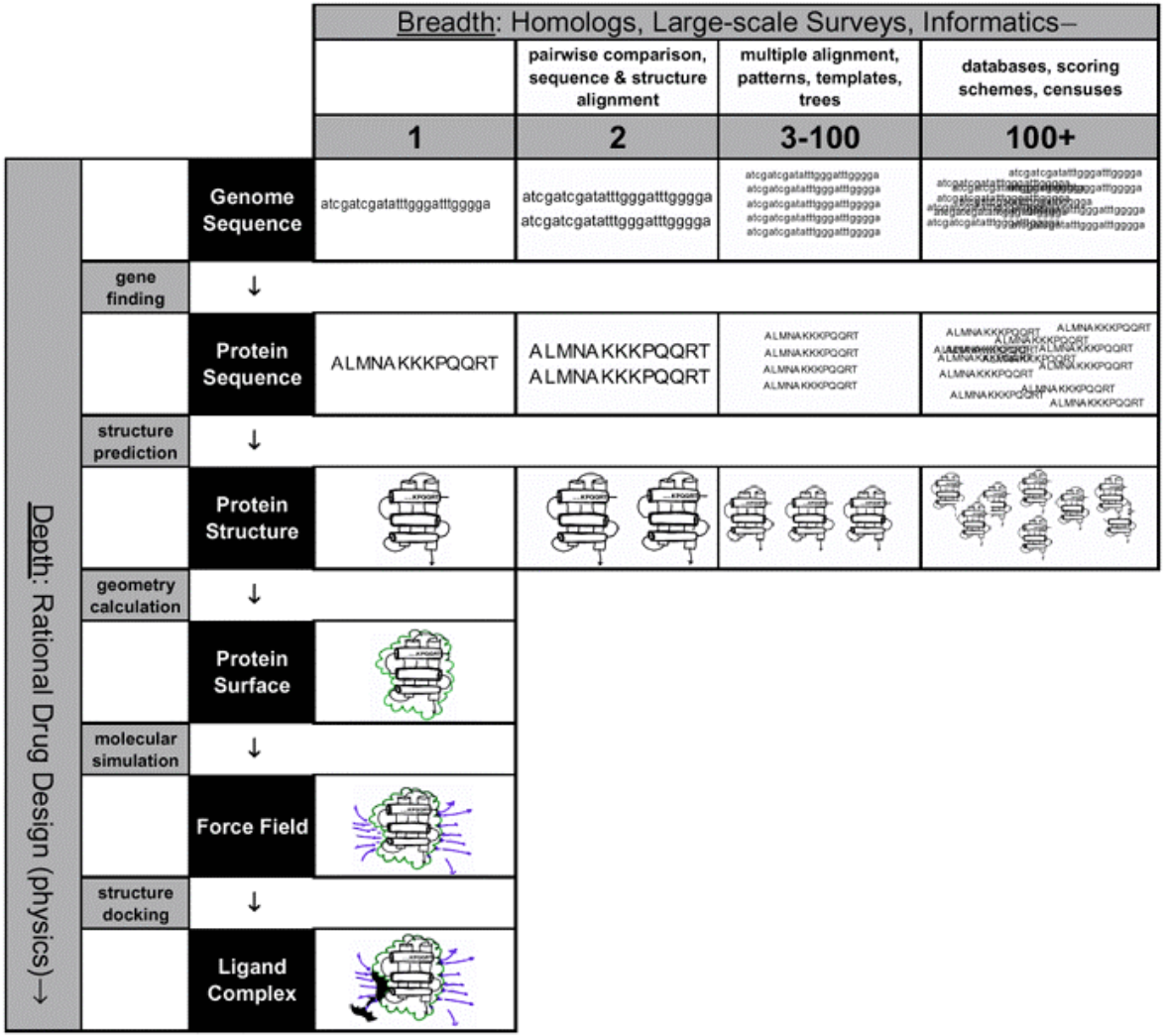
- Overall Occurrence of a Certain Feature in the Genome
  - ◊ e.g. how many kinases in Yeast
- Compare Organisms and Tissues
  - ◊ Expression levels in Cancerous vs Normal Tissues
- Databases, Statistics

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

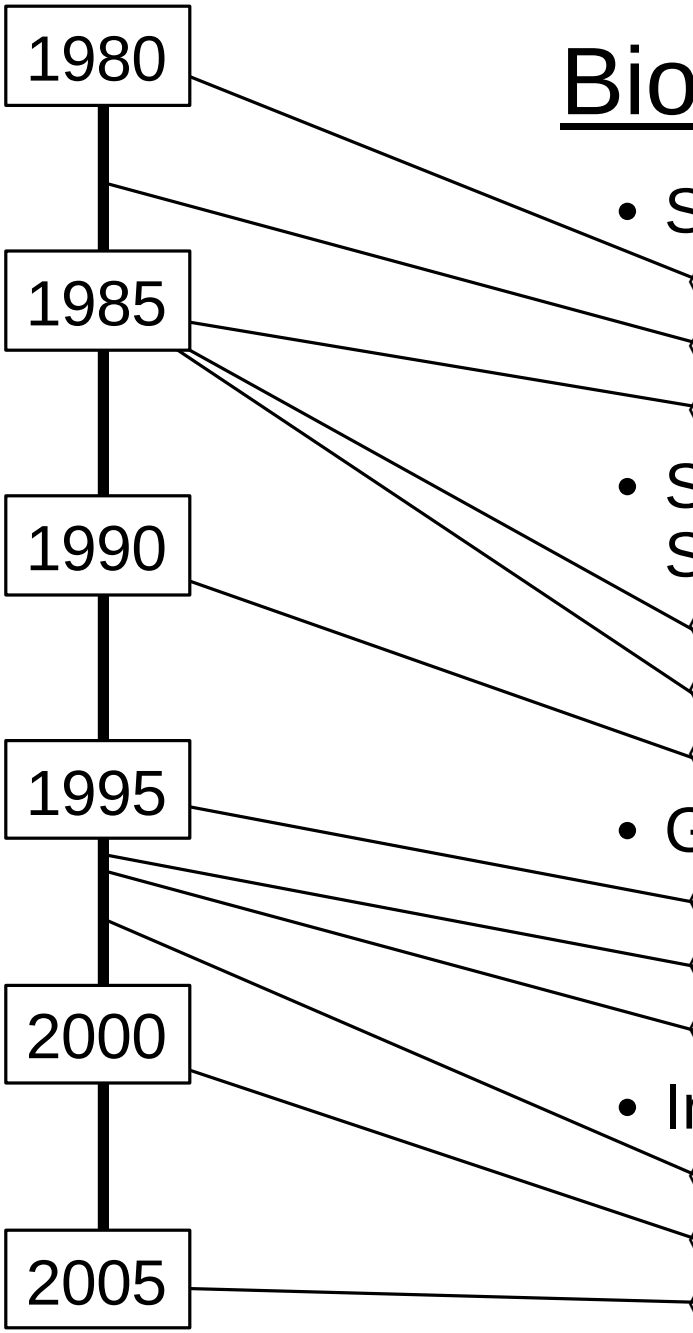




Bioinformatics Schematic



Bioinformatics - History



- Single Structures
  - Modeling & Geometry
  - Forces & Simulation
  - Docking
- Sequences, Sequence-Structure Relationships
  - Alignment
  - Structure Prediction
  - Fold recognition
- Genomics
  - Dealing with many sequences
  - Gene finding & Genome Annotation
  - Databases
- Integrative Analysis
  - Expression & Proteomics Data
  - Datamining
  - Simulation again….