LBNL-47001

ERNEST ORLANDO LAWRENCE BERKELEY NATIONAL LABORATORY

Computational Biology and High Performance Computing 2000

Horst D. Simon, Manfred D. Zorn, Sylvia J. Spengler,
Brian K. Shoichet, Craig Stewart, Inna L. Dubchak,
and Adam P. Arkin

National Energy Research
Scientific Computing Division

October 2000

To be presented at
Supercomputing 2000,
Dallas, TX,
November 6–10, 2000,
and to be published in
the Proceedings

REFERENCE COPY
Does Not
Circulate
Bldg. 50 Library - Ref.
Lawrence Berkeley National Laboratory

LBNL-47001
Copy 1

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory
is an equal opportunity employer.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBNL-47001

Computational Biology and High Performance Computing 2000

Horst D. Simon, Manfred D. Zorn, Sylvia J. Spengler, Brian K. Shoichet,
Craig Stewart, Inna L. Dubchak, and Adam P. Arkin

National Energy Research Scientific Computing Division
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

October 2000

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, Mathematical, Information, and Computational Sciences Division, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Computational Biology and High Performance Computing 2000

Tutorial M4 a.m.

November 6, 2000

SC'2000, Dallas, Texas

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Abstract

BERKELEY LAB

The pace of extraordinary advances in molecular biology has accelerated in the past decade due in large part to discoveries coming from genome projects on human and model organisms. The advances in the genome project so far, happening well ahead of schedule and under budget, have exceeded any dreams by its protagonists, let alone formal expectations. Biologists expect the next phase of the genome project to be even more startling in terms of dramatic breakthroughs in our understanding of human biology, the biology of health and of disease. Only today can biologists begin to envision the necessary experimental, computational and theoretical steps necessary to exploit genome sequence information for its medical impact, its contribution to biotechnology and economic competitiveness, and its ultimate contribution to environmental quality. High performance computing has become one of the critical enabling technologies, which will help to translate this vision of future advances in biology into reality. Biologists are increasingly becoming aware of the potential of high performance computing. The goal of this tutorial is to introduce the exciting new developments in computational biology and genomics to the high performance computing community.

Introduction

Horst Simon
HDSimon@lbl.gov
NERSC

Computational Biology and High Performance Computing

- **Presenters:**
  - Horst D. Simon
    - Director, NERSC
  - Manfred Zorn
    - Co-Head, Center of Bioinformatics and Computational Genomics, NERSC
  - Sylvia J. Spengler
    - Co-Head, Center of Bioinformatics and Computational Genomics, NERSC and Program Director, NSF
  - Craig Stewart
    - Director, Research & Academic Computing, Indiana University
  - Inna Dubchak
    - Staff Scientist, NERSC
- **Organizer:**
  - Manfred D. Zorn
- November 6, 2000

Tutorial Outline

- 8:30 a.m. - 12:00 p.m.
  - Introduction to Biology
  - Overview Computational Biology
  - DNA sequences
- 1:30 p.m. - 5:00 p.m.
  - Protein Sequences
  - Phylogeny
  - Specialized Databases

Tutorial Outline: Morning

8:30 a.m. - 8:45 a.m.	Introduction
8:45 a.m. - 10:00 a.m.	Biology
10:00 a.m. - 10:30 a.m.	BREAK
10:30 a.m. - 12:00 p.m.	Working with DNA

Tutorial Outline

- Introduction
- Brief Introduction into Biology
- DNA
  - What is DNA and how does it work?
  - What can you do with it?
- Proteins
  - What are proteins?
  - What do we need to know?
- Phylogeny
- Specialized Databases

Slide Credits

- Adam Arkin, LBNL
- Brian Shoichet, NorthWestern Univ.
- Teresa Head-Gordon, LBNL
- Sylvia J. Spengler, LBNL
- Manfred Zorn, LBNL
- Dodson-Hoagland: "The Way Life Works"
- National Museum of Health
  http://www.accessexcellence.org/
- B. Alberts et al. : "Essential Cell Biology"
  http://www.essentialcellbiology.com/
- L. Stryer: Biochemistry
- Genome Annotation Consortium
- Bob Robbins, FHCRC

Revolutionary Experimental Efforts in Biology

Sequence → Structure → Function

Sequence

Genome projects
Microbial organisms
C. elegans
Fruitfly
Human

Structure

Structural Genomics Initiative
High throughput effort underway
NIH, new beamlines
LBNL: ALS

Function

Functional Annotation Initiatives
Gene deletion projects
Yeast two-hybrid screening
Gene expression micro-arrays
In vivo GFP protein (kinetics)

Computational Biology White Paper

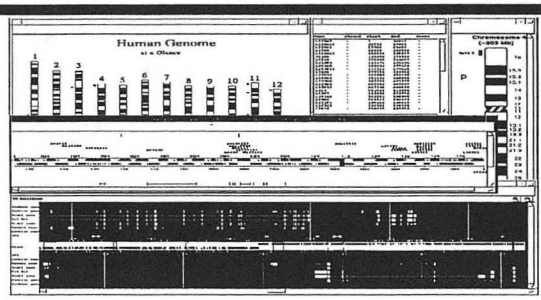
http://cbcg.lbl.gov/ssi-csb

A technical document to define areas of biology exhibiting computational problems of scale

Organization:
- Introduction to biological complexity and needs for advanced computing (1)
- Scientific areas (2-6)
- Computing hardware, software, CSET issues (7)
- Appendices

For each scientific chapter:
- illustrate with state of the art application (current generation hpc platform)
- define algorithmic kernals
- deficiencies of methodologies
- define what can be accomplished with 100 teraflop computing

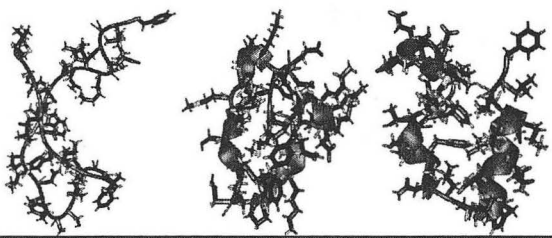
High-Throughput Genome Sequence Assembly, Modeling, and Annotation



The Genome Channel Browser to access and visualize current data flow, analysis and modeling. (Manfred Zorn, NERSC)

- Genome sequencing and annotation ⟶ Bioinformatics
  - 100,000 human genes; genes from other organism
  - Structure/functional annotation at the sequence level
- Computation to determine regions of a genome that might yield new folds
  - Experimental Structural Genomics Initiative
  - Functional annotation at the structure level by experiment

Low Resolution Fold Topologies to High Resolution Structure



One microsecond simulation of a fragment of the protein, Villin. Duan & Kollman, Science 1998

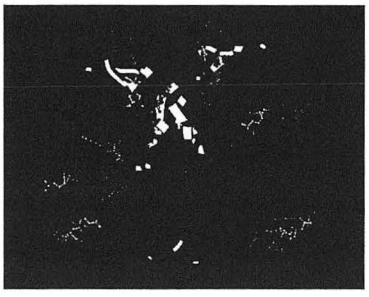
Low Resolution Structures from Predicted Fold Topology

- Fold class gives some idea of biological function, but….

Higher Resolution Structures with Biochemical Relevance

- Drug design, bioremediation, diseases of new pathogen

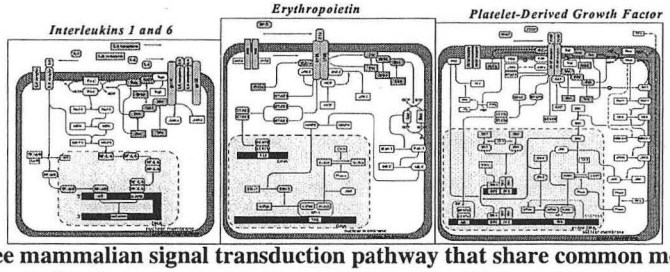
Simulating Molecular Recognition/Docking



Changes in the structure of DNA that can be induced by proteins.
Through such mechanisms proteins regulate genes, repair DNA, and carry out other cellular functions.

- Improvements in Methodology and Algorithms of Higher Resolution Structure
  - Breaking down size, time, lengthscale bottlenecks (IT^2, algorithms, teraflop computing)
- Protein, DNA recognition, binding affinity, mechanism with which drugs bind to proteins
  - Simulating two-hybrid yeast experiments
  - Protein-protein and Protein-nucleic acid docking

Modeling the Cellular Program



Three mammalian signal transduction pathway that share common molecular elements (i.e. they cross-talk). From the Signaling PAthway Database (SPAD) (http://www.grt.kyushu-u.ac.jp/spad/)

- Integrating Computational/Experimental Data at all levels
  - Sequence, structural functional annotation (Virtually all biological initiatives)
- Simulating biochemical/genetic networks to mode cellular decisions
  - Modeling of network connectivity (sets of reactions: proteins, small molecules, DNA)
  - Functional analysis of that network (kinetics of the interactions)

The Need for Advanced Computing for Computational Biology

Computational Complexity arises from inherent factors:

- 100,000 gene products just from human; genes from many other organisms
- Experimental data is accumulating rapidly
- N^2, N^3, N^4, etc. interactions between gene products
- Combinatorial libraries of potential drugs/ligands
- New materials that elaborate on native gene products from many organisms

Algorithmic Issues to make it tractable

- Objective Functions
- Optimization
- Treatment of Long-ranged Interactions
- Overcoming Size and Time scale bottlenecks
- Statistics

Introduction to Biology

Sylvia Spengler
SJSpengler@lbl.gov
NERSC

Biology

Biology

Cells

Proteins

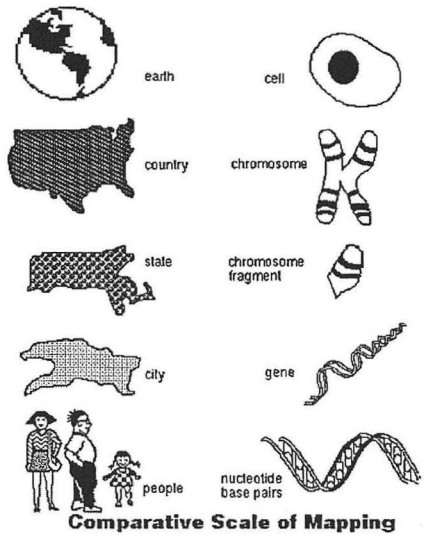
DNA

DNA

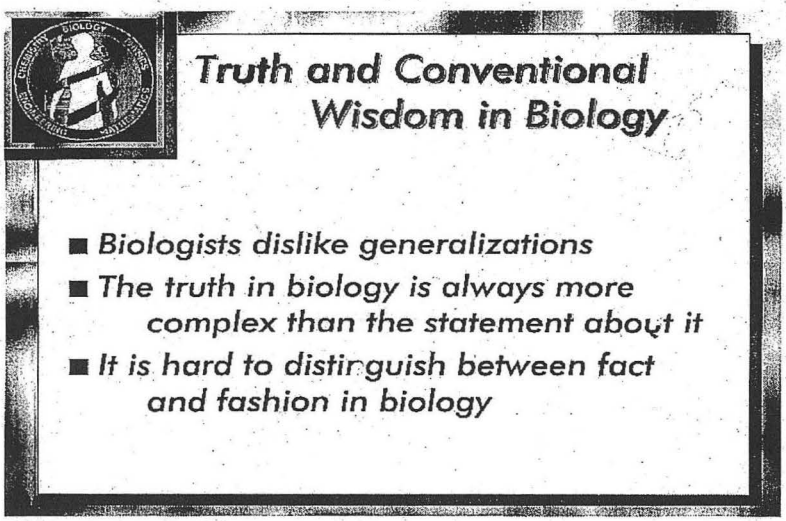
Proteins

Cells

Scale



Comparative Scale of Mapping



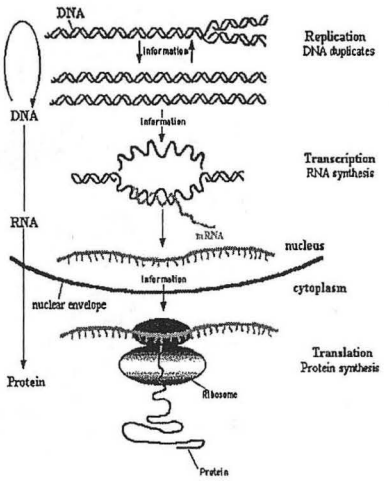
Truth and Conventional Wisdom in Biology

- *Biologists dislike generalizations*
- *The truth in biology is always more complex than the statement about it*
- *It is hard to distinguish between fact and fashion in biology*

Central Dogma

The fundamental dogma of molecular biology is that genes act to create phenotypes through a flow of information form DNA to RNA to proteins, to interactions among proteins (regulatory circuits and metabolic pathways), and ultimately to phenotypes.

Collections of individual phenotypes constitute a population.



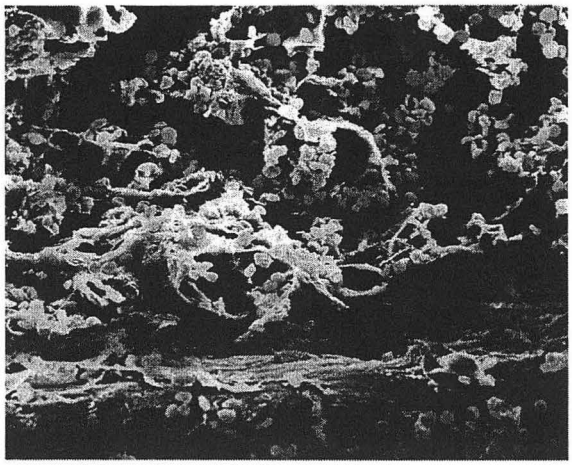
The Central Dogma of Molecular Biology

Biology is Special

Life is characterized by

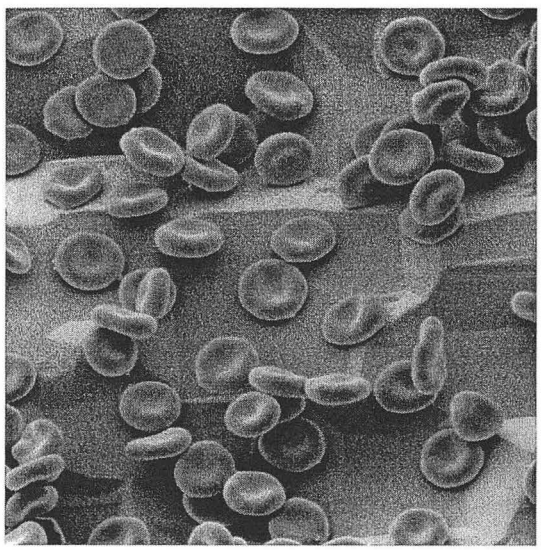
- *Individuality*
- *Historicity*
- *Contingency*
- *high (digital) information content*

No law of large numbers, since every living thing is genuinely unique.



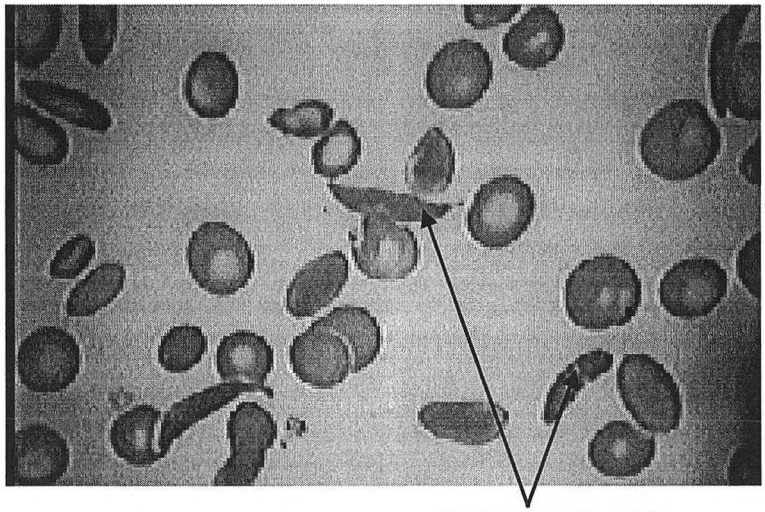
Computational Biology
@ SC 2000

Chocolate Mints?



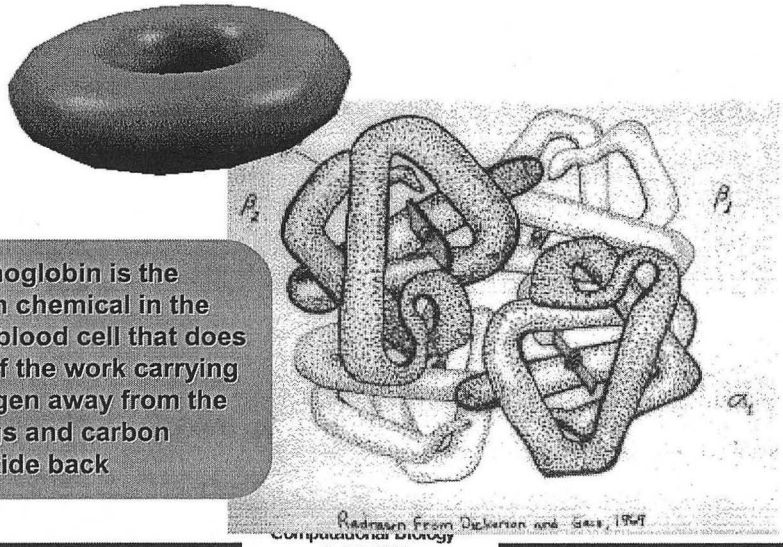
Computational Biology
@ SC 2000

Diagnosis - Blood Smear



Sickle red cells

Red Blood Cells - Hemoglobin

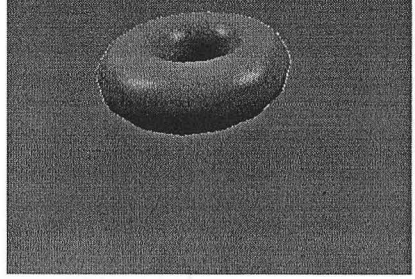


Hemoglobin is the main chemical in the red blood cell that does all of the work carrying oxygen away from the lungs and carbon dioxide back

Normal vs. Sickle Hemoglobin

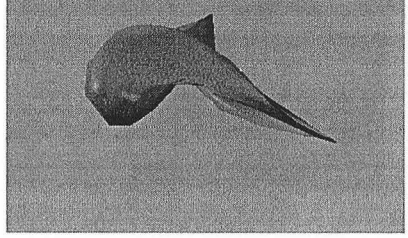
Normal

- disc-Shaped
- soft(like a bag of jelly)
- easily flow through small blood vessels
- lives for 120 days

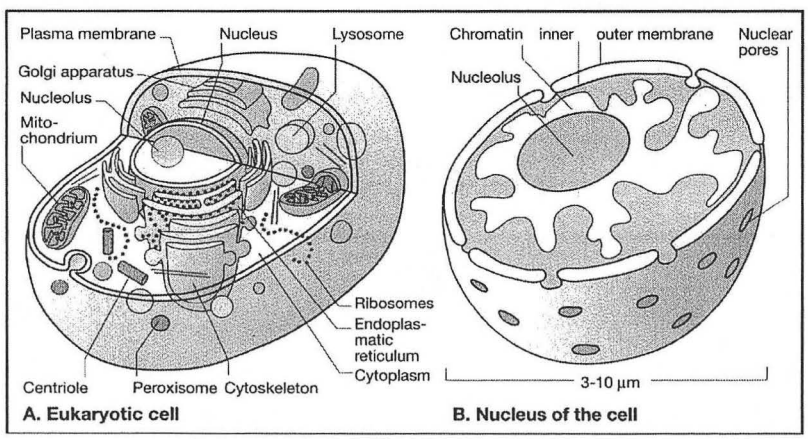


Sickle

- sickle-Shaped
- hard (like a piece of wood)
- often get stuck in small blood vessels
- lives for 20 days or less

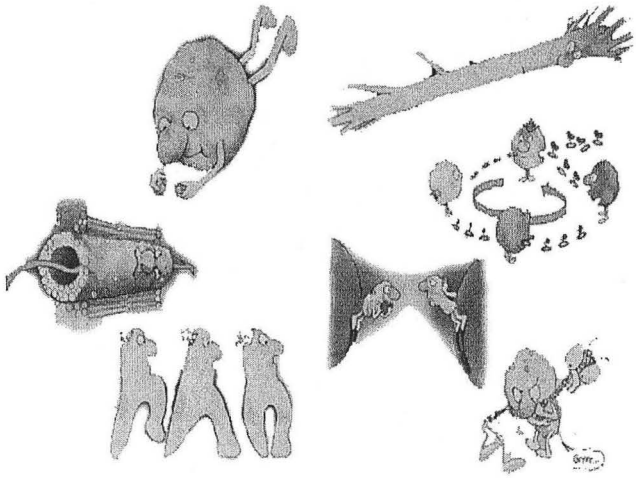


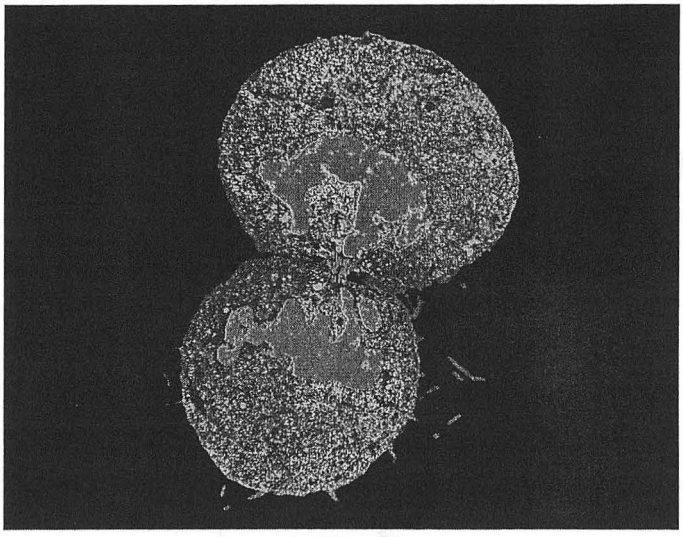
Cell Structure



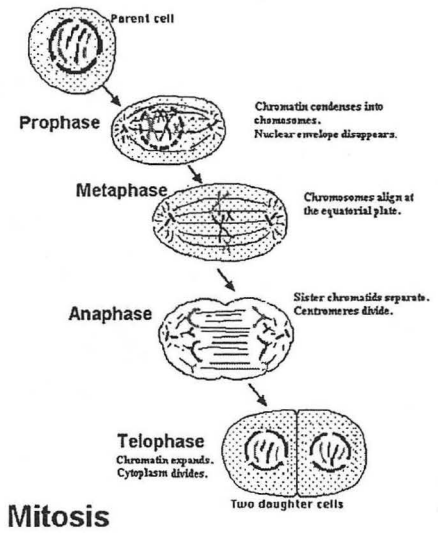
ZBD9806-01631.TIF

Protein Functions

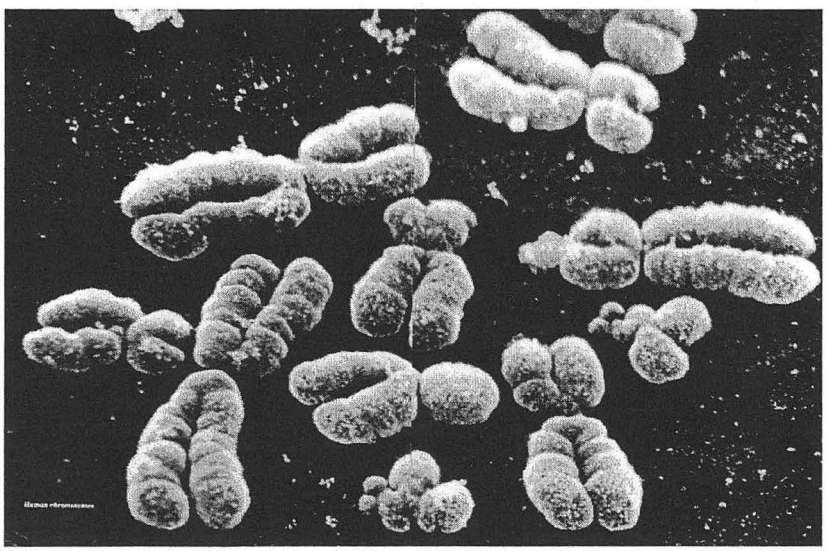




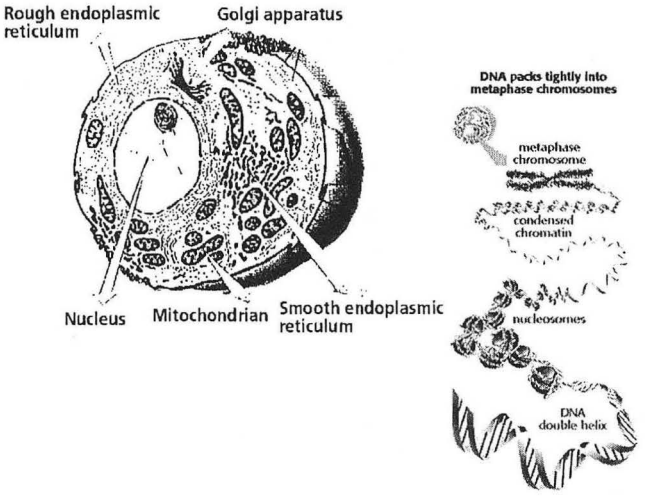
Cell Division



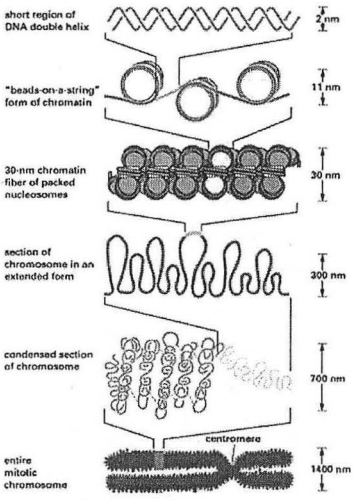
Chromosomes



Basic Biology



Scale

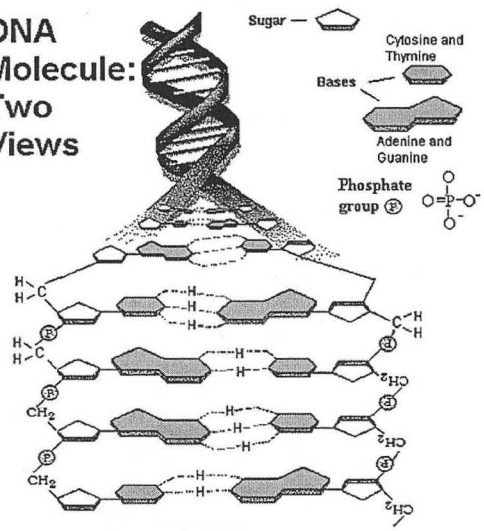


NET RESULT: EACH DNA MOLECULE HAS BEEN PACKAGED INTO A MITOTIC CHROMOSOME THAT IS 50,000x SHORTER THAN ITS EXTENDED LENGTH

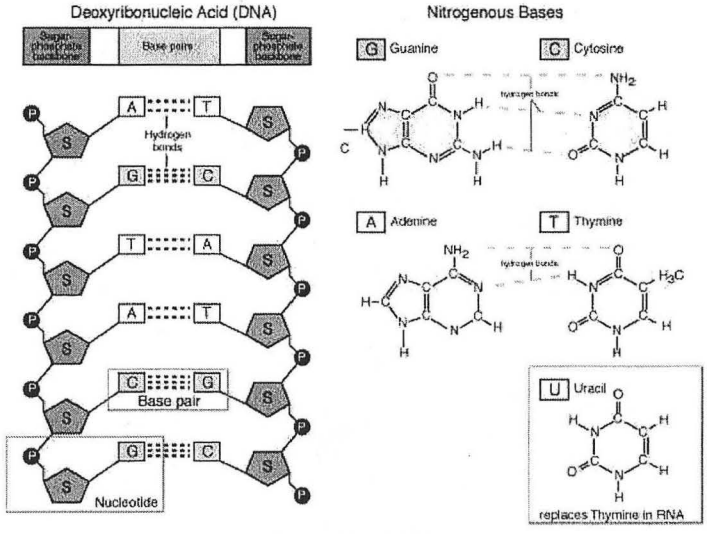
DNA

Two Views

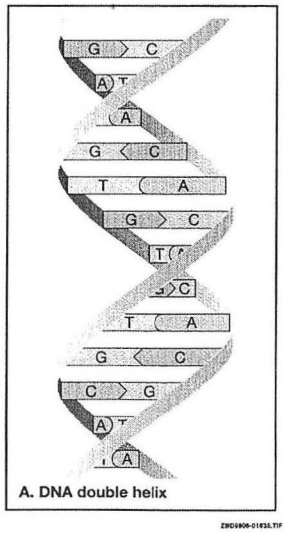
DNA Molecule: Two Views



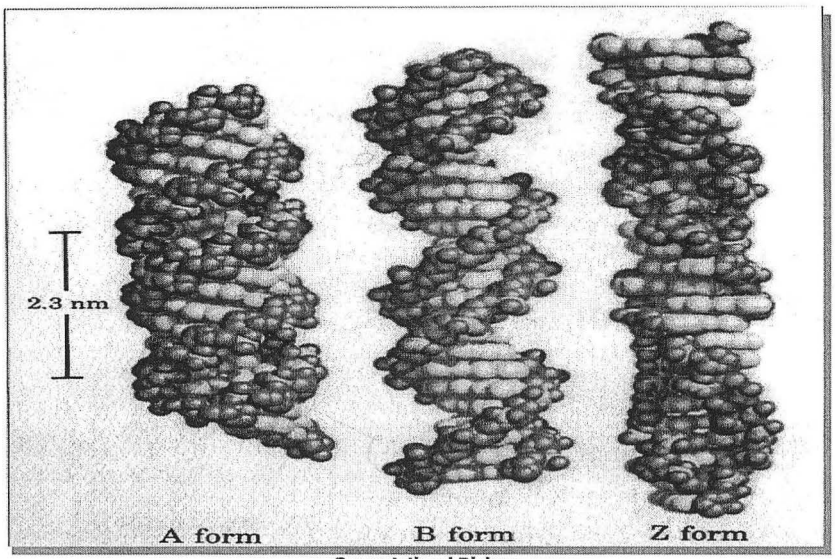
Four Bases



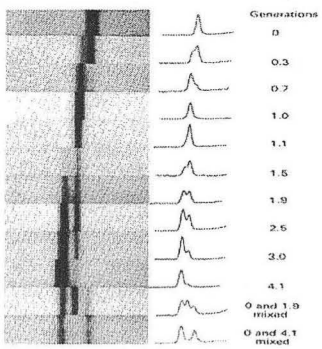
Double Helix

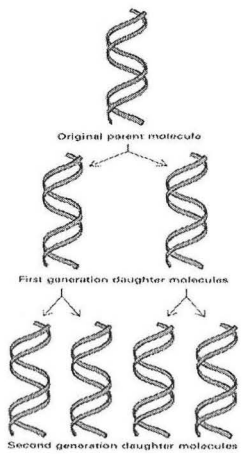


DNA



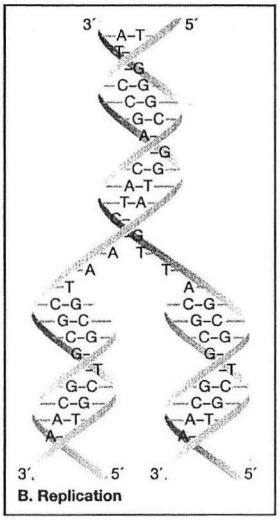
Semi-conservative Replication



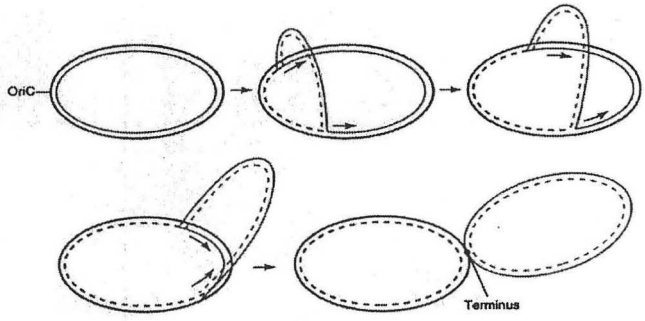


T-23

Replication

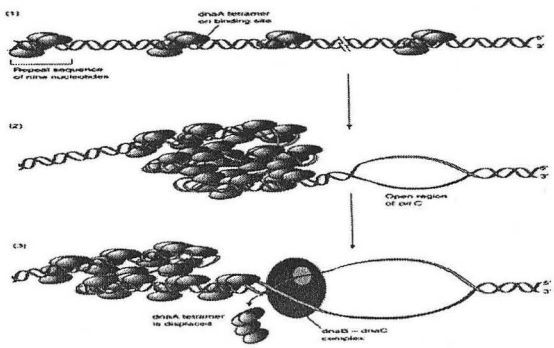


Computational Biology
@ SC 2000

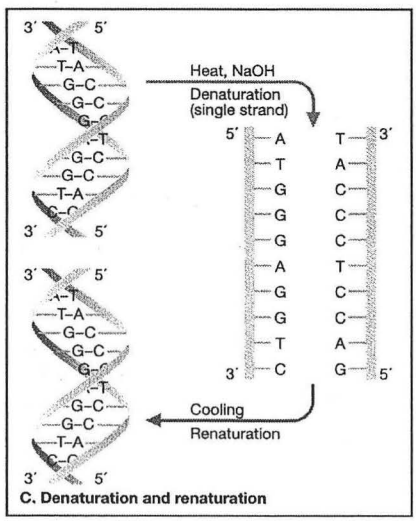


Computational Biology
@ SC 2000

DNA Replication

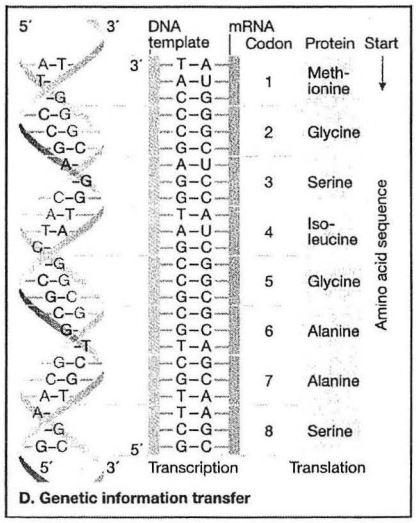


Hybridisation



C. Denaturation and renaturation

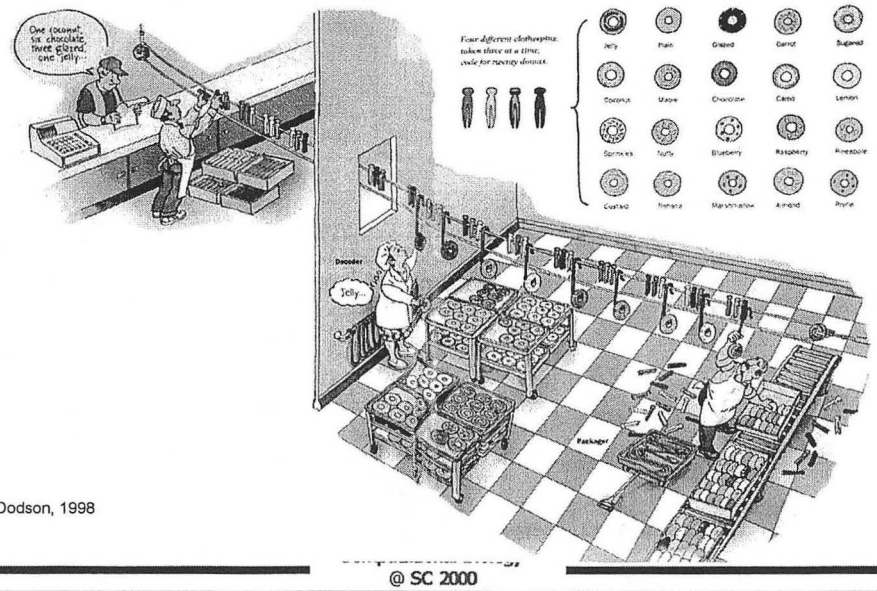
Information Transfer



D. Genetic information transfer

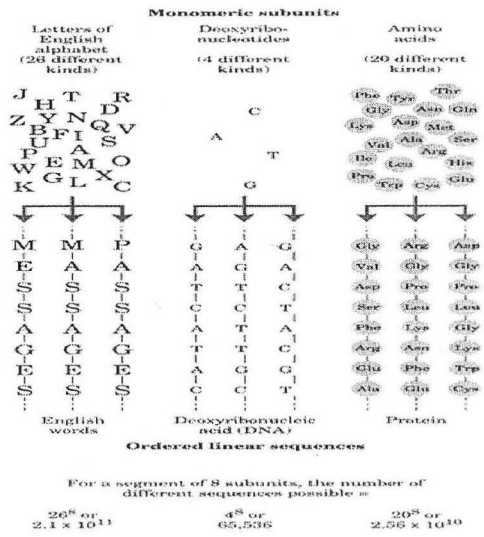
ZB09806-01438.TIF

DNA Codes

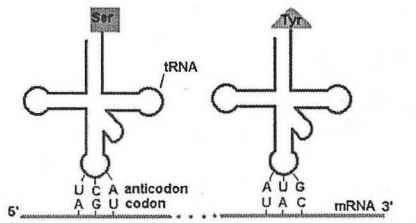


Dodson, 1998

Monomeric sub-units



Genetic Code

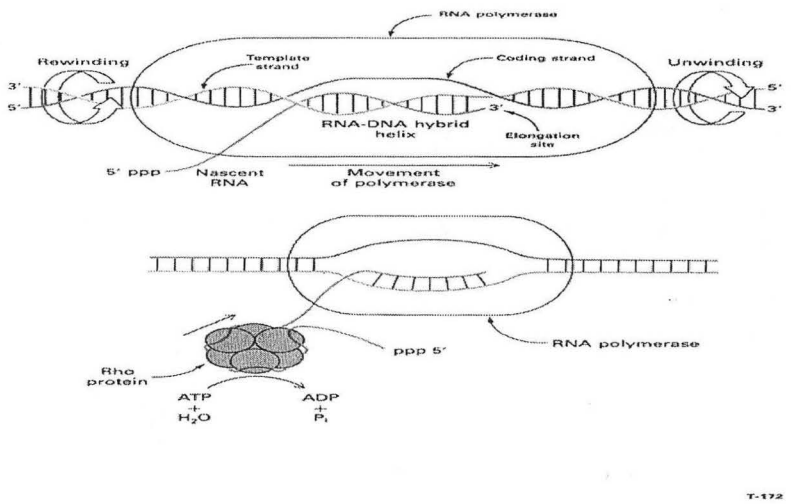


2nd base in codon

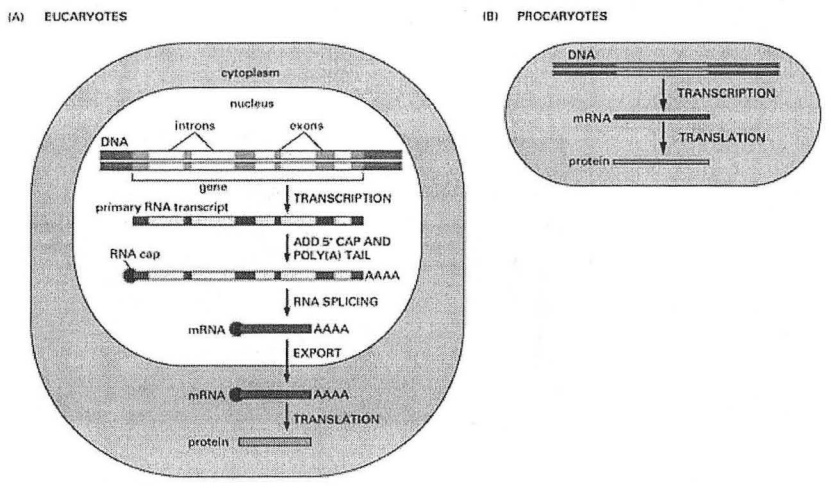
1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

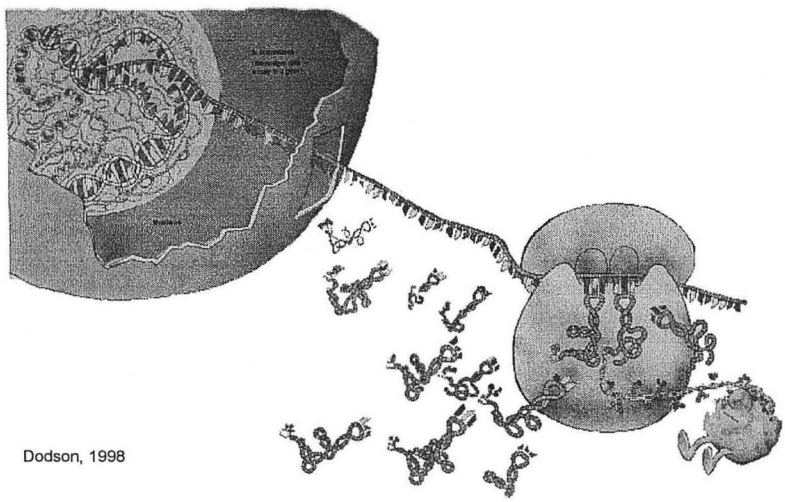
Transcription



Translation

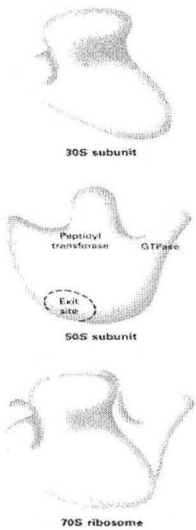


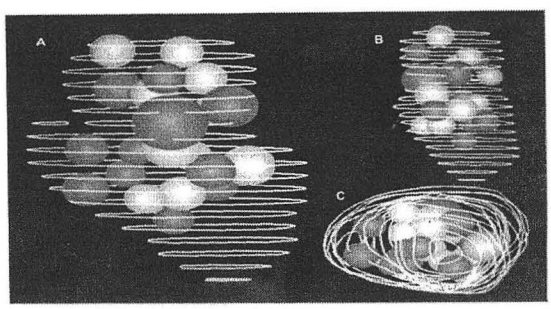
Protein Construction



Dodson, 1998

Ribosome





Ribosome

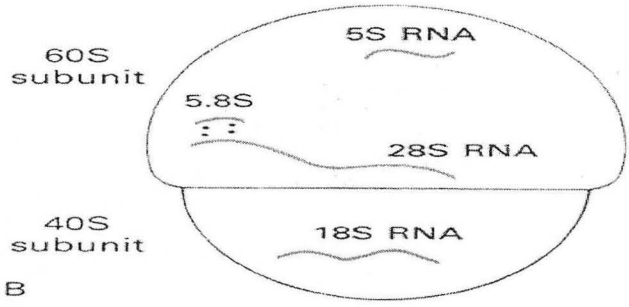


Figure 30-37
(A) Electron micrograph of eucaryotic ribosomes. [Courtesy of Dr. Miloslav Bublik.] (B) Schematic diagram of a eucaryotic ribosome.

RNA Base Pairs

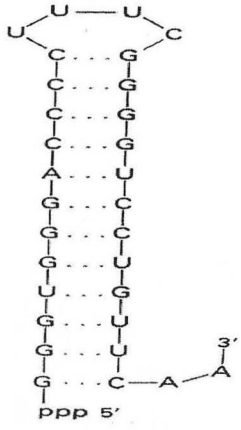


Figure 5-2
RNA can fold back on itself to form double-helical regions.

16S rRNA

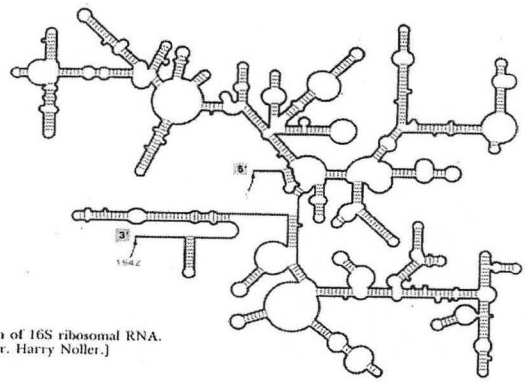


Figure 30-18
Folding pattern of 16S ribosomal RNA.
[Courtesy of Dr. Harry Noller.]

Small Subunit rRNA

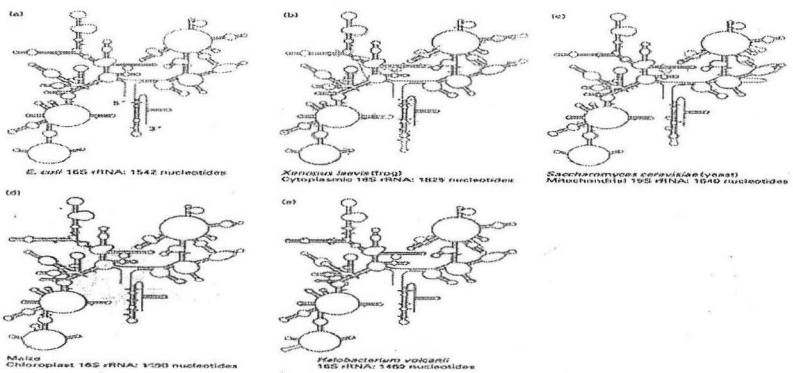
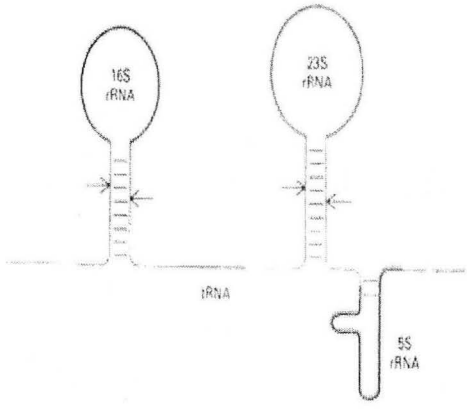


Figure 26-19 a, b, c, d, e
Darnell, Lodish, Baltimore: *MOLECULAR CELL BIOLOGY*, Second Edition
© 1990, Scientific American Books, Inc.

T-119

Cleavage by RNase III

Figure 30-19
The three ribosomal RNA molecules are derived from primary transcripts that also contain at least one tRNA molecule. Arrows mark the sites of cleavage by RNase III.



tRNA Structure

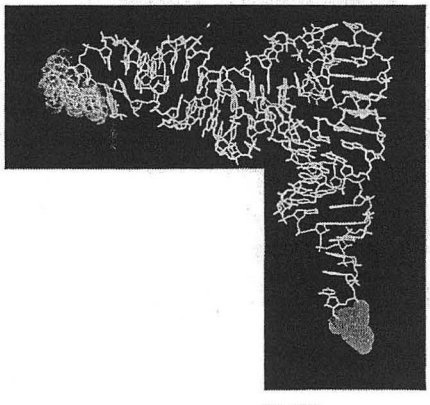
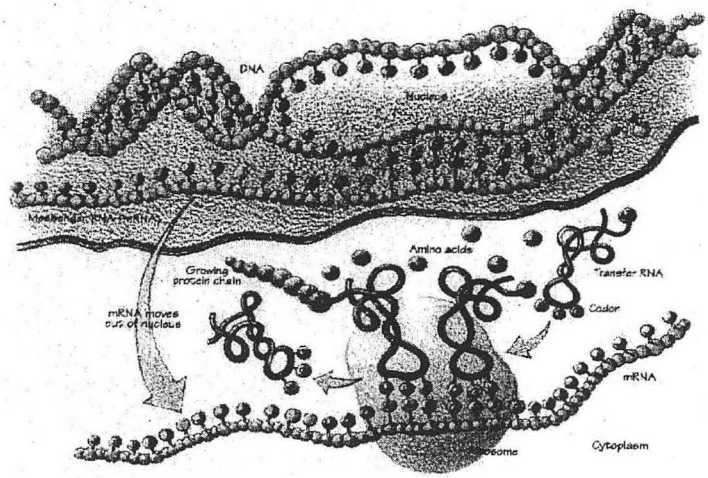


Figure 30-1
Transfer RNAs, the central molecules of protein synthesis, serve as adaptors between the four-base language of nucleic acids and the twenty-amino-acid language of proteins. The van der Waals outline of the anticodon is shown in red and the outline of the attachment site for the amino acid in blue.

Protein Synthesis



Initiation

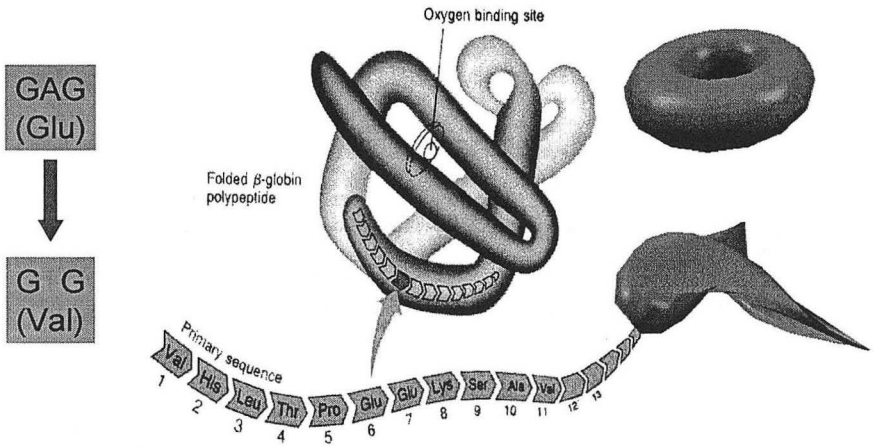
```
AGCACGAGGGGAAAUCUGAUGGAACGCUAC   E. coli trpA
UUUGGAUGGAGUGAAACGAUGGCGAUUGCA   E. coli araB
GGUAACCAGGUAACAACCAUGCGAGUGUUG   E. coli thrA
CAAUUCAGGGUGGUGAAUGUGAAACCAGUA   E. coli lacI
AAUCUUGGAGGCUUUUUUAUGGUUCGUUCU   φX174 phage A protein
UAACUAAGGAUGAAAUGCAUGUCUAAGACA   Qβ phage replicase
UCCUAGGAGGUUUGACCUAUGCGAGCUUUU   R17 phage A protein
AUGUACUAAGGAGGUUGUAUGGAACAACGC   λ phage cro
```

Pairs with 16S rRNA — Pairs with initiator tRNA

```
3'OH        G
 A          A
 U          U
            C
UCCUCCA
5'                                                  3'
GAUUCCUAGGAGGUUUGACCUAUGCGAGCUUUUAGU—Messenger RNA
                      fMet—Arg—Ala—Phe—Ser—Polypeptide
```

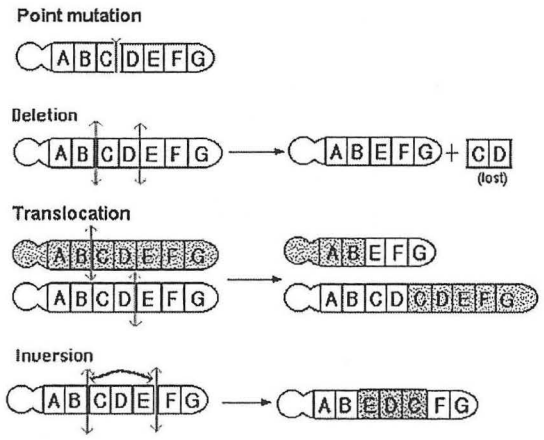
3′ end of 16S ribosomal RNA

Sickle Mutation



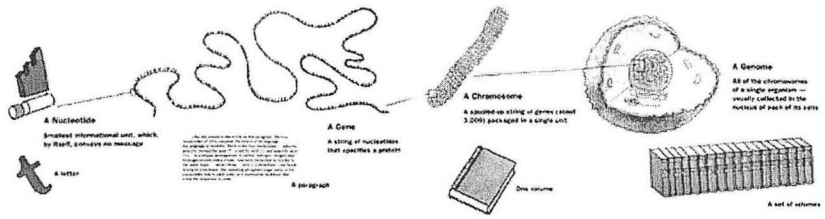
In sickle-cell hemoglobin, the Glu at position 6 is replaced by Val

Mutations



Mutations of Chromosomes

Scale



Morning Break

Nucleomics

Manfred Zorn
MDZorn@lbl.gov
NERSC

Genome Project Timeline

- 1984
  - Department of Energy and Intl. Commission on Protection Against Environmental Mutagens and Carcinogens in Alta, Utah.
- 1986
  - DOE announces Human Genome Initiative
- 1987
  - NIH Director establishes Office of Genome Research
- 1988
  - NRC Mapping and Sequencing the Human Genome
  - Berkeley Lab launches Human Genome Center
- 1990 Human Genome I

Genome Timeline cont'd

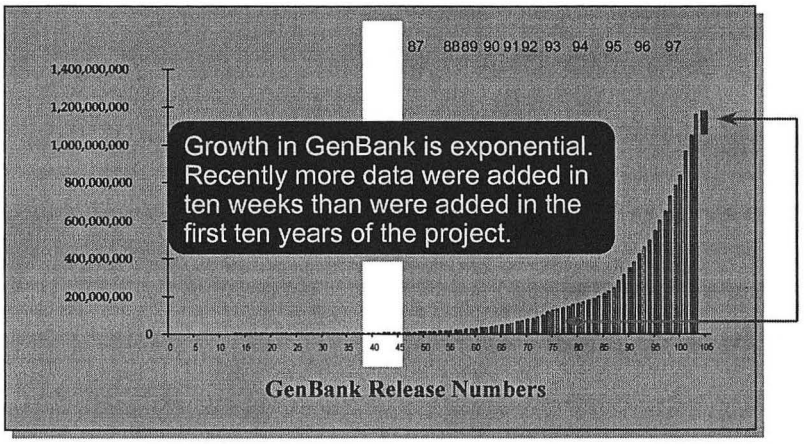
- September 1994
  - First complete map of all human chromosomes one year ahead of schedule.
- May 1995
  - First genome sequenced: H. inf.
- May 1998
  - Celera announces commercial project
  - Public effort regroups to five major centers
- June 2000
  - Joint announcement by NHGRI - Celera

We're done!

Genome Projects

1995 H. influenzae	2 Mb
1996 S. cerevisiae	12 Mb
1997 E. coli	5 Mb
1998 C. elegans	100 Mb
1999 Human Chromosome 22	34 Mb
2000 D. melanogaster	140 Mb
2000 H. sapiens	3,000 Mb

Base Pairs in GenBank



Growth in GenBank is exponential. Recently more data were added in ten weeks than were added in the first ten years of the project.

DNA Sequencing

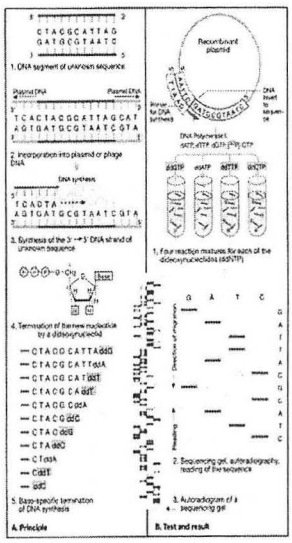
Read base code from storage medium!

- **Read length: About 600 bases at once**
- **Reader capacity**
  - **100 lanes in parallel in about 2-5 hours**
  - **1000 lanes in parallel in about 2 hours**

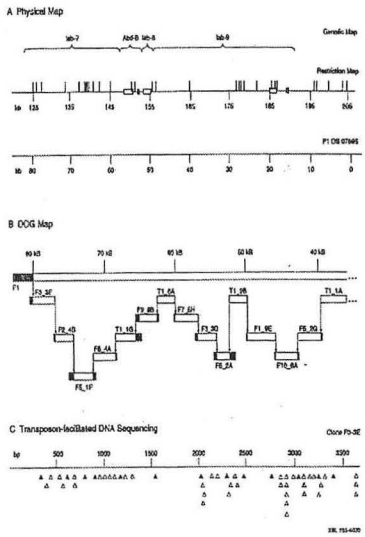
Sequencing: "bird's eye view"

- Prepare DNA
  - about a trillion DNA molecules
- Do the sequencing reactions
  - synthesize a new strand with terminators
- Separate fragments
  - by time, length = constant
- Sequence determination
  - automatic reading with laser detection systems

Sequencing



Mapping



Sequencing Strategies

Any genome is larger than amount of sequence that can be generated in a single step.

- **Shotgun**
- **Directed**
- **Finishing**

Shotgun

- Break DNA into manageable pieces
- Sequence each piece
- Use sequence to reassemble original DNA

Uniform process
Easily automatable

Coverage

$$\text{Coverage} = \frac{\text{Number} \times \text{Size of clone}}{\text{Genome size}}$$

$$\text{Expected gaps} \sim \text{Number } e^{-\text{coverage}}$$

Mapping project (Olson et al. 1986):
N=4,946
L=15,000
G=20,000,000
1,422 contigs vs. 1,457 predicted

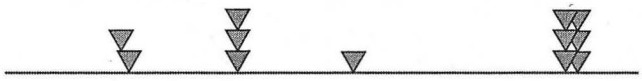
Lander-Waterman 1988

Directed

- Break DNA into manageable pieces
- Map pieces into tiling path
- Repe

Two separate processes: mapping and sequencing
More difficult to automate
Hard to integrate map information into assembly

- Transposon mediated sequencing



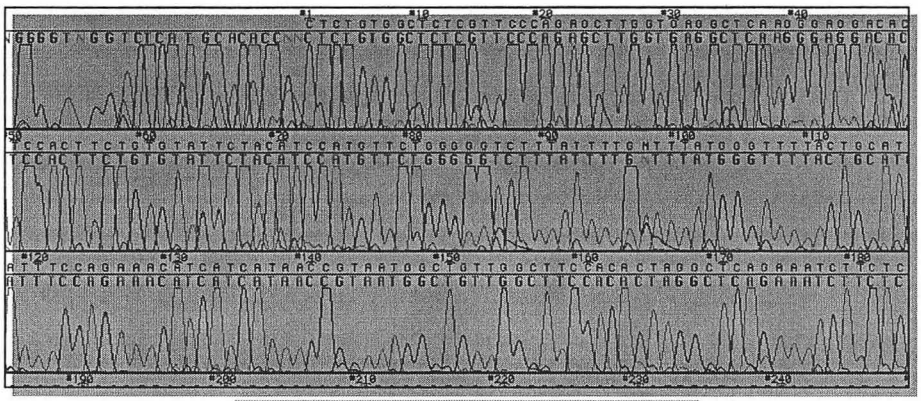
- Use maps to assemble original DNA

Finishing

- Special cases that drop out of the pipeline
- Gap closing
- Difficult stretches

- Primer walking
- Different strains, vectors, chemistry
- Creative solutions, ...

Sequence Traces



Good quality sequence needs about 10X Coverage

Base Calling

- Machine records intensities in each channel
- Vendor software translates values into smooth signal for each base
- Base calling software "calls" the sequence

- Modern base callers use peak shape, size, and spacing as well as heuristics to improve quality of calls, i.e., fewer N's and better confidence.
- Quality values carry base quality to the assembly step.

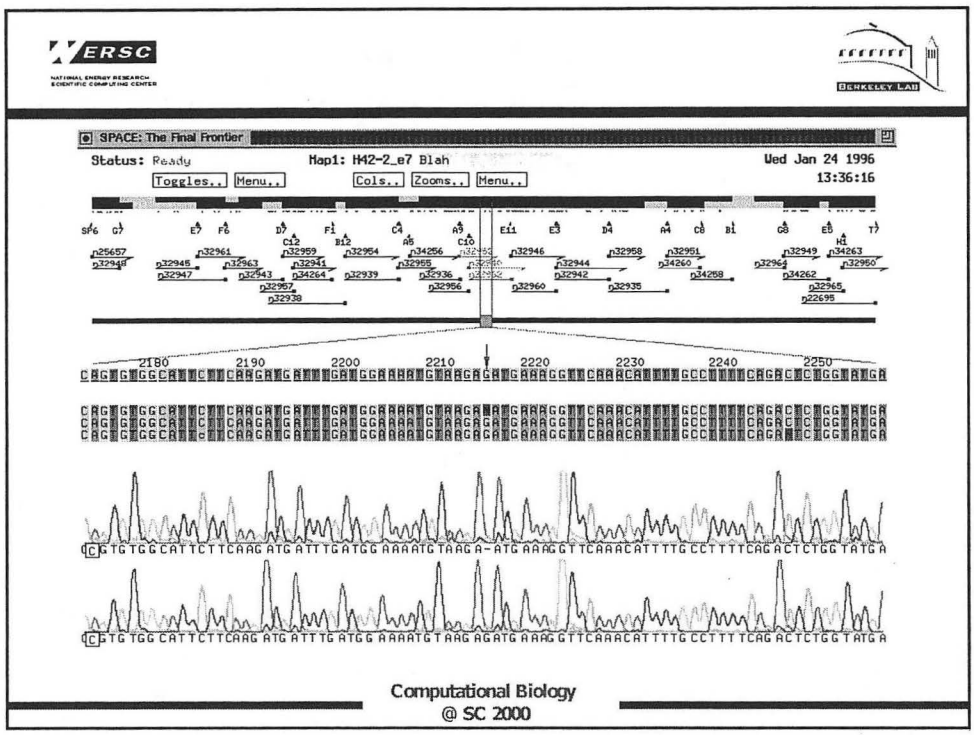
Phred - Base-caller

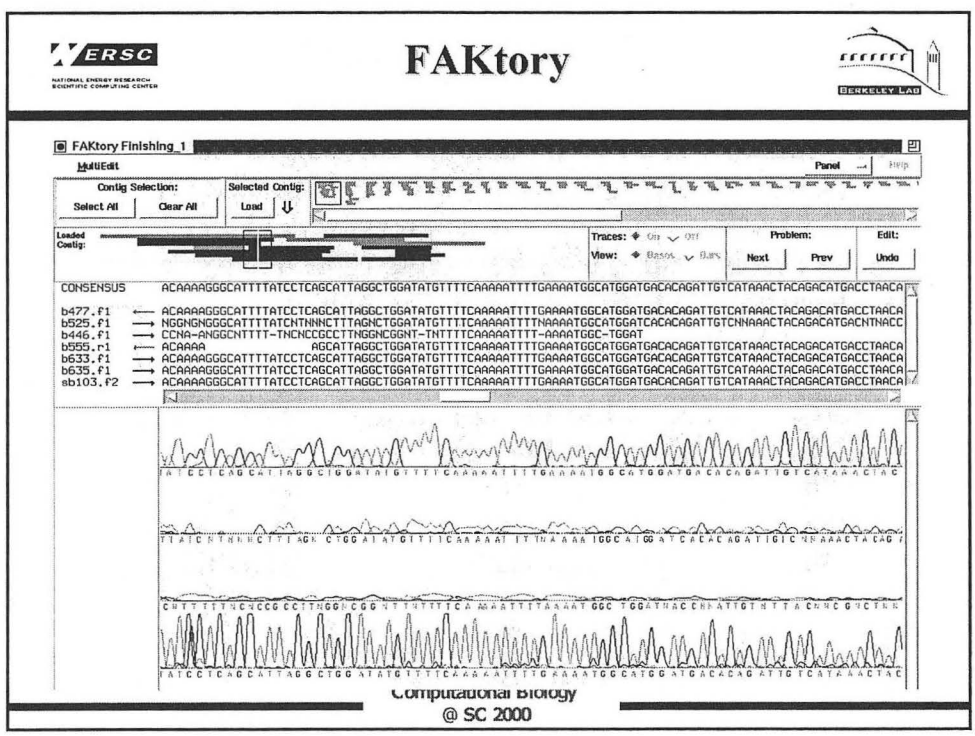
- Developed by Phil Green and Brent Ewing
- Better base calling accuracy
  - 40-50% lower error rates than ABI software on large test data sets
- Error probabilities for each base call
  - More accurate consensus sequences
  - Automatic identification of areas that require "finishing" efforts
  - Identification of repeat sequences in during assembly

Phred's quality scores

After calling bases, Phred examines the peaks around each base call to assign a quality score to each base call. Quality scores range from 4 to about 60, with higher values corresponding to higher quality. The quality scores are logarithmically linked to error probabilities.

Quality score	Probability of wrong call	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%


SPACE: The Final Frontier
Status: Ready
Map1: M42-2_e7 Blah
Wed Jan 24 1996
13:36:16
Toggles.. Menu.. Cols.. Zooms.. Menu..
Computational Biology
@ SC 2000


FAKtory
FAKtory Finishing_1
MultiEdit
Contig Selection:
Select All
Clear All
Selected Contig:
Load
Loaded Contig:
Traces:
View:
Problem:
Next
Prev
Edit:
Undo
CONSENSUS
Computational Biology
@ SC 2000

Assembly

Putting humpty-dumpty together again!

- **Overlap**
  - **Find overlapping fragments**
- **Layout**
  - **Order and orientation of fragments**
- **Consensus**
  - **Determining the consensus sequence**
- **Use of constraints**

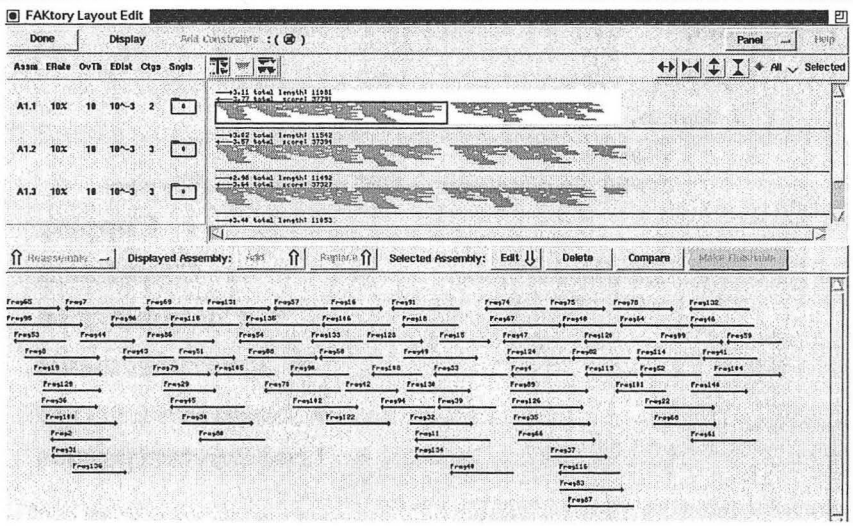
Assembly Features

- **Repeats,**
  - **repeats,**
    - **repeats,**
      - **Repeats**
      - **200 bp Alu repeat every ~4,000 bp with 5% -15% error**
- **Clipping**
- **Orientation**
- **Contamination**
- **Rearrangements**
- **Sequencing errors**
- **True Polymorphisms**

Phrap - Assembler

- Fast assemblies
  - Projects with several hundred to two thousand reads typically take only minutes
- Accurate consensus sequences from mosaic
  - Examines all individual sequences at a given position, and generally uses the highest quality sequence to build the consensus.
- Consensus quality estimates
  - Quality information of individual sequences yields the quality of the consensus sequence
  - Other available information about sequencing chemistry (dye terminator or dye primer) and confirmation by "other strand" reads used in estimating the consensus quality.

FAKtory Layout



More assembly

- Finishing: closing gaps
- Building chromosomes from large contigs that are consistent with map information

What is a Gene?

- Definition: An inheritable trait associated with a region of DNA that codes for a polypeptide chain or specifies an RNA molecule which in turn have an influence on some characteristic phenotype of the organism.

Abstract concept that describes a complex phenomenon

What is Annotation?

- Definition: Extraction, definition, and interpretation of features on the genome sequence derived by integrating computational tools and biological knowledge.

Identifiable features in the sequence

How does an annotation differ from a gene?

- Many annotations describe features that constitute a gene.
- Other annotations may not always directly correspond in this way, e.g., an STS, or sequence overlap

DNA Analysis

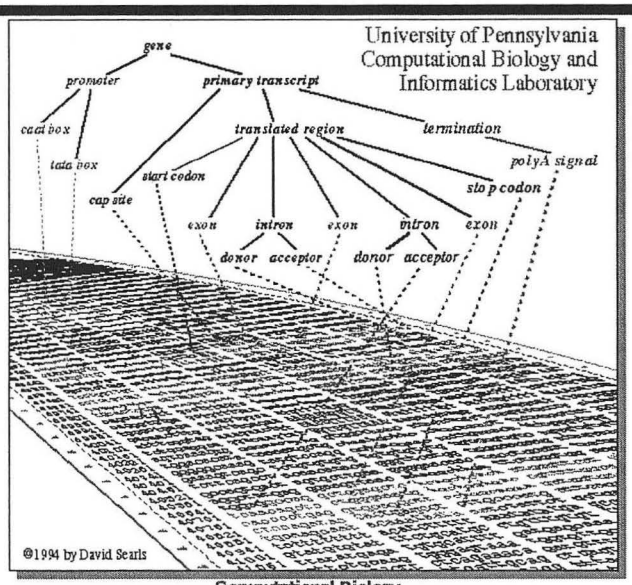
- Heuristics
- Statistics
- Artistics

DNA Analysis

Disassemble the base code!

- Find the genes
  - Heuristic signals
  - Inherent features
  - Intelligent methods
- Characterize each gene
  - Compare with other genes
  - Find functional components
  - Predict features

What is a Gene?

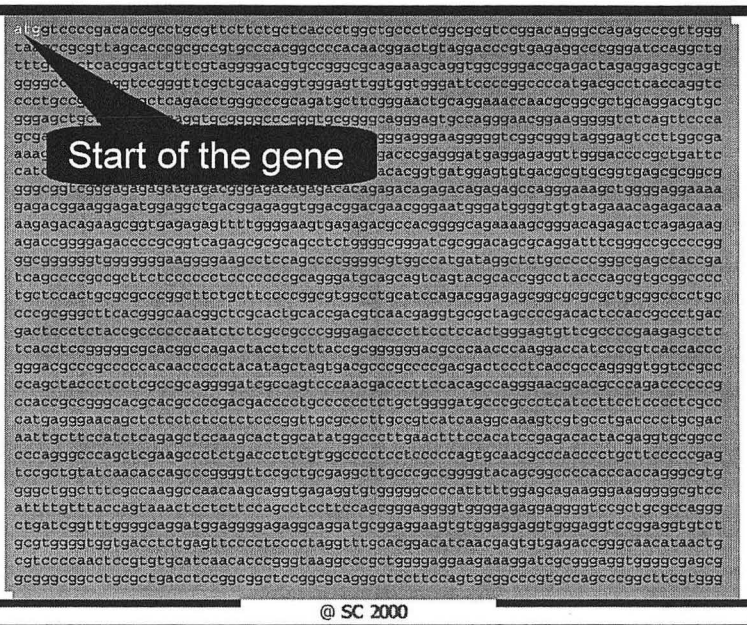


Heuristic Signals

DNA contains various recognition sites for internal machinery

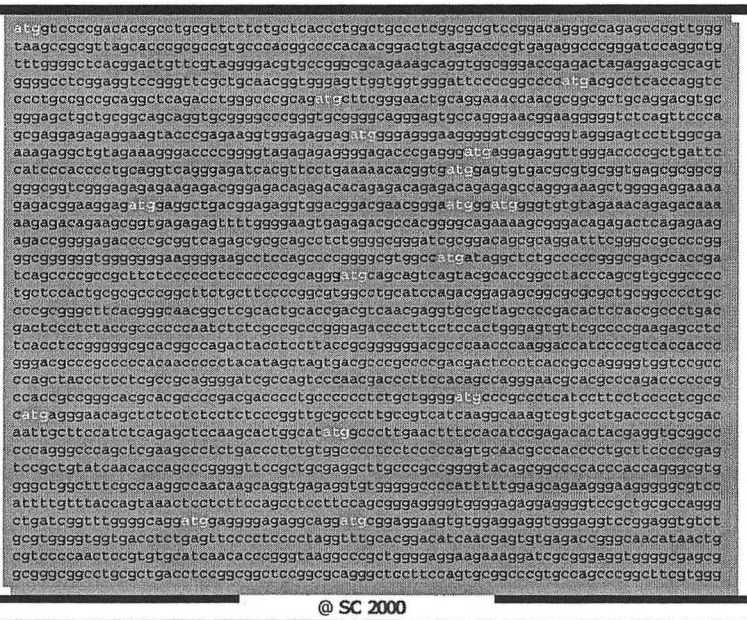
- **Promoter signals**
- **Transcription start signals**
- **Start Codon**
- **Exon, Intron boundaries**
- **Transcription termination signals**

Heuristic Signals



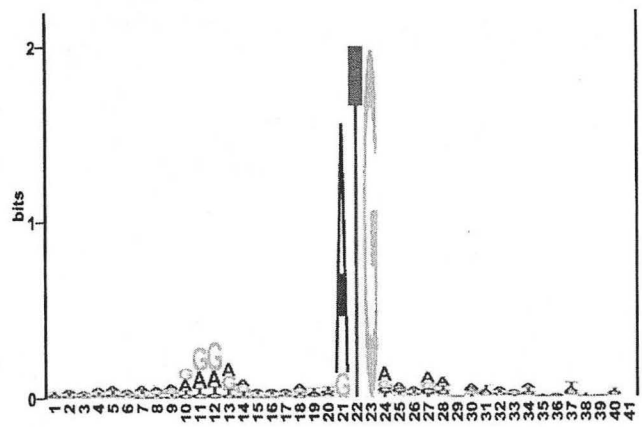
@ SC 2000

Heuristic Signals



@ SC 2000

Start Codon

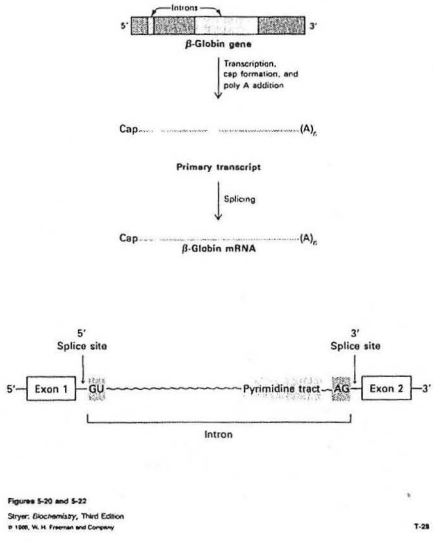


Inherent Features

DNA exhibits certain biases that can be exploited to locate coding regions

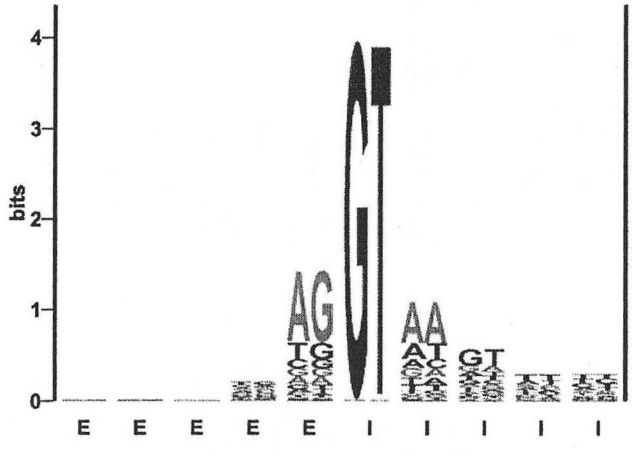
- **Uneven distribution of bases**
- **Codon bias**
- **CpG islands**
- **In-phase words**
- **Encoded amino acid sequence**
- **Imperfect periodicity**
- **Other global patterns**

Splicing

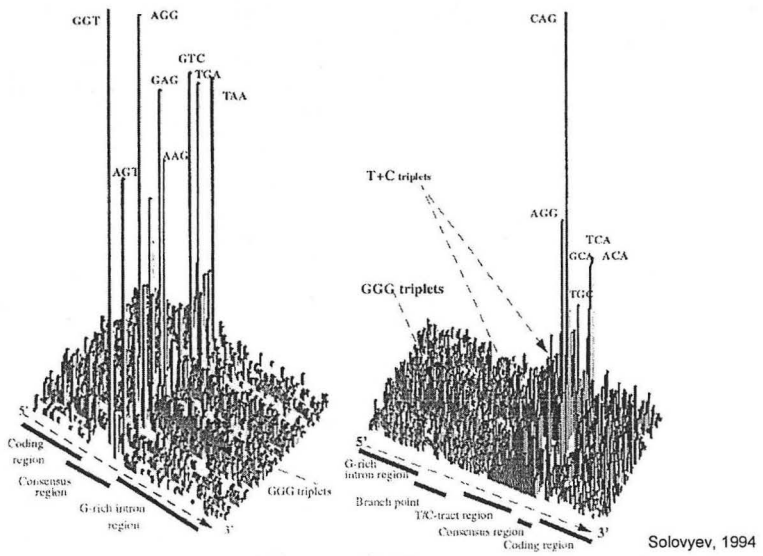


Donor Splice Site

Plate IV: A Logo of Donor Splice Sites from the Dicot Plant *A. thaliana* (cress). See page 34 for full discussion.



Inherent Features



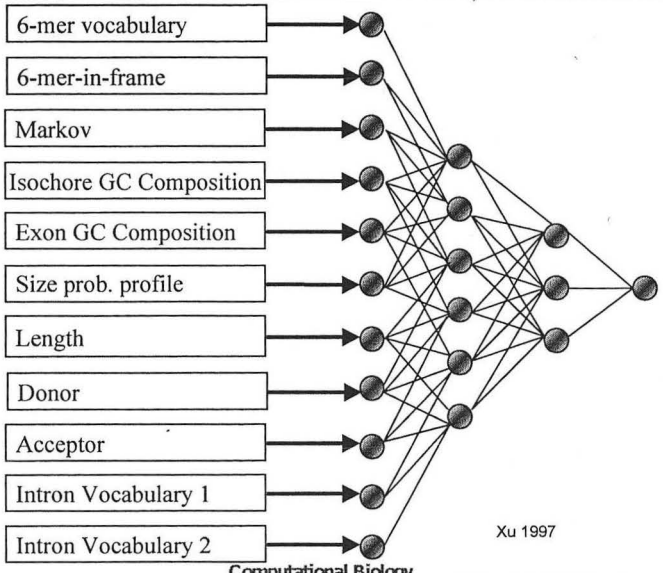
Solovyev, 1994

Intelligent Methods

Pattern recognition methods weigh inputs and predict gene location

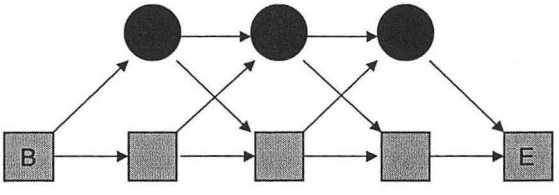
- **Neural Networks**
- **Hidden Markov Models**
- **Stochastic Context-Free Grammer**

Neural networks



Xu 1997

Hidden Markov Models



Silent states

Production states

Characterize a Gene

Collect clues for potential function

- Comparison with other known genes, proteins
- Predict secondary structure
- Fold classification

- Gene Expression
- Gene Regulatory Networks
- Phylogenetic comparisons
- Metabolic pathways

Comparison with other sequences

- Dynamic programming
  - Needleman - Wunsch
  - Smith - Waterman
  - Evolution

- Speed vs. sensitivity
  - Hashing
  - Statistical considerations
  - Suffix trees

Terminology

- Homology
  - Common ancestry
  - Sequence (and usually structure) conservation
  - Homology is not a measurable quantity, but can be inferred, under suitable conditions
- Identity
  - Objective and well defined
  - Can be quantified by several methods:
    - Percent
    - The number of identical matches divided by the length of the aligned region
- Similarity
  - Most common method used
  - Not so well defined
  - Depends on the parameters used (alphabet, scoring matrix, etc.)

Alignment

- An alignment is an arrangement of two sequences opposite one another
- It shows where they are different and where they are similar
  We want to find the optimal alignment - the most similarity and the least differences

Alignment

- Alignments have two aspects:
  - Quantity: To what degree are the sequences similar (percentage, other scoring method)
  - Quality: Regions of similarity in a given sequence

How is an alignment done?

- When we compare sequences, we take two strings of letters (nucleotides or amino acids) and align them.
- Where the characters are identical, we give them a positive score, and where they differ, a negative value.
- We count the identical and nonidentical characters, and give the alignment a score (usually called the quality)

Dynamic Programming

- Sequence A $A = (A_1, \ldots A_m)$
- Sequence B $B = (B_1, \ldots B_n)$
- Substitution $\omega(A_i, B_j)$
- Deletion $\omega(A_i, \Delta)$
- Insertion $\omega(\Delta, B_j)$
- Matrix Element

$$H_{i,j} = \max \begin{Bmatrix} H_{i-1,j-1} + \omega_{A_i,B_j} \\ H_{i,j-1} + \omega_{A_i,\Delta} \\ H_{i-1,j} + \omega_{\Delta,B_j} \end{Bmatrix}$$

Differences in the sequence can be caused by deletions or insertions in the DNA, or by point mutations. These changes can be seen at the protein level as well (changes in the translation of the protein

This scheme works fine as long as you assume that all possible mutations occur at the same frequency. However, nature doesn't work this way. It has been found that in DNA, transitions occur more often than transversions.

Scoring Matrices

- Identity scoring
- Genetic code scoring
- Physical chemical similarities
- Observed substitutions
  - Dayhoff matrix (PAM)
  - BLOSUM

The Gap Penalty

Consider the two following alignments:

```
VITKLGTCVGS     VITKLGTCVGS
VIT...TCVGS    V.TK.GTCV.S
```

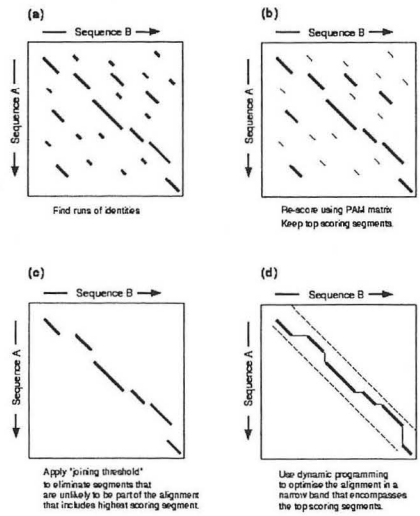
According to the algorithm these 2 cases will get the same gap penalty. However nature is different. In most cases insertions/deletions are longer than a single residue, even for very homologous sequences.

- To compensate for this, and to differentiate between cases like the one above, the gap penalty is made up of two factors:
  - The gap creation penalty - subtracted from the alignment quality whenever a gap is opened.
  - The gap extension penalty - subtracted from the alignment quality according to the length of the gap.

- Thus we have:
  - Quality = matches - (mismatches + gap penalty)
  - Gap penalty = gap creation penalty + (gap extension penalty X gap length)

FASTA

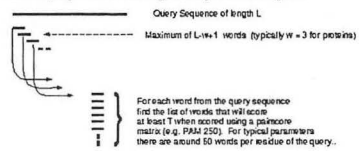
FASTA Algorithm



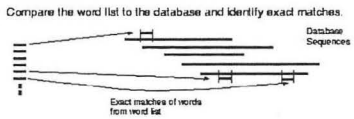
BLAST

BLAST Algorithm

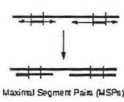
(1) For the query find the list of high scoring words of length w.



(2) Compare the word list to the database and identify exact matches.



(3) For each word match, extend alignment in both directions to find alignments that score greater than score threshold S.

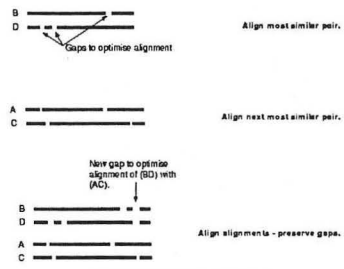


Multiple Alignments

Steps in Multiple Alignment

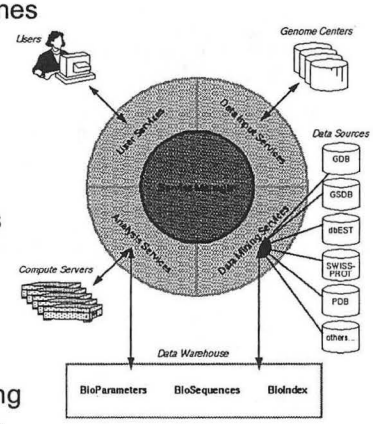
(A) Pairwise Alignment

(B) Multiple alignment following the tree from A.

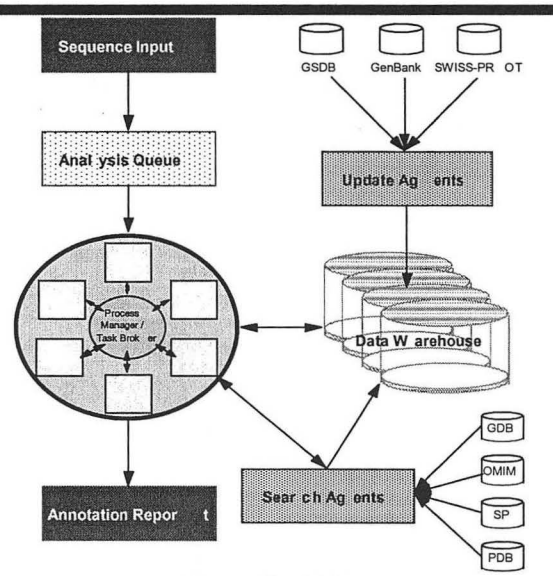


Large-scale Genome Annotation

- Multi-laboratory Project
- Standard Annotation of Genomes
  - Genome Channel
  - Genome Catalog
- Comprehensive integration of
  - Analysis tools
  - Data management systems
  - Data mining
  - User services
- Extensible Framework
  - High-performance computing
  - Data integration technology
  - Artificial intelligence



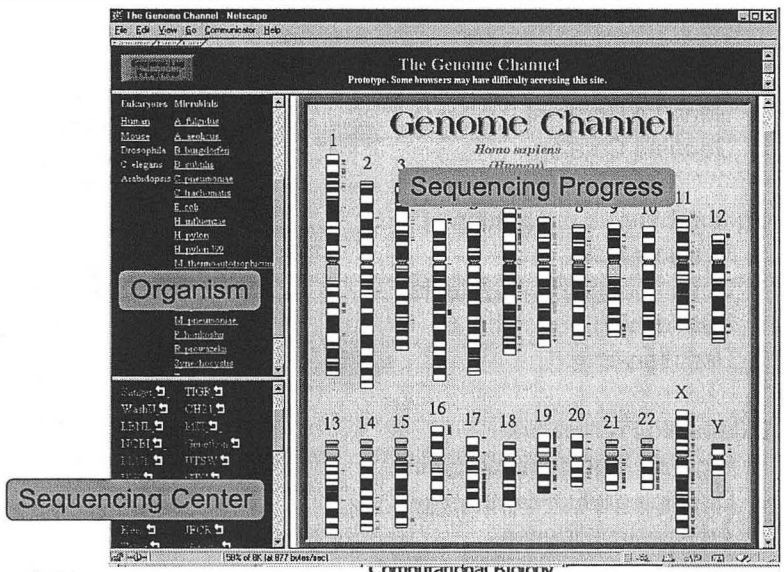
Annotation Pipeline



Computational Biology
@ SC 2000

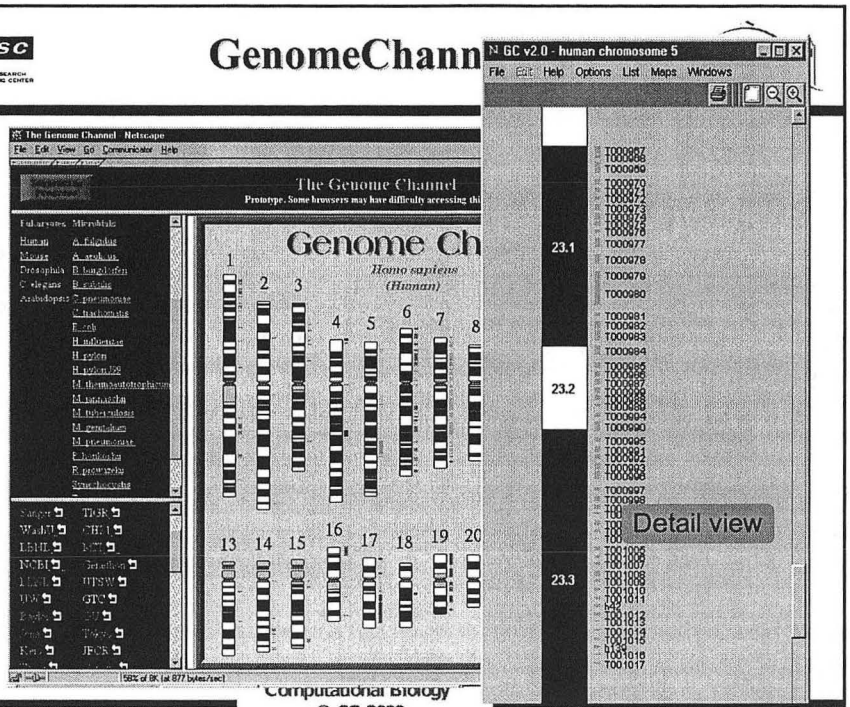
Data Sour ces

GenomeChannel



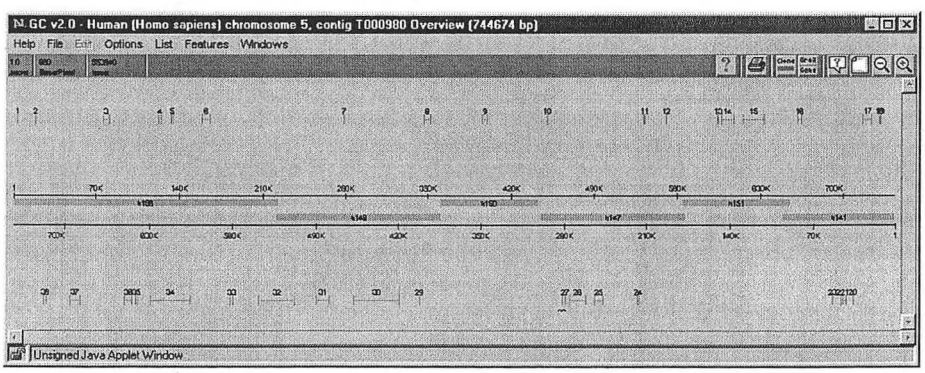
Computational Biology
@ SC 2000

GenomeChann



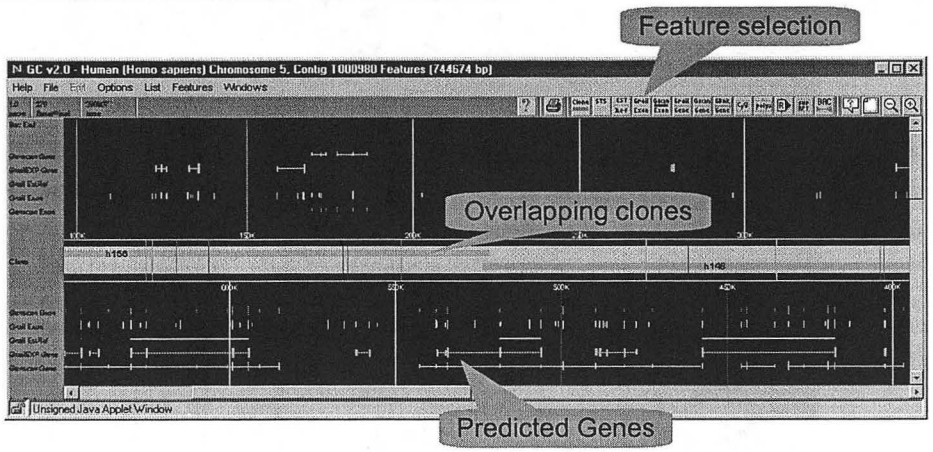
Computational Biology
@ SC 2000

A Contig Overview



Computational Biology
@ SC 2000

Feature Display



Gene Summary Report

Genome Channel Report - Netscape

File Edit View Go Communicator Help

Genome Channel | Gene Summary Report | Help

human Chromosome 5, Contig T000980, GrailEXP Gene 32

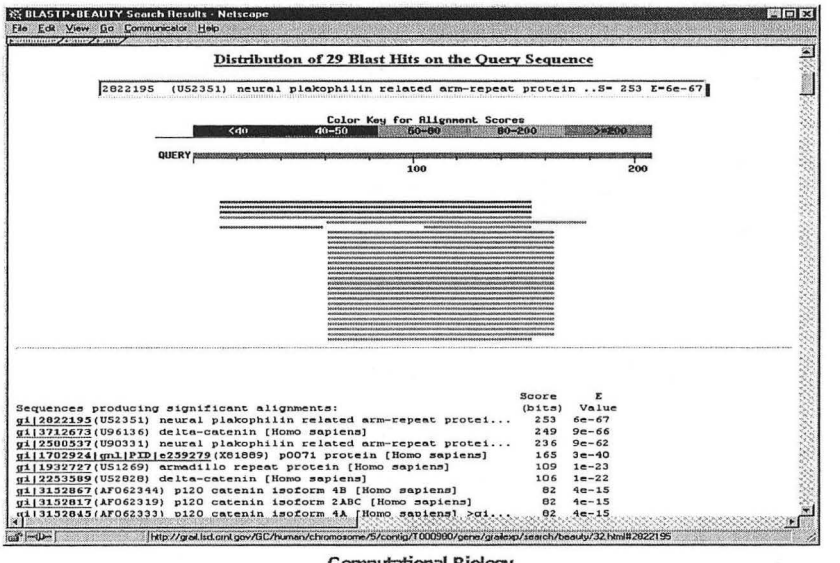
Links: Protein

```
TYPE          gene
SIZE          31475 bp
ORGANISM*     Homo sapiens
CHROMOSOME*   5
MAP*          5q23.1
GCID          GC05241232 (chr.5.ctg.T000980.gene.grailexp.32)
SIMILARITY    (U52351) neural plakophilin related arm-repeat protei...
FROM_ACC
FROM_NID
SEQ_SOURCE
FEATURES*           Location/Qualifiers
     gene           join(<1..111,12191..12394,28123..28277,28805..28942,
                    31445..31475>)
                    /similarity="(U52351) neural plakophilin related arm-repeat protei..." (blast_score=
                    /evidence="not experimental"
                    /translation=MGTDELDGLLCGEANGKDAESSGCWGKKKKKKKSQDQKFALLF
                    FFRQWDGVGPLPDCAEPPKGIQMLWHPSIVKPYLTLLSECSNPDTLEGAAGALQNL
                    AAGSWKWSVYIRAAVRKEKGLPILVELLRIDNDRVVCAVATALRNMALDVRNKELI
                    GHPVLGPEIKSISKTDRKPCPCGVYIQEKEMFKKQTKQNNHKMELSKGWEDAKAKA
                    D*
     exon           (1..111)
                    /EST=T77214
     exon           (12191..12394)
                    /EST=T77214
     exon           (28123..28277)
     exon           (28805..28942)
     exon           (31445..31475)

BASE COUNT    9381 a     6543 c     6139 g     9412 t
ORIGIN
        1 atgggcacgg acgagctgga cgggctactc tgtggcgagg ccaatggcaa ggatgctgag
       61 agctctgggt gctggggcaa gaagaagaag aaaaagaaat cccaagatca ggtgatgcag
      121 gctgctttgca gctgcatgca attatcccct tcctaattgc aactgtaata tatcctcaat
```

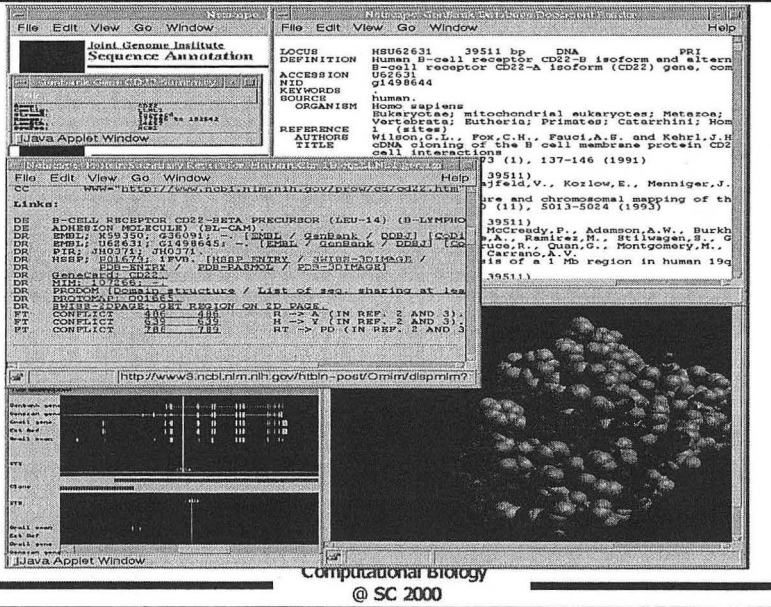
http://compbio.ornl.gov/cgi-bin/PrinRpt.pl?human,5,T000980,gene,grailexp,32

BEAUTY - Gene Search Results



Computational Biology
@ SC 2000

Reports and Links



Computational Biology
@ SC 2000

Navigate from human chromosome

Bring up two Java Genome Channel Viewers

Genome

Chromosome [Genes]

Genome Catalog - human / Chromosome - 19 - Netscape

Genome Catalog
An Annotated Genome Sequence Framework

Chromosomes 2 3 4 5 6 7 8 9 10 12 13 14 15 16 17 19 20 21 22 X Y

Species / Human

Chromosome 19 Overview

Genome Channel

[Overview] [Contig] [Clone] [Band] [Map]
Gene Models: [GenBank] [Genscan] [GRAIL_EXP]

Organism viewer

Chromosome viewer

Clone ID
Clone ID
Contig ID
Gene ID
Protein Homology

Search Chromosome

Sequences

Query just on Chromosome

SNP Mining from Clone Overlaps

Clone overlap:
AF064865
AF042091
overlap 9,338 variant bases 36
approx. 1 SNP per 250 bp

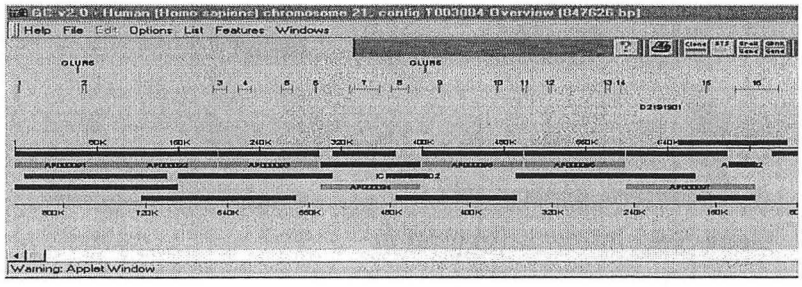
Example

```
AF064865: 157047 agggcttatcagtgtcgctgttgaccttggccacctggctaaggtggtgcctgccaggtt 157106
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF042091: 6961   agggcttatcagtgtcgctgttgaccttggccacctggctaaggtggtgcctgccaggtt 7020

AF064865: 157107 tctccactggaaagcttctctttccatgttgtcctttctggaaggaagtcgctctgcaaa 157166
                 ||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||||
AF042091: 7021   tctccactggaaagcttctctttccatgtcgtcctttctggaaggaagtcgctctgcaaa 7080

AF064865: 157167 gcccacacataaggagtgagagttatgcttcatcttcttgaggtggtatatctacataaa 157226
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AF042091: 7081   gcccacacataaggagtgagagttatgcttcatcttcttgaggtggtatatctacataaa 7140
```

SNP Mining from Clone Overlaps



Coverage includes clones from different sources
1 SNP per 250 bases
160,000 SNPs in 408 Mb dataset

What's supercomputing got to do with it?

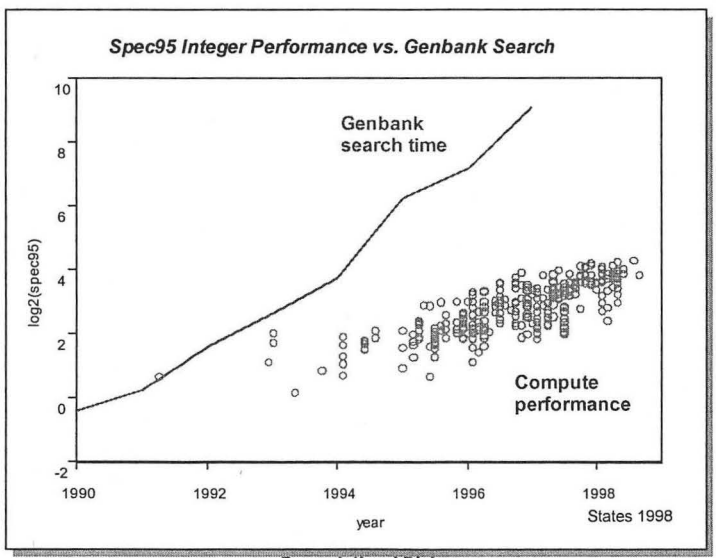
- **Complexity of the information**
- **Amount of data**
- **Most applications are trivially parallel**

Layers of Information

The same base sequence contains many layered instructions!

- Chromosome structure and function
  - Telomers, centromers
- Gene Regulatory information
  - Enancers, promoters
- Instructions for gene structure
- Instructions for protein
- Instructions for protein post-processing and localization

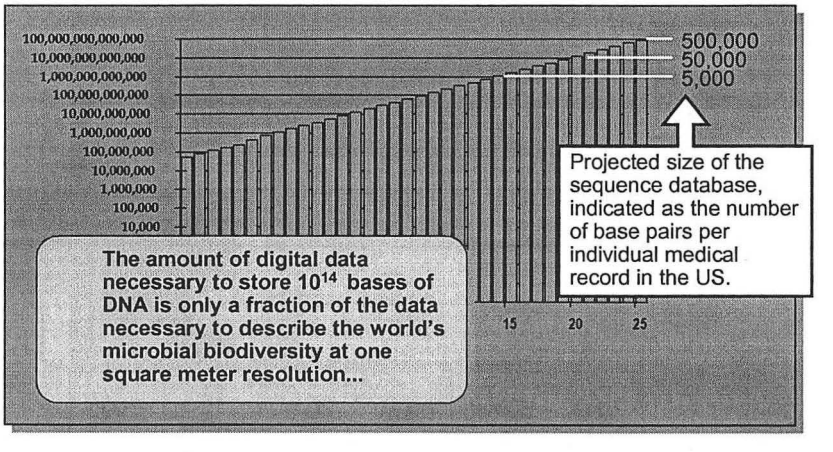
Moore's Law and Genomics



CPU Requirements

- Current annotation
  - 250 Mbases DNA yield ~125 Gbytes of data
  - It takes ~ 7.5 days on 20 workstations ~3,600nhr
- Celera Sequencing
  - Assembly of 1.7 Million reads in 25 hrs
  - Annotation 8-10 Mbases per months with 6 FTE
  - Assembly of Human Genome: expected ~ 3 months

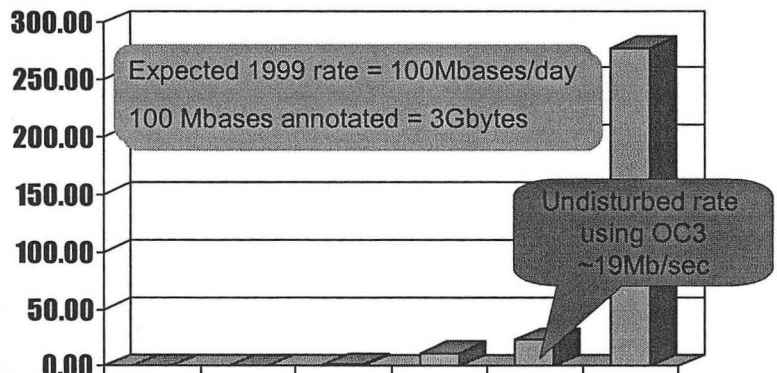
Projected Base Pairs



Sequence Assembly

- Complexity
  - Adding a day's read of 100 Mb to a billion base pairs of contig would require 100 Pops operations
  - A 1 Tops machine would take about one day to process 100 Mbases

Data Transfer

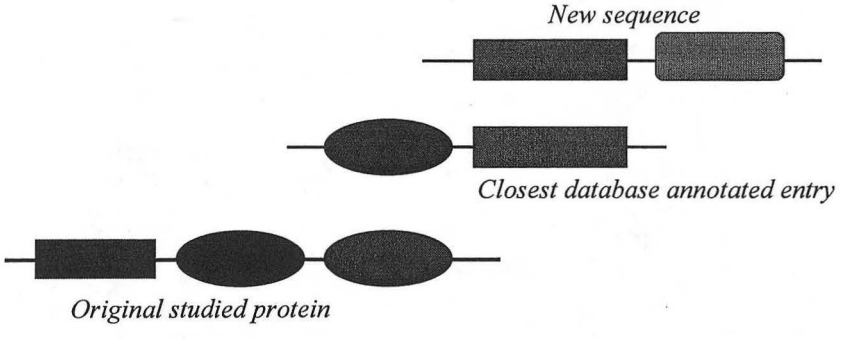


	year	month	week	day	12 hours	1 hour
Mbytes/sec	0.03	0.39	1.65	11.60	23.10	277.70

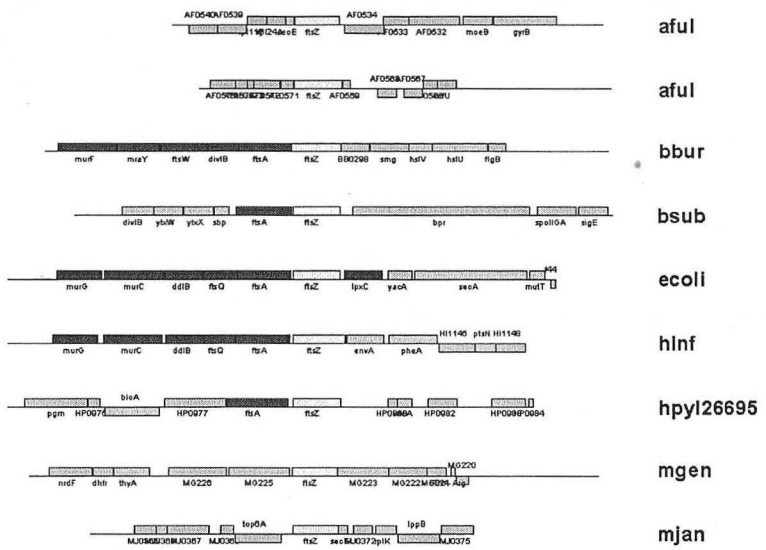
Challenges

- Discovering new biology
- Lack of software integration
- Beginning to build high-performance applications
- Shortage of personnel

Inherited Annotation Problems in Multi-Domain Proteins

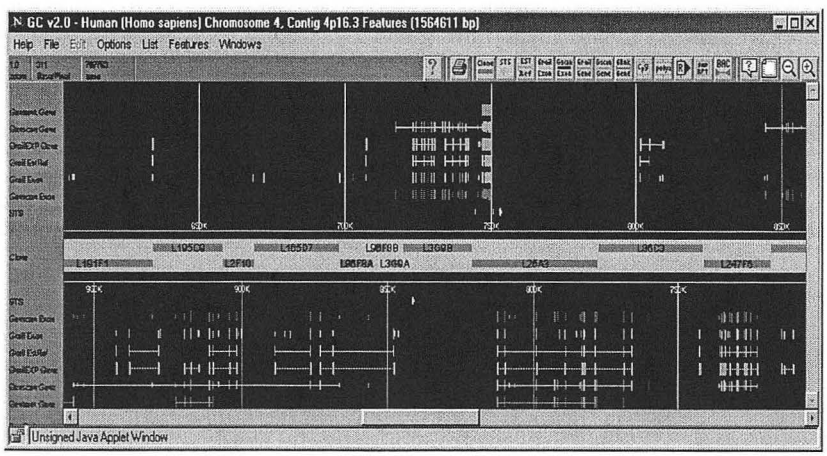


Comparative Genome Analysis



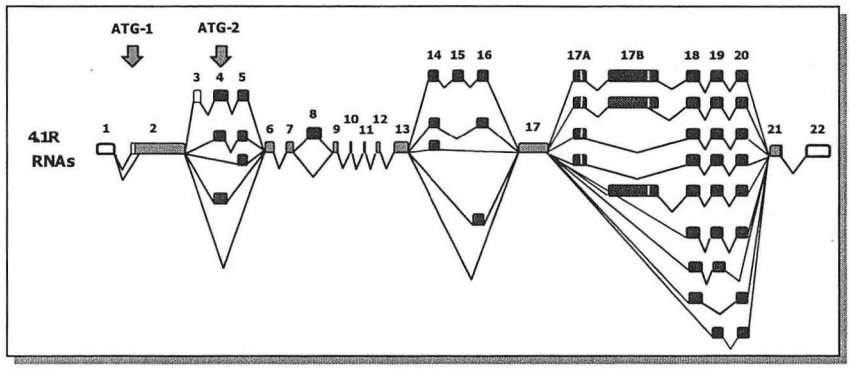
Computational Biology
@ SC 2000

Alternatively Spliced ?



Computational Biology
@ SC 2000

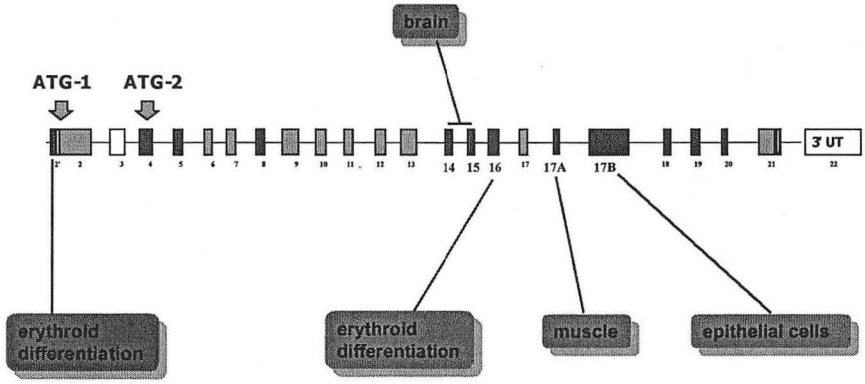
One Gene - Many Proteins



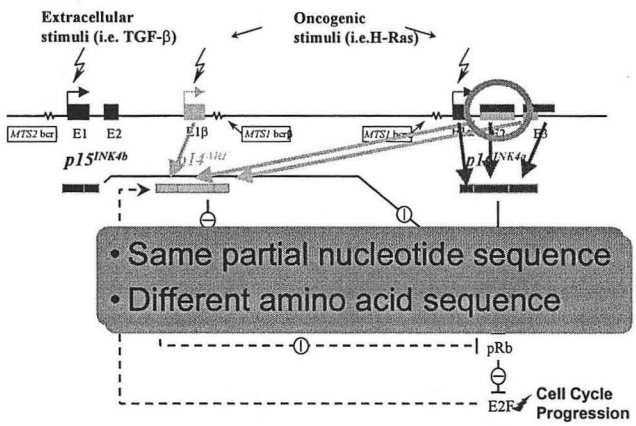
Conboy 1998

As many as 30% of human genes, in particular structural genes, may be alternatively spliced.

One Gene - Many Proteins



9p21 Gene Cluster is a Nexus of the Rb and p53 Pathways



Credits

- NERSC / LBNL
  - John Conboy
  - Donn Davy
  - Inna Dubchak
  - Sylvia Spengler
  - Denise Wolf
  - Eric P. Xing
  - Manfred Zorn
- ORNL
  - Ed Uberbacher
  - Richard Mural
  - Phil LoCascio
  - Sergey Petrov
  - Manesh Shah
  - Morey Parang

Computational Biology and High Performance Computing 2000

Tutorial M4 p.m.

November 6, 2000

SC'2000, Dallas, Texas

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Tutorial Outline

BERKELEY LAB

- 8:30 a.m. - 12:00 p.m.
  - Introduction to Biology
  - Overview Computational Biology
  - DNA sequences
- 1:30 p.m. - 5:00 p.m.
  - Protein Sequences
  - Phylogeny
  - Specialized Databases

Tutorial Outline: Afternoon

- 1:30 p.m. - 2:00 p.m. Working with Proteins
- 2:00 p.m. - 3:00 p.m. Phylogeny
- 3:00 p.m. - 3:30 p.m. BREAK
- 3:30 p.m. - 4:30 p.m. Specialized Databases
- 4:30 p.m. - 5:00 p.m. Genetic Networks

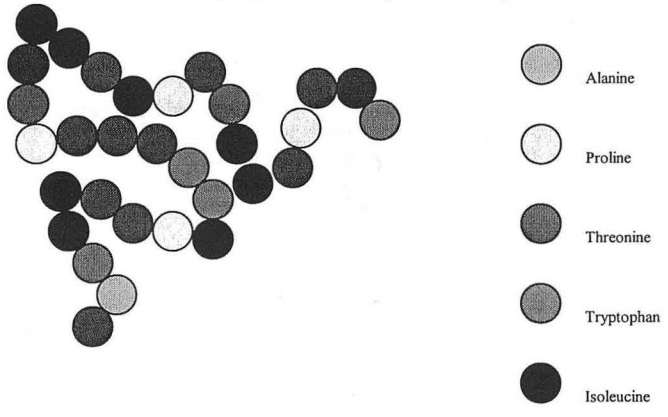
Proteins

Manfred Zorn
MDZorn@lbl.gov
NERSC

Proteins

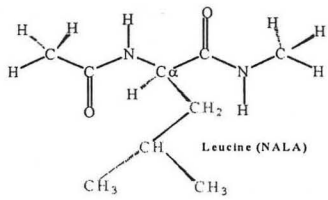
What is a protein?

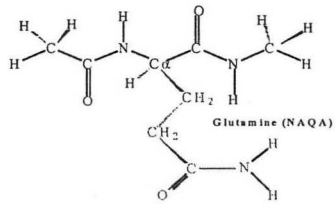
A biopolymer which is distinct from a heteropolymer in one very important way
It's 3-D structure is uniquely tailored to perform a specific function



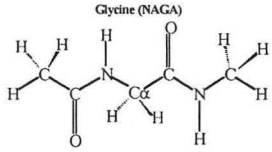
NMR, X-ray and electron crystallography solve structures slowly (1/2-3 yrs.)

The "Beads" are Chemically Complex Structures

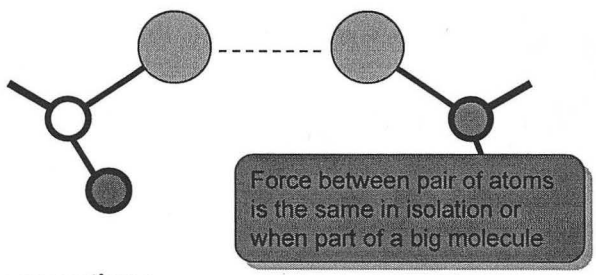




Glycine (NAGA)

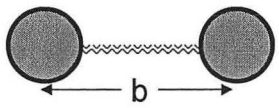


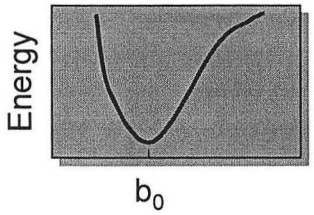
Forces Between Atoms



- **Basic assumptions:**
  - **Energy contributions are strictly additive**
  - **Energy is independent of neighbors; transferability**
  - **Quantum mechanics is insignificant as long as no bonds are broken**

Bond Stretching Forces



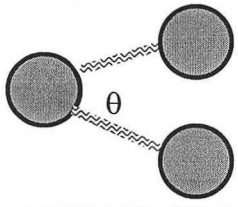


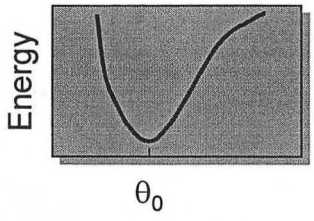
Equilibrium length ~ 0.1-0.2nm

$$U(b) = K_b(b - b_0)^2$$

K_b spring force constant ~ 500kcal/mole Å^2

Bond Angle Forces

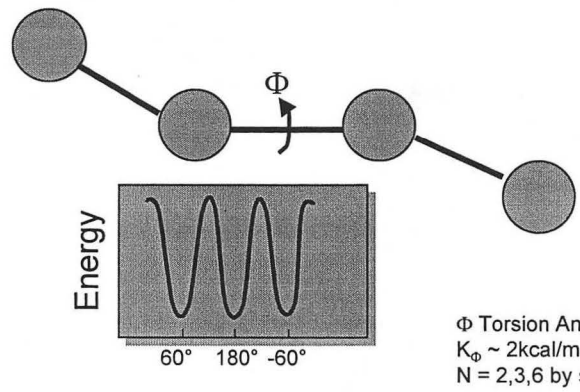




$$U(\theta) = K_\theta(\theta - \theta_0)^2$$

K_θ spring force constant

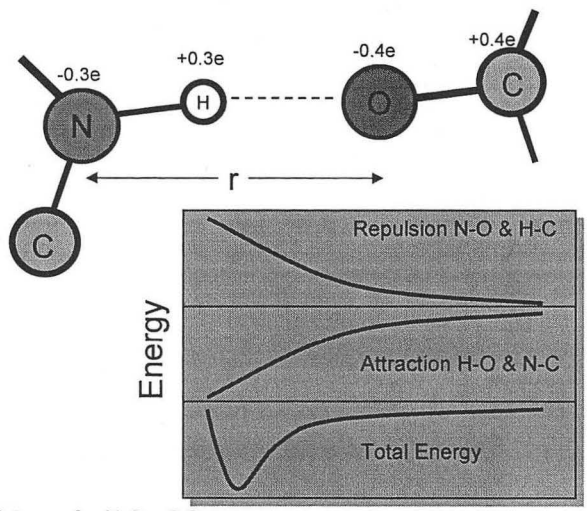
Bond Twisting Forces



Φ Torsion Angle
K_Φ ~ 2kcal/mole
N = 2,3,6 by symmetry

$$U(\Phi) = K_\Phi\left[1 - \cos(n\Phi_i + \delta)\right]$$

Hydrogen Bonds

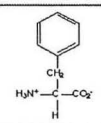
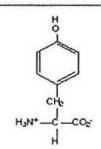
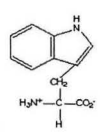


Optimum distance for N-O = 0.3nm
Net interaction ~ -5kcal/mole

Scale of Interactions

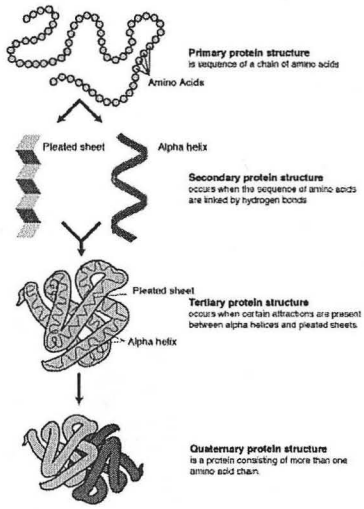
Interaction	*Energy (kcal/mole)*
Van der Waals (in water)	-0.1
Hydrogen bond (in water)	-1.0
Torsion barrier (single bond)	~+3.0
Torsion barrier (double bond)	+20.0
Bond breakage	+100.0
Change bond angle by 10°	+2.0
Stretch bond length by 10pm (0.1Å)	+2.5
Thermal energy 300K	0.6

Aromatic Amino Acids

Amino Acid	pK_a's[2]	Pro Structure[1]	Chemical Structure[4]	3-DStructure[1]
Phenylalanine, Phe, F No charge absorbs UV hydrophobic (25) Molec. Wt. 147 Mole % 3.5	N=9.13 C=1.83 pI=5.48	α =1.16 β =1.33 t =0.59		
Tyrosine, Tyr, Y weak charge absorbs UV hydrogen bonding nonhydropathic (0.08) Molec Wt. = 163 Mole % = 3.5	N=9.11 C=2.20 R=10.07 pI=5.66	α =0.74 β =1.45 t =0.76		
Tryptophan, Trp, W largest amino acid rarest amino acid no charge absorbs UV hydrogen bonding hydrophobic (15) Molec Wt. = 186 Mole % = 1.1	N=9.39 C=2.38 pI=5.89	α =1.02 β =1.35 t =0.65		

Copyright Charles S. Grasser 1996

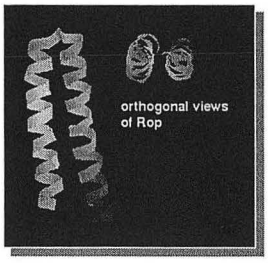
Protein Structure



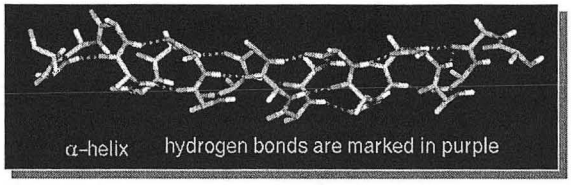
Secondary Structure

- Alpha-helix
- Beta-sheet
- Coil

Alpha Helix

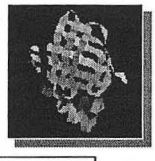


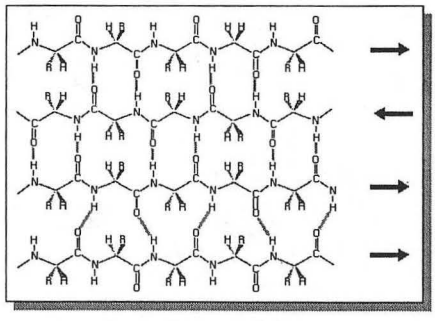
- Alpha-helix
  - Right-handed alpha helix
  - 3.6 amino acids per turn
  - Most abundant (35%)



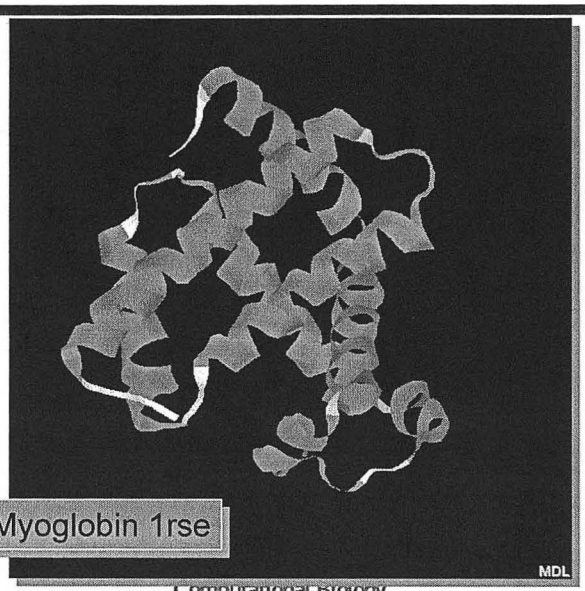
Beta-Sheet

- Beta-sheet
  - Parallel - antiparallel
  - 25% of proteins





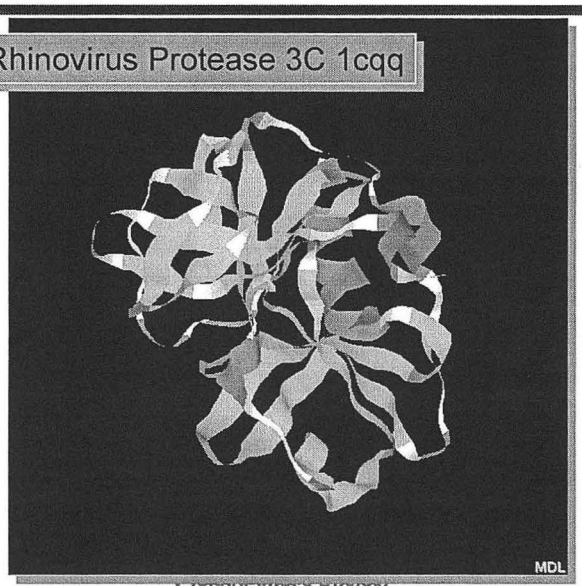
Alpha Helix



Human Myoglobin 1rse

Beta sheets

Human Rhinovirus Protease 3C 1cqq



SCOP: Structural Classification of Proteins

- 1. All alpha proteins (a)
- 2. All beta proteins (b)
- 3. Alpha and beta proteins (a/b)
  - Mainly parallel beta sheets (beta-alpha-beta units)
- 4. Alpha and beta proteins (a+b)
  - Mainly antiparallel beta sheets (segregated alpha and beta regions)
- 5. Multi-domain proteins (alpha and beta)
  - Folds consisting of two or more domains belonging to different classes
- 6. Membrane and cell surface proteins and peptides
  - Does not include proteins in the immune system
- 7. Small proteins
  - Usually dominated by metal ligand, heme, and/or disulfide bridges
- 8. Coiled coil proteins
- 9. Low resolution protein structures
- 10. Peptides
- 11. Designed proteins

SCOP Classifications

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	128	197	296
All beta proteins	87	158	251
Alpha and beta proteins (a/b)	93	153	323
Alpha and beta proteins (a+b)	168	237	345
Multi-domain proteins	25	25	32
Membrane and cell surface proteins	11	17	19
Small proteins	52	72	102
Total	564	859	1368

SCOP: Structural Classification of Proteins. 1.53 release
11410 PDB Entries (1 Jul 2000).
26219 Domains.
Copyright © 1994-2000 The scop authors / scop@mrc-lmb.cam.ac.uk
September 2000

Protein Fold Recognition, Structure Prediction, and Folding

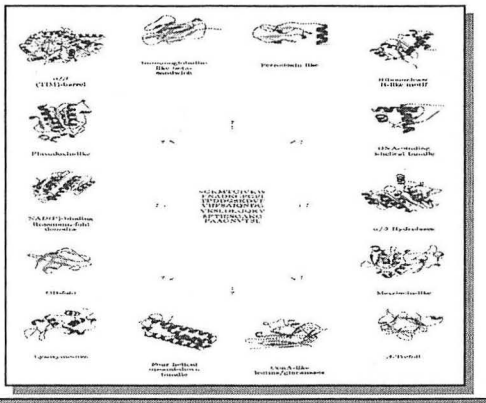
- Drawing analogies with known protein structures
  - Sequence homology, Structural Homology
  - Inverse Folding, Threading
- Ab initio folding: the ability to follow kinetics, mechanism
  - robust objective function
  - severe time-scale problem
  - proper treatment of long-ranged interactions
- Ab initio prediction: the ability to extrapolate to unknown folds
  - multiple minima problem
  - robust objective function
  - Stochastic Perturbation and Soft Constraints
- Simplified Models that Capture the Essence of Real Proteins
  - Lattice and Off-Lattice Simulations
  - Off-Lattice Model that Connect to Experiments: Whole Genomes?

Protein Fold Predictions: Neural Network Structure Classifications

- Protein fold predictor based on global descriptors of amino acid sequence
- Empirical prediction using a database of known folds in machine learning
- Databases
  - 3D-ALI (83 folds)
  - SCOP (used ~120 folds)
- Representation of protein sequence in terms of physical, chemical, and structural properties of amino acids
- Feed forward neural network for machine learning

Protein Fold Recognition: Threading

Sequence Assignments to Protein Fold Topology
(David Eisenberg, UCLA)



Take a sequence with unknown structure and align onto structural template of a given fold
Score how compatible that sequence is based on empirical knowledge of protein structure
Right now 25-30% of new sequences can be assigned with high confidence to fold class
100,000's of sequences and 10,000's of structures (each of order 10^2-10^3amino acids long)

Protein Fold Recognition: Threading

Computational Approach:

Dynamic programming: capable of finding optimal alignments if

optimal alignments of subsequences can be extended to optimal alignments of whole

objective functions that are one-dimensional $E=\Sigma V_i + \Sigma V_{gap}$

Complexity: all to all comparison of sequence to structure scales as L^2
Whole human genome: 10^{13} flops

Improve Objective function:

Take into account structural environment

3D→1D: dynamic programming, L^2

Build pairwise or multi-body objective function

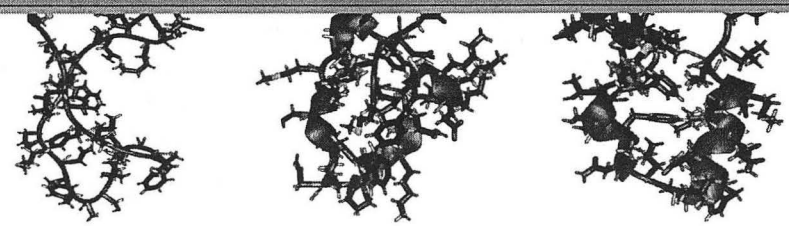
NP-hard if: variable-length gaps and model nonlocal effects such as distance dependence

Recursive dynamic programming, Hidden markov models, stochastic grammers

Complexity: all to all comparison of sequence to structure scales as L^3
Whole human genome: ~10^{16} flops

Computational Protein Folding

One microsecond simulation of a fragment of the protein, Villin. (Duan & Kollman, Science 1998)



✓ **robust objective function**
all atom simulation with molecular water present: some structure present

✓ **severe time-scale problem**
required 10^9 energy and force evaluations: parallelization (spatial decomposition)

- **proper treatment of long-ranged interactions**
✗ cut-off interactions at 8Å, poor by known simulation standards

- **Statistics (1 trajectory is anecdotal)**
✗ Many trajectories required to characterize kinetics and thermodynamics

Computational Protein Folding

(1) Size-scaling bottlenecks: Depends on complexity of energy function, V

Empirical (less accurate): cN^2; ab initio (more accurate): CN^3 or worse ; $c \ll C$

empirical force field used

"long-ranged interactions" truncated so cM^2 scaling; $M < N$

spatial decomposition, linked lists

(2) Time-Scale of motions bottlenecks (Δt)

$$r_i(t+\Delta t) = 2r_i(t) - r_i(t-\Delta t) + \frac{f_i(t)}{m_i}\frac{(\Delta t)^2}{2!} + O\left[(\Delta t)^4\right] \quad v_i(t) = \frac{r_i(t+\Delta t) - r_i(t-\Delta t)}{2\Delta t} + O\left[(\Delta t)^3\right]$$

$$f_i = m_i a_i = -\nabla_i V(r_1, r_2, \ldots r_N)$$

Use timestep commensurate with fastest timescale in your system

bond vibrations: 0.01Å amplitude: 10^{-15} seconds (1fs)

Shake/Rattle bonds (2fs)

Multiple timescale algorithms (~5fs) (not used here)

Ab Initio Protein Structure Prediction

Primary Squence and an Energy function → Tertiary structure

Empirical energy functions:

(1) Detailed, Atomic description: leads to enormous difficulties!

$$V_{MM} = \sum_{i}^{\#\,Bonds} k_b (b_i - b_o)^2 + \sum_{i}^{\#\,Angles} k_\theta (\theta_i - \theta_o)^2 + \sum_{i}^{\#\,Impropers} k_\tau (\tau_i - \tau_o)^2 +$$

$$\sum_{i}^{\#\,dihedrals} k_\phi \left[1 + \cos(n\phi + \delta)\right] + \sum_{i}^{\#\,atoms} \sum_{i<j}^{\#\,atoms} \left\{ \frac{q_i q_j}{r_{ij}} + \varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} + \sum_{i}^{\#\,atoms} \Delta\sigma A$$

(1) Multiple minima problem is fierce

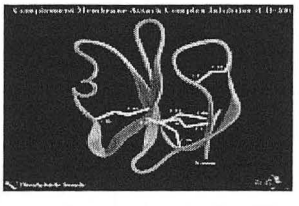
Find a way to effectively overcome the multiple minima problem

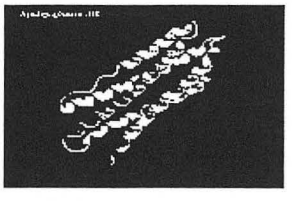
(2) Objective Functions: Replaceable algorithmic component?

Global energy minimum should be native structure, misfolds higher in energy

The Objective (Energy) Function

Empirical Protein Force Fields: AMBER, CHARMM, ECEPP
"gas phase"





CATH protein classification: http://pdb.pdb.bnl.gov/bsm/cath

α-helical sequence/ β-sheet structure **β-sheet sequence/a-helical structure**

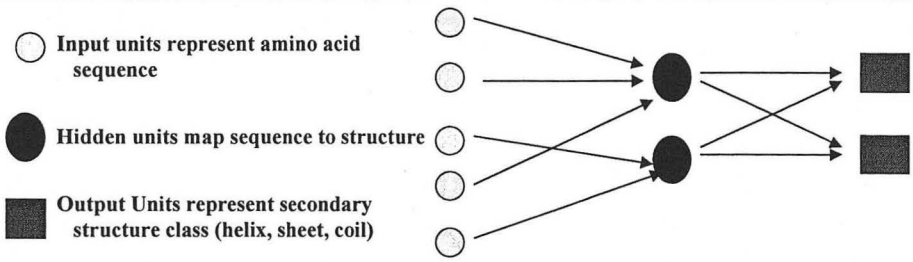
Energies the same! Makes energy minimization difficult!

Add penalty for exposing hydrophobic surface: favors more compact structures

$E_{native\ folds} < E_{misfolds}$ for a few test cases

Solvent accessible surface area functions: Numerically difficult to use in optimization

Neural Networks for 2° Structure Prediction



→ Weights are optimizable variables that are trained on database of proteins

Poorly designed networks result in overfitting, inadequate generalization to test set

Neural network design

- input and output representation
- number of hidden neurons
- weight connection patterns that detect structural features

Neural Network Results

No sequence homology through multiple alignments

Train			Test		
Total predicted correctly = 66%			Total predicted correctly = 62.5%		
Helix:	51%	C_a=0.42	Helix:	48%	C_a=0.38
Sheet:	38%	C_b=0.39	Sheet:	28%	C_b=0.31
Coil:	82%	C_c=0.36	Coil:	84%	C_c=0.35

Network with Design: Yu and Head-Gordon, Phys. Rev. E 1995

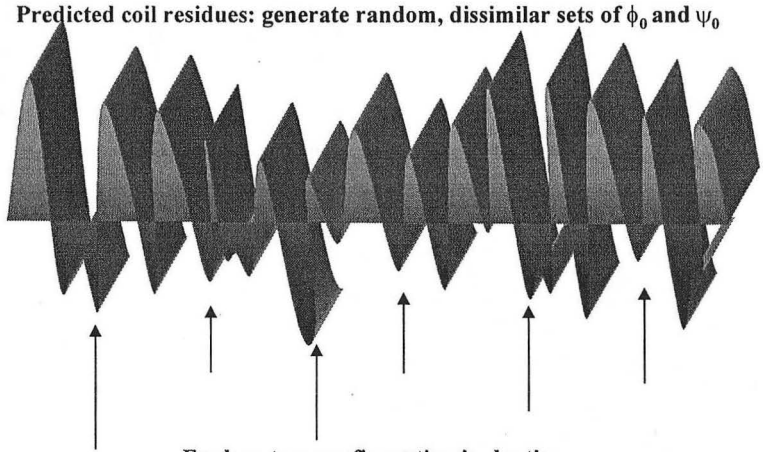
Train			Test		
Total predicted correctly = 67%			Total predicted correctly = 66.5%		
Helix:	66%	C_a=0.52	Helix:	64%	C_a=0.48
Sheet:	63%	C_b=0.46	Sheet:	53%	C_b=0.43
Coil:	69%	C_c=0.43	Coil:	73%	C_c=0.44

Combine networks of Yu and Head-Gordon with multiple alignments

Neural Networks Used To Guide Global Optimization Methods

Generate expanded tree of configurations

Predicted coil residues: generate random, dissimilar sets of ϕ_0 and ψ_0



Explore tree configuration in depth:

Global Optimization in sub-space of coil residues: walk through barriers, move downhill

Hierarchical Parallel Implementation of Global Optimization Algorithm

Static vs. Dynamic Load Balancing of Tasks

Central Processor

↓

GOPT1	GOPT2	GOPT3	GOPT4	GOPT5
↓	↓	↓	↓	↓
$W_{1,1} \rightarrow W_{1,11}$	$W_{2,1} \rightarrow W_{2,11}$	$W_{3,1} \rightarrow W_{3,11}$	$W_{4,1} \rightarrow W_{4,11}$	$W_{5,1} \rightarrow W_{5,11}$

Central Processor: Assigns starting coordinates to GOPT's

Task time is highly variable

GOPT's: Divide up sub-space into N regions for global search

Task time is variable

Workers: Generate sample points; find best minimizer in region

(Number of workers depends on sub-space)

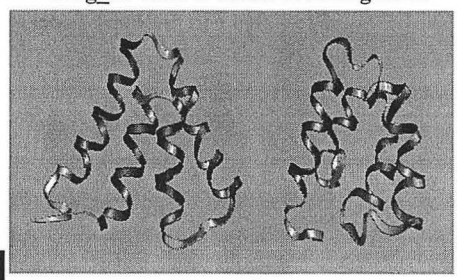
Dynamical load balancing of tasks: reassigning GOPT/workers to GOPT/workers

Gain in efficiency of a factor of 5-10

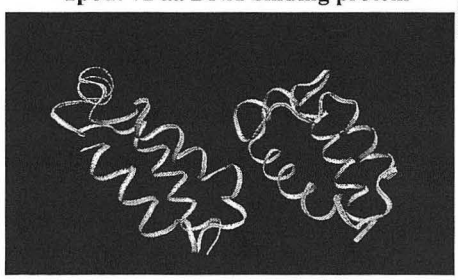
Global Optimization Predictions of α-Helical Proteins

Crystal (left), Prediction (right)
R.M.S. 7.0Å

2utg_A: 70aa α-chain of uteroglobin:



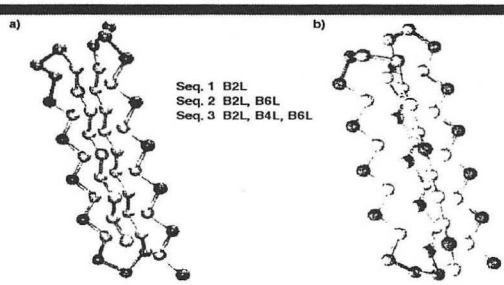
1pou: 72 aa DNA binding protein



Prediction (left) and crystal (right)
R.M.S. 6.3Å

Still have not reached crystal energy yet!

Simplified Models for Simulating Protein Folding



Simplifies the "real" energy surface topology sufficiently that you can do

(1) Statistics ✓
Can do many trajectories to converge kinetics and thermodynamics

(2) severe time-scale problem✓
characterize full folding pathway: mechanism, kinetics, thermodynamics

(3) proper treatment of long-ranged interactions ✓
all interactions are evaluated; no explicit electrostatics

(4) robust objective function?
good comparison to experiments

Acknowledgements

Teresa Head-Gordon, Physical Biosciences Division, LBNL

Silvia Crivelli, Physical Biosciences and NERSC Divisions, LBNL

Betty Eskow, Richard Byrd, Bobby Schnabel, Dept. Computer Science, U. Colorado

Jon M. Sorenson, NSF Graduate Fellow, Dept. Chemistry UCB

Greg Hura, Graduate Group in Biophysics, UCB

Alan K. Soper, Rutherford Appleton Laboratory, UK

Alexander Pertsemlidis, Dept. of Biochemistry, U. Texas Southwestern Medical Center

Robert M. Glaeser, Mol. & Cell Biology, UCB and Life Sciences Division, LBNL

Funding Sources:

AFOSR, DOE (MICS), DOE/LDRD (LBNL), NIH, NERSC for cycles

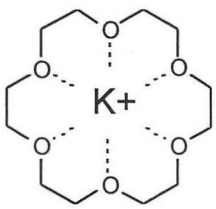
Structure-Based Drug Discovery

Brian K. Shoichet, Ph.D
Northwestern University, Dept of MPBC
303 E. Chicago Ave, Chicago, IL 60611-3008
Nov 15, 1999

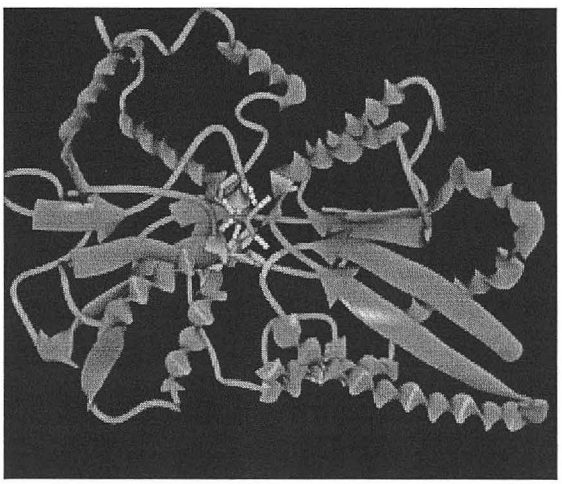
Problems in Structure-Based Inhibitor Discovery & Design

- Balance of forces in binding
  - Energies in condensed phases
    - interaction energies
    - desolvation
- Problem scales badly with degrees of freedom
  - Configuration
    - configs α (prot-features)^4 X (lig-features)^4
  - Conformation
    - Ligand & Protein, confs α 3^{lbonds} X 3^{pbonds}
- Sampling chemical space (scales *very* badly)
- Defining binding sites

The Pros & Cons of Proteins

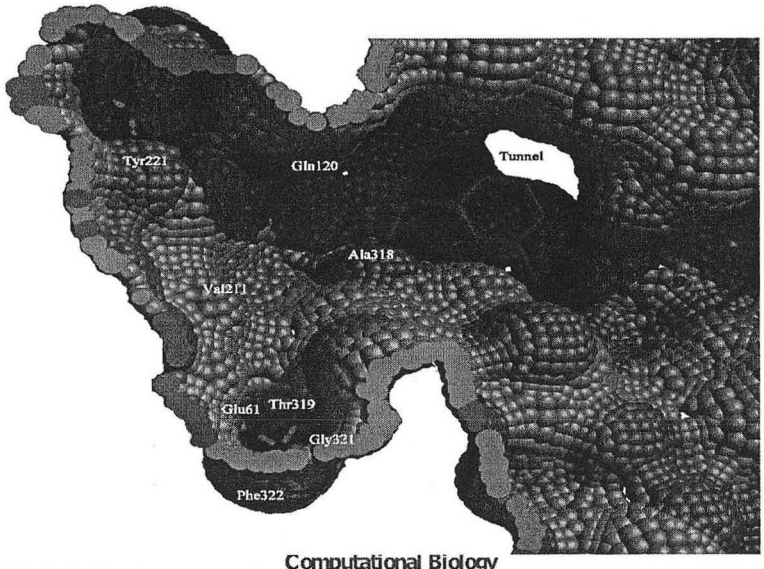


18 - Crown-6



sulfate binding protein

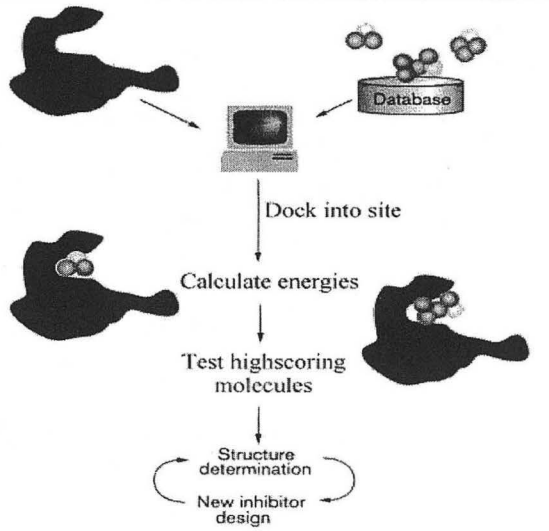
Conserved Residues, Ordered Structure, Function Unknown



Inhibitor Discovery or Design?

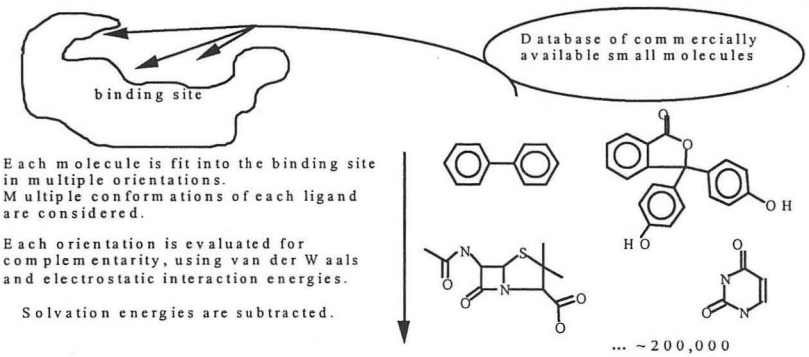
- **Design ligands**
  - Ludi (Bohm)
  - Grow (Moon & Howe)
  - Builder (Roe & Kuntz)
  - MCSS-Hook (Miranker & Karplus)
  - SMOG (DeWitte & Shaknovitch)
  - Others...
- **Discover Ligands**
  - DOCK (Kuntz, et al., Shoichet)
  - CAVEAT (Bartlett)
  - Monte Carlo (Hart & Read)
  - AutoDock (Goodsell & Olson)
  - SPECITOPE (Kuhn et al)
  - Others...

Screening Databases by Molecular Docking



© Chemistry & Biology, 1996

Database Screening Using DOCK

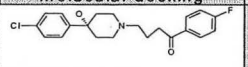
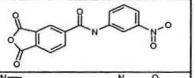
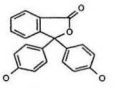
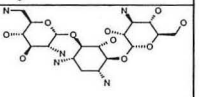
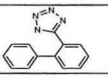
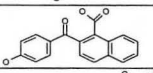
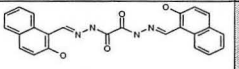
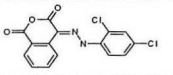


The inhibition constants of the best fitting molecules are established in an enzyme assay

Inhibitor-receptor complex structures are determined.

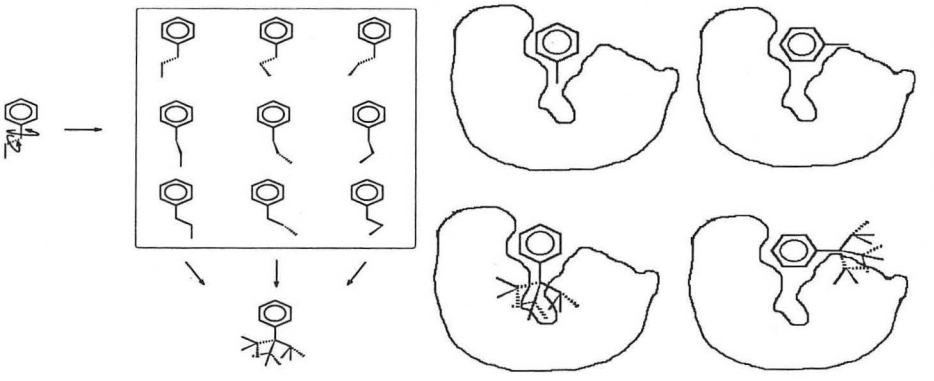
New interactions with the enzyme are targeted.

Novel Ligand Discovery Using Molecular Docking

Receptor	Lead from molecular docking	Receptor	Lead from molecular docking
HIV protease		HGXPRTase	
thymidylate synthase		RNA	
hemagglutinin		Zn β-lactamase	
cercarial elastase		Thrombin	
malarial protease		AmpC β-lactamase	
CD4-gp120	unpublished	thymidylate synthase	
		HGXPRTase	unpublished

Computational Biology
@ SC 2000

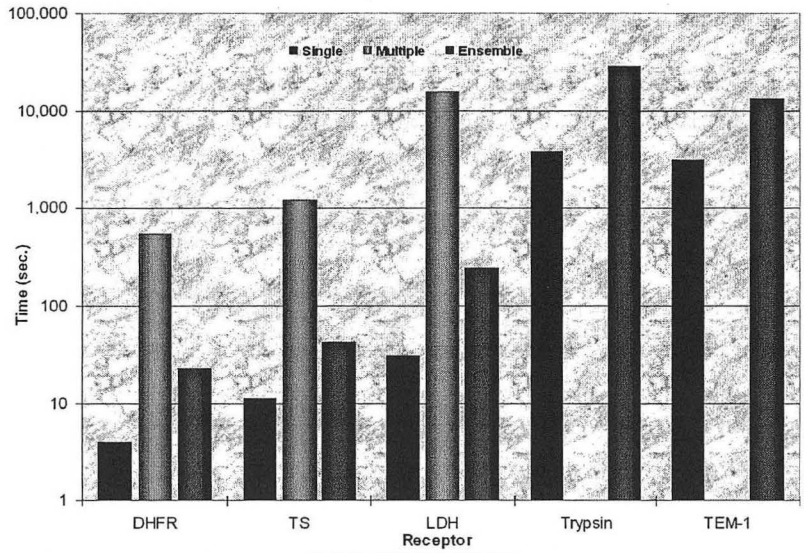
Ligand Flexibility: Conformational Ensembles



Generate an ensemble dock it into the site

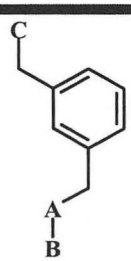
Computational Biology
@ SC 2000

Conformational Ensembles vs. Brute Force

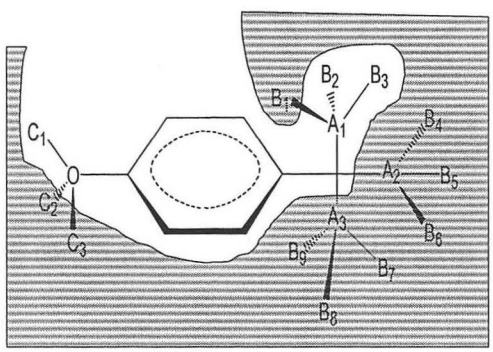


Computational Biology
@ SC 2000

Hierarchical Docking



Flexible docking:	Hierarchical docking:
27 confs	27 confs
x3 atoms	3C + 3A + 9B
81 atom positions	15 atom positions



Computational Biology
@ SC 2000

Unmet Challenges

- Better Scoring
  - context dependent desolvation
  - receptor desolvation
  - better force-fields
- Receptor Flexibility
- Cominatorial Chemistry

Computational Phylogenetics

Craig Stewart
stewart@iu.edu
Indiana University

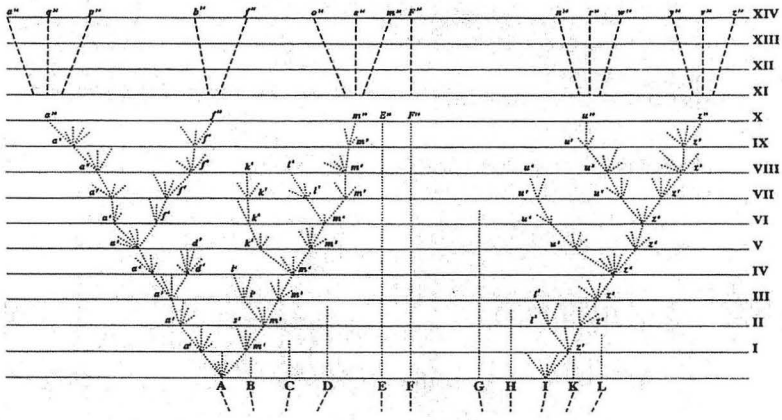
Outline

- Evolution & Phylogenetics
- Why is this an HPC problem?
- Alignment (brief)
- Summary of methods and software for phylogenetics
- One example in detail: Maximum Likelihood analysis with fastDNAml
- Some interesting results and challenges for the future
- Caveat: this is an introduction, not an exhaustive review.

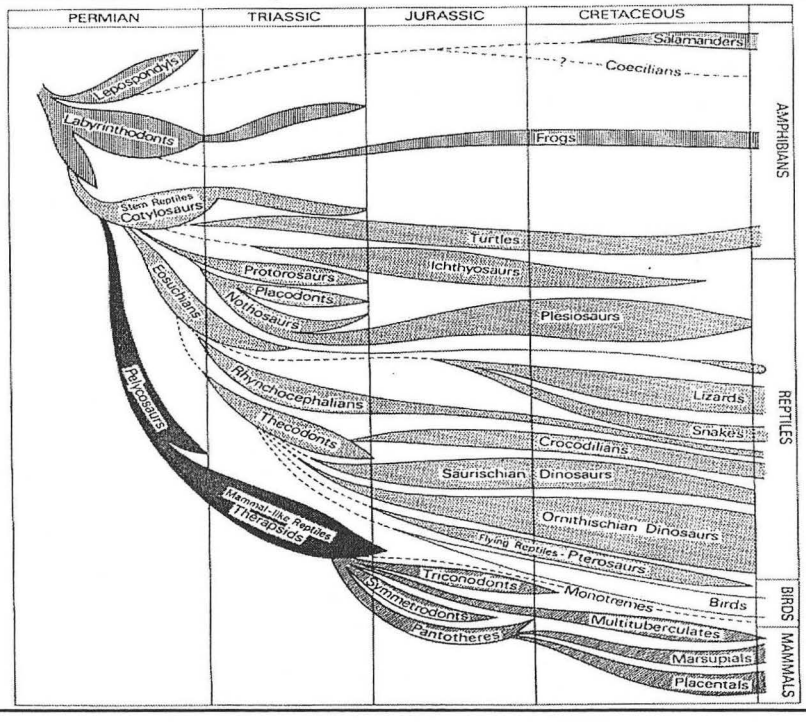
Phylogeny

- Evolution is an explicitly historical branch of biology, one in which the subjects are active players in the historical changes.
- A phylogeny, or phylogenetic tree, is a way of depicting evolutionary relationships among organisms, genes, or gene products.
- Modern evolutionary theory began with Darwin's *Origin of Species,* which included one figure – an evolutionary tree

Origin of Species, Figure 1

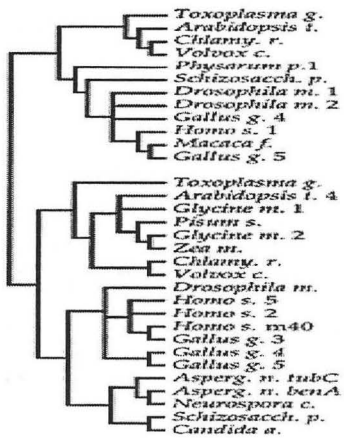


Computational Biology
@ SC 2000



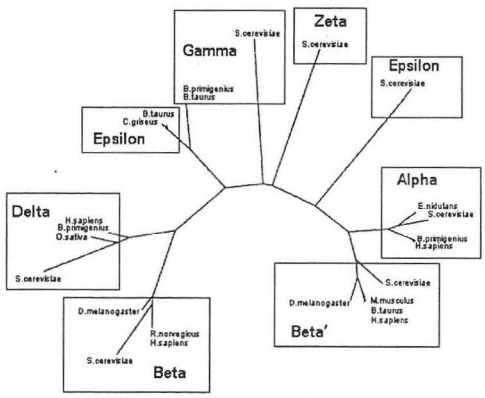
Building Phylogenetic Trees

- Goal: an objective means by which phylogenetic trees can be estimated in tolerable amounts of wall-clock time, producing phylogenetic trees with measures of their uncertainty



Basic Evolutionary Biology

- All evolutionary changes are described as bifurcatin trees
  - evolutionary relationships among genes or gene products (trees of paralogues)
  - evolutionary relationships among organisms (trees of orthologues)



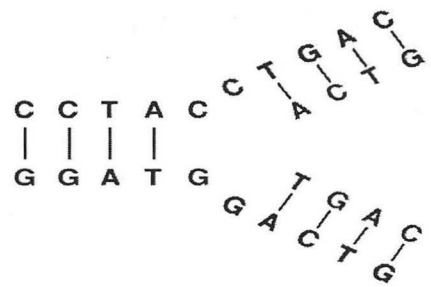
Why

- Curiosity: Anyone who as a child wandered through the dinosaur section of a natural history museum understands the inherent intellectual attraction of evolutionary biology
- Theoretical uses: testing hypotheses in evolutionary biology
- Practical uses:
  - Medicine
  - Environmental management (biodiversity maintenance)

Reconstructing history from DNA sequences

- DNA changes over time; much of this change is not expressed
- Changes in unexpressed DNA can be modeled as Markov processes
- By comparing similar regions of DNA from different organisms (or different genes) one can infer the phylogenetic tree and evolutionary history that seems the best explanation of the current situation

DNA replication



Purines: Adenine & Guanine
Pyrimidines: Thymine & Cytosine

Changes in genetic information over time

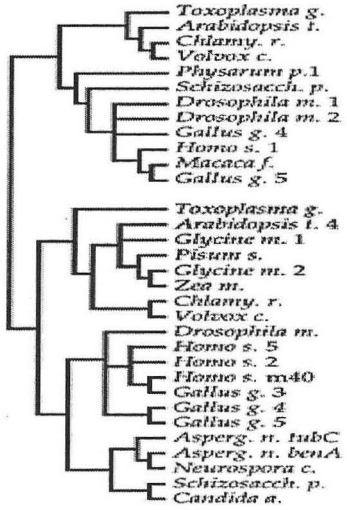
- **Point mutations**
  DNA – sequences of the 4 nucleotides
  CCTCTGAC
  vs
  TCTCCGAC
  Protein – sequences of the 20 amino acids
  GSAQVKGHGKK
  vs
  GNPKVKAHGKK

- **Insertions and deletions**
  DNA CCTCT+GAC
  vs
  CCTCTTGAC

Sequences available

- DNA (sequences are series of the base molecules; aligned sequences will also contain +s for gaps)
- Amino acid sequences (series of letters indicating the 20 amino acids). Computational challenges more severe than with DNA sequences.
- RNA
- The availability of data at present exceeds the ability of researchers to analyze it!

Why is tree-building a HPC problem?



- The number of bifurcating unrooted trees for n taxa is (2n-5)!/ (n-3)! 2n-3
- for 50 taxa the number of possible trees is ~1074; most scientists are interested in much larger problems
- The number of rooted trees is (2n-5)!

Alignment

- To build trees one compares and relates 'similar' segments of genetic data. Getting 'similar' right is absolutely critical!
- Methods:
  - dynamic programming
  - Hidden Markov Models
  - Pattern matching
- Some alignment packages:
  - BLAST
    http://www.ncbi.nlm.nih.gov/BLAST/
  - FASTA
    http://gcg.nhri.org.tw/fasta.html
  - MUSCA http://www.research.ibm.com/bioinformatics/home

Matching cost function

```
GCTAAATTC
 ++  x x
GC  AAGTT
```

- Penalize for mismatches, for opening of gap, and for gap length
- This approach assumes independence of loci: good assumption for DNA, some problems with respect to amino acids, significant problems with RNA

Example of aligned sequences

```
Thermotoga    ATTTGCCCCA GAAATTAAAG CAAAAACCCC AGTAAGTTGG GGATGGCAAA
Tthermophi    ATTTGCCCCA GGGGTTCCCG CAAAAACCCC AGTAAGTTGG GGATGGCAGG
Taquaticus    ATTTGCCCCA GGGGTTCCCG CAAAAACCCC AGTAAGTTGG GGATGGCAGG G
deinon        ATTTGCCCCA GGGATTCCCG CAAAAACCCC AGTAAGTTGG GGATGGCAGG G
Chlamydi      ATTTTCCCCA GAAATTCCCG AAAAAACCCC AATAAATTGG GGATGGCAGG
flexistips    ATTTTCCCCA CAAAAAAAAG AAAAAACCCC AGTAAGTTGG GGATGGCAGG
borrelia-b    ATTTGCCCCA GAAGTTAAAG CAAAAACCCC AATAAGTTGG GGATGGCAGG
bacteroide    ATTTGCCCCA GAAATTCCCG CAAAAACCCC AGTAAATTGG GGATGGCAGG GG
Pseudom       ATTTGCCCCA GGGATTCCCG CAAAAACCCC AGTAAGTTGG GGATGGCAGG G
ecoli-----    GTTTTCCCCA GAAATTCCCG CAAAAACCCC AGTAAGTTGG GGATGGCAGG
salmonella
    ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
    +++
shewanella    GTTTGCCCCA GCCATTCCCG TAAAAACCCC AGTAAGTTGG GGATGGCAGG
bacillus--    ATTTGCCCCA GAAATTCCCG CAAAAACCCC AGCAAATTGG GGATGGCAGG G
myco-gentl    ATTTGCCCCG GAAATTCCCG CAAAAACCCC AGTAAGTTGG GGATGGCAAA
```


Phylogenetic methodologies

- Define a specific series of steps to produce the 'best' tree
  - Pair-group cluster analyses
  - Fast, but tend not to address underlying evolutionary mechanisms
- Define criteria for comparing different trees and judging which is better. Two steps:
  - Define the objective function (evolutionary biology)
  - Generate and compare trees (computation)
- All of the techniques described produce an unrooted tree.
- The trees produced likewise describe relationships among extant taxa, not the progress of evolution over time.

Distance-based Tree-building methods

- Aligned sequences are compared, and analysis is based on the differences between sequences, rather than the original sequence data.
- Less computationally intensive than character-based methods
- Tend to be problematic when sequences are highly divergent

Distance-based Tree building methods, 2

- Cluster analysis. Most common variant is Unweighted Pair Group Method with Arithmetic Mean (UPGMA) – join two closest neighbors, average pair, keep going. Problematic when highly diverged sequences are involved
- Additive tree methods – built on assumption that the lengths of branches can be summed to create some measure of overall evolution.
  - Fitch-Margoliash (FM) – minimizes squared deviation between observed data and inferred tree.
  - Minimum evolution (ME) – finds shortest tree consistent with data
- Of the distance methods, ME is the most widely implemented in computer programs

Character-based methods

- Use character data (actual sequences) rather than distance data
- Maximum parsimony. Creates shortest tree – one with fewest changes. Inter-site rate heterogeneity creates difficulties for this approach.
- Maximum likelihood. Searches for the evolutionary model that has the highest likelihood value given the data. In simulation studies ML tends to outperform others, but is also computationally intensive.

Rooting trees

- If the assumption of a constant molecular clock holds, then the root is the midpoint of the longest span across the tree.
- Sometimes done by including an 'outgroup' in the analysis
- Remember that the trees produced from sequence data are fundamentally different than a historical evolutionary tree

Evaluating trees

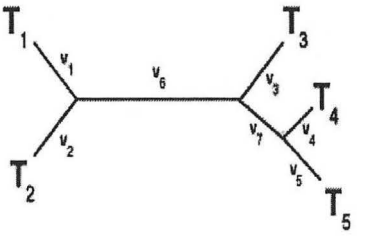
- Once a phylogenetic tree has been produced by some means, how do you test whether or not the tree represents evolutionary change, or just the results of a mathematical technique applied to a set of random data? These methods below can be used to perform a statistical significance test.
- Significance tests for MP trees:
  - Skewness tests. MP tree lengths produced from random data should be symmetric; tree lengths produced from data sets with real signal should be skewed.
- Significance tests for distance, MP, and ML trees:
  - Bootstrap. Recalculate trees using multiple samples from same data with resampling.
  - Jackknife. Recalculate trees using subsampling
- All of these methods are topics of active debate

Phylogenetic software

- Phylip. (J. Felsenstein). Collection of software packages that cover most types of analysis. One of the most popular software collections. Free.
- PAUP. (D. Swofford). Parsimony, distance, and ML methods. Also one of the most popular software collections. Not free, but not expensive.
- PAML. (Ziheng Yang). Maximum likelihood methods for DNA and proteins. Not as well suited for tree searching, but performs several analyses not generally available. Free.
- fastDNAml. (G. Olsen). Maximum likelihood method for DNA; becoming one of the more popular ML packages. MPI version available soon; well suited to tree searching in large data sets. Free.

More on Maximum Likelihood methods

- Typical statistical inference: calculate probability of data given the hypothesis.
- Tree, branch lengths, and associated likelihood values all calculated from the data.
- Likelihood values used to compare trees and determine which is best.



Stochastic change of DNA

- Markov process, independent for each site: 4 x 4 matrix for DNA, 20 x 20 for amino acids

	A	C	G	T
A	p(A->A)	p(A->C)	p(A->G) ...	
C	p(C->A)	p(C->C)	p(C->G) ...	
G	.			
T	.			

- Transitions more probable than transversions.
- Must account for heterogeneity in substitution rates among sites (DNArates – Olsen)

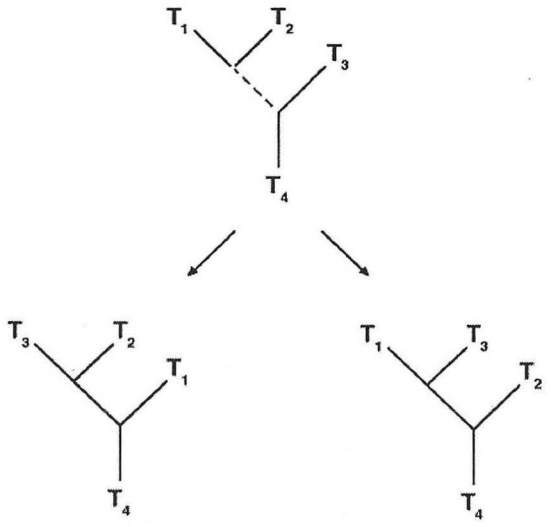
fastDNAml

- Developed by Gary Olsen
- Derived from Felsensteins's PHYLIP programs
- One of the more commonly used ML methods
- The first phylogenetic software implemented in a parallel program (at Argonne National Laboratory, using P4 libraries)
- Olsen, G.J.,et al.1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Computer Applications in Biosciences 10: 41-48
- MPI version produced in collaboration with Indiana University will be available soon

fastDNAml algorithm

- Compute the optimal tree for three taxa (chosen randomly) - only one topology possible
- Randomly pick another taxon, and consider each of the 2i-5 trees possible by adding this taxon into the first, three-taxa tree.
- Keep the best (maximum likelihood tree)
- Local branch rearrangement: move any subtree to a neighboring branch (2i-6 possibilities)
- Keep best resulting tree
- Repeat this step until local swapping no longer improves likelihood value

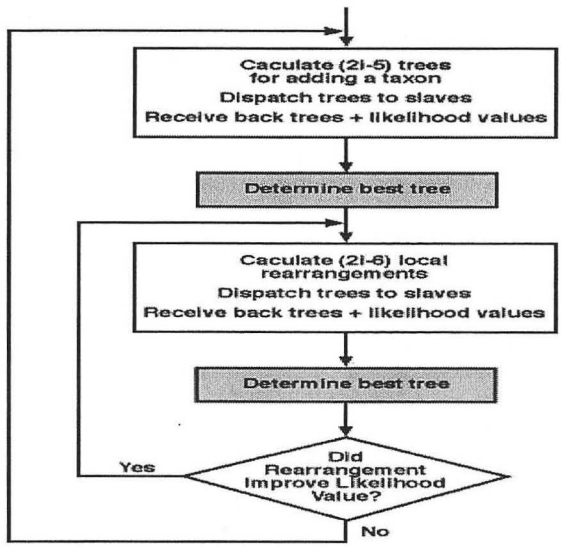
Local branch rearrangement diagram



fastDNAml algorithm con't: Iterate

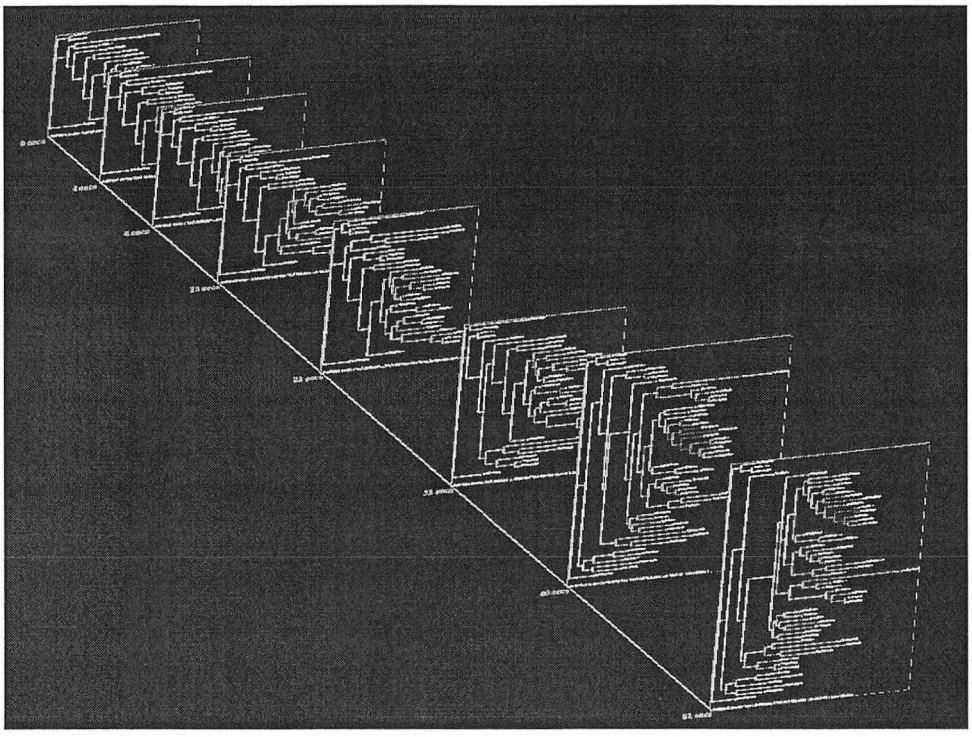
- Get sequence data for next taxon
- Add new taxa (2i-5)
- Keep best
- Local rearrangements (2i-6)
- Keep best
- Keep going....
- When all taxa have been added, perform a full tree check

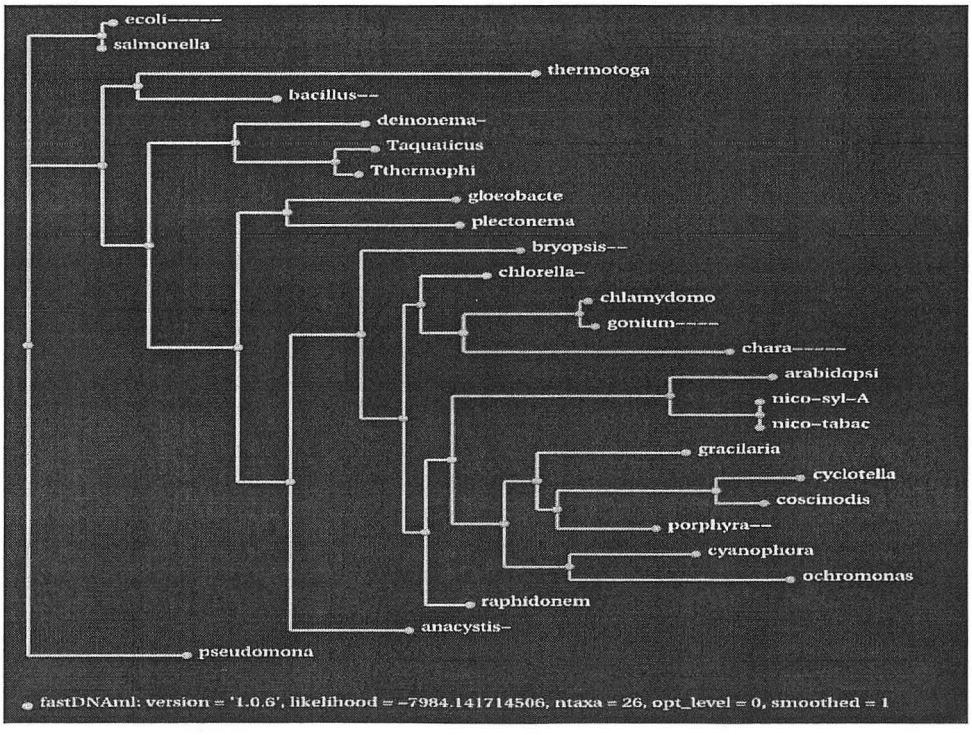
Overview of parallel program flow



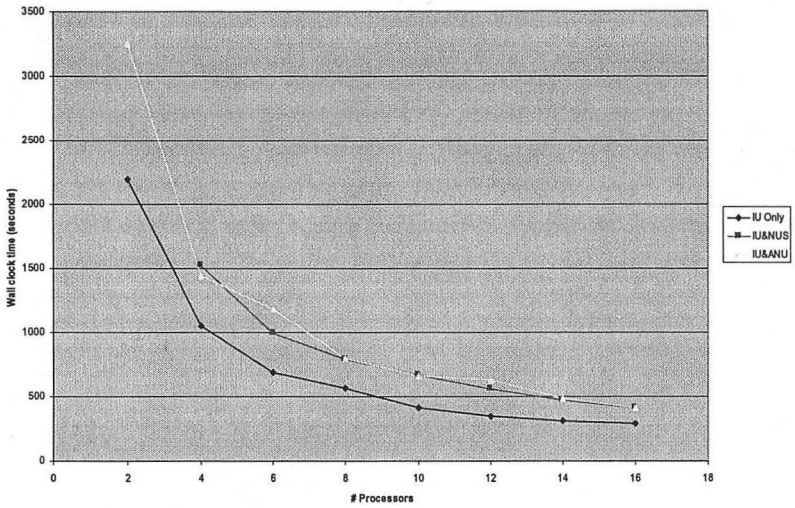
Because of local effects....

- Where you end up sometimes depends on where you start
- This process searches a huge space of possible trees, and is thus dependent upon the randomly selected initial taxa
- Can get stuck in local optimum, rather than global
- Must do multiple runs with different randomizations of taxon entry order, and compare the results
- Similar trees and likelihood values provide some confidence, but still the space of all possible trees has not been searched extensively




ecoli-----
salmonella
thermotoga
bacillus--
deinonema-
Taquaticus
Tthermophi
gloeobacte
plectonema
bryopsis--
chlorella-
chlamydomo
gonium----
chara-----
arabidopsi
nico-syl-A
nico-tabac
gracilaria
cyclotella
coscinodis
porphyra--
cyanophora
ochromonas
raphidonem
anacystis-
pseudomona
fastDNAml: version = '1.0.6', likelihood = -7984.141714506, ntaxa = 26, opt_level = 0, smoothed = 1

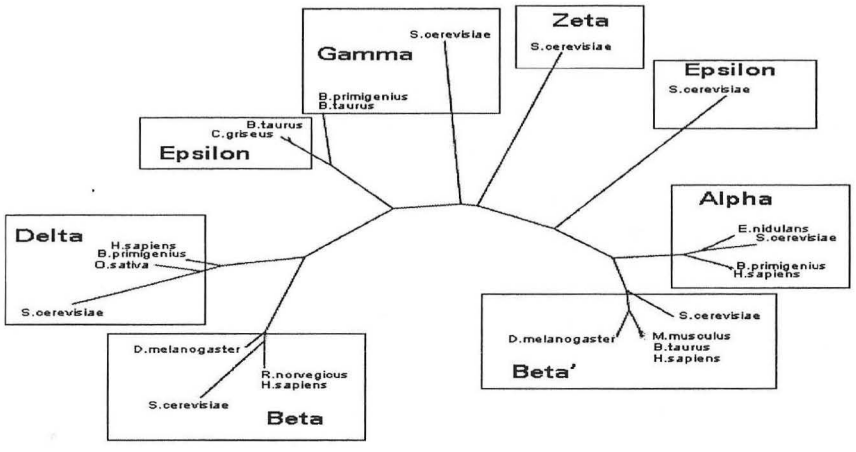
Performance of fastDNAml



Applications & Interesting examples

- Better understanding of evolution (Ceolocanths, cyanobacterial origin of plastids)
- Maintenance of biodiversity
- Medicine & molecular biology
  - our cousins, the fungi
  - Cytoplasmic coat proteins
  - HIV

Cytoplasmic Coat Proteins



HIV

- Where did HIV come from, and how recent is it?
- Korber, et al. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789. (Online at www.sciencemag.org/cgi/content/full/288/5472/1789)
- Used completed HIV sequences from 159 individuals with known sampling dates (including one from 1959)
- Used a general-reversible (REV) base substitution model, accounting for different site-specific rates of evolution and base frequencies biased in favor of adenosine. Used modified version of fastDNAml.
- Used SIV as an outgroup
- Last common ancestor of main group of HIV-1 was 1931 (95% confidence interval: 1915-1941). Supports hypothesis that HIV has been around for some time and simply took a while to be common enough to be noticed.

Challenges for future

- HPC implementations of more phylogenetic techniques
- Better treatment of insertions and deletions (indels)
- Algorithms for more thorough searching of treespaces in incremental tree building processes (keep best n trees and keep looking)
- Techniques for not shaking the whole tree (that is, adding a taxa to a tree in a fashion that acknowledges damping of effect as you travel away from altered part of tree)
- Use of high-throughput techniques

Acknowledgements

- The phylogeny depicted in slide 5 is taken from E. Colbert. 1965. The age of reptiles. W.W. Norton, NY, NY.
- Some of the tree diagrams were adapted from Olsen *et al.* 1994.
- Les Teach [IU] created all other graphics for this talk.
- IU's work on parallel versions of fastDNAml has been facilitated by Shared University Research grants from IBM, Inc.
- IU's work with fastDNAml would be impossible without our collaboration with Gary Olsen, U. of Illinois, the creator of this program.

References

- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368-376
- Baxevanis, A.D., and B.F.F. Ouellette. 1998. Bioinformatics: a practical guide to the analysis of genes and proteins.Wiley-Interscience, NY.
- Swofford, D.L., and G.J. Olsen. Phylogeny reconstruction. pp. 411-501 IN D.M. Nillis & C. Mority (eds). Molecular systematics. Sinauer Associates, Sunderland, MA.
- Durbin, R. et al. 1998. Biological sequence analysis. Cambridge University Press, Cambridge, UK.
- www.ucmp.berkely.edu/subway/phylogen
- evolution.genetics.washington.edu/phylip/software
- http://www.indiana.edu/uits/~rac

urls for phylogenetic software

- **Phylip** evolution.genetics.washington.edu/phylip/software.html
- **PAUP**
  www.lms.si.edu/PAUP/index.html
- **PAML**
  abacus.gene.ucl.ac.uk/software/paml.html
- **fastDNAml**
  geta.life.uiuc.edu/~gary/

Afternoon Break

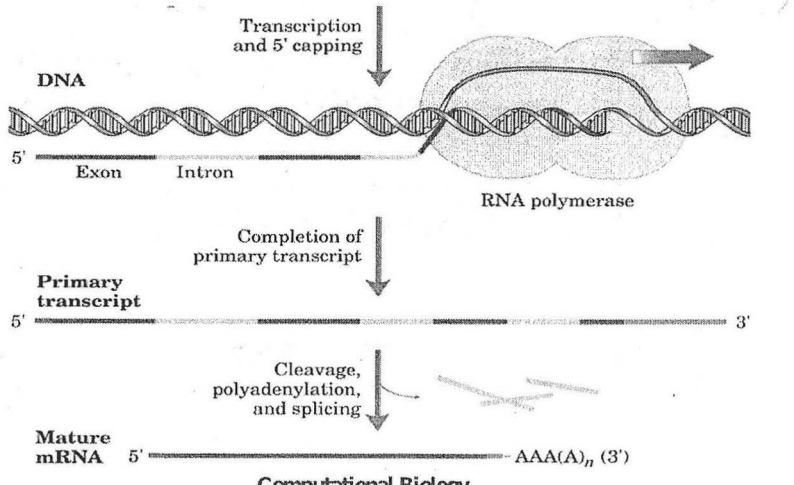
Specialized biological databases and their role in building models of regulation

Inna Dubchak
ILDubchak@lbl.gov
NERSC

Overview of alternative splicing

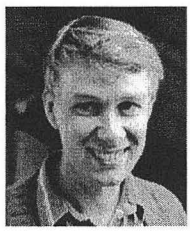
- What is alternative splicing?
- What is possible to do computationally to better understand this complicated phenomenon?
  - Frequency of alternative splicing
  - Specialized databases
  - Search for regulatory elements

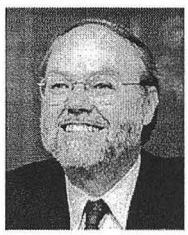
PROCESSING mRNA



The Nobel Prize in Physiology or Medicine 1993

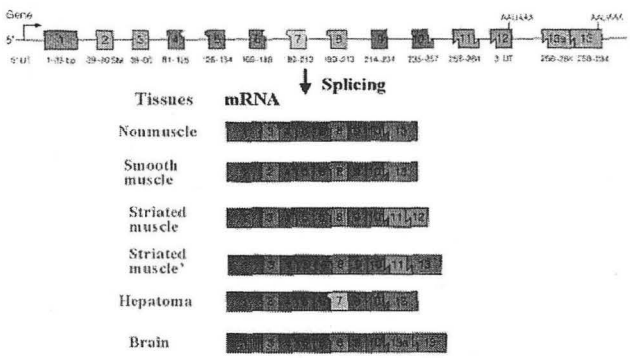
The Nobel Assembly at the Karolinska Institute in Stockholm, Sweden, has awarded the Nobel Prize in Physiology or Medicine for 1993 jointly to Richard J. Roberts and Phillip A. Sharp for their discovery of split genes.





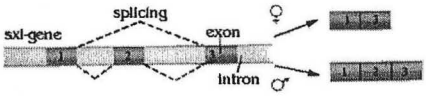
a-Tropmyocin pre-mRNA

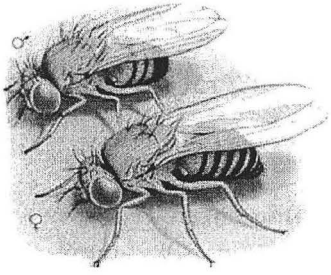
Alternative Splicing of a-tropomyocin pre-mRNA



Gender in Drosophila

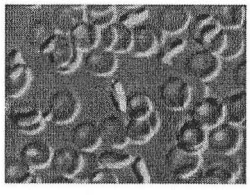
- A percursor-RNA may often be matured to mRNAs with alternative structures. An example where alternative splicing has a dramatic consequence is somatic sex determination in the fruit fly Drosophila melanogaster.
- In this system, the female-specific sxl-protein is a key regulator. It controls a cascade of alternative RNA splicing decisions that finally result in female flies.
- Sex in Drosophila is largely determined by alternative splicing





Splicing and diseases

- Splicing errors cause thalassemia
- Thalassemia, a form of anemia common in the Mediterranian countries, is caused by errors in the splicing process.
- Normal red blood cells contain correctly spliced beta-globin, an important component in hemoglobin that takes up oxygen in the lungs.



Information on alternative splicing in public databases:

- Swiss-Prot (protein) database is well curated, but the information content is incomplete with reference to alternative splicing and does not allow for automatic retrieval of such entries.
- Swiss-Prot entries just state the fact that a particular protein is one of the products of alternative splicing.
- Some entries contain the information on the limited number of isoforms.

Clustering procedure

Similarity analysis of two sequences

- Gene families
  multiple similar genes exist due too duplication and divergence of genes.

↓

- Short similar fragments, a lot of mutations

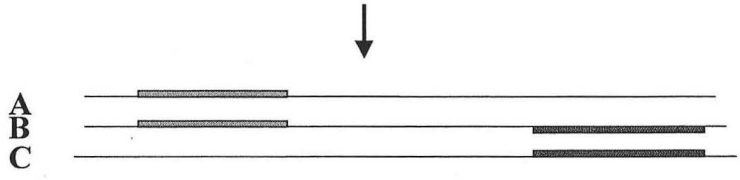
- Alternative splicing
  one gene but primary transcript spliced in more than one way

↓

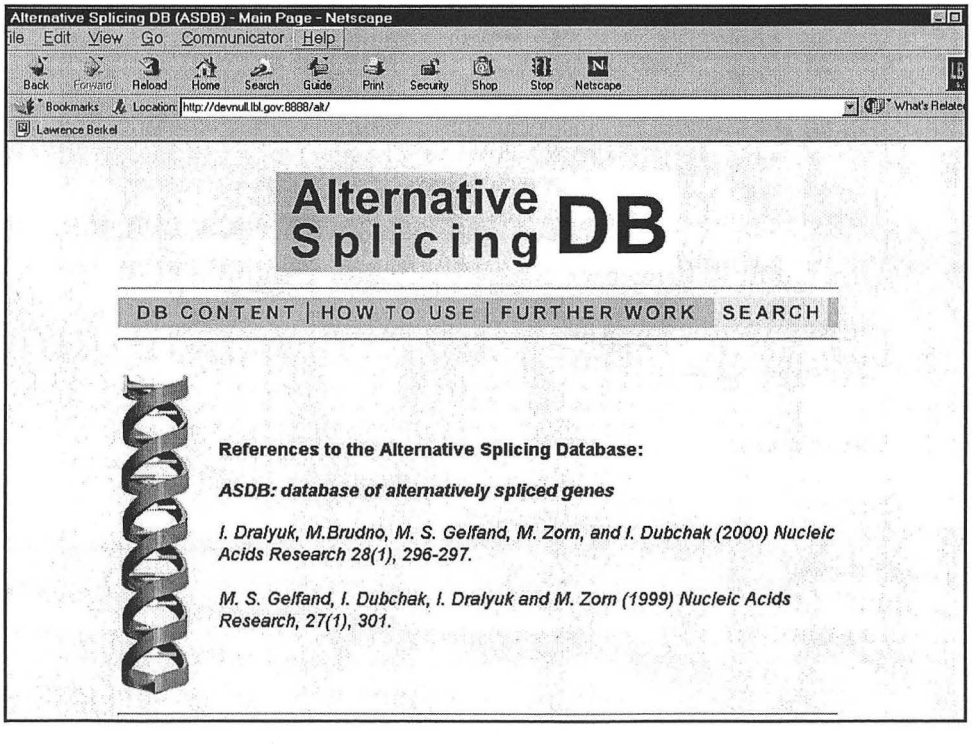
- Relatively long identical fragments

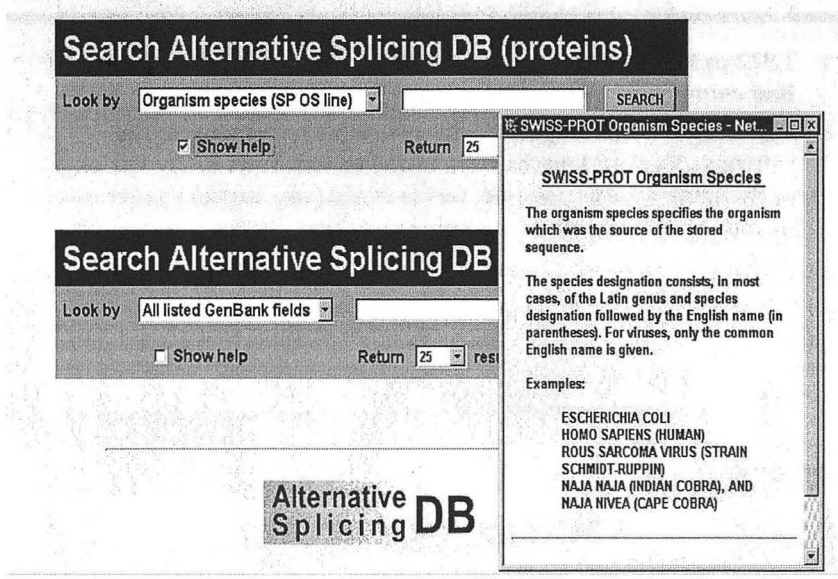
Clustering procedure

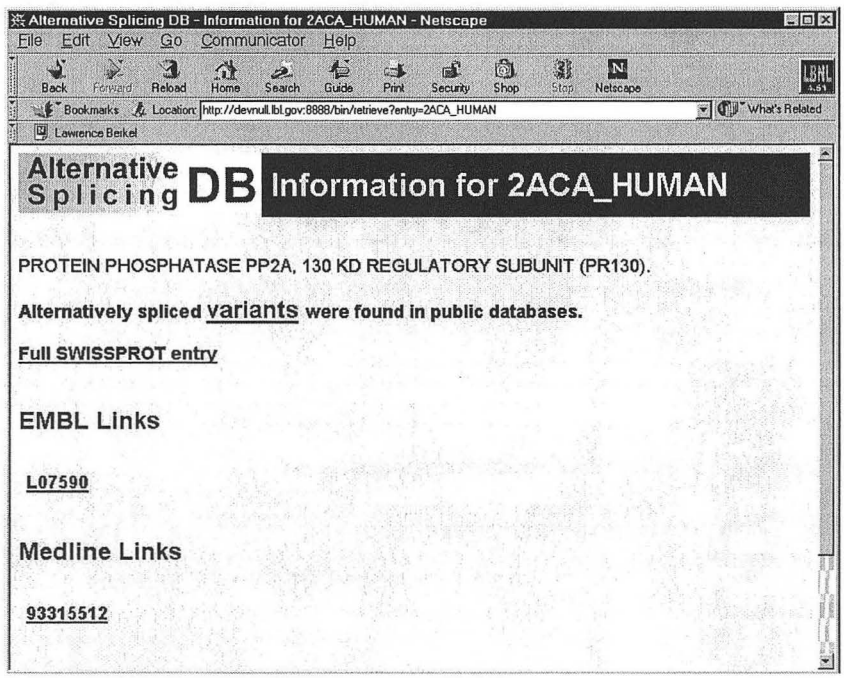
- 1,922 protein sequences were compared all-against-all in order to find common sequence fragments.
- The length of this fragment was a variable parameter in the software. Various lengths were tested to cluster as many variants of the same gene as possible, but to avoid false clusters generated by too short fragments.

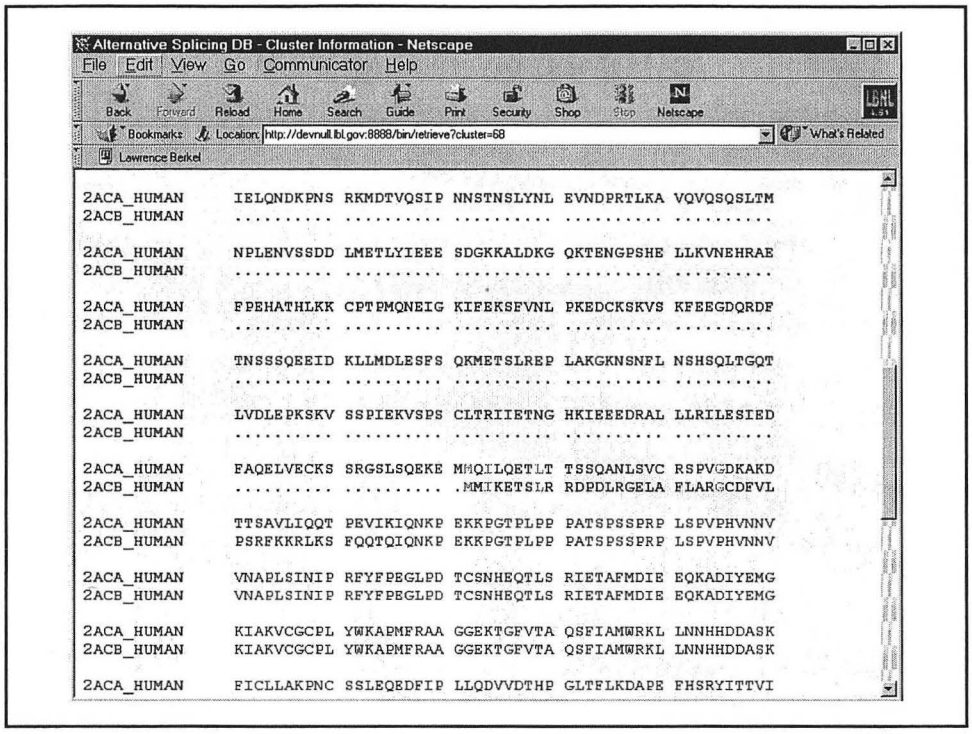


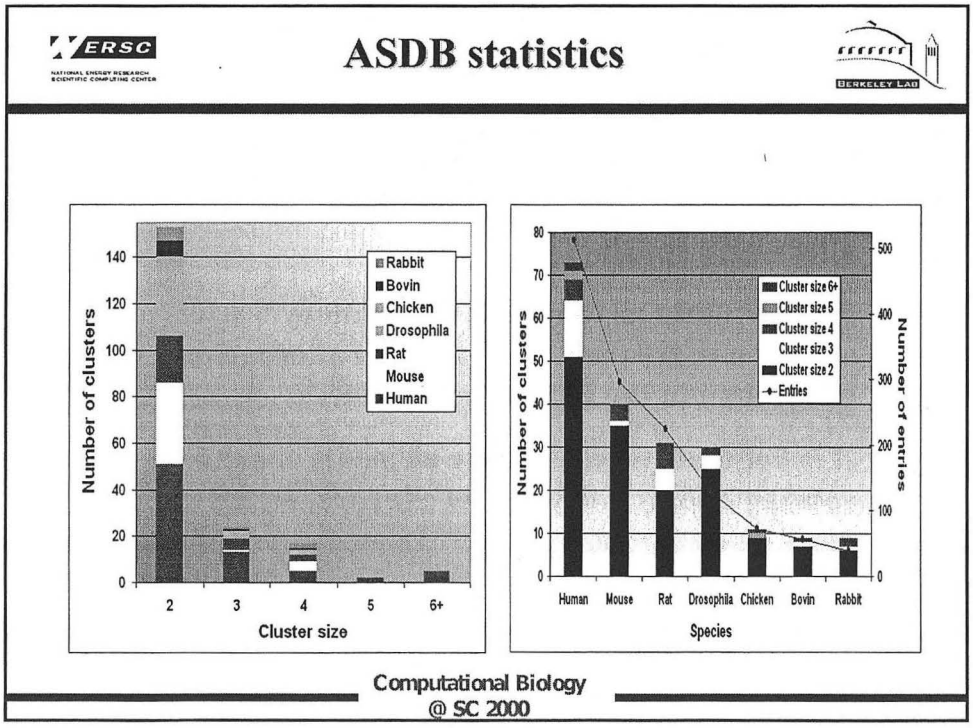
~240 clusters of isoforms



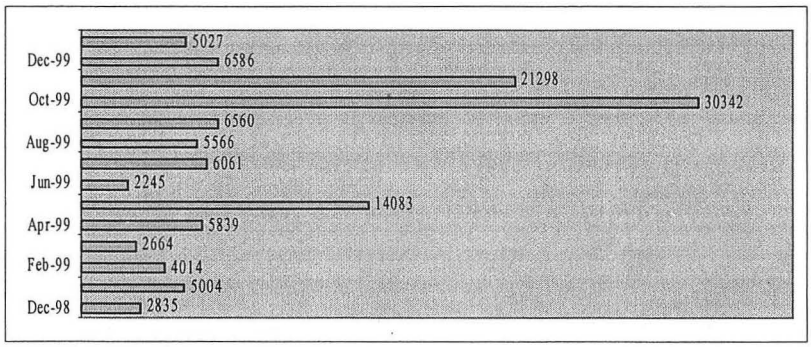

Search Alternative Splicing DB (proteins)
Look by
Organism species (SP OS line)
SEARCH
Show help
Return
25
SWISS-PROT Organism Species - Net...
SWISS-PROT Organism Species
The organism species specifies the organism which was the source of the stored sequence.
The species designation consists, in most cases, of the Latin genus and species designation followed by the English name (in parentheses). For viruses, only the common English name is given.
Examples:
ESCHERICHIA COLI
HOMO SAPIENS (HUMAN)
ROUS SARCOMA VIRUS (STRAIN SCHMIDT-RUPPIN)
NAJA NAJA (INDIAN COBRA), AND
NAJA NIVEA (CAPE COBRA)
Search Alternative Splicing DB
Look by
All listed GenBank fields
Show help
Return
25
Alternative Splicing DB


Alternative Splicing DB - Information for 2ACA_HUMAN - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Shop Stop Netscape
Bookmarks Location: http://devnull.lbl.gov:8888/bin/retrieve?entry=2ACA_HUMAN
What's Related
Lawrence Berkel
Alternative Splicing DB
Information for 2ACA_HUMAN
PROTEIN PHOSPHATASE PP2A, 130 KD REGULATORY SUBUNIT (PR130).
Alternatively spliced variants were found in public databases.
Full SWISSPROT entry
EMBL Links
L07590
Medline Links
93315512


Alternative Splicing DB - Cluster Information - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Shop Stop Netscape
Bookmarks Location: http://devnull.lbl.gov:8888/bin/retrieve?cluster=68
What's Related
Lawrence Berkel
2ACA_HUMAN IELQNDKPNS RKMDTVQSIP NNSTNSLYNL EVNDPRTLKA VQVQSQSLTM
2ACB_HUMAN ....
2ACA_HUMAN NPLENVSSDD LMETLYIEEE SDGKKALDKG QKTENGPSHE LLKVNEHRAE
2ACB_HUMAN ....
2ACA_HUMAN FPEHATHLKK CPTPMQNEIG KIFEKSFVNL PKEDCKSKVS KFEEGDQRDF
2ACB_HUMAN ....
2ACA_HUMAN TNSSSQEEID KLLMDLESFS QKMETSLREP LAKGKNSNFL NSHSQLTGQT
2ACB_HUMAN ....
2ACA_HUMAN LVDLEPKSKV SSPIEKVSPS CLTRIIETNG HKIEEEDRAL LLRILESIED
2ACB_HUMAN ....
2ACA_HUMAN FAQELVECKS SRGSLSQEKE MMQILQETLT TSSQANLSVC RSPVGDKAKD
2ACB_HUMAN .... .MMIKETSLR RDPDLRGELA FLARGCDFVL
2ACA_HUMAN TTSAVLIQQT PEVIKIQNKP EKKPGTPLPP PATSPSSPRP LSPVPHVNNV
2ACB_HUMAN PSRFKKRLKS FQQTQIQNKP EKKPGTPLPP PATSPSSPRP LSPVPHVNNV
2ACA_HUMAN VNAPLSINIP RFYFPEGLPD TCSNHEQTLS RIETAFMDIE EQKADIYEMG
2ACB_HUMAN VNAPLSINIP RFYFPEGLPD TCSNHEQTLS RIETAFMDIE EQKADIYEMG
2ACA_HUMAN KIAKVCGCPL YWKAPMFRAA GGEKTGFVTA QSFIAMWRKL LNNHHDDASK
2ACB_HUMAN KIAKVCGCPL YWKAPMFRAA GGEKTGFVTA QSFIAMWRKL LNNHHDDASK
2ACA_HUMAN FICLLAKPNC SSLEQEDFIP LLQDVVDTHP GLTFLKDAPE FHSRYITTVI


ERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER
ASDB statistics
BERKELEY LAB
Number of clusters
Cluster size
2
3
4
5
6+
Rabbit
Bovin
Chicken
Drosophila
Rat
Mouse
Human
Number of clusters
Number of entries
Species
Human
Mouse
Rat
Drosophila
Chicken
Bovin
Rabbit
Cluster size 6+
Cluster size 5
Cluster size 4
Cluster size 3
Cluster size 2
Entries
Computational Biology
@ SC 2000

ASDB usage during 1999



Study of Regulation

- No systematic surveys to address the relative importance of such elements in the regulation of alternative splicing.
- It is unknown as to whether regulatory words occur more frequently adjacent to alternative exons than in the rest of the genome.
- It is not clear whether these elements enhance splicing of only a limited set of exons, or have a more general role.

Alternative Splicing Regulation

- A number of genomic sequence regulatory elements have been identified outside of traditional splice sites.
- The concept of splicing "enhancers" and "silencers" that promote or inhibit splicing at neighboring splice sites is well established.
- Many alternative exons are probably regulated by a combination of silencers and enhancers.

Data Collection

- Automated processing of GenBank/Medline
- Manual analysis of abstracts & articles
- Collecting the sample

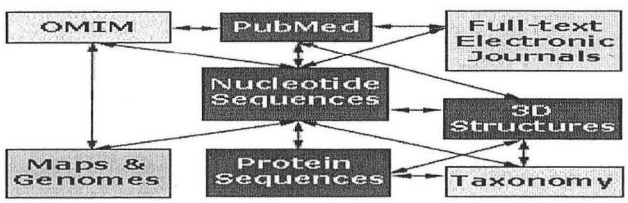
BiSyCLES Search Options

- BiSyCLES searches in the two databases, then establishes which of the retrieved entries are linked
  - Medline: +"alternative splicing," tissue, muscle, brain, neuro*, heart, regul*, enhancer, silencer
  - Genbank: +"alternative splicing" +"complete CDS"

- Results:
  - ~300 abstracts
  - ~50 relevant papers

BiSyCLES: Biological System for Cross-Linked Entry Search

- GenBank contains genomic data but little annotation
- Medline (PubMed) contains abstracts from journals but no genomic data
- NCBI's Entrez system keeps links between related entries in its databases



Word Counting

- To calculate the confidence value of a particular word we select random subsets of a large dataset of constitutively spliced exons (1,504 exons; Burset & Guigo, 1996) equal in size to our alternative dataset.
- We then calculate the fraction of these subsets in which the word is over-represented at a higher rate than in the alternative set.
- (Over-representation is calculated as difference of frequencies)

Known Regulatory Elements

enhancers	reference
UGCAUG	Huh & Hynes, 1994; Hedjran et al., 1997; Modafferi & Black, 1997; Kawamoto, 1996; Carlo et al., 1996
CUG repeat	Ryan et al., 1996; Philips et al., 1998
(A/U)GGG	Sirand-Pugnet et al., 1995a
GGGGCUG	Carlo et al., 1996
silencers	
UUCUCU	Chan & Black, 1995; Chan & Black, 1997; Ashiya & Grabowski, 1997

Short summary

- In the simple cases of splicing, introns are always introns and exons are always exons
- During alternative splicing, within the same RNA, sequences can be recognized as either intron or exon under different conditions and the concept of exons and introns becomes rather empirical
- RNAs are not spliced differently in the same cell at the same time but in different cells or in the same cell types at different times in development or under different conditions
- A variety of patterns of alternate splicing have been observed.

Evolutionarily conserved non-coding DNA sequences

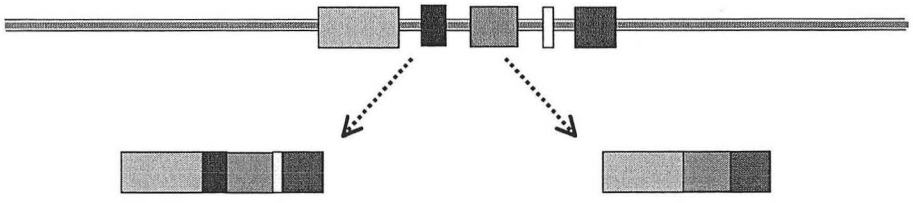
- Discovering them in DNA sequence
- Tools for their visualization
- Biological importance

Non-coding Sequences

Non-Coding

~ 5% coding
~ 95% non-coding

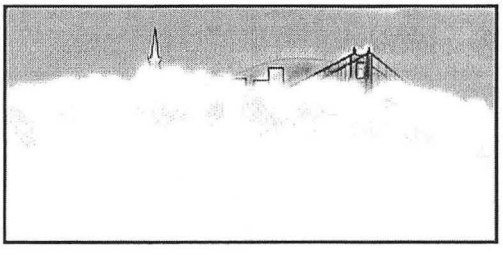
Gene A



Protein A

Protein A'

Information in Sequence



Conserved Human/Mouse Sequences in 830 kb Region

> 40 bp and > 90%
OR > 60 bp and > 80%
OR > 100 bp and > 70%

411

270
Transcribed

141
Non-Transcribed

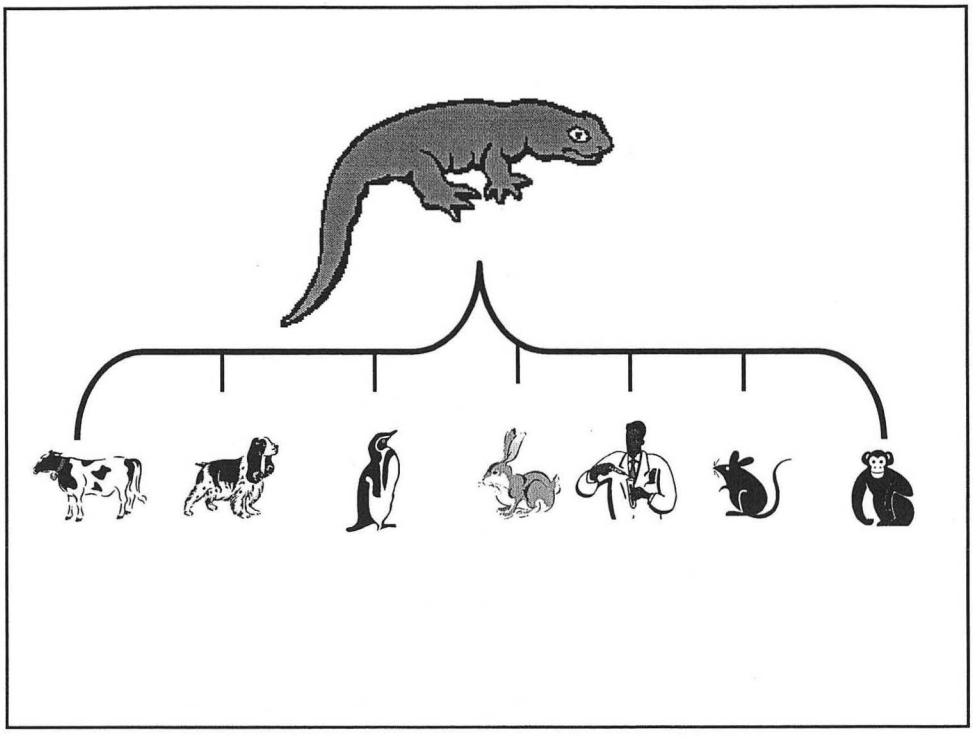
Introns
59%

< 1 k from gene
8%

> 1 kb from gene
33%

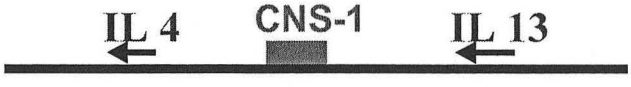
90 Elements in 1 Megabase

Are most conserved noncoding sequences "functional" or are they a product of passive evolution?



Analysis of CNS-1

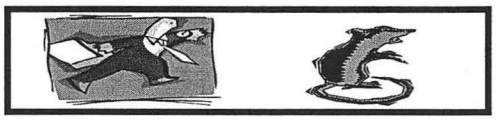
- Present in other species:
  - Cow (86%)
  - Dog (81%)
  - Rabbit (73%)
- Genomic position conserved in human, mouse, dog and baboon



- Single copy in the human genome

Evolutionarily Conserved Non-Coding Sequences

Identification



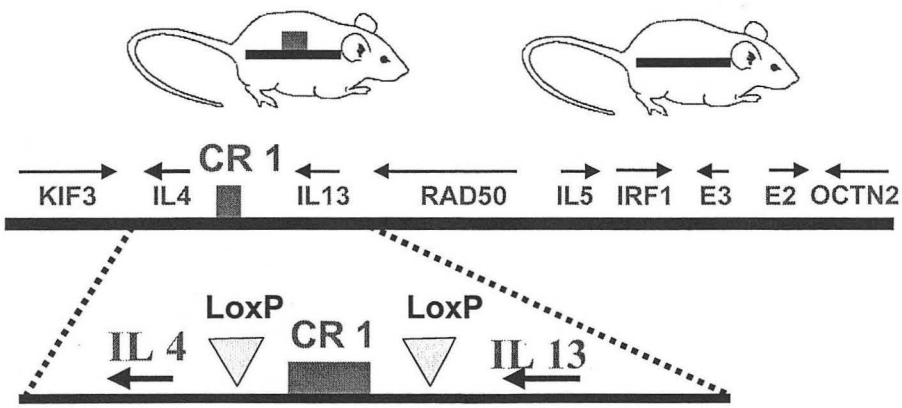
Verification

Analysis

Computational Biology
@ SC 2000

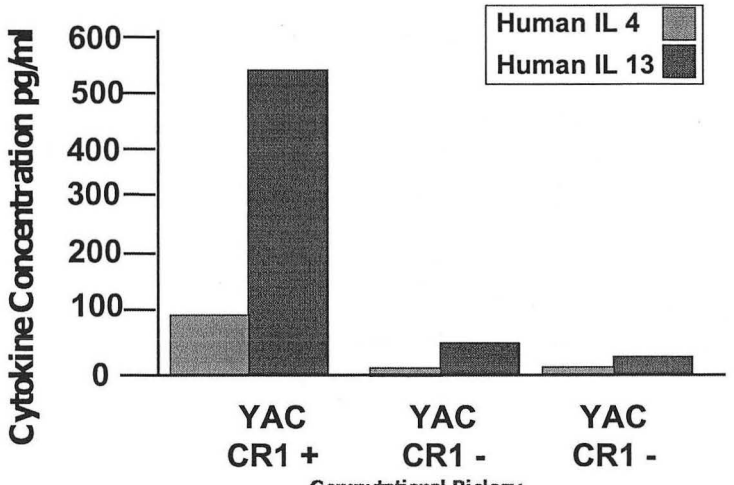
Functional Analysis of CR 1

Generate Human 5q31 YAC Transgenic Mice

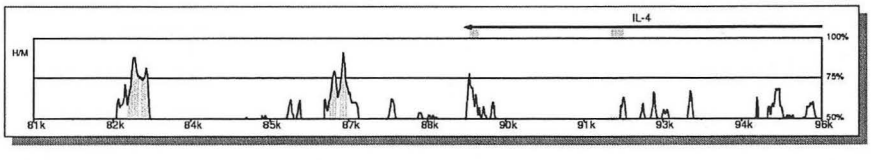


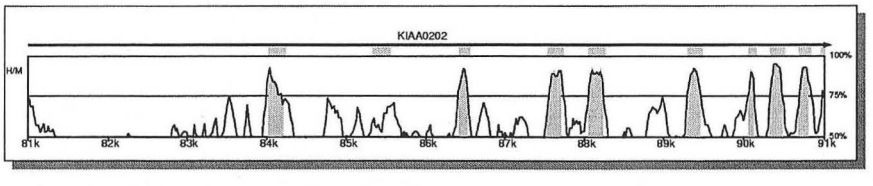
Computational Biology
@ SC 2000

Human IL4 and IL13 Production in YAC Transgenics Containing and Lacking CR1

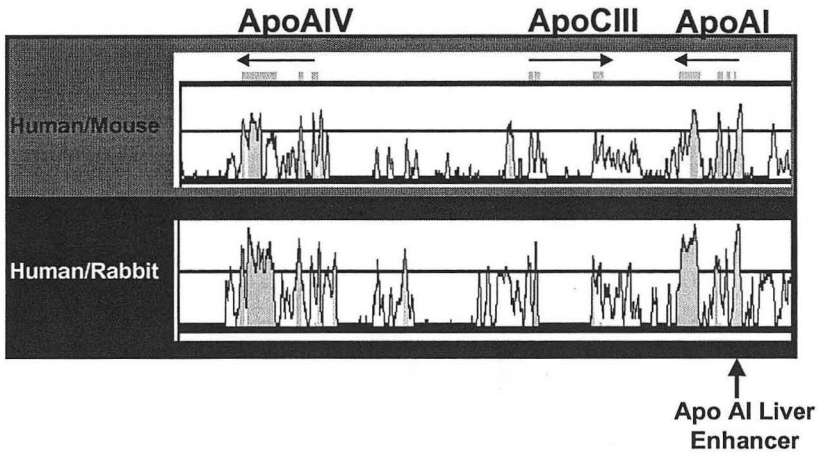


Vista (Visual Tool for Alignment)

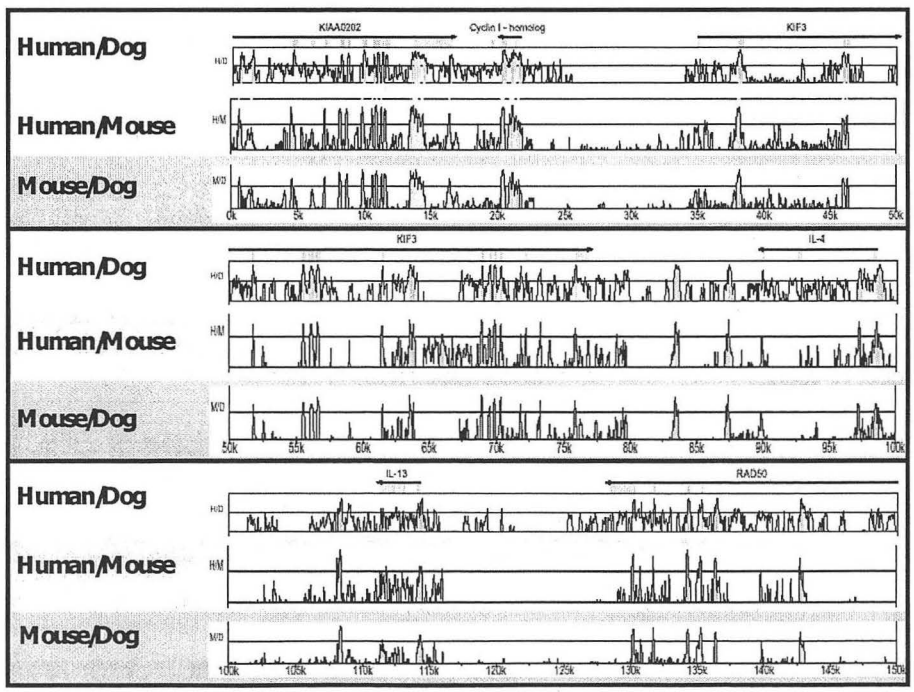


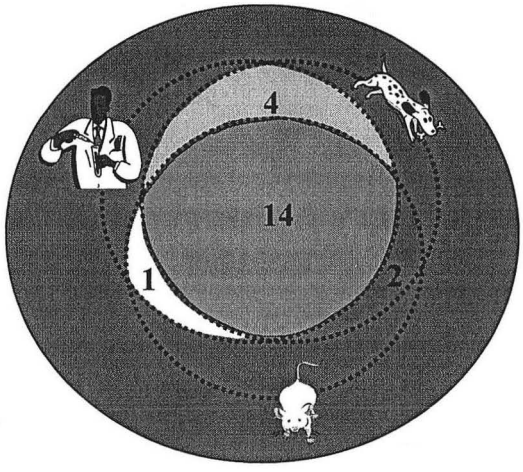


Comparative Genomic Sequence Analysis of Human/Mouse/Rabbit ApoAI, CIII, AIV Cluster

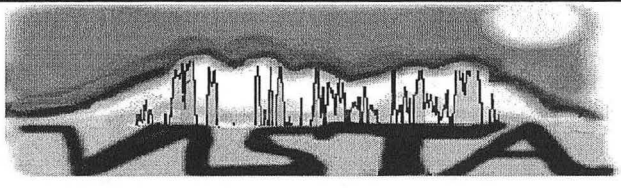


Computational Biology
@ SC 2000





http://www-gsd.lbl.gov/vista/



Welcome to theVISTA, or VISualization Tool for Alignments home page

VISTA is an integrated system for global alignment and visualization, designed for comparative genomic analysis.

1. *The visual output is clean and simple, allowing the user to easily identify conserved regions.*
2. *Similarity scores are displayed for the entire sequence, thus allowing for the identification of shorter conserved regions, or regions with gaps.*

Gene Regulatory Networks and Cellular Processes

Adam Arkin
APArkin@lbl.gov
LBNL

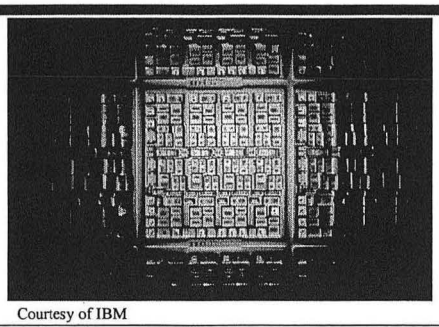
NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

BERKELEY LAB

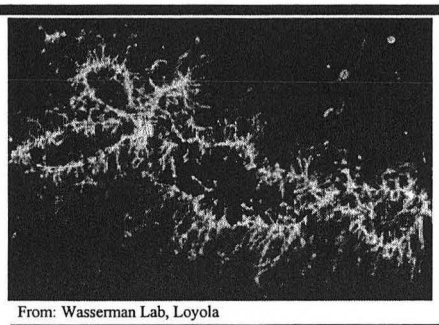
Cells

Computational Biology
@ SC 2000

Engineering of Cellular Circuitry



Courtesy of IBM



From: Wasserman Lab, Loyola

Asynchronous Digital Telephone Switching Circuit

Full knowledge of parts list
Full knowledge of "device physics"
Full knowledge of interactions

No one fully understands how this circuit works!!
Its just too complicated.

Designed and prototyped on a computer (SPICE analysis)
Experimental implementation fault tested on computer

Asynchronous Analog Biological Switching Circuit

Partial knowledge of parts list
Partial knowledge of "device physics"
Partial knowledge of interactions

No one fully understands how this circuit works!!
Its just too complicated.

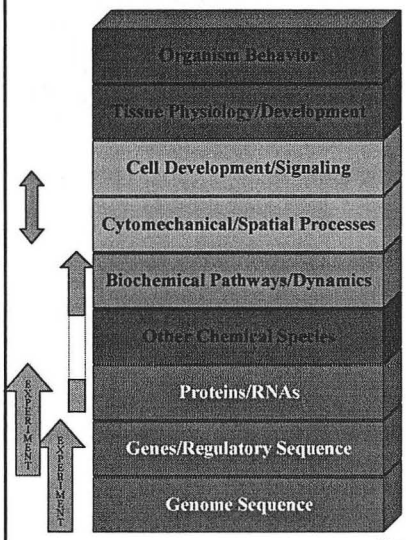
We ***need*** a SPICE-like analysis for biological systems

A foundation for cell network analysis

In analogy to the steps necessary to allow design, control and diagnosis in electronics we must perform the following (non-sequential) tasks:

- Compile a list of parts that make up your system.
- Determine "device physics" for the parts and their interactions.
- Group subsets of parts based on function, physics and interactions
- Create databases of the above and the "circuits" that are to be considered for analysis
- Create quantitative simulations and analyses of the dynamics of complex networks.

Analysis of Cell Function

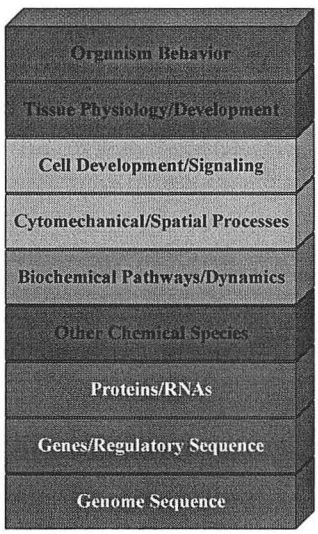


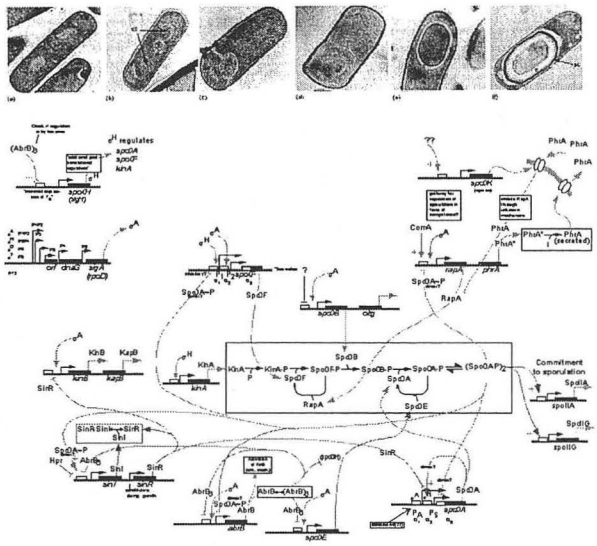
The challenge is to integrate data from all levels to produce a description of cellular function.

- **There are challenges in:**
  - **Systematization and structuring of data**
  - **Serving and query this data**
  - **Representing the data**
  - **Building multiscale, multi-resolution models**
  - **Dynamic and static analysis of these models**
- **Pay-off in**
  - **Industrial bioengineering**
  - **Rational pharmaceutical design**
  - **Basic biological understanding**

Computational Biology
@ SC 2000

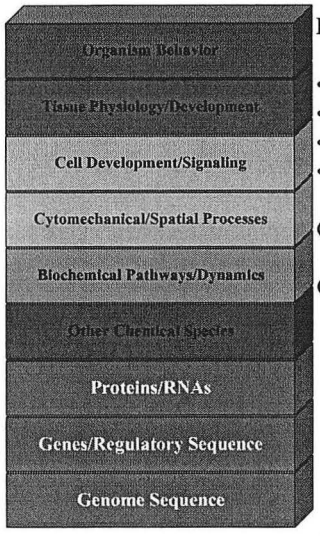
Complexities of Cellular Function





Computational Biology
@ SC 2000

Heterogeneity of Data

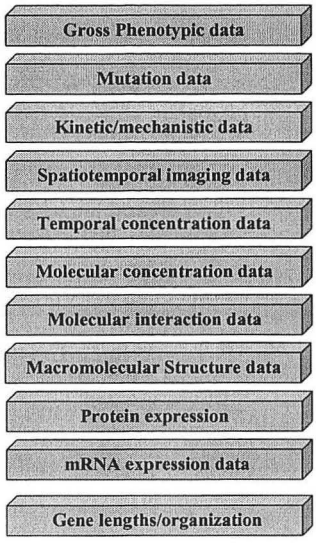


Data are:

- Qualitative>Quantitative
- Collected at many levels
- Of heterogeneous structure
- Of heterogeneous availability

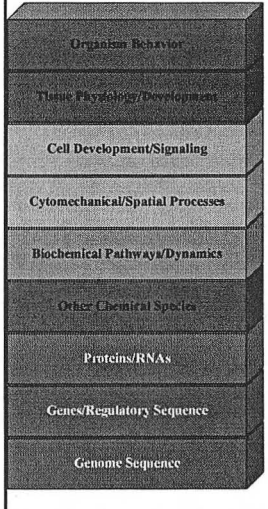
Challenge:

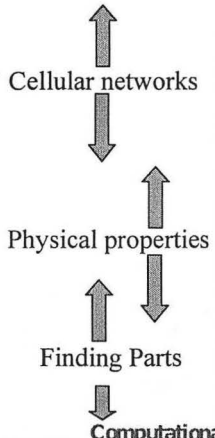
Optimal use of available data to make predictions about cell function and failure.

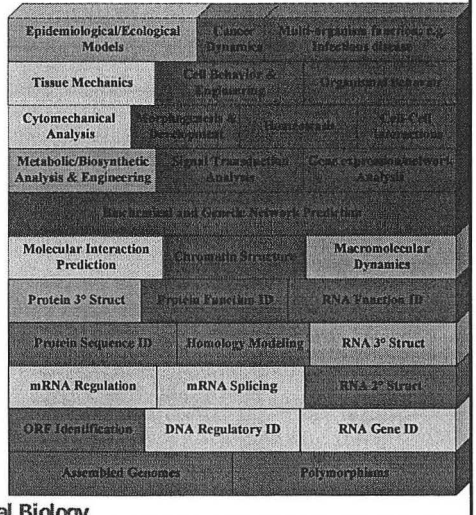


Computational Biology
@ SC 2000

Tools for "multilevel" analysis







Computational Biology
@ SC 2000

Why now?

•Genome projects are providing a large (but partial) list of parts

•New measurement technologies are helping to identify further components, their interactions, and timings

- Gene microarrays
- Two-Hybrid library screens
- High-throughput capillary electrophoresis arrays for DNA, proteins and metabolites
- Fluorescent confocal imaging of live biological specimens
- High-throughput protein structure determination

•Data is being compiled, systematized, and served at an unprecedented rate
- Growth of GenBank and PDB > polynomial
- Proliferation of databases of everything from sequence to confocal images to literature

•The tools for analyzing these various sorts of data are also multiplying at an astounding rate

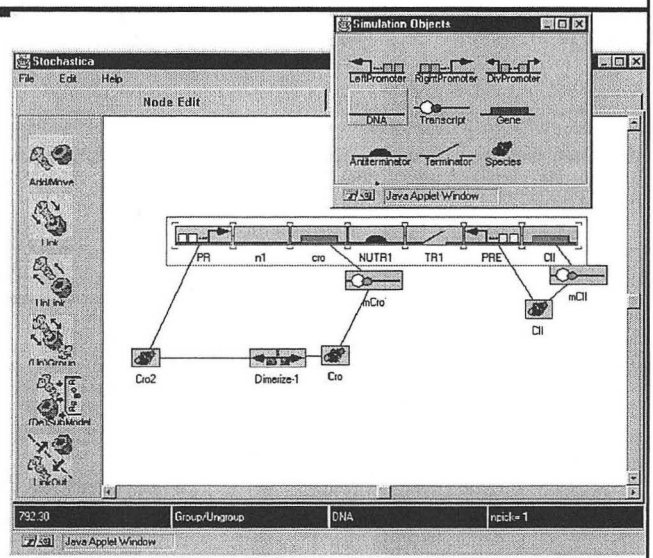
SPICE Tools for Biology?

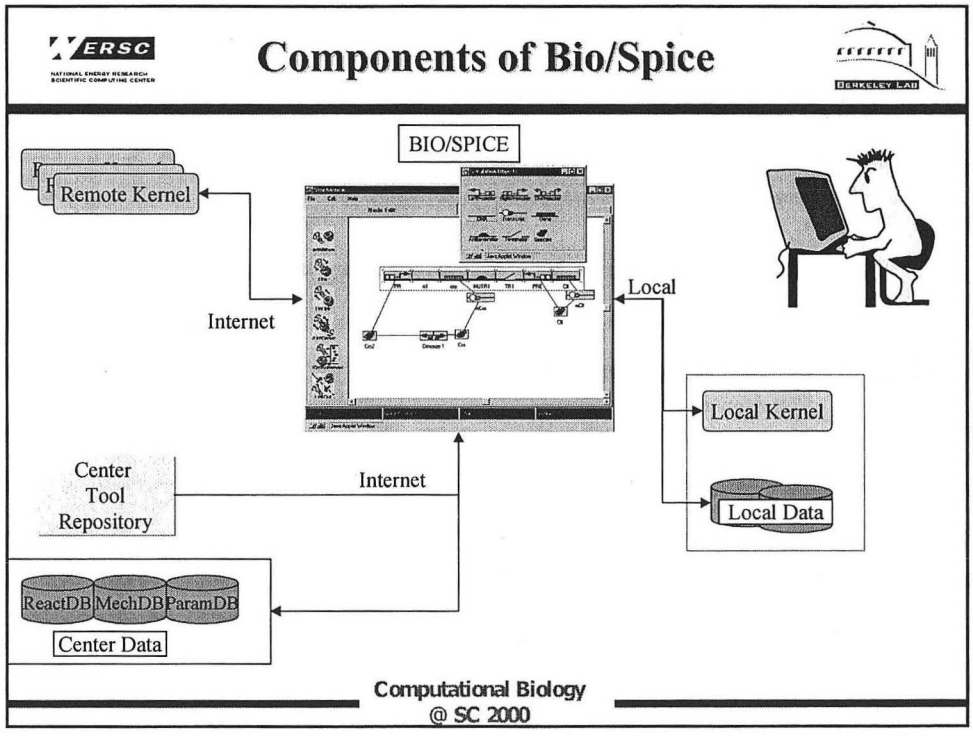
Bio/Spice: A Web-Servable, Biologist-Friendly, database, analysis and simulation interface was developed into a true beta product.

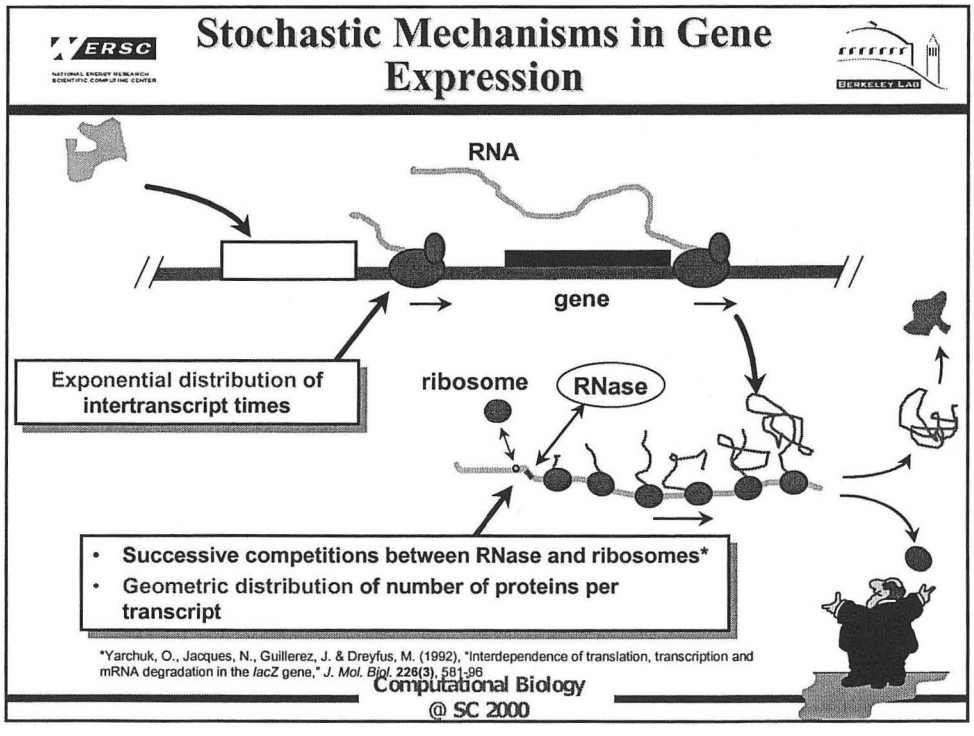
Interfaces to ReactDB, MechDB, and ParamDB.

With Kernel, performs basic: flux-balance analysis, stochastic and deterministic kinetics, Scientific Visualization of results.

Notebook/Kernel design optimized for distributed computing.




Components of Bio/Spice
BIO/SPICE
Remote Kernel
Internet
Local
Local Kernel
Local Data
Center Tool Repository
Internet
ReactDB
MechDB
ParamDB
Center Data
Computational Biology @ SC 2000


Stochastic Mechanisms in Gene Expression
RNA
gene
Exponential distribution of intertranscript times
ribosome
RNase
• Successive competitions between RNase and ribosomes*
• Geometric distribution of number of proteins per transcript
*Yarchuk, O., Jacques, N., Guillerez, J. & Dreyfus, M. (1992), "Interdependence of translation, transcription and mRNA degradation in the lacZ gene," J. Mol. Biol. 226(3), 581-96
Computational Biology @ SC 2000

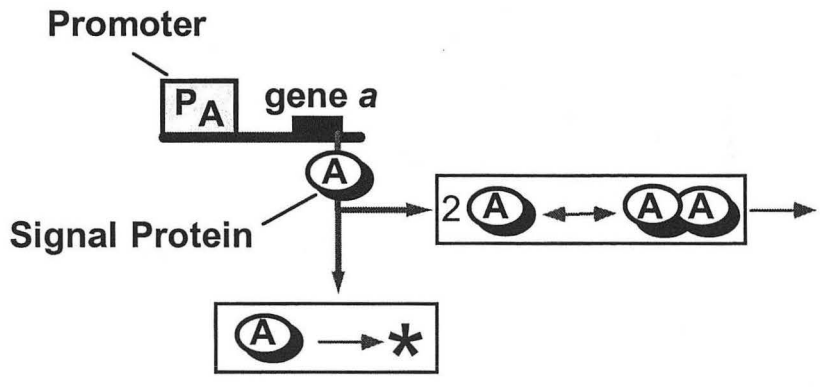
Some Stochastic Cellular Phenomena

- Lineage commitment in human hemopoiesis
- Random, bimodal eukaryotic gene transcription in
  - Activated T cells
  - Steroid hormone activation of mouse mammary tumor virus
  - HIV-1 virus
- Clonal variation in:
  - Bacterial chemotactic responses
  - Cell cycle timing
- E. coli type-1 pili expression
  - Enhances virulence
- Changing cell surface protein expression
  - For immune response avoidance
- Bacteriophage l lysis/lysogeny decision

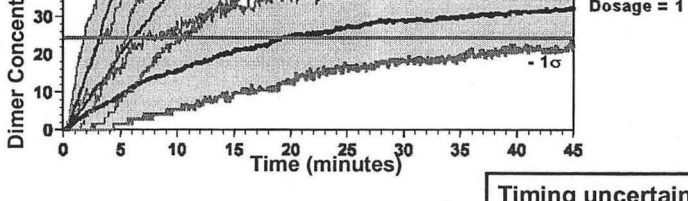
Where Noise Comes From

- Random environmental influences
- Mutations
- Asymmetric partitioning at cell division
- Stochastic mechanisms in gene expression
  - Stochastic timing of gene expression
  - Random variation in time for signal propagation
  - Random variation total protein production

A simple example

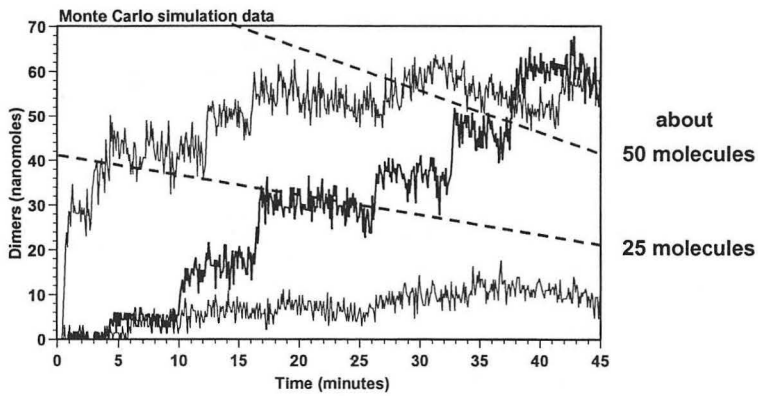


Time to Effectivity



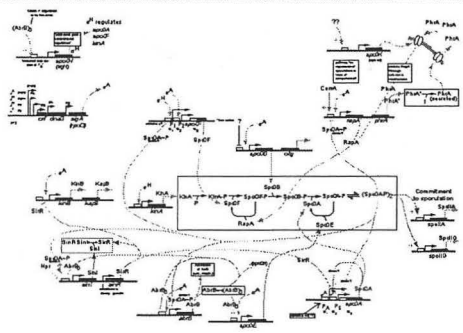
Timing uncertainty reduced by:
- Higher gene dosage
- Strong promoter
- Multiple promoters
- Lower effectivity threshold
- Slower cell growth

Signal Growth in Three Cells



- One gene
- Growing cell, 45 minutes division time
- Average ~60 seconds between transcripts
- Average 10 proteins/transcript:

Complexities of Cellular Function



This is approximately 1/3 of just the initiation of the sporulation program from *Bacillus subtilis*.

There are over 100 proteins, 40 genes, 300 reactions for which data is available.

The total data on just this process is a tens of Gb and it is incomplete.
Microarray and microscope data are added 100 Mb per week.
Model builders need to query this data and arrange it for simulation.
Simulations must be run under many different condition and hypotheses.

The Need for Advanced Computing

- **Data Handling:**

  The total data necessary for network analysis is huge. By nature it will be distributed and heterogeneous

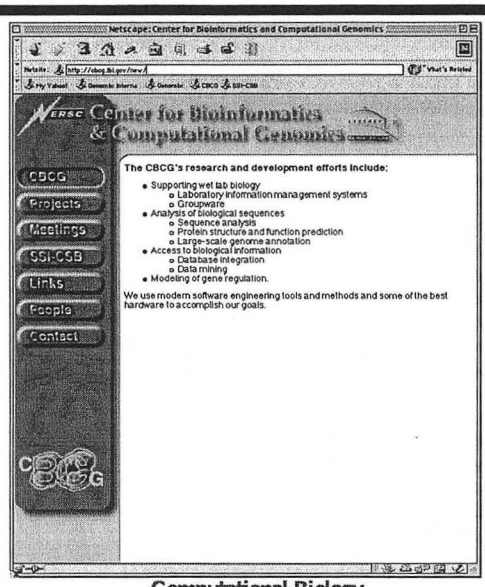
  We need:
  - **Database standard and new query types**
  - **Means of secure,fast transmission of information**
  - **Means of quality control on data input**

- **Tool integration:**
  - **Centralization of computational biology tools and standards**
  - **Ability to use tools together to generate good network hypotheses**
  - **Good quality ratings on Tool outputs**

- **Advanced Simulation Tools:**
  - **Fast, distributed algorithms for dynamical simulation**
  - **Mixed mode systems (differential, Markov, algebraic, logical)**
  - **Spatially distributed systems**

The End

http://cbcg.lbl.gov



Computational Biology
@ SC 2000

Ernest Orlando Lawrence Berkeley National Laboratory
One Cyclotron Road | Berkeley, California 94720

Prepared for the U.S. Department of Energy under Contract No. DE-AC03-76SF00098