

Computational Biology and High Performance Computing

Tutorial M4 p.m.

**November 6, 2000
SC'2000, Dallas, Texas**

- **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

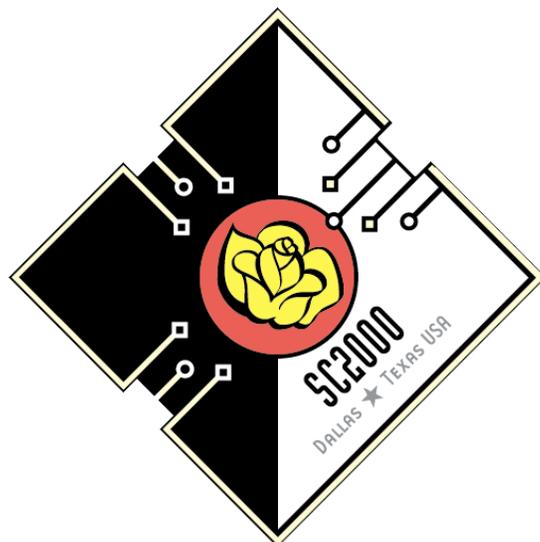


Tutorial Outline:

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**



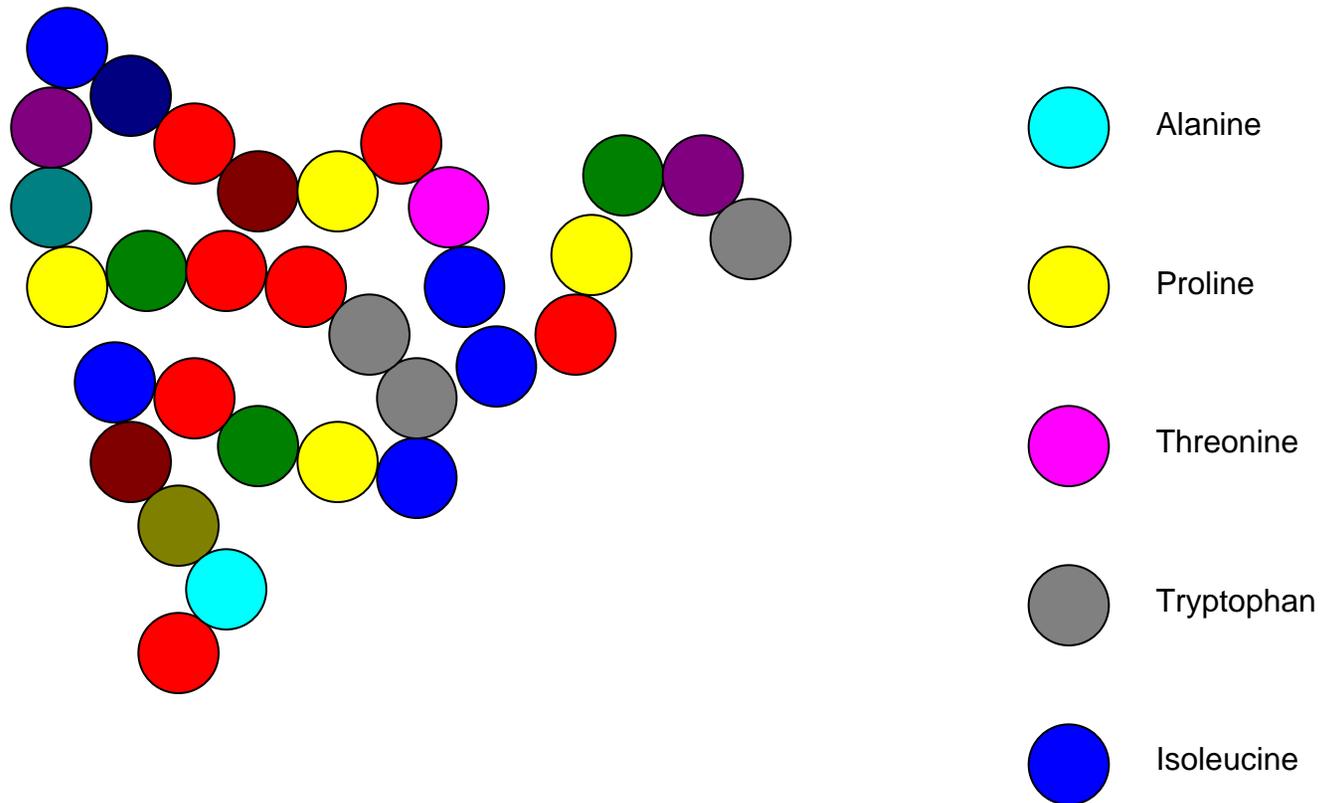
Proteins

Manfred Zorn
MDZorn@lbl.gov
NERSC

What is a protein?

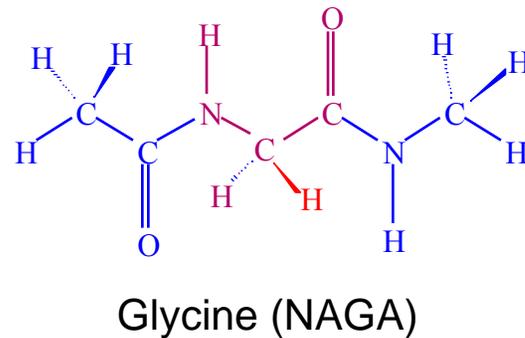
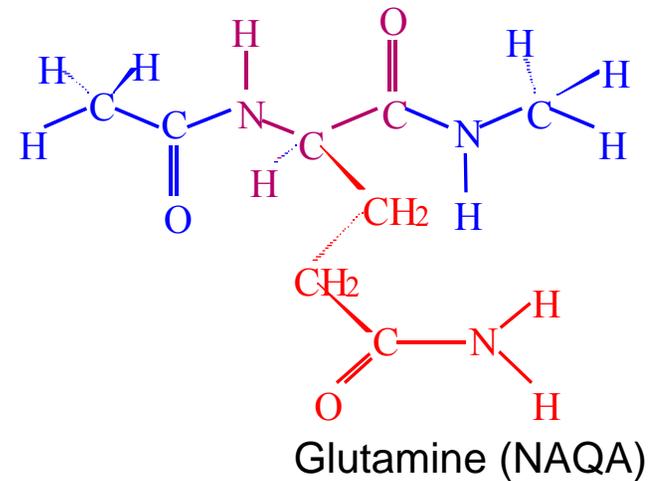
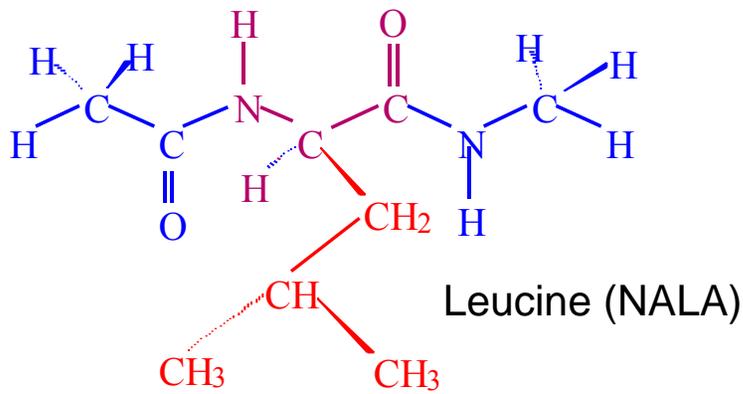
A biopolymer which is distinct from a heteropolymer in one very important way

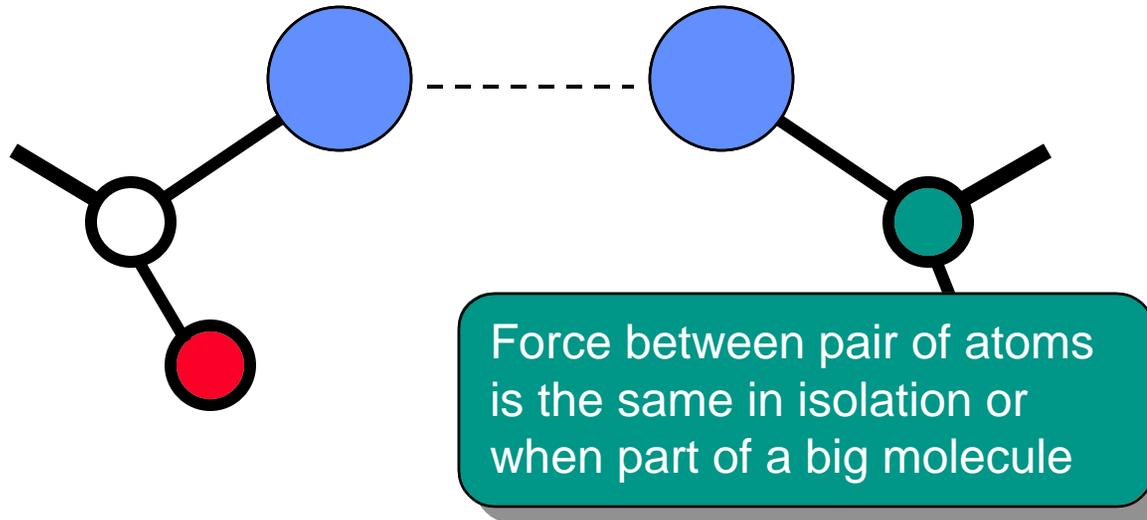
It's 3-D structure is uniquely tailored to perform a specific function



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

The "Beads" are Chemically Complex Structures

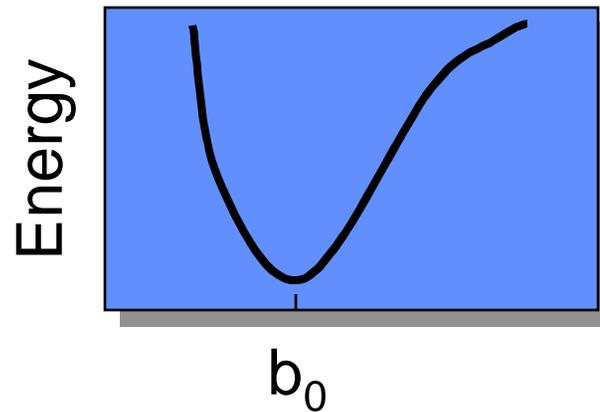
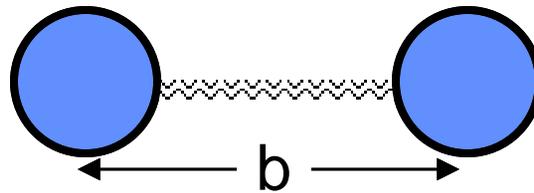




■ Basic assumptions:

- ✓ 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

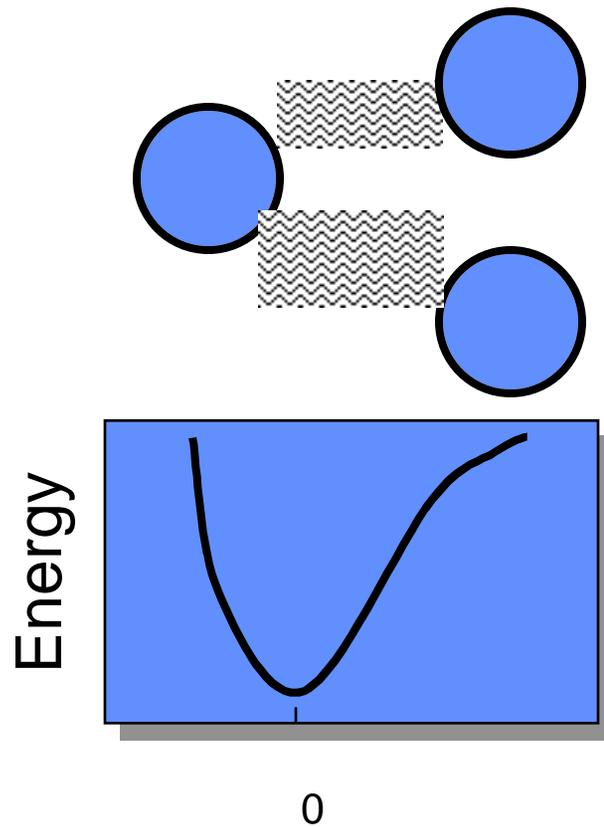


Equilibrium length $\sim 0.1\text{-}0.2\text{nm}$

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

K_b spring force constant $\sim 500\text{kcal/mole \AA}^2$

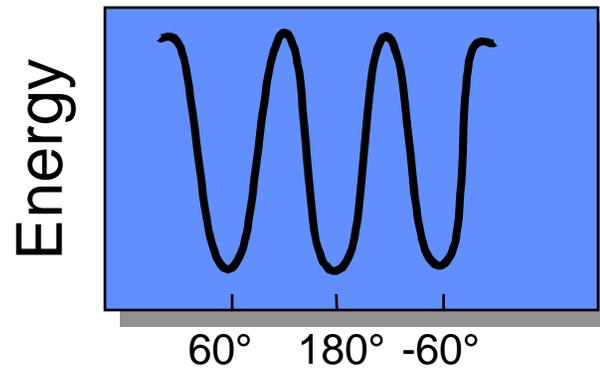
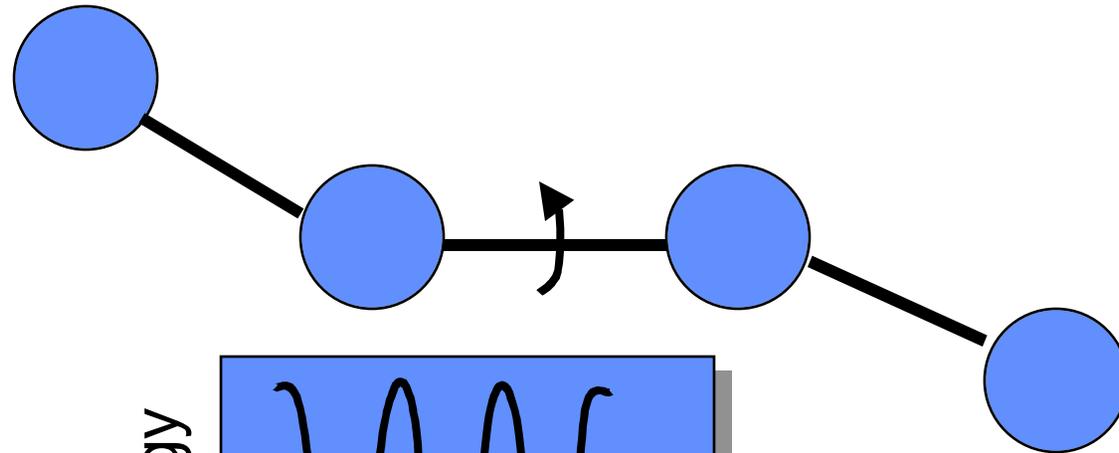
Bond Angle Forces



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

K spring force constant

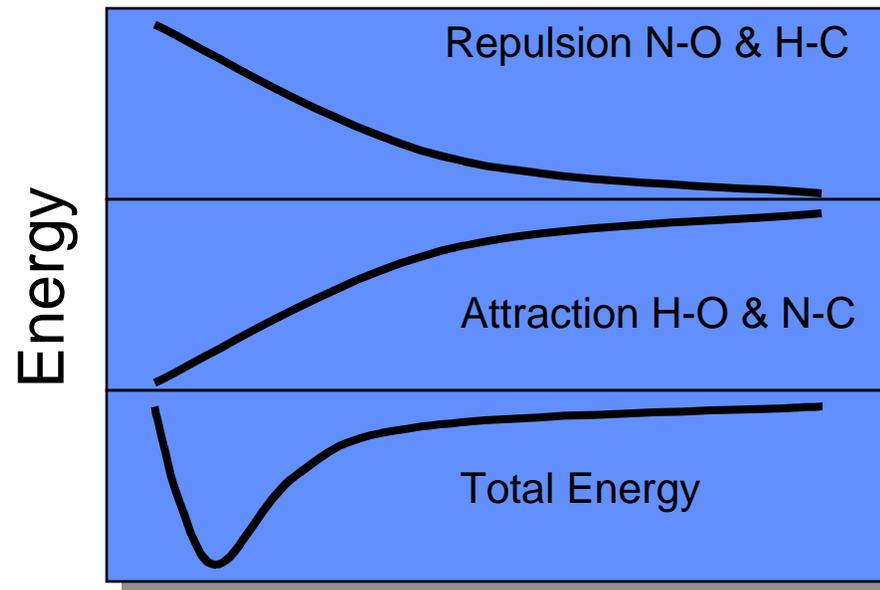
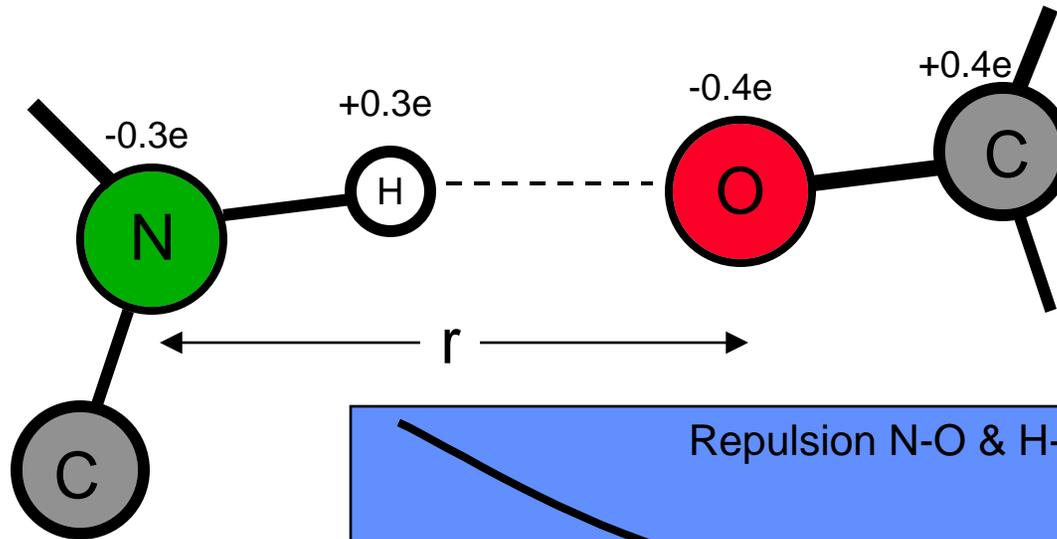
Bond Twisting Forces



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

$$U(\theta) = K \left[1 - \cos(n\theta + \delta) \right]$$

Hydrogen Bonds



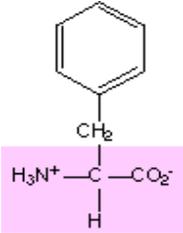
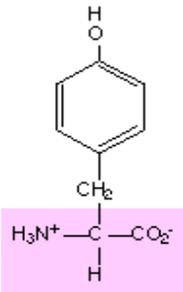
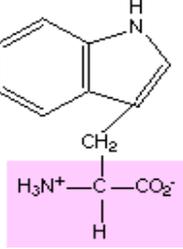
Optimum distance for N-O = 0.3nm
Net interaction ~ -5kcal/mole

N-O separation (r)

Scale of Interactions

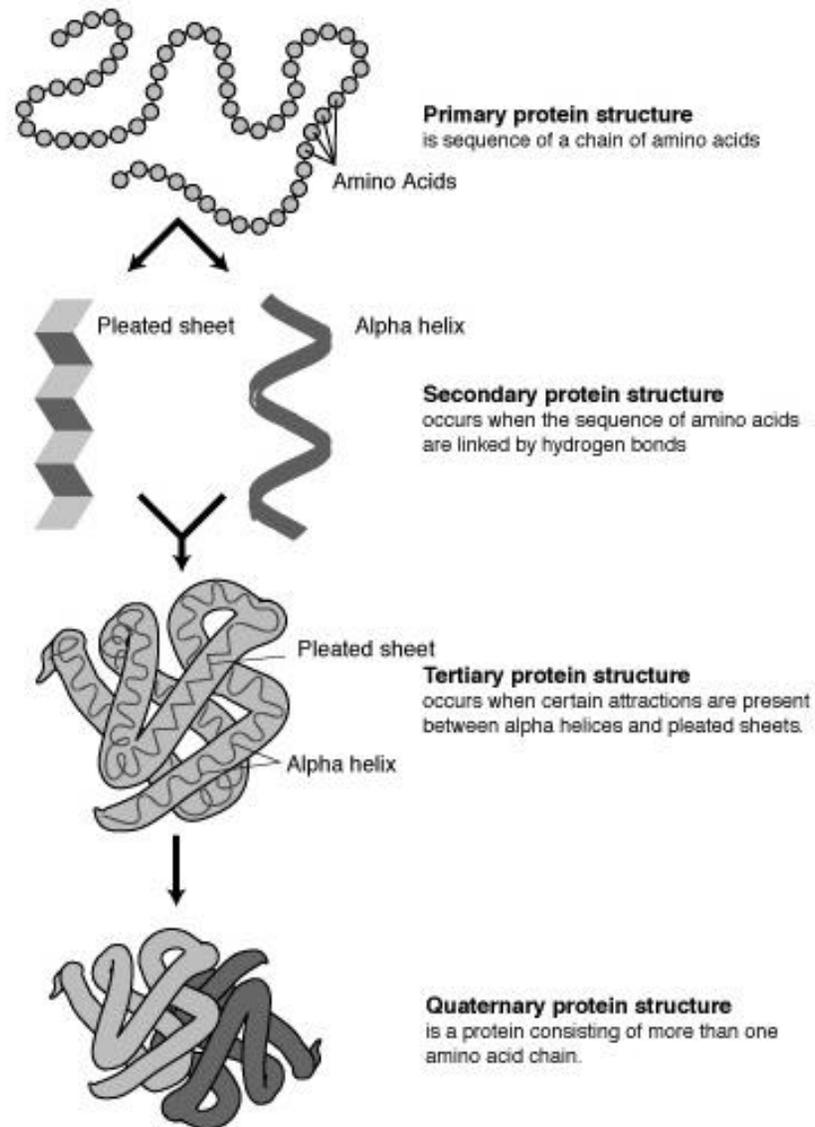
<i>Interaction</i>	<i>Energy</i> (kcal/mole)
Van der Waals (in water)	-0.1
Hydrogen bond (in water)	-1.0
Torsion barrier (single bond)	~+3.0
Torsion barrier (double bond)	+20.0
Bond breakage	+100.0
Change bond angle by 10°	+2.0
Stretch bond length by 10pm (0.1Å)	+2.5
Thermal energy 300K	0.6

Aromatic Amino Acids

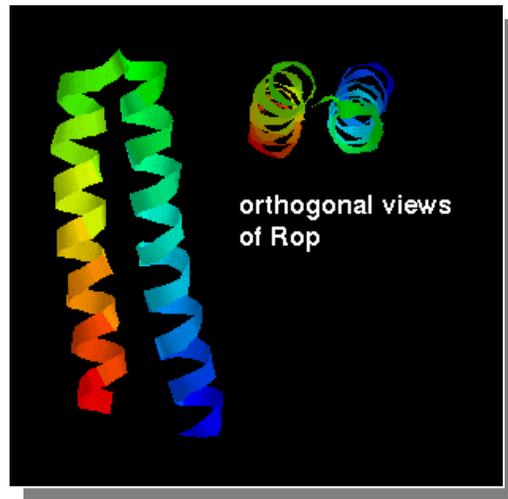
Amino Acid	pK _a 's ²	Pro Structure ³	Chemical Structure ⁴	3-D Structure ⁵
Phenylalanine, Phe, F No charge absorbs UV hydrophobic (2.5) Molec. Wt. = 147 Mole % = 3.5	N=9.13 C=1.83 pI=5.48	a =1.16 β =1.33 t =0.59		
Tyrosine, Tyr, Y weak charge absorbs UV hydrogen bonding not hydrophilic (0.08) Molec. Wt. = 163 Mole % = 3.5	N=9.11 C=2.20 R=10.07 pI=5.66	a =0.74 β =1.45 t =0.76		
Tryptophan, Trp, W largest amino acid rarest amino acid no charge absorbs UV hydrogen bonding hydrophobic (1.5) Molec. Wt. = 186 Mole % = 1.1	N=9.39 C=2.38 pI=5.89	a =1.02 β =1.35 t =0.65		

Copyright© [Charles S. Gasser](#) 1996

Protein Structure

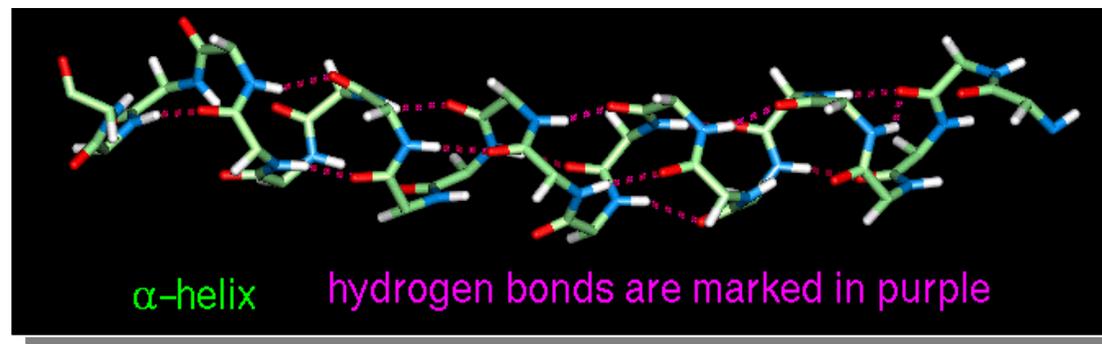


- **Alpha-helix**
- **Beta-sheet**
- **Coil**

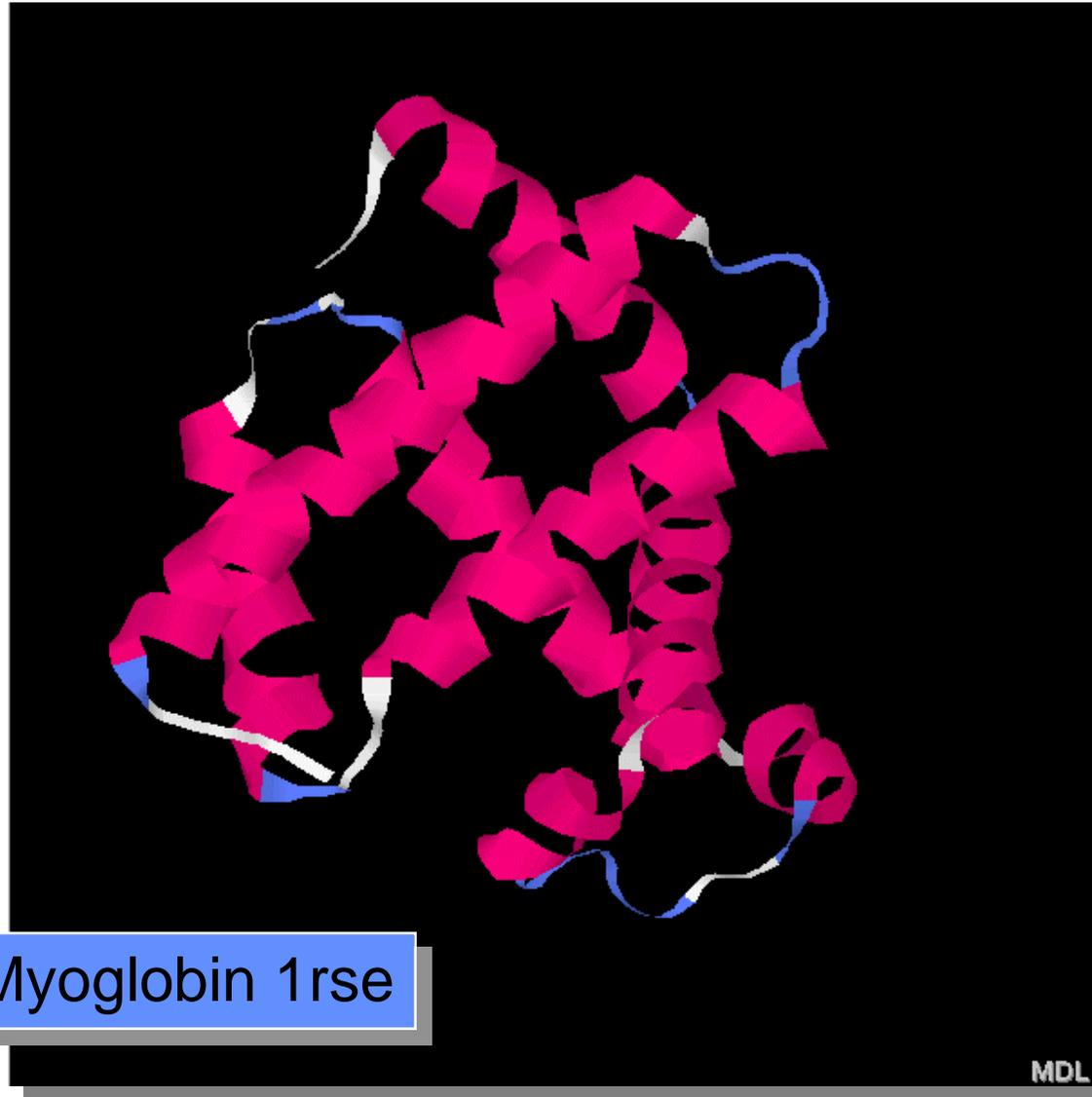


■ Alpha-helix

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



Alpha Helix

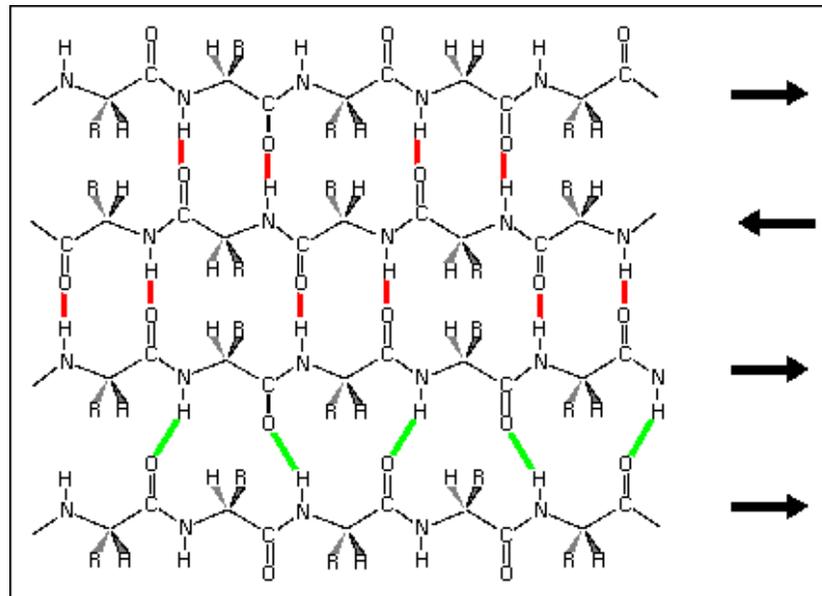
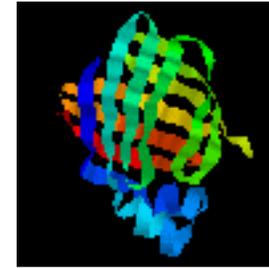


Human Myoglobin 1rse

MDL

■ Beta-sheet

- ✓ Parallel - antiparallel
- ✓ 25% of proteins



Beta sheets

Human Rhinovirus Protease 3C 1cqq



MDL

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

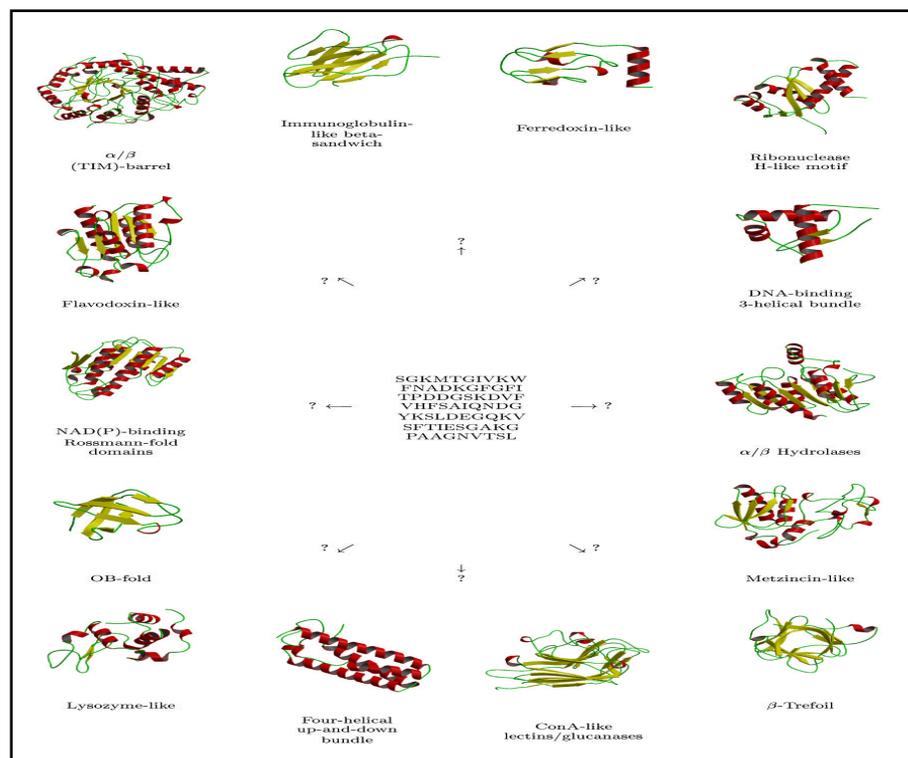
September 2000

- **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 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 - 10^3 amino 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 $E = V_i + V_{\text{gap}}$

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 \Rightarrow dynamic programming, L^2

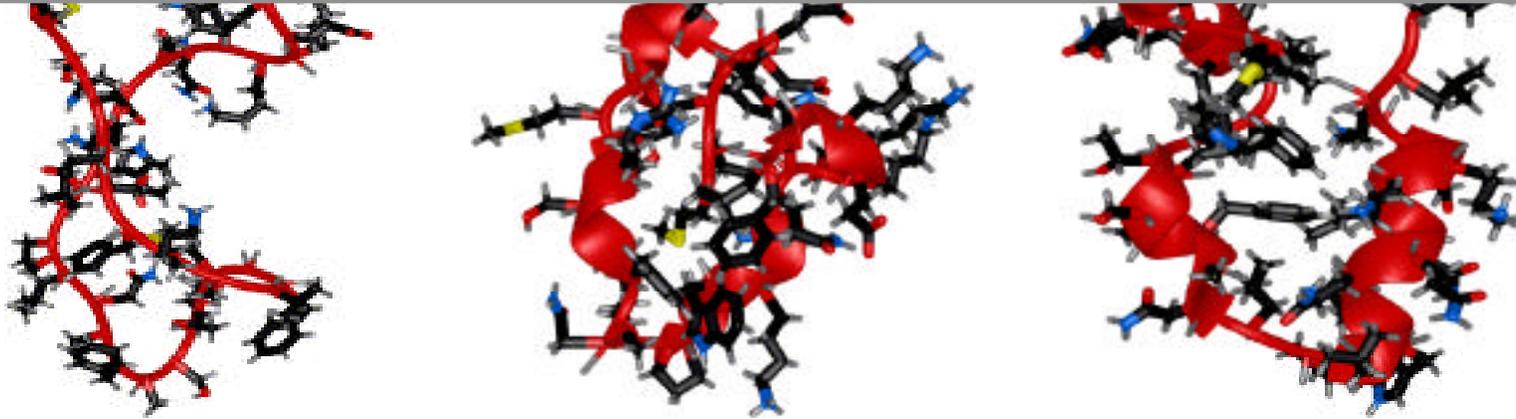
Build pairwise or multi-body objective function

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

Recursive dynamic programming, Hidden markov models, stochastic grammars

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

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 8\AA , poor by known simulation standards
- ✗ Statistics (1 trajectory is anecdotal)
Many trajectories required to characterize kinetics and thermodynamics

(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)

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

$$f_i = m_i a_i = - \frac{\partial V(r_1, r_2, \dots, r_N)}{\partial r_i}$$

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)

Primary Sequence and an Energy function → Tertiary structure

Empirical energy functions:

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

$$\begin{aligned}
 V_{MM} = & \sum_i^{\# \text{ Bonds}} k_b (b_i - b_o)^2 + \sum_i^{\# \text{ Angles}} k_\theta (\theta_i - \theta_o)^2 + \sum_i^{\# \text{ Impropers}} k_\tau (\tau_i - \tau_o)^2 + \\
 & \sum_i^{\# \text{ dihedrals}} k_\phi [1 + \cos(n\phi + \delta)] + \sum_{i < j}^{\# \text{ atoms} \# \text{ atoms}} \frac{q_i q_j}{r_{ij}} + \epsilon_{ij} \frac{\sigma_{ij}^{12}}{r_{ij}} - \frac{\sigma_{ij}^6}{r_{ij}} + \sum_i^{\# \text{ atoms}} \sigma A
 \end{aligned}$$

(1) Multiple minima problem is fierce

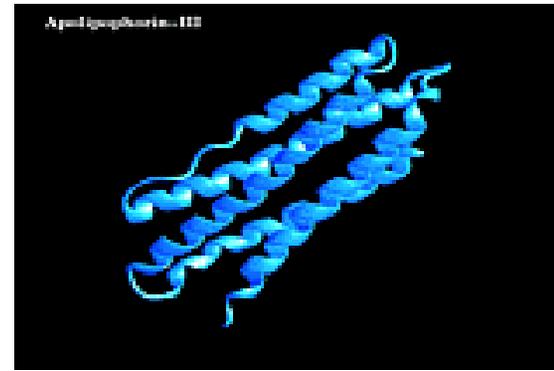
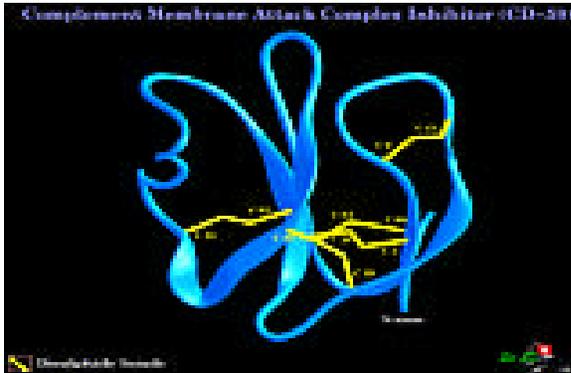
Find a way to effectively overcome the multiple minima problem

(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/ α -helical structure

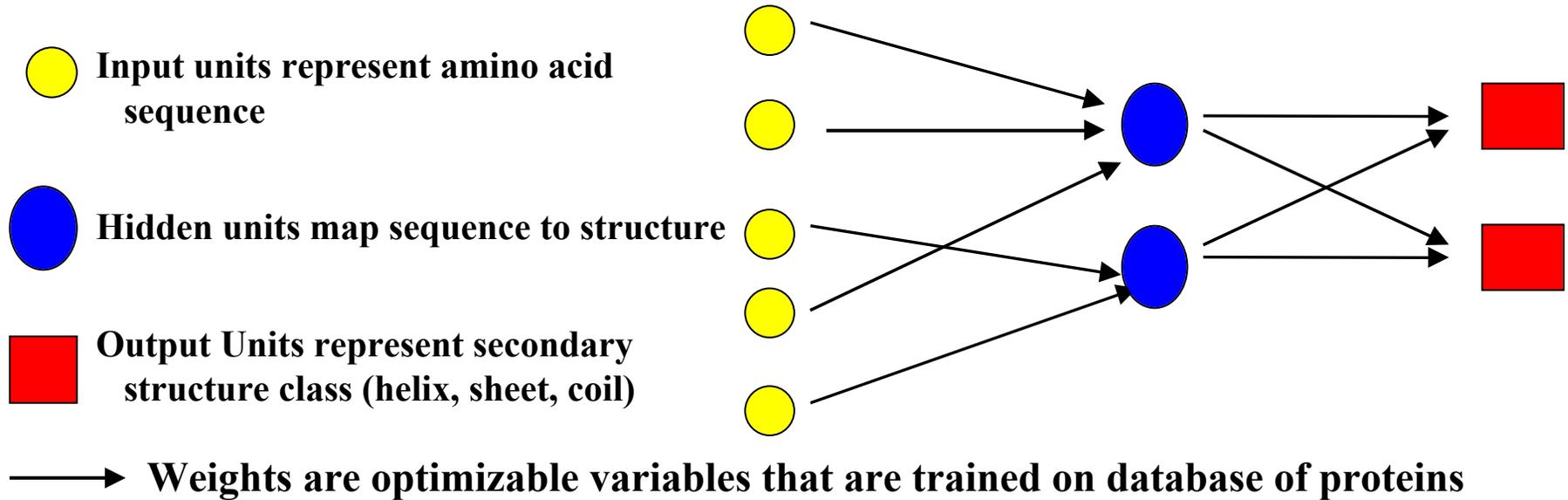
Energies the same! Makes energy minimization difficult!

Add penalty for exposing hydrophobic surface: favors more compact structures

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

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

Neural Networks for 2^o Structure Prediction



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

No sequence homology through multiple alignments

Train

Total predicted correctly = 66%

Helix: 51% $C_a=0.42$

Sheet: 38% $C_b=0.39$

Coil: 82% $C_c=0.36$

Test

Total predicted correctly = 62.5%

Helix: 48% $C_a=0.38$

Sheet: 28% $C_b=0.31$

Coil: 84% $C_c=0.35$

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

Train

Total predicted correctly = 67%

Helix: 66% $C_a=0.52$

Sheet: 63% $C_b=0.46$

Coil: 69% $C_c=0.43$

Test

Total predicted correctly = 66.5%

Helix: 64% $C_a=0.48$

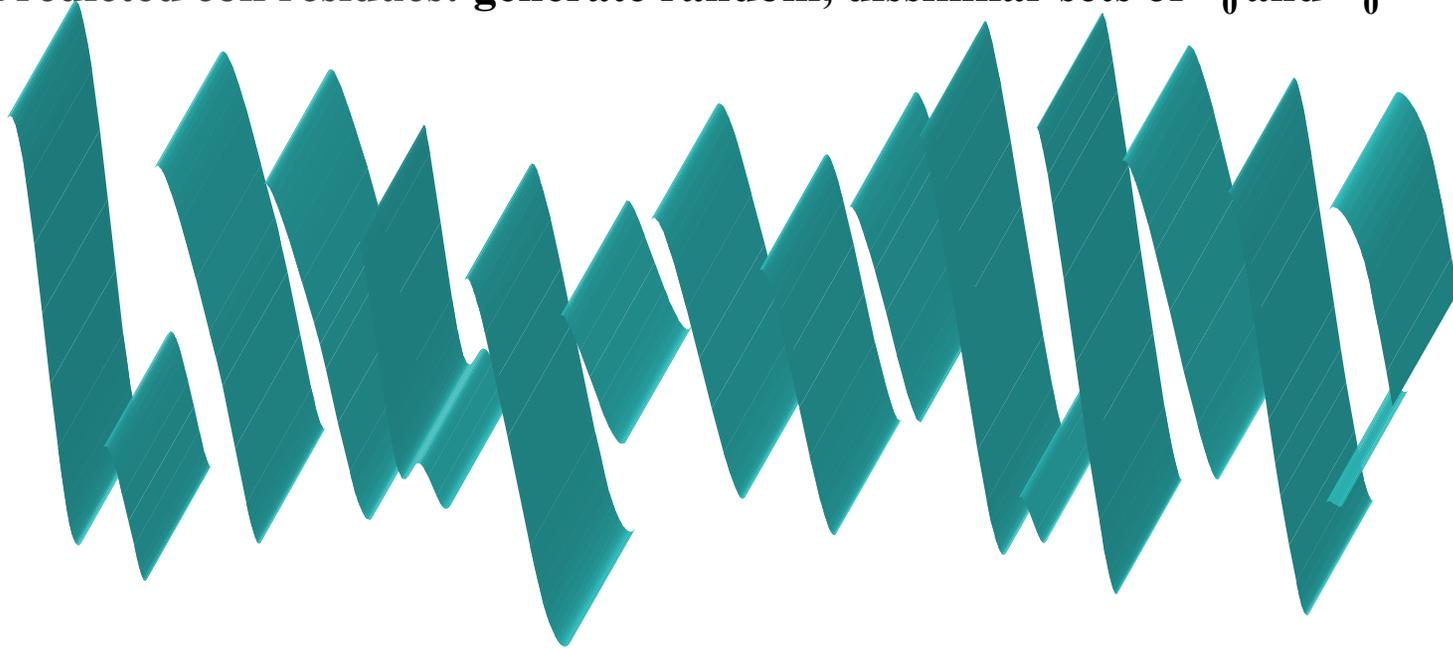
Sheet: 53% $C_b=0.43$

Coil: 73% $C_c=0.44$

Combine networks of Yu and Head-Gordon with multiple alignments

Generate expanded tree of configurations

Predicted coil residues: generate random, dissimilar sets of θ_0 and ϕ_0

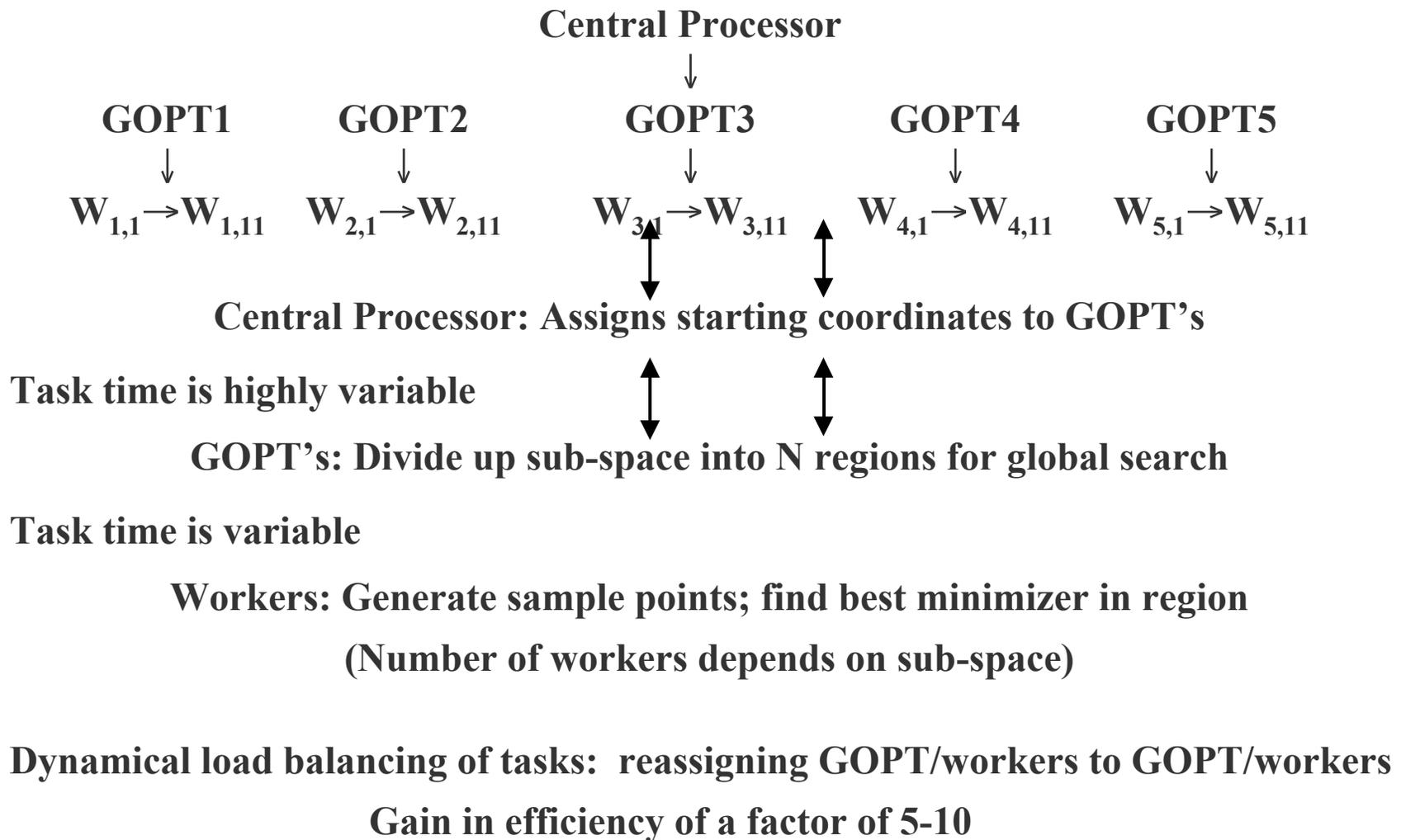


Explore tree configuration in depth:

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

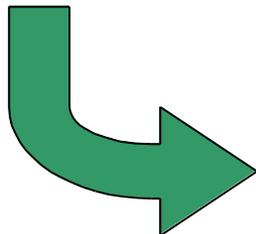
Hierarchical Parallel Implementation of Global Optimization Algorithm

Static vs. Dynamic Load Balancing of Tasks

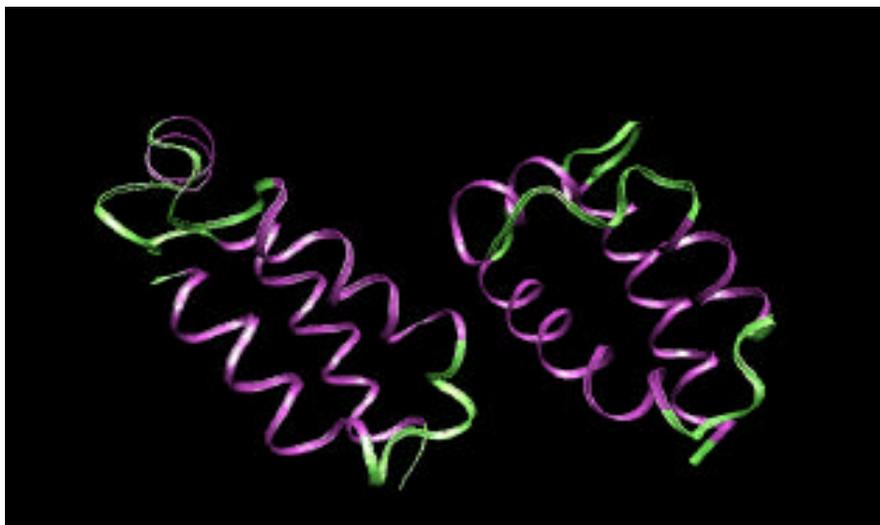


Crystal (left), Prediction (right)

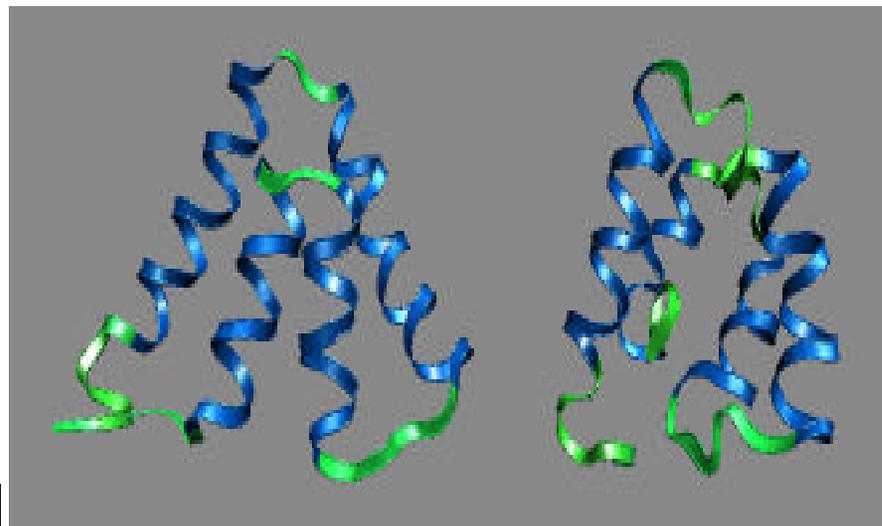
R.M.S. 7.0Å



1pou: 72 aa DNA binding protein

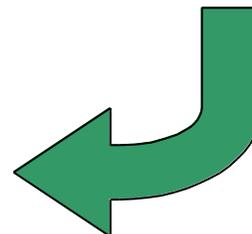


2utg_A: 70aa α -chain of uteroglobin:



Prediction (left) and crystal (right)

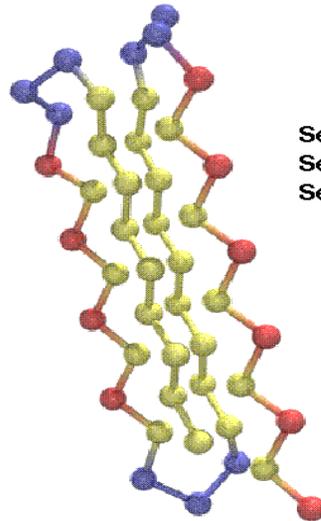
R.M.S. 6.3Å



Still have not reached crystal energy yet!

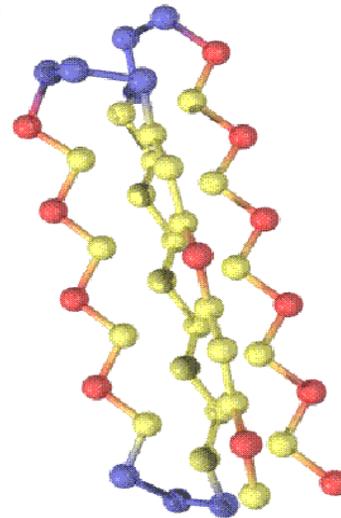
Simplified Models for Simulating Protein Folding

a)



Seq. 1 B2L
Seq. 2 B2L, B6L
Seq. 3 B2L, B4L, B6L

b)



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**

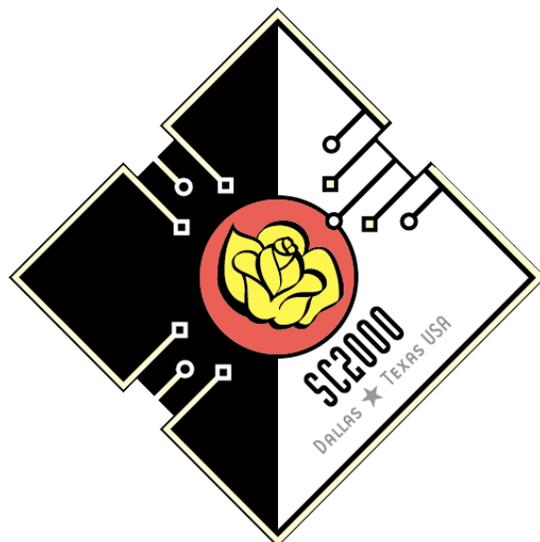
Jon M. Sorenson, NSF Graduate Fellow, Dept. Chemistry UCB

Greg Hura, Graduate Group in Biophysics, UCB

Alan K. Soper, Rutherford Appleton Laboratory, UK

**Alexander Pertsemididis, Dept. of Biochemistry, U. Texas Southwestern Medical
Center**

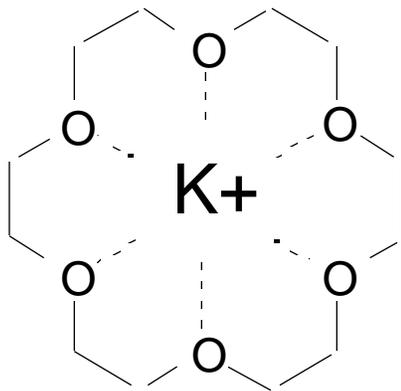
Robert M. Glaeser, Mol. & Cell Biology, UCB and Life Sciences Division, LBNL



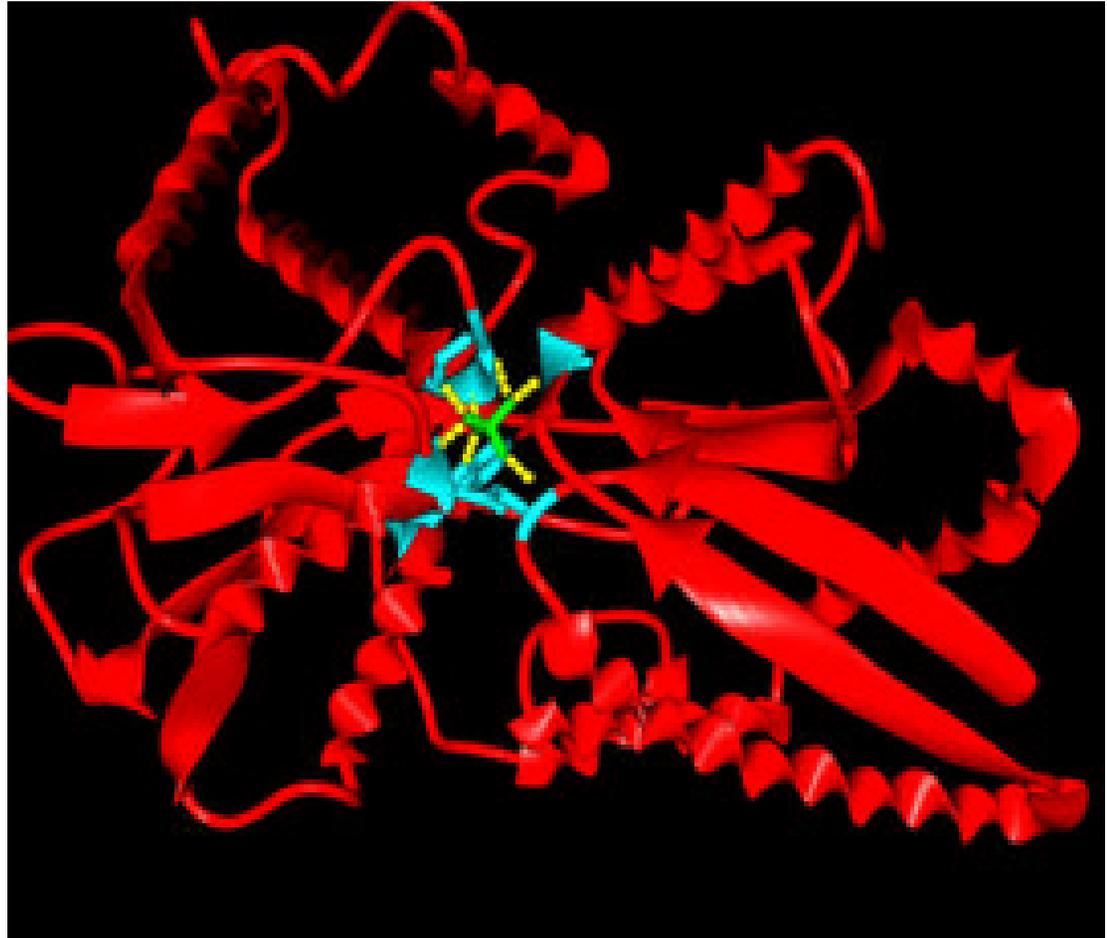
Structure-Based Drug Discovery

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

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

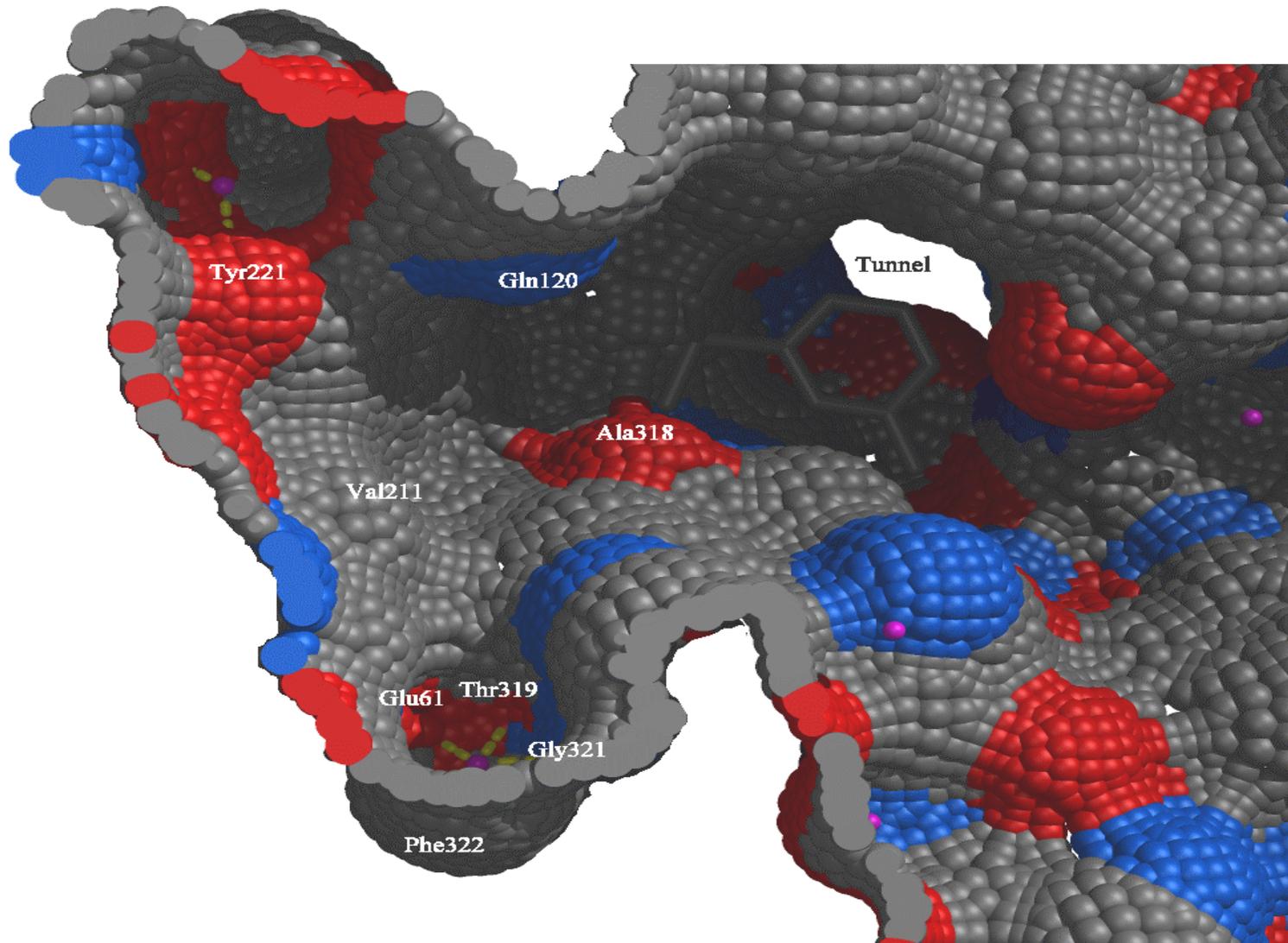


18 - Crown-6



sulfate binding protein

Conserved Residues, Ordered Structure, Function Unknown



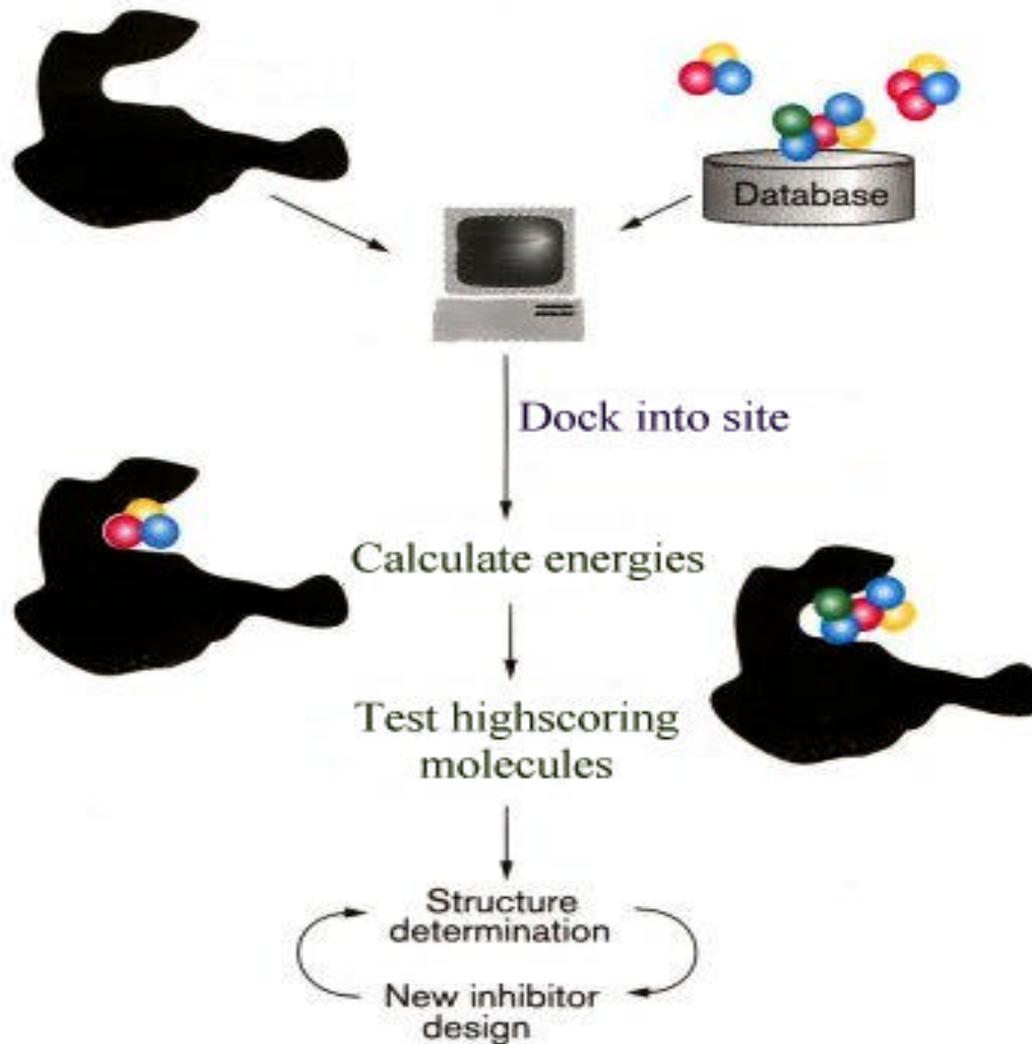
■ Design ligands

- Ludi (Bohm)
- Grow (Moon & Howe)
- Builder (Roe & Kuntz)
- MCSS-Hook (Miranker & Karplus)
- SMOG (DeWitte & Shakhovitch)
- Others...

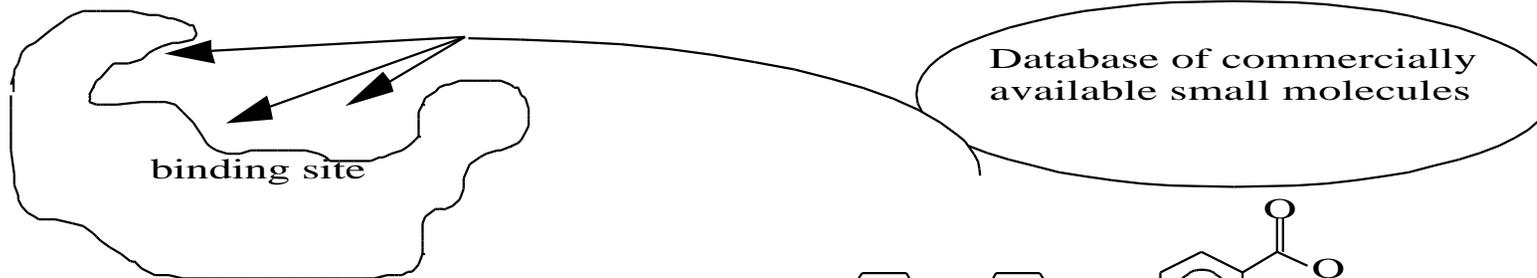
■ Discover Ligands

- DOCK (Kuntz, et al., Shoichet)
- CAVEAT (Bartlett)
- Monte Carlo (Hart & Read)
- AutoDock (Goodsell & Olson)
- SPECITOPE (Kuhn et al)
- Others...

Screening Databases by Molecular Docking



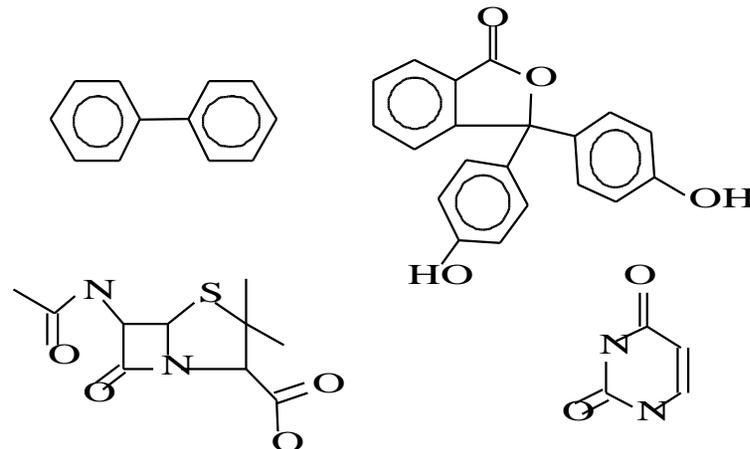
© Chemistry & Biology, 1996



Each molecule is fit into the binding site in multiple orientations. Multiple conformations of each ligand are considered.

Each orientation is evaluated for complementarity, using vander Waals and electrostatic interaction energies

Solvation energies are subtracted.

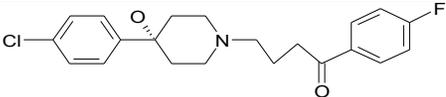
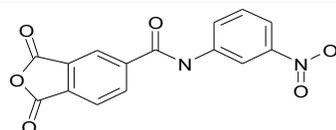
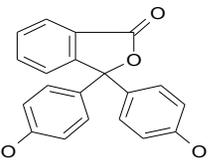
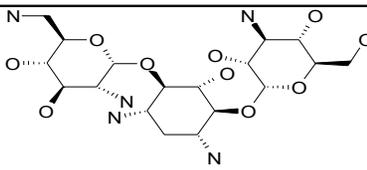
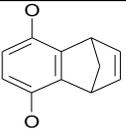
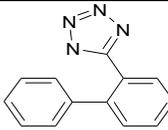
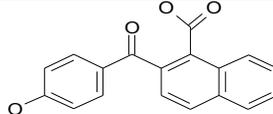
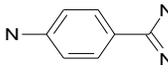
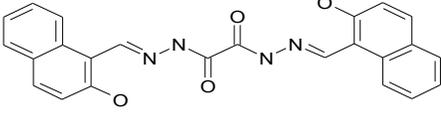
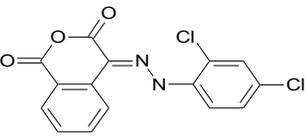
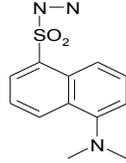


...~200,000 compounds

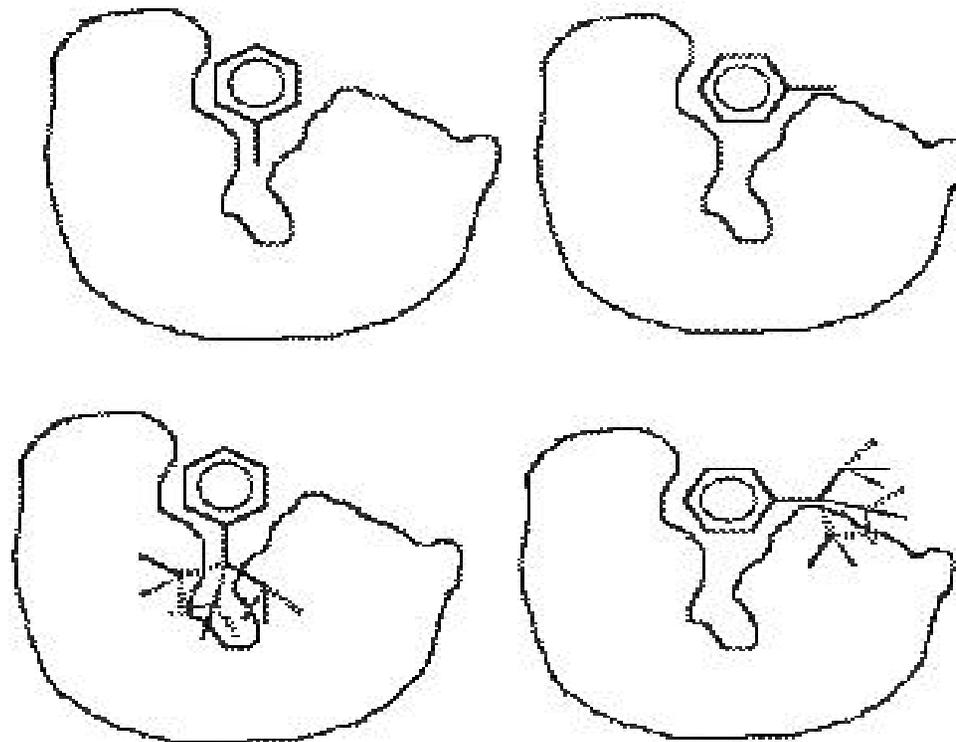
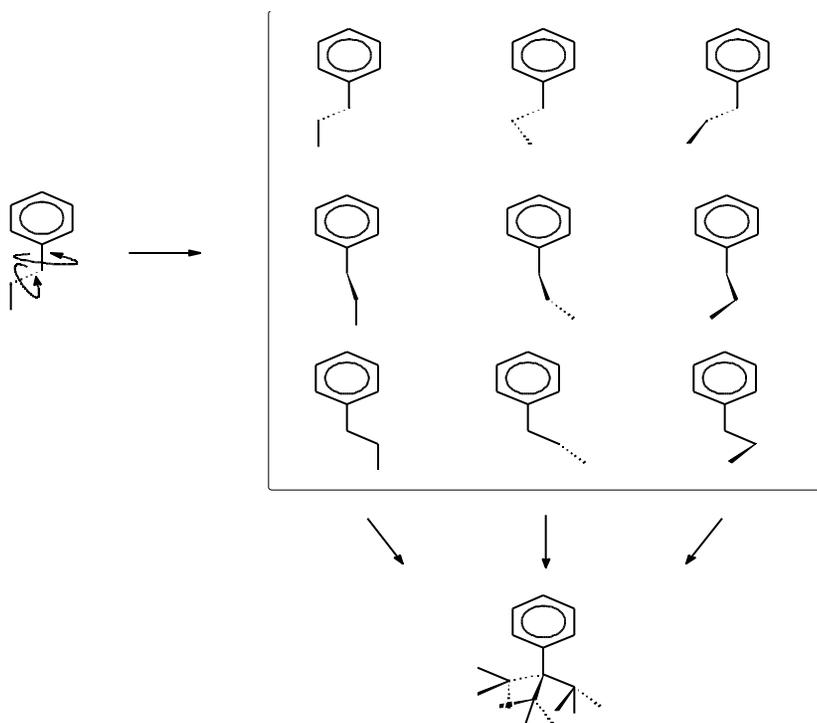
The inhibition constants of the best fitting molecules are established in an enzyme assay

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 -lactamase	
cercarial elastase		Thrombin	
malarial protease		AmpC -lactamase	
CD4-gp120	unpublished	thymidylate synthase	
		HGXPRtase	unpublished

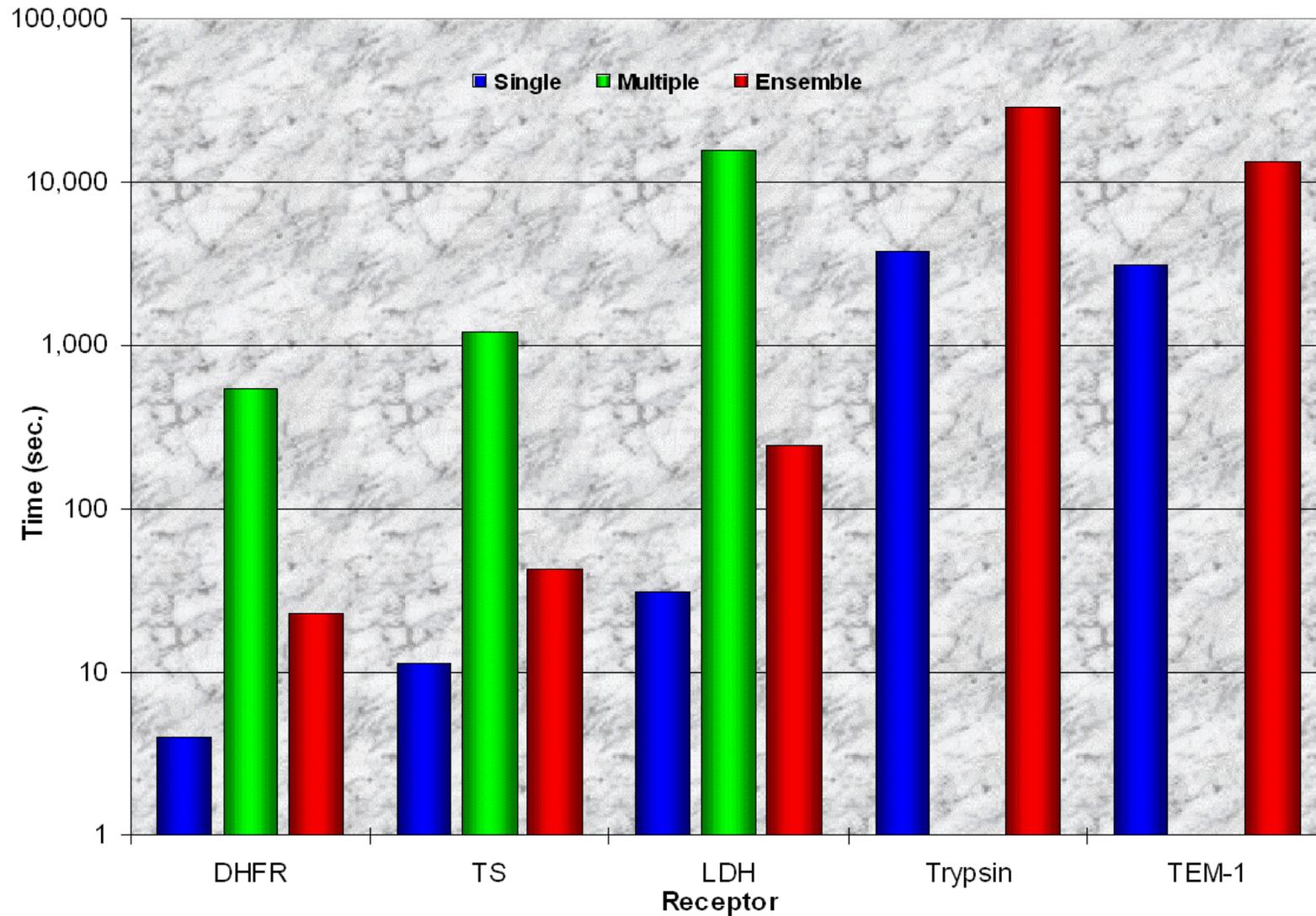
Ligand Flexibility: Conformational Ensembles



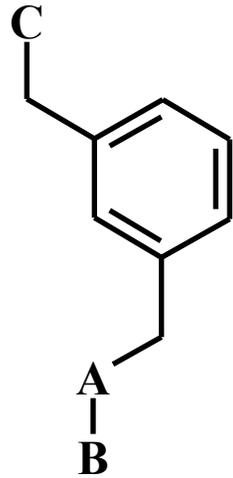
Generate an ensemble

dock it into the site

Conformational Ensembles vs. Brute Force

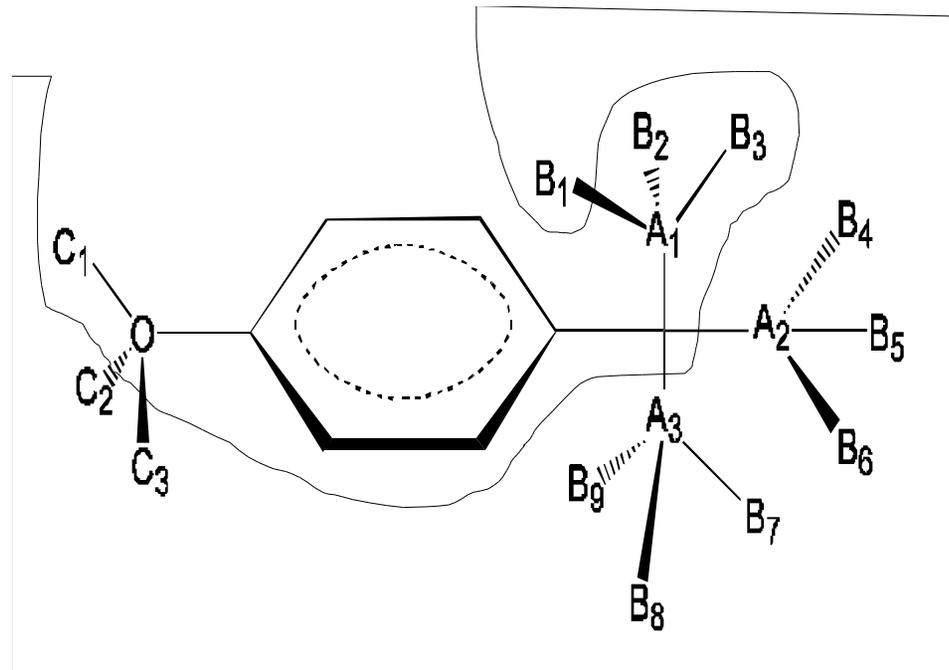


Hierarchical Docking



Flexible docking:
27 confs
x3 atoms
81 atom positions

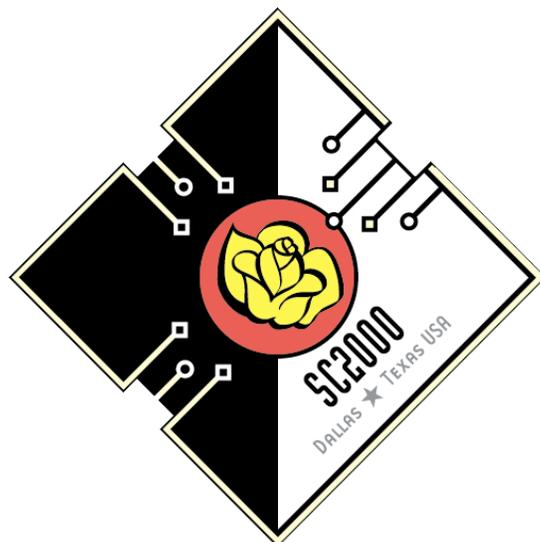
Hierarchical docking:
27 confs
3C + 3A + 9B
15 atom positions



- **Better Scoring**
 - **context dependent desolvation**
 - **receptor desolvation**
 - **better force-fields**

- **Receptor Flexibility**

- **Cominatorial Chemistry**



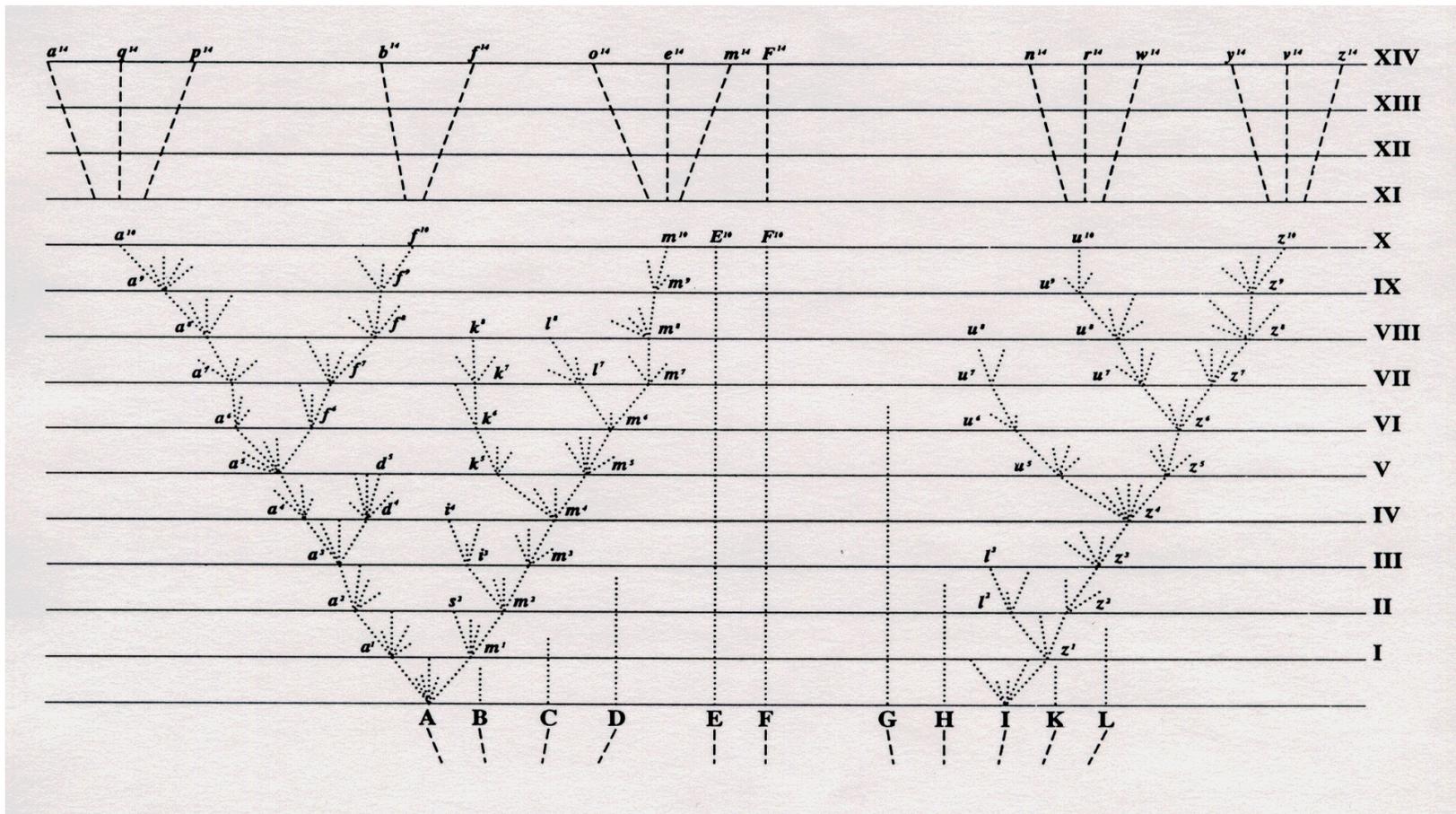
Computational Phylogenetics

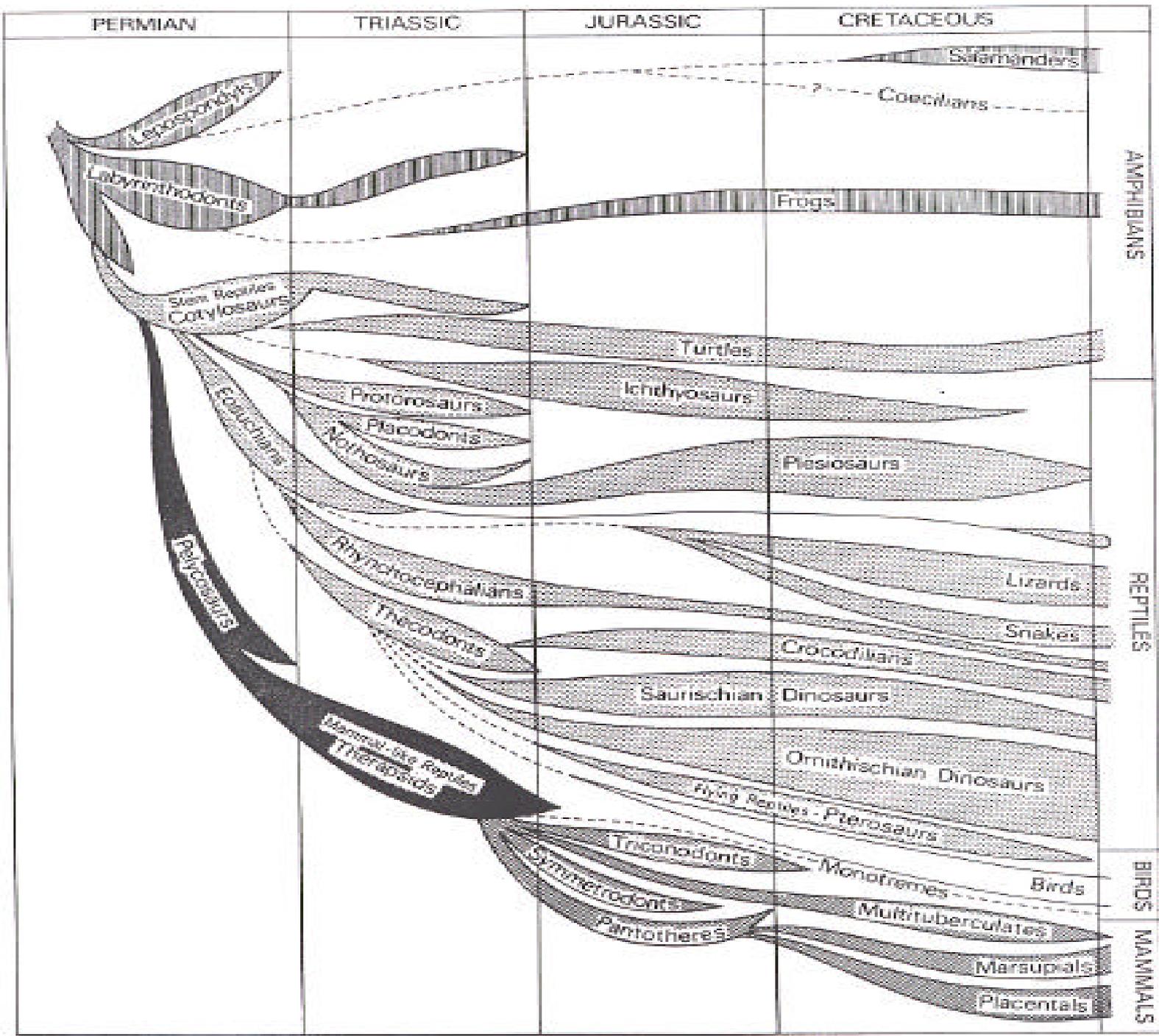
Craig Stewart
stewart@iu.edu
Indiana University

- **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 fastDNAmI**
- **Some interesting results and challenges for the future**
- **Caveat: this is an introduction, not an exhaustive review.**

- **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**

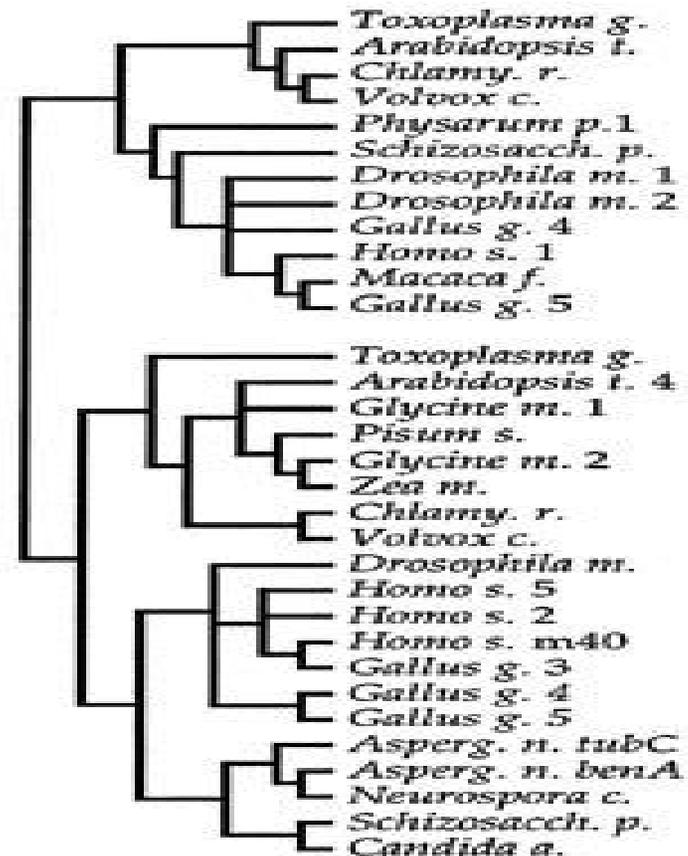
Origin of Species, Figure 1



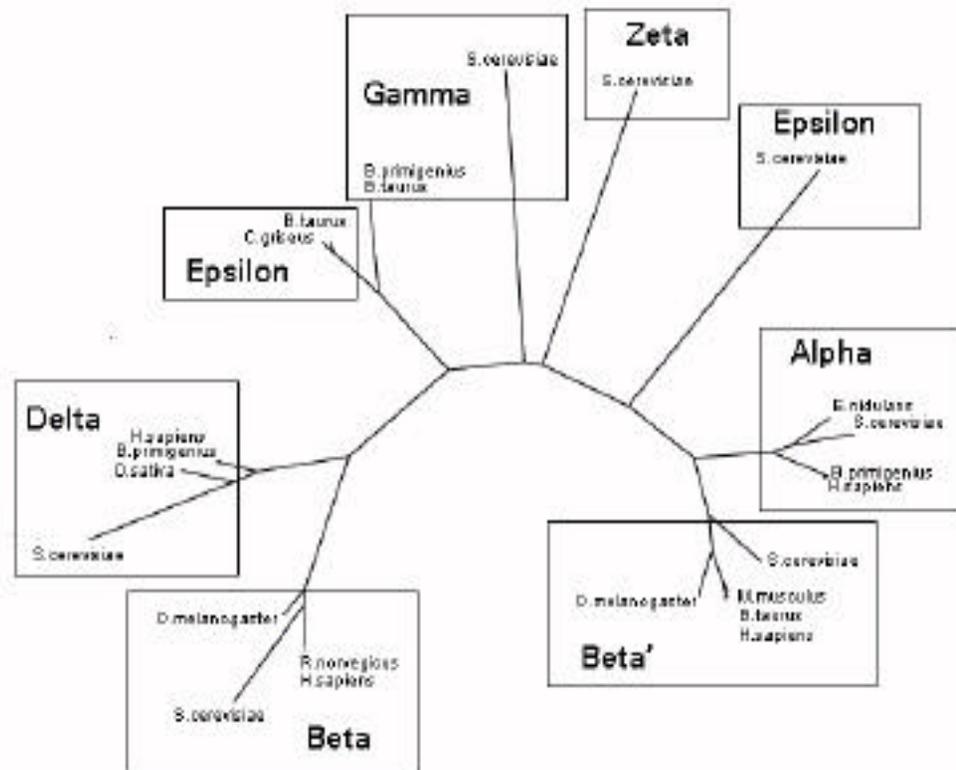


Building Phylogenetic Trees

- **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**



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

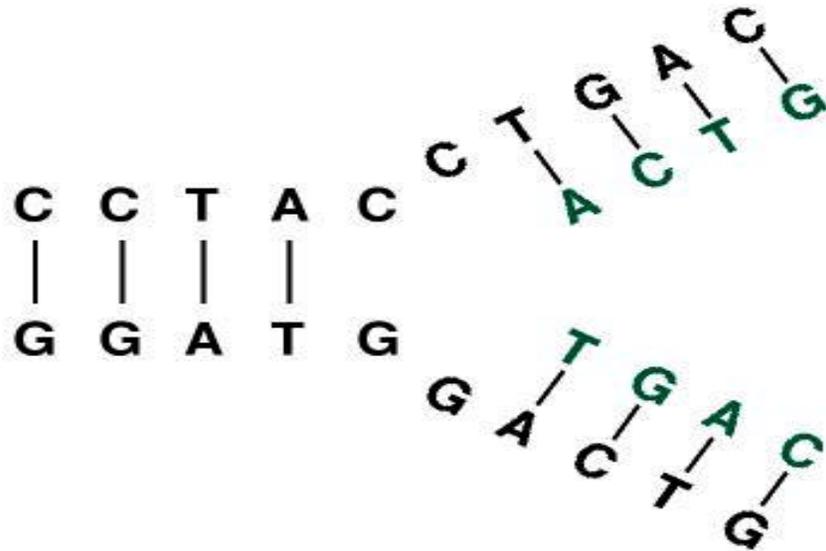


- **Curiosity:** Anyone who as a child wandered through the dinosaur section of a natural history museum understands the inherent intellectual attraction of evolutionary biology
- **Theoretical uses:** testing hypotheses in evolutionary biology
- **Practical uses:**
 - **Medicine**
 - **Environmental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

Pyrimidines:

Adenine & Guanine

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

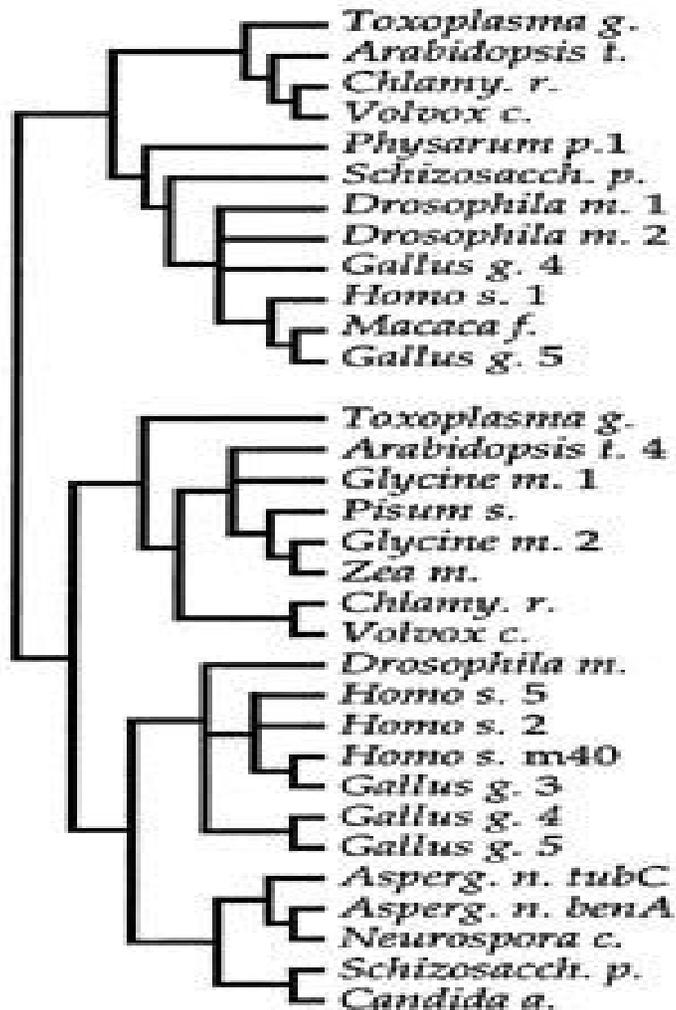
DNA CCTCT+GAC

VS

CCTCTTGAC

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

- **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>

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	CAAAAAAAG	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**

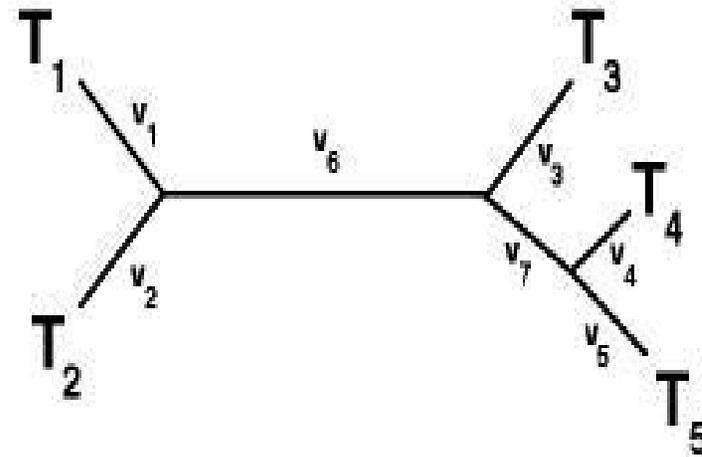
Evaluating trees

- 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

- **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.
- **fastDNAmI.** (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	$p(A \rightarrow A)$	$p(A \rightarrow C)$	$p(A \rightarrow G) \dots$	
C	$p(C \rightarrow A)$	$p(C \rightarrow C)$	$p(C \rightarrow G) \dots$	
G	.			
T	.			

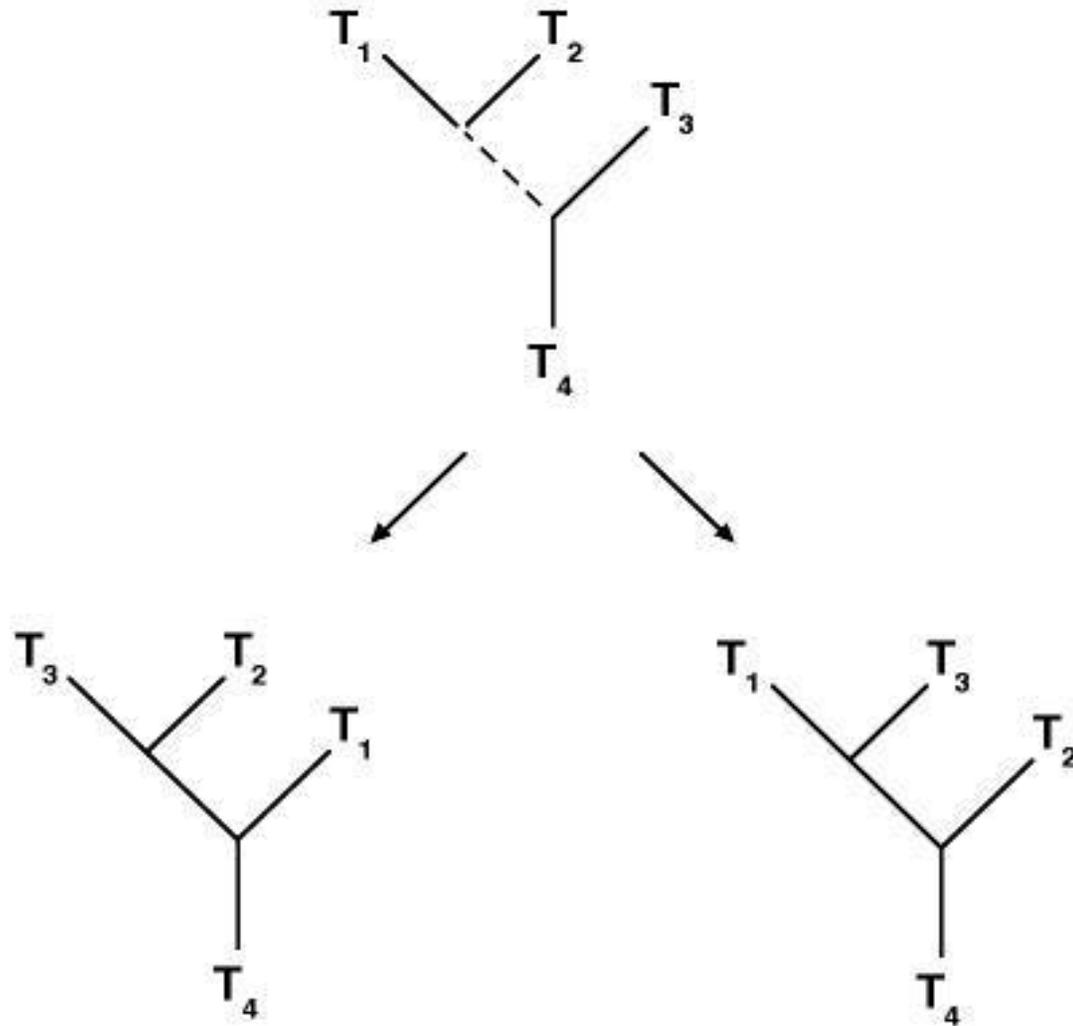
- **Transitions more probable than transversions.**
- **Must account for heterogeneity in substitution rates among sites (DNARates – Olsen)**

- **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. fastDNAmI: 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**

fastDNAmI 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-taxon 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



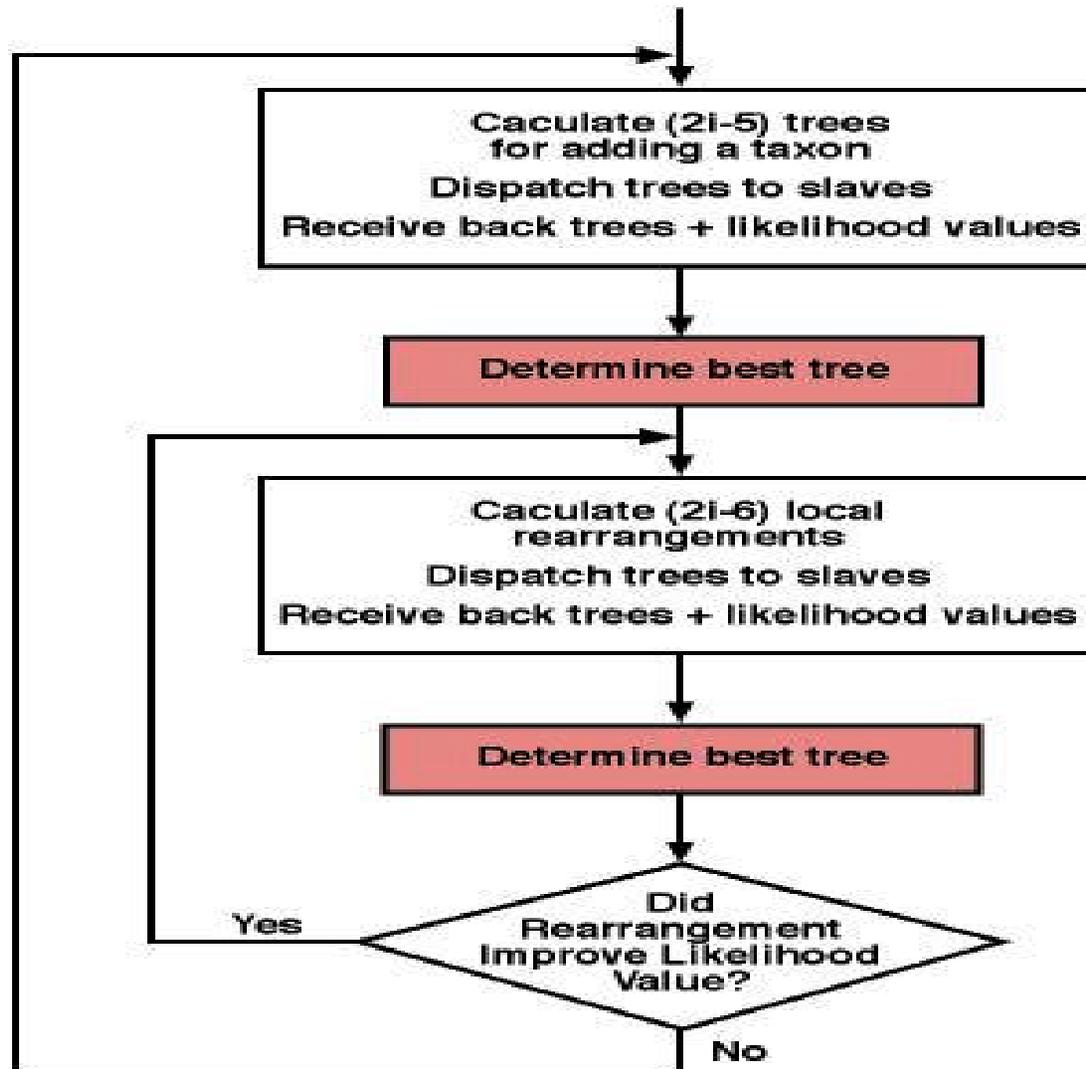


fastDNAmI algorithm con't: Iterate



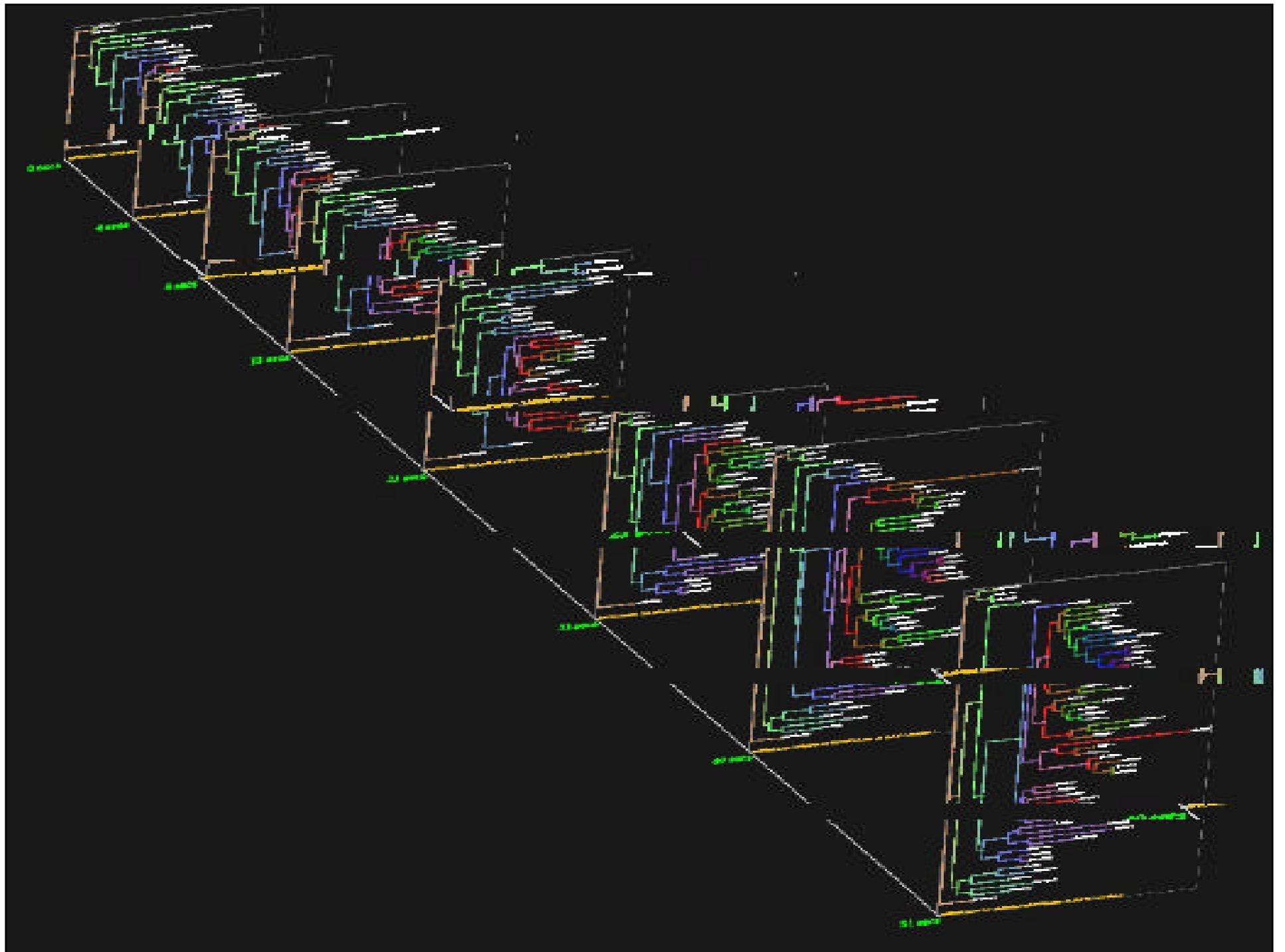
- **Get sequence data for next taxon**
- **Add new taxa ($2i-5$)**
- **Keep best**
- **Local rearrangements ($2i-6$)**
- **Keep best**
- **Keep going....**
- **When all taxa have been added, perform a full tree check**

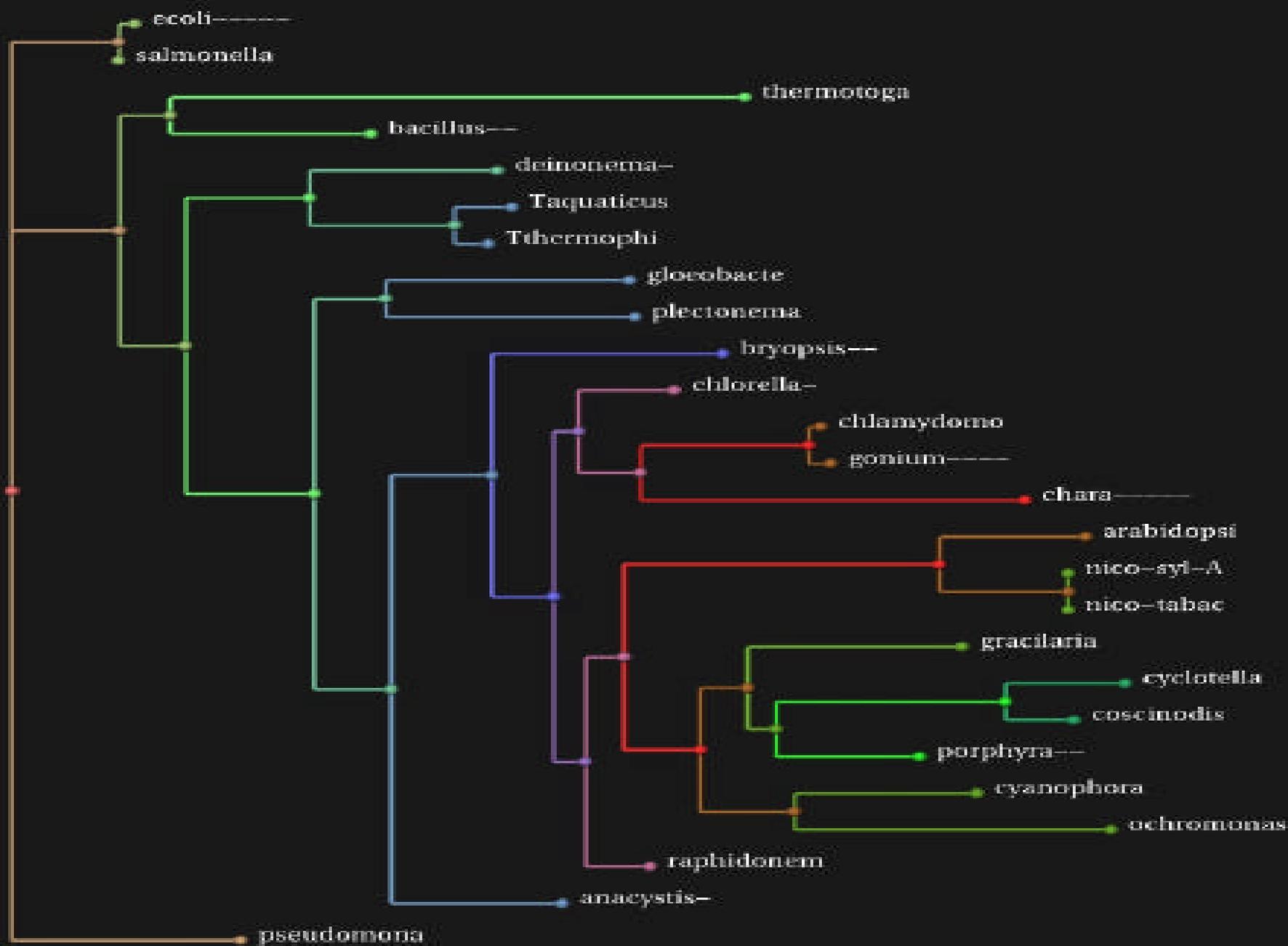
Overview of parallel program flow



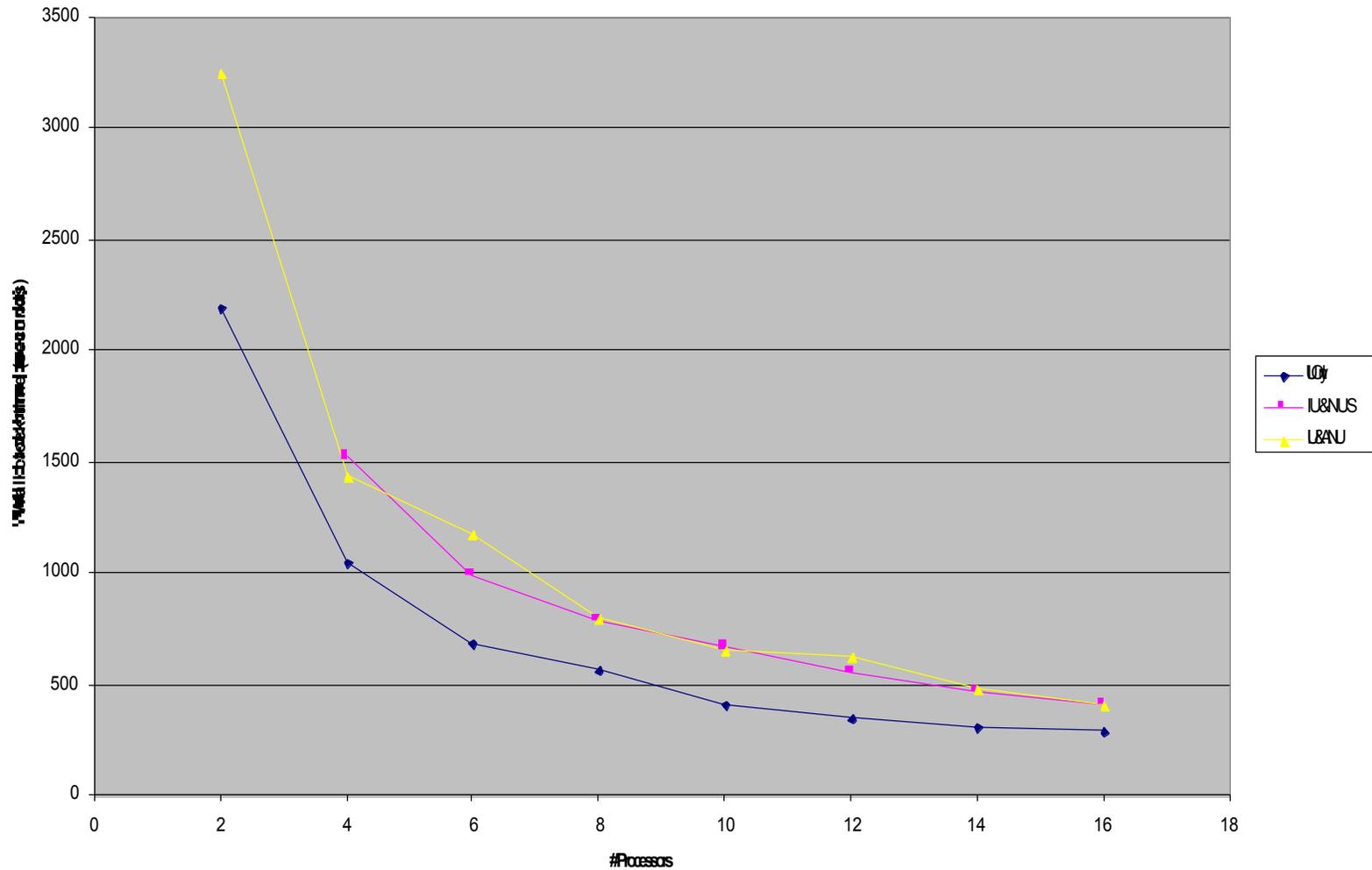
Because of local effects....

- **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**





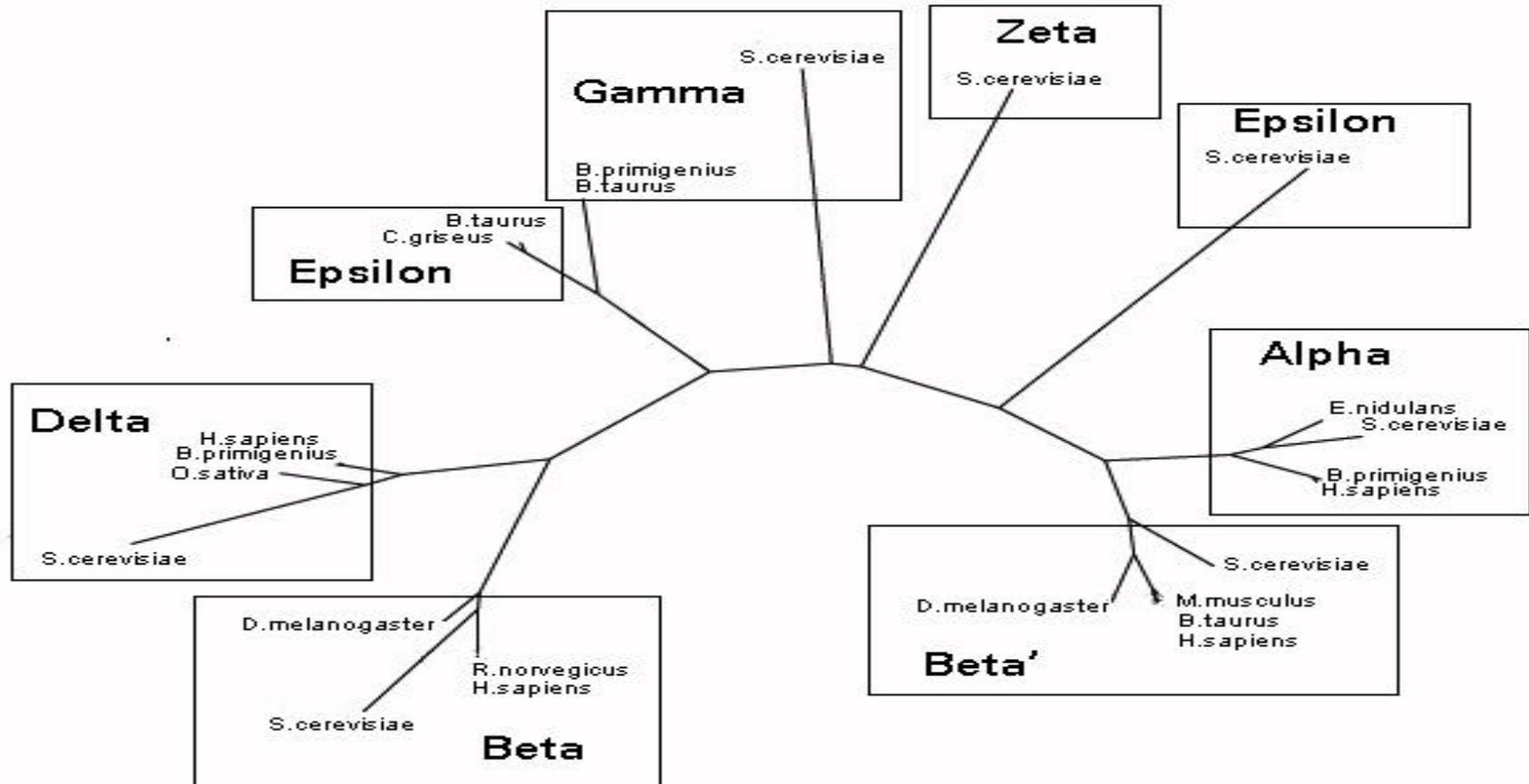
fastDNAml: version = '1.0.6', likelihood = -7984.141714506, ntaxa = 26, opt_level = 0, smoothed = 1



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**

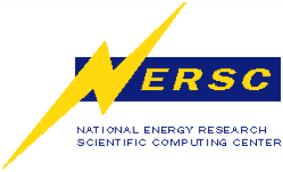
Cytoplasmic Coat Proteins



- **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 fastDNAmI.**
- **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 fastDNAmI has been facilitated by Shared University Research grants from IBM, Inc.**
- **IU's work with fastDNAmI would be impossible without our collaboration with Gary Olsen, U. of Illinois, the creator of this program.**

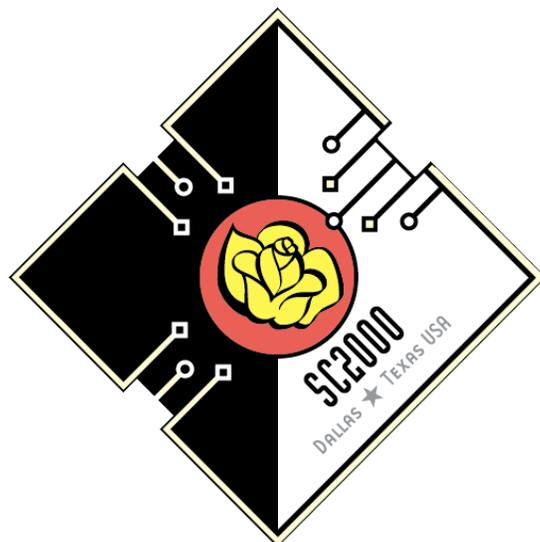
- **Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368-376**
- **Baxevanis, A.D., and B.F.F. Ouellette. 1998. *Bioinformatics: a practical guide to the analysis of genes and proteins*. Wiley-Interscience, NY.**
- **Swofford, 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.berkeley.edu/subway/phylogen**
- **evolution.genetics.washington.edu/phylip/software**
- **<http://www.indiana.edu/uits/~rac>**



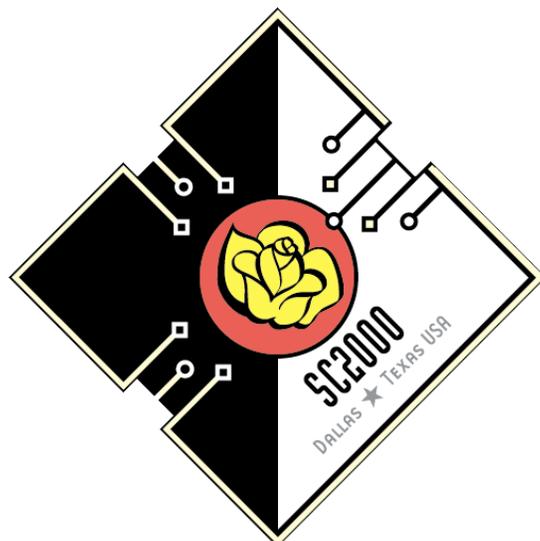
URLs for phylogenetic software



- **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
- **fastDNAmI**
geta.life.uiuc.edu/~gary/



Afternoon Break



Specialized biological databases and their role in building models of regulation

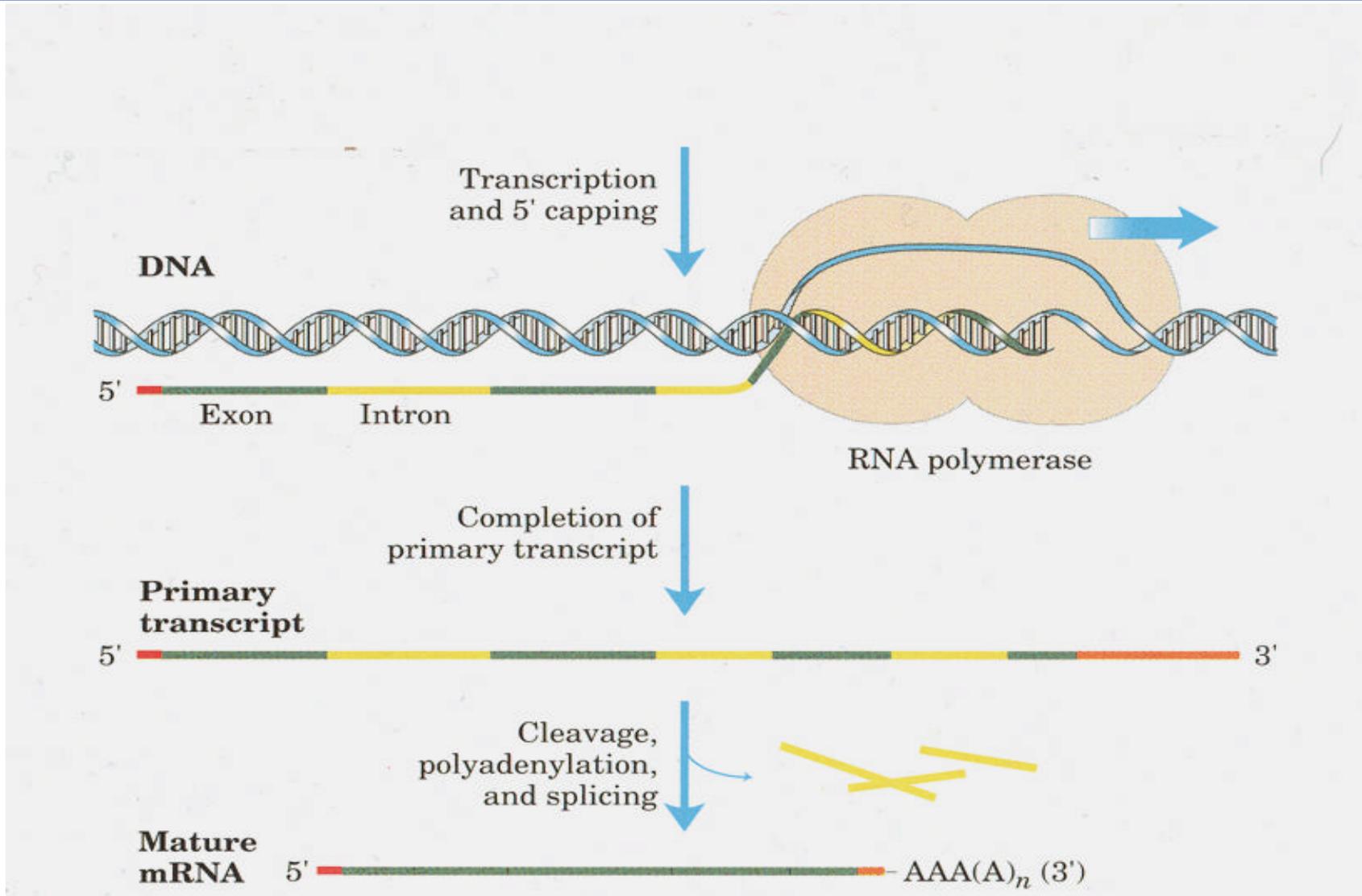
Inna Dubchak
ILDubchak@lbl.gov
NERSC

Overview of alternative splicing

- **What is alternative splicing?**

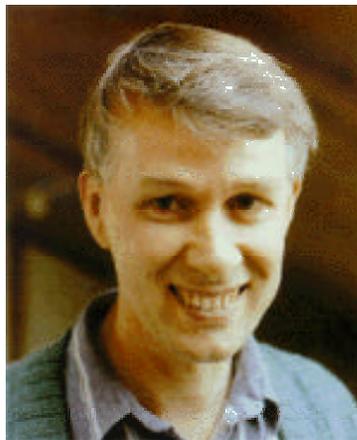
- **What is possible to do computationally to better understand this complicated phenomenon?**
 - **Frequency of alternative splicing**
 - **Specialized databases**
 - **Search for regulatory elements**

PROCESSING mRNA



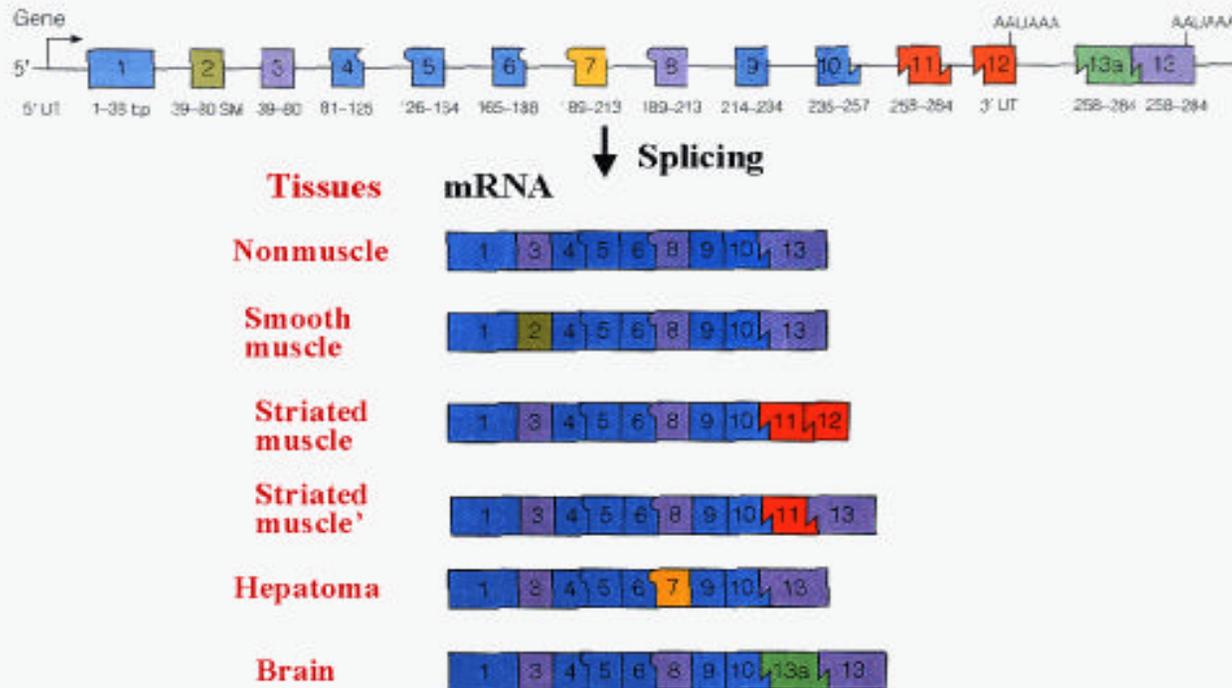
The Nobel Prize in Physiology or Medicine 1993

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.

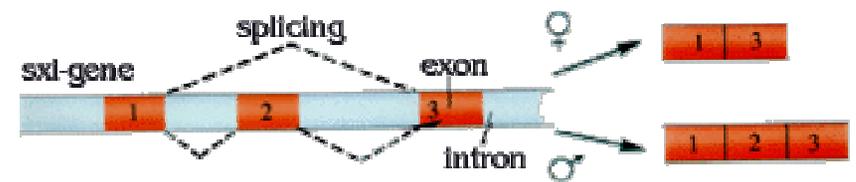


α-Tropomyosin pre-mRNA

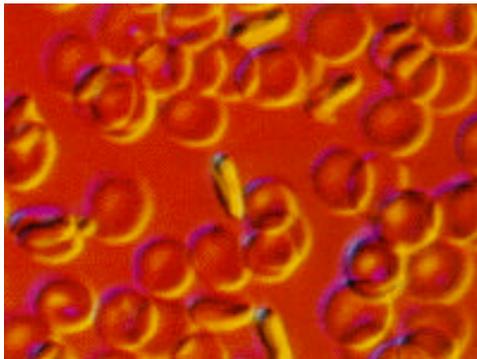
Alternative Splicing of α-tropomyosin pre-mRNA



- A precursor-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 errors cause thalassemia**
- **Thalassemia, a form of anemia common in the Mediterranean 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.**

Similarity analysis of two sequences

- **Gene families**
multiple similar genes
exist due to 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



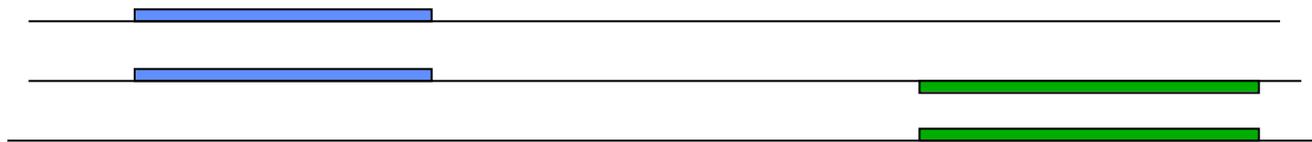
- **Relatively long identical
fragments**

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.**



A
B
C



~ 240 clusters of isoforms

Alternative Splicing DB

[DB CONTENT](#) | [HOW TO USE](#) | [FURTHER WORK](#) | [SEARCH](#)



References to the Alternative Splicing Database:

ASDB: database of alternatively spliced genes

I. Dralyuk, M. Brudno, M. S. Gelfand, M. Zorn, and I. Dubchak (2000) Nucleic Acids Research 28(1), 296-297.

M. S. Gelfand, I. Dubchak, I. Dralyuk and M. Zorn (1999) Nucleic Acids Research, 27(1), 301.

Search Alternative Splicing DB (proteins)

Look by

Show help

Return

Search Alternative Splicing DB

Look by

Show help

Return res

Alternative
Splicing **DB**

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)

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



Bookmarks Location: <http://devnull.lbl.gov:8888/bin/retrieve?cluster=68> What's Related

Lawrence Berkel

```
2ACA_HUMAN      IELQNDKPNR  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  LLRILESID
2ACB_HUMAN      .....

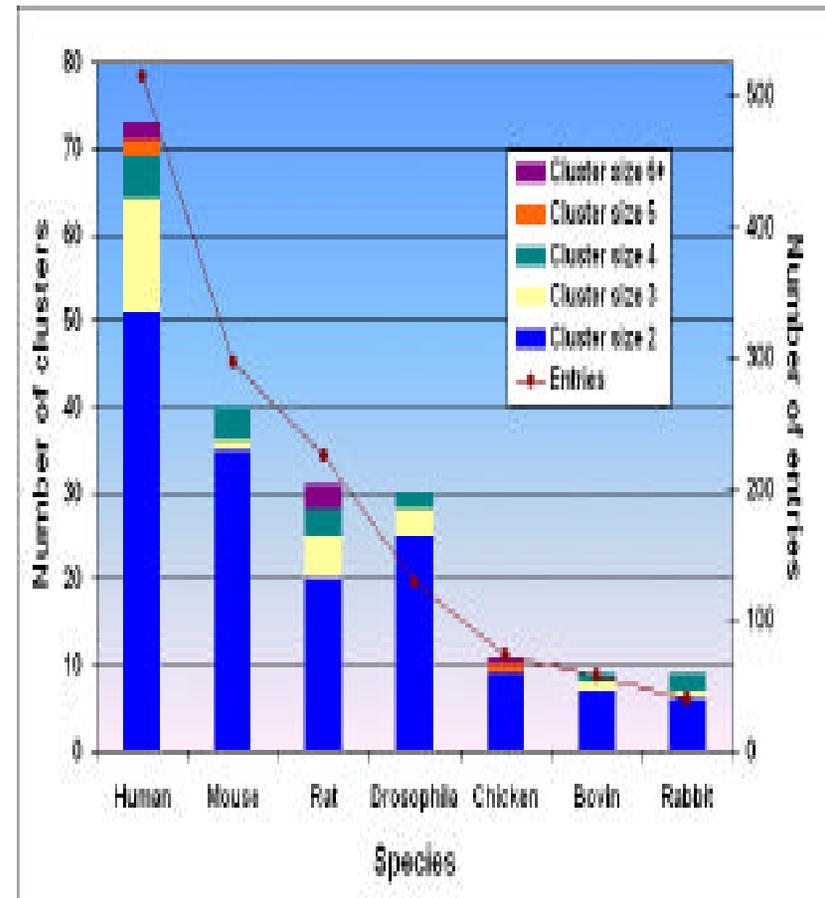
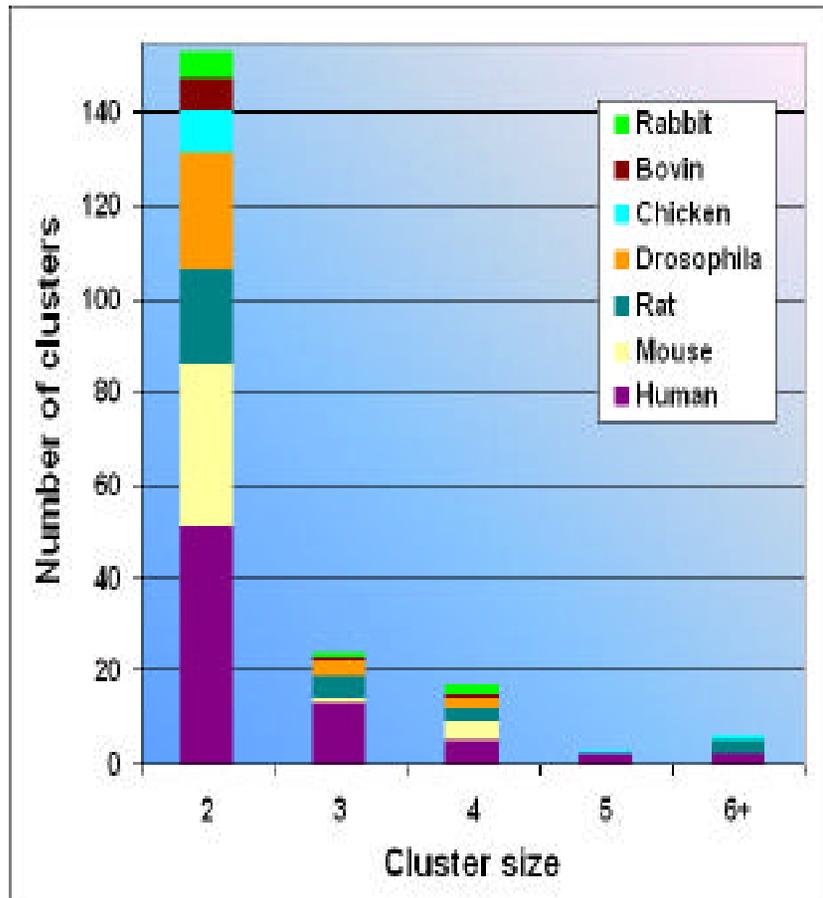
2ACA_HUMAN      FAQELVECKS  SRGSLSQEKE  MMQILQETLT  TSSQANLSVC  RSPVGDKAKD
2ACB_HUMAN      ..... .MMIKETSLR  RDPDLRGELA  FLARGCDFVL

2ACA_HUMAN      TTSAVLIQQT  PEVIKIQNKP  EKKPGTPLPP  PATSPSSPRP  LSPVPHVNNV
2ACB_HUMAN      PSRFKRLKS  FQQTQIQNKP  EKKPGTPLPP  PATSPSSPRP  LSPVPHVNNV

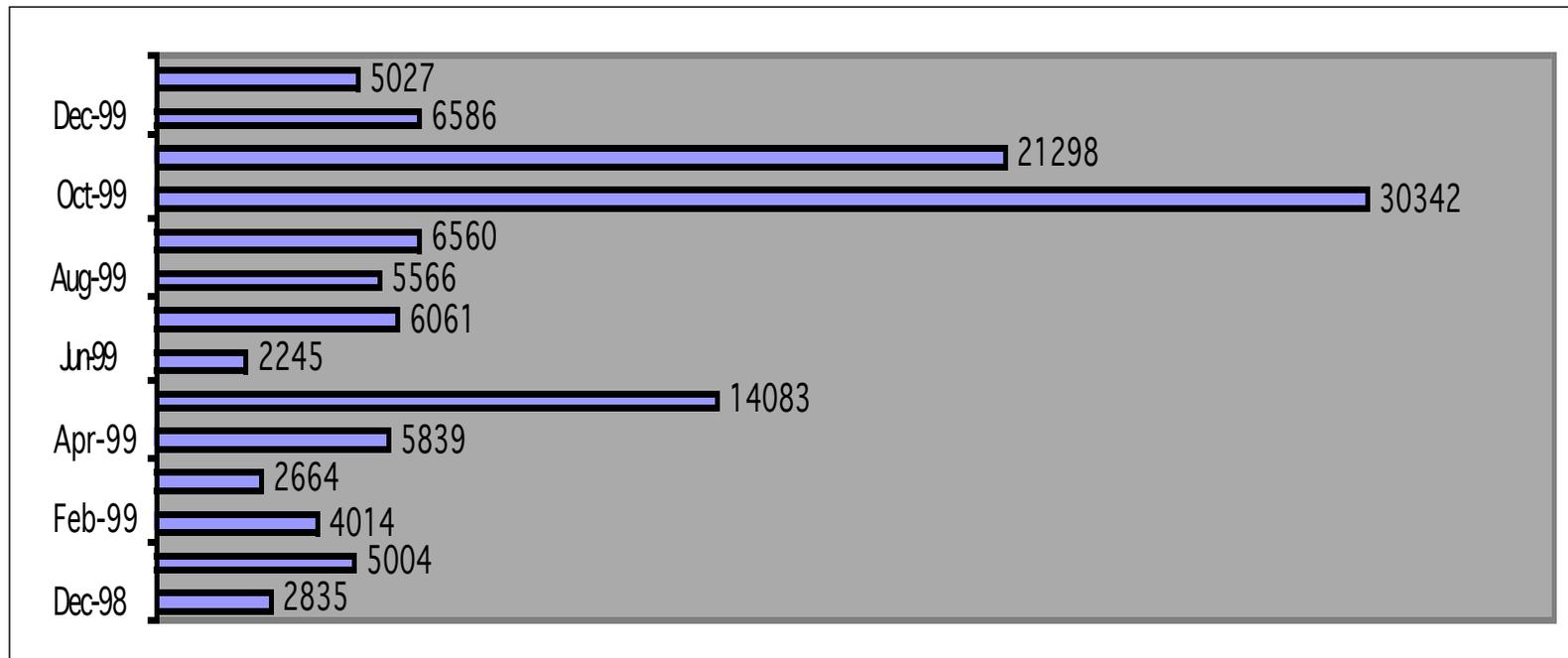
2ACA_HUMAN      VNAPLSINIP  RFYFPEGLPD  TCSNHEQTLS  RIETAFMDIE  EQKADIYEMG
2ACB_HUMAN      VNAPLSINIP  RFYFPEGLPD  TCSNHEQTLS  RIETAFMDIE  EQKADIYEMG

2ACA_HUMAN      KIAKVCGCPL  YWKAPMFAA  GGEKTGFVTA  QSFIAMWRKL  LNNHHDDASK
2ACB_HUMAN      KIAKVCGCPL  YWKAPMFAA  GGEKTGFVTA  QSFIAMWRKL  LNNHHDDASK

2ACA_HUMAN      FICLLAKPNC  SSLEQEDFIP  LLQDVVDTHP  GLTFLKDAPE  FHSRYITTVI
```



ASDB usage during 1999



- **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

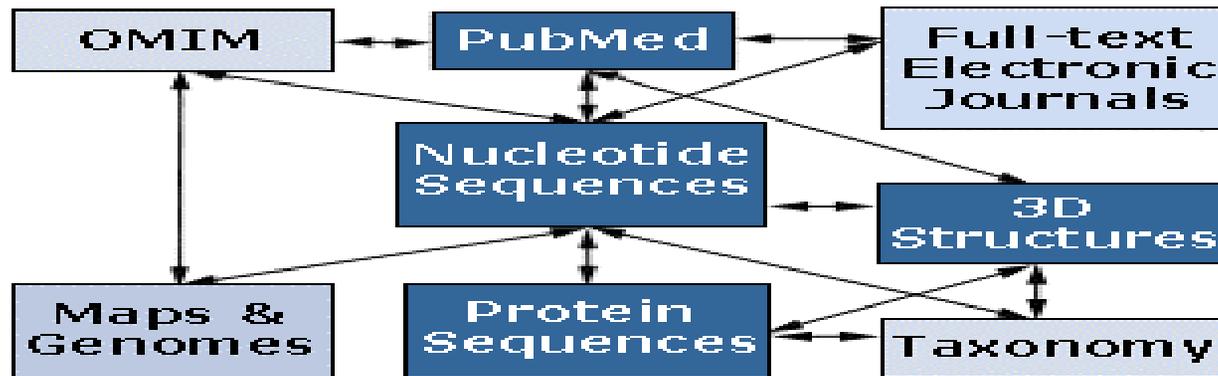
- A number of genomic sequence regulatory elements have been identified outside of traditional splice sites.
- The concept of splicing "**enhancers**" and "**silencers**" that promote or inhibit splicing at neighboring splice sites is well established.
- Many alternative exons are probably regulated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

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

- **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

- 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

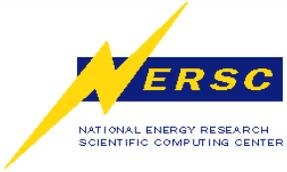
- **To calculate the confidence value of a particular word we select random subsets of a large dataset of constitutively spliced exons (1,504 exons; Burset & Guigo, 1996) equal in size to our alternative dataset.**
- **We then calculate the fraction of these subsets in which the word is over-represented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

Known Regulatory Elements

<u>enhancers</u>	<u>reference</u>
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
<u>silencers</u>	
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

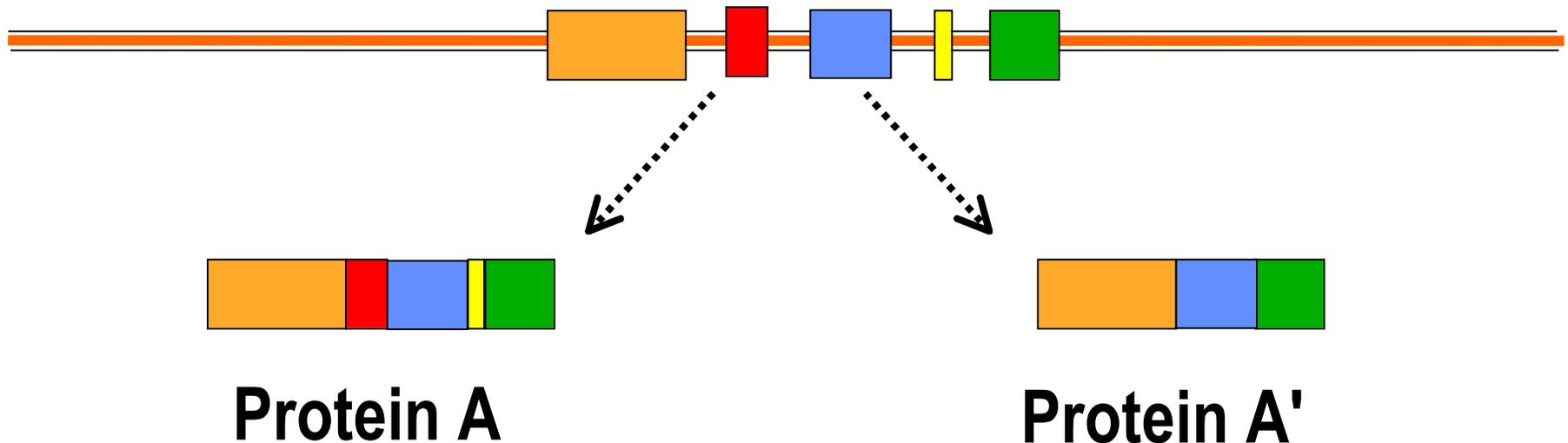


- **Discovering them in DNA sequence**
- **Tools for their visualization**
- **Biological importance**

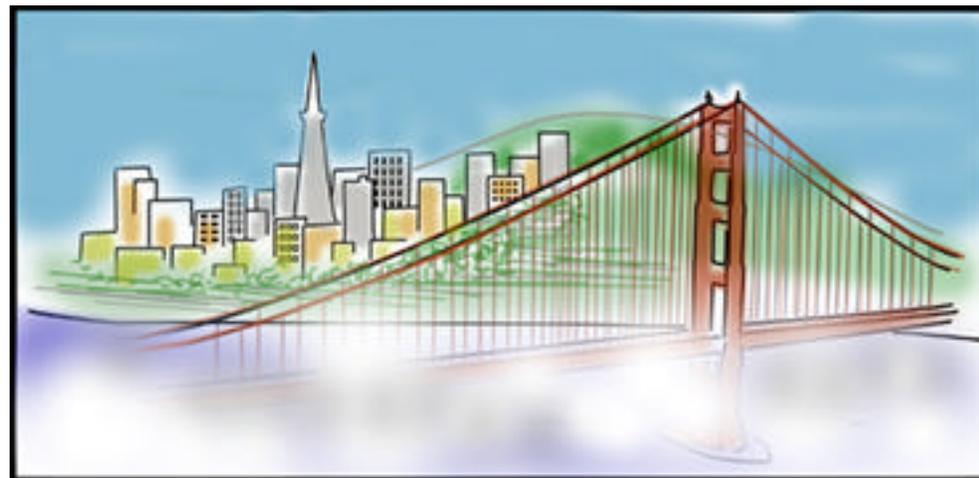
 Non-Coding

~ 5% coding
~ 95% non-coding

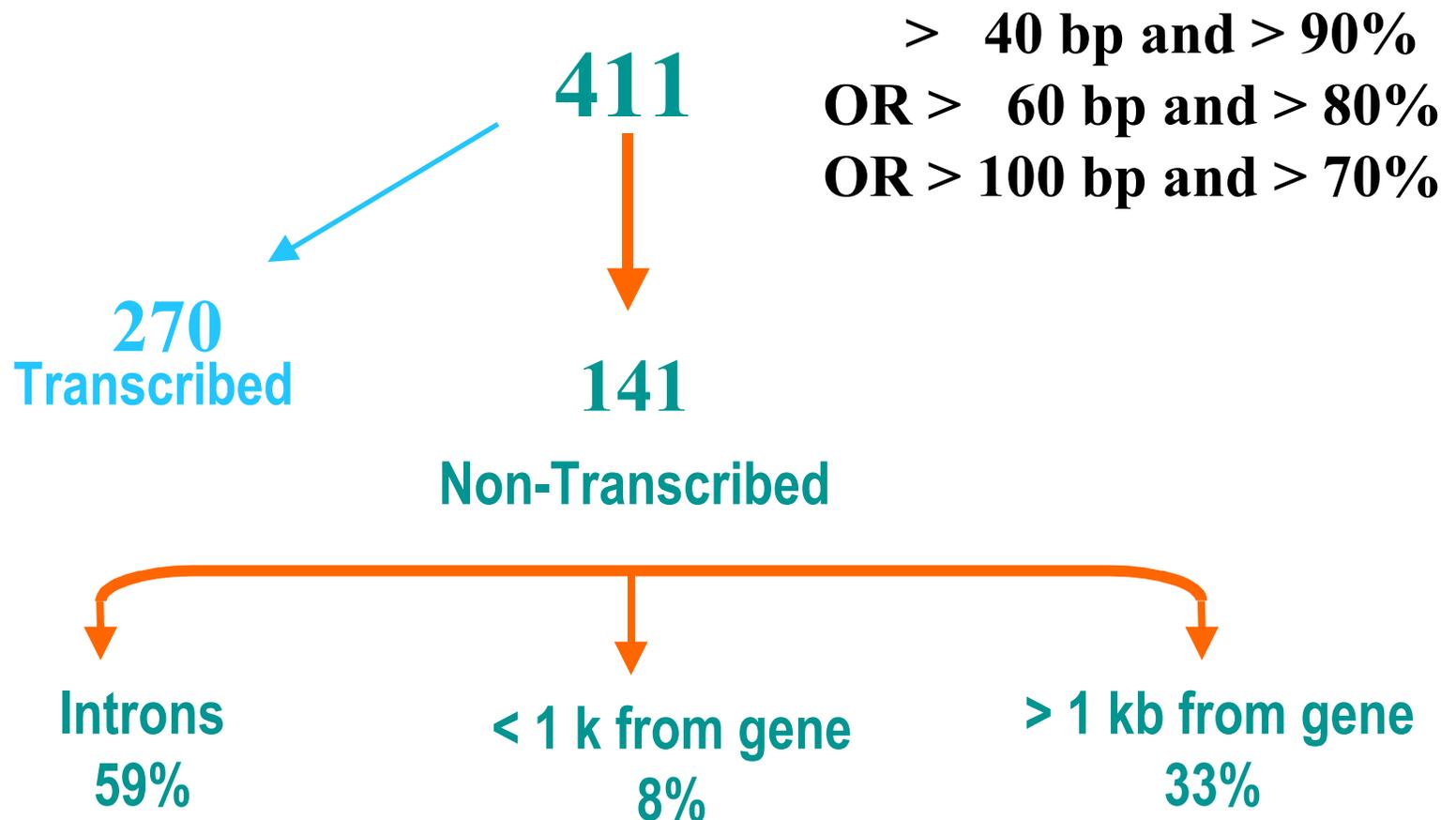
Gene A



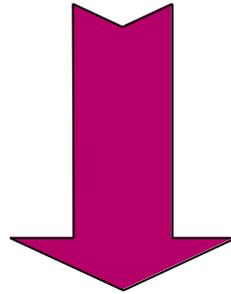
Information in Sequence



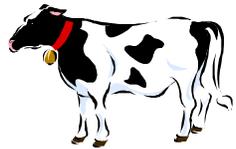
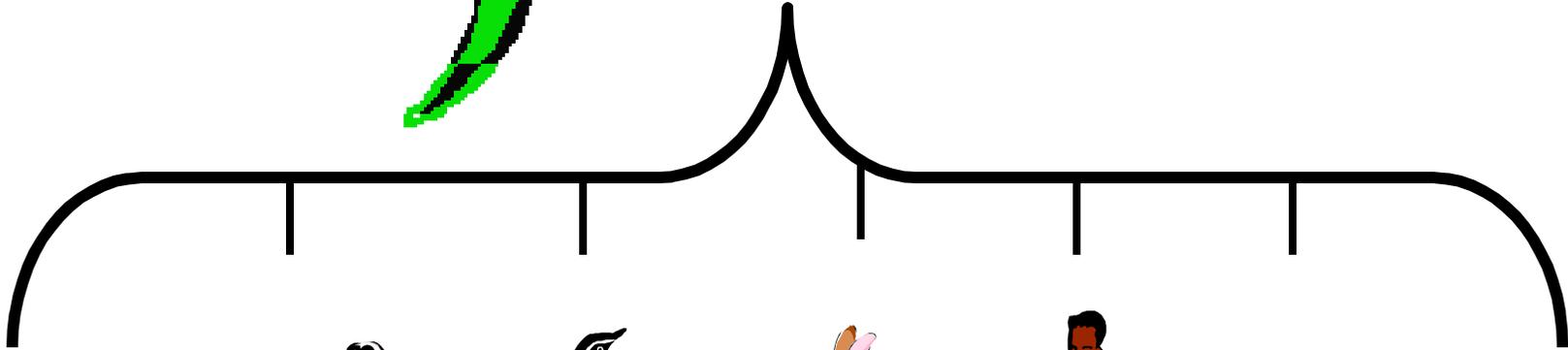
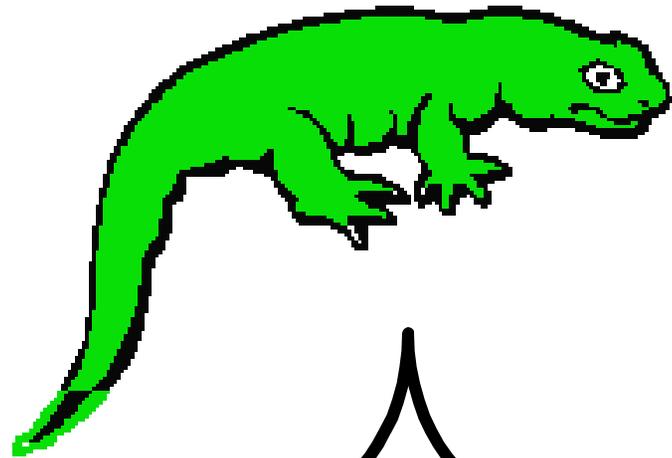
Conserved Human/Mouse Sequences in 830 kb Region



90 Elements in 1 Megabase



**Are most conserved
noncoding sequences
“functional” or are
they a product of
passive evolution?**

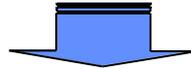
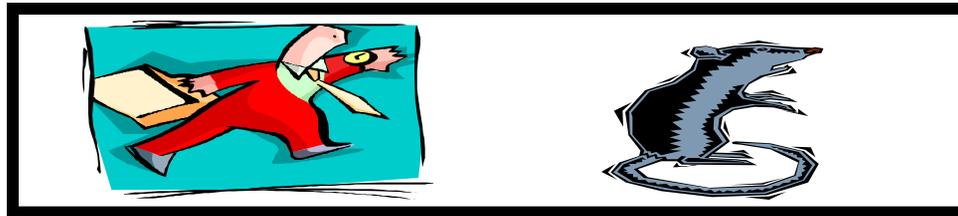


- Present in other species:
 - ◆ Cow (86%)
 - ◆ Dog (81%)
 - ◆ Rabbit (73%)
- Genomic position conserved in human, mouse, dog and baboon



- Single copy in the human genome

Identification



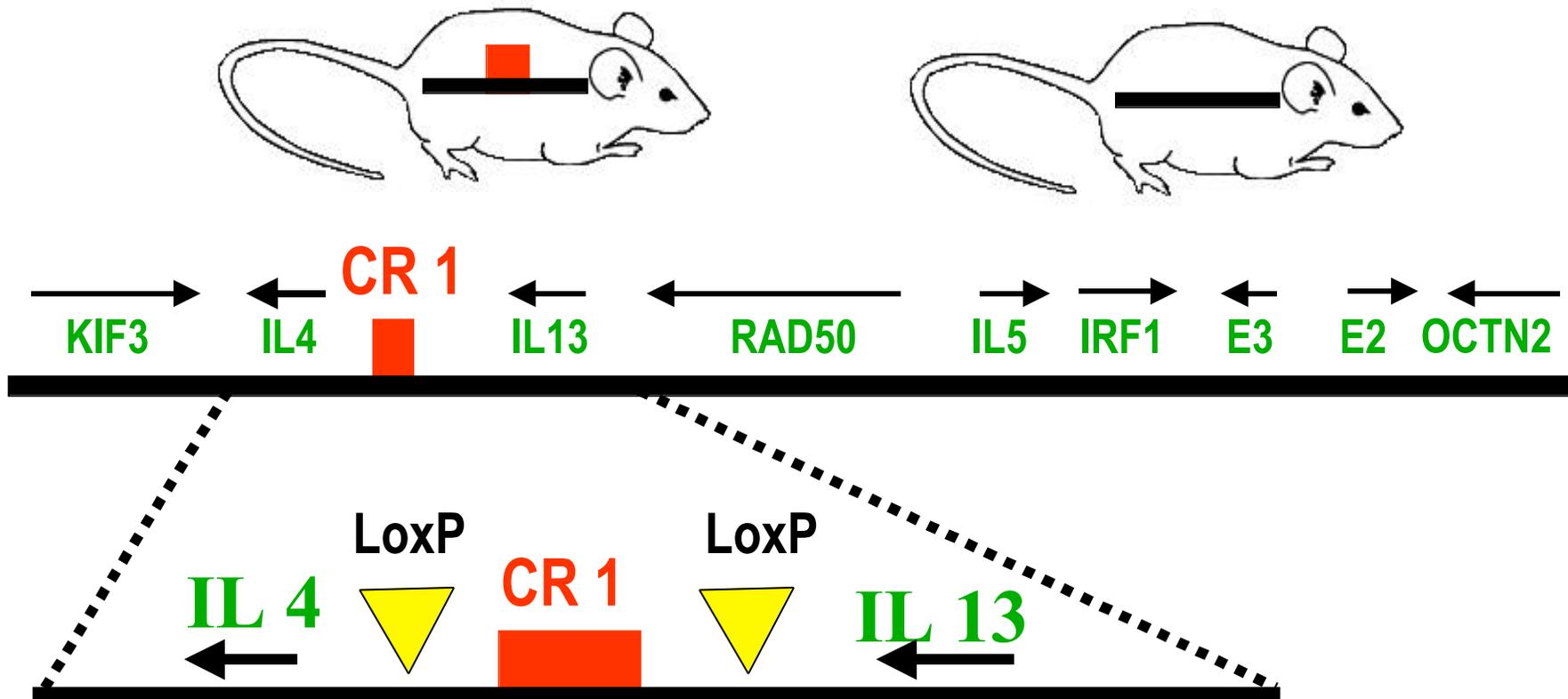
Verification

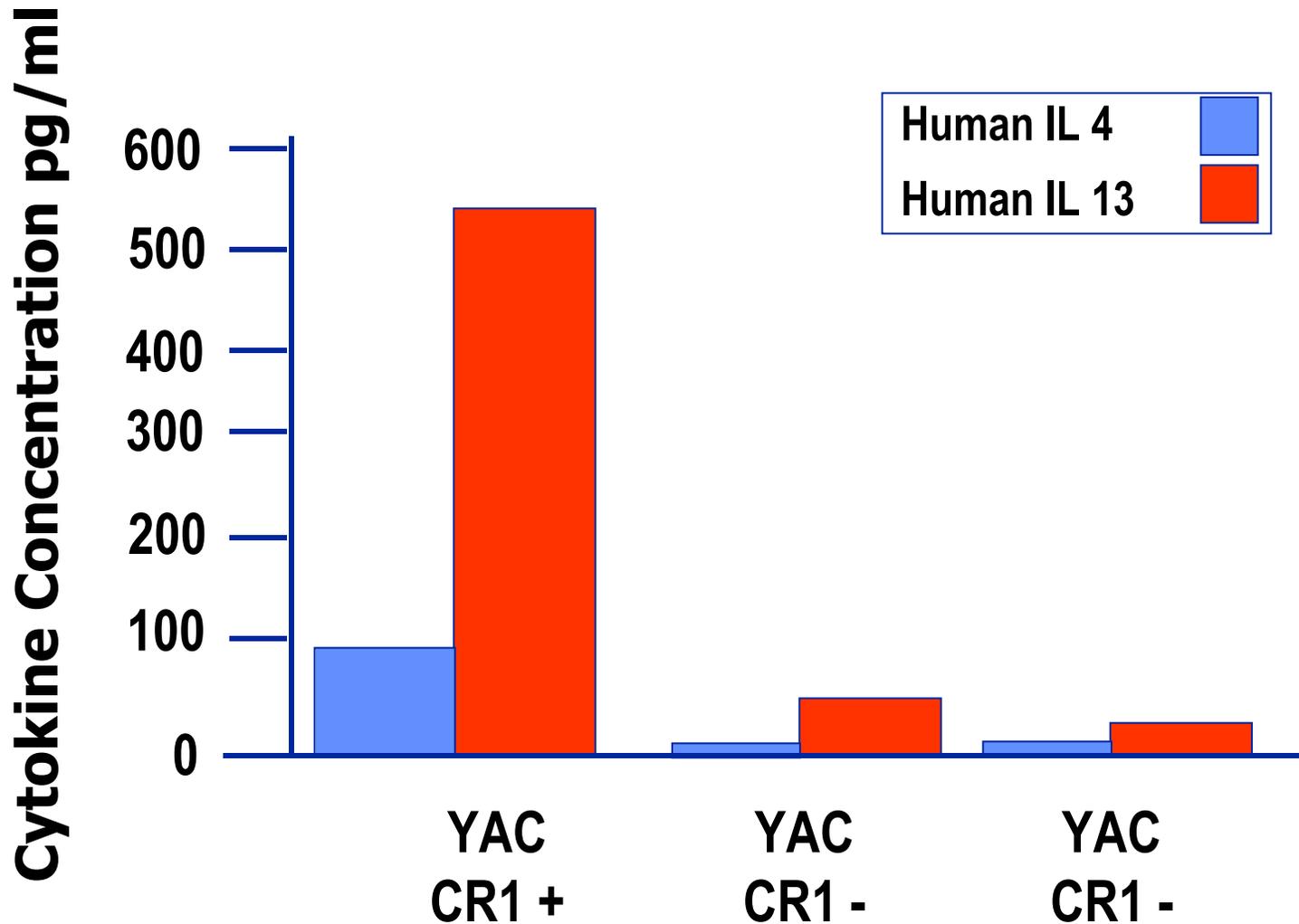


Analysis

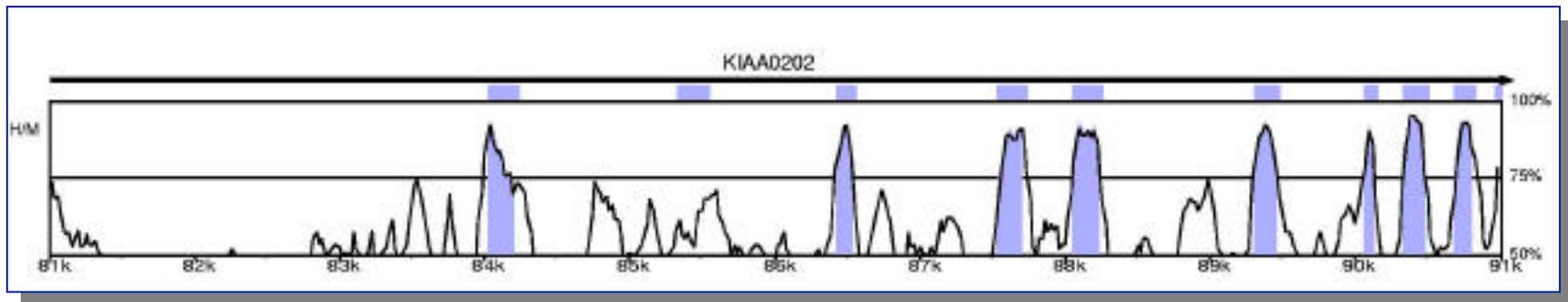
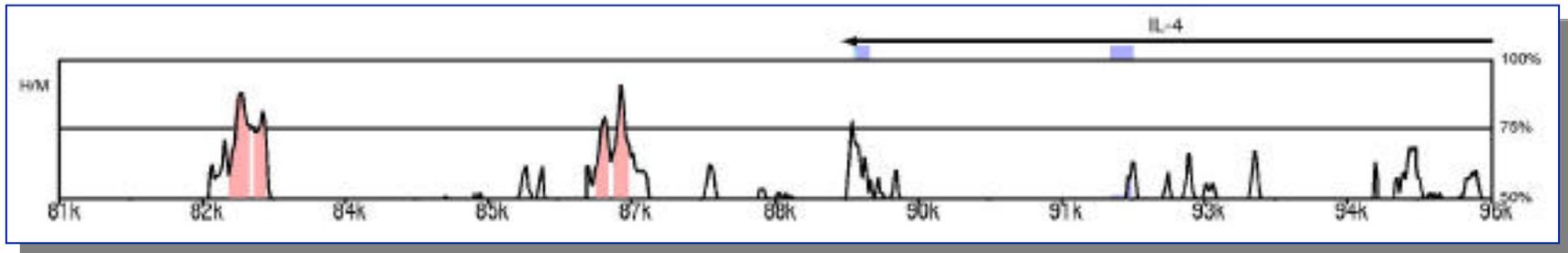


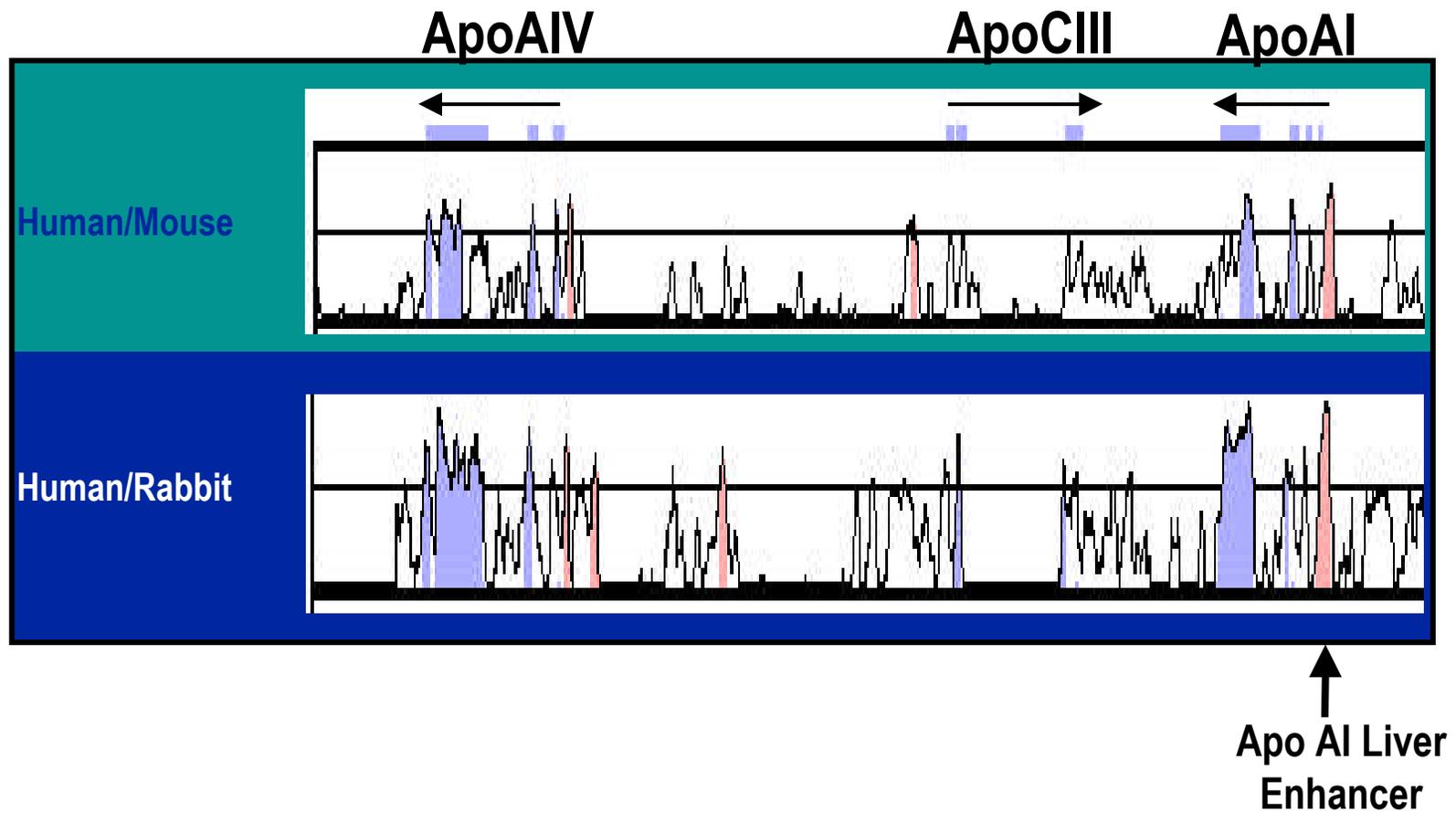
Generate Human 5q31 YAC Transgenic Mice



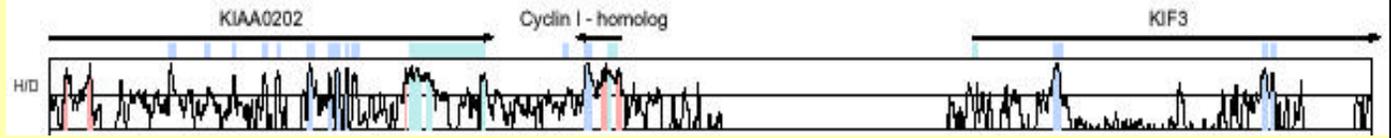


Vista (Visual Tool for Alignment)

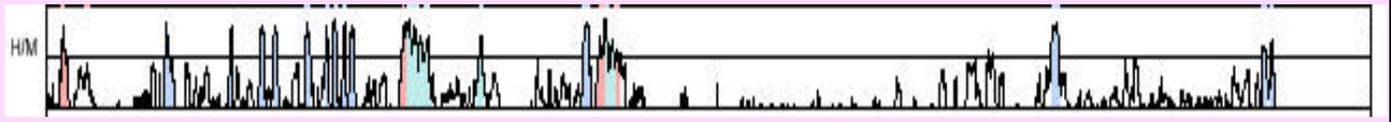




Human/Dog



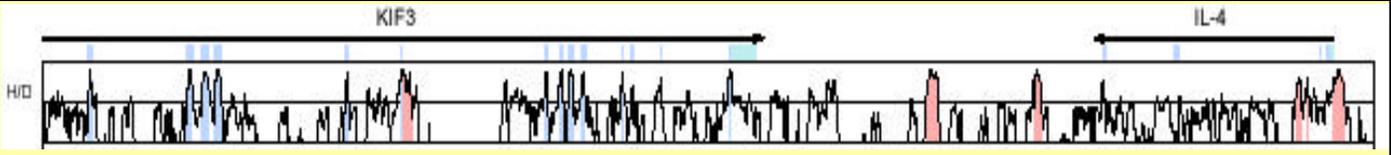
Human/Mouse



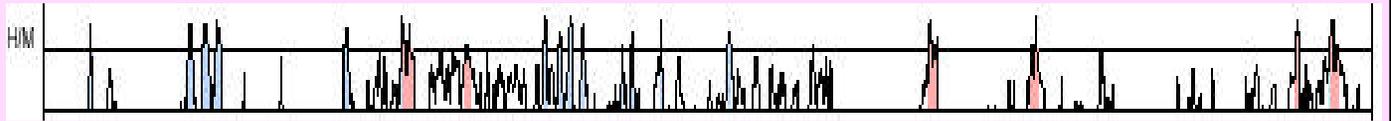
Mouse/Dog



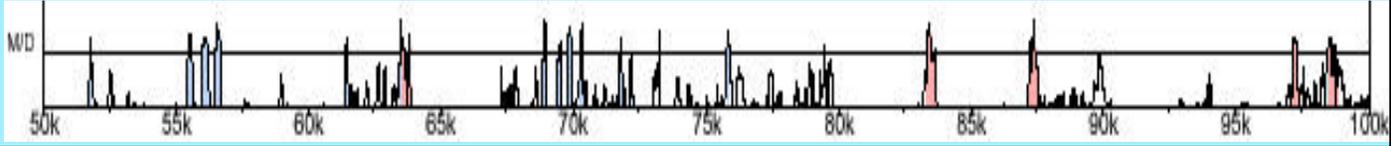
Human/Dog



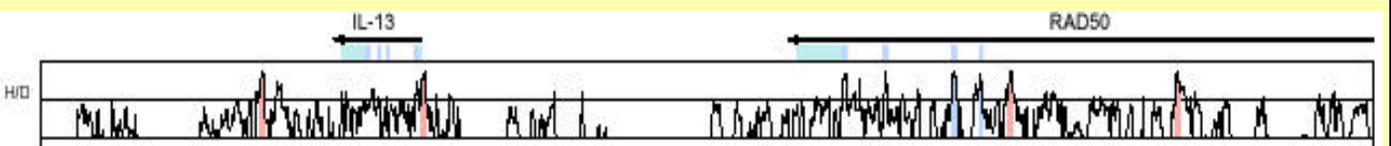
Human/Mouse



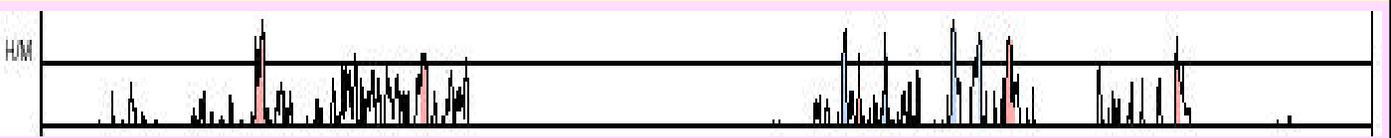
Mouse/Dog



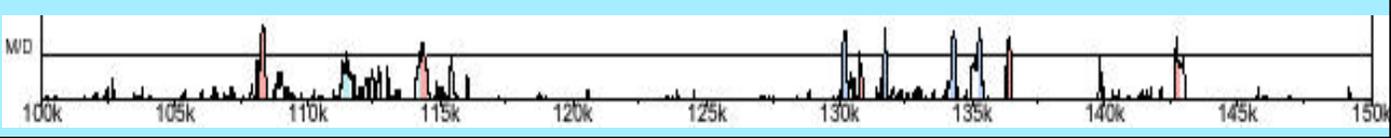
Human/Dog

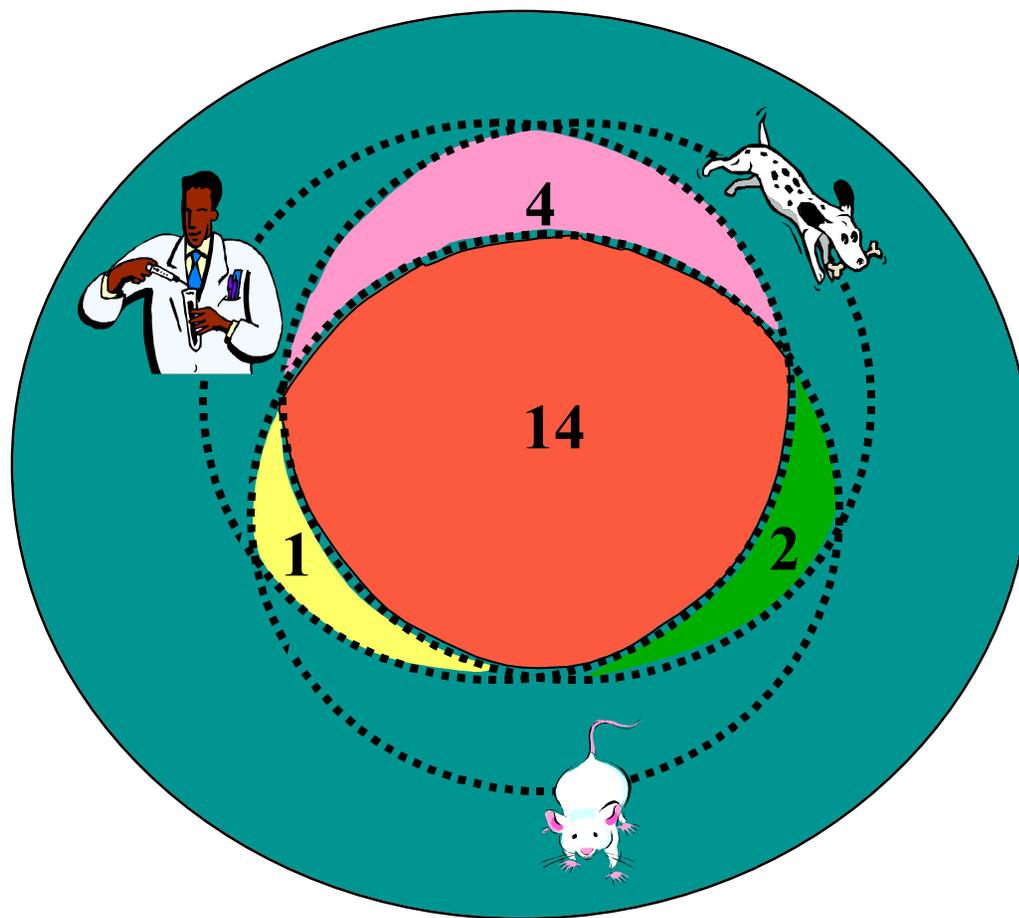


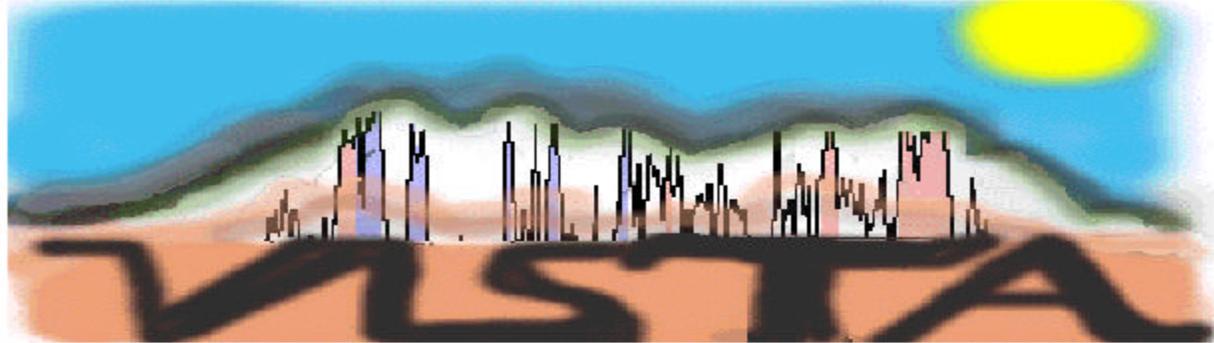
Human/Mouse



Mouse/Dog



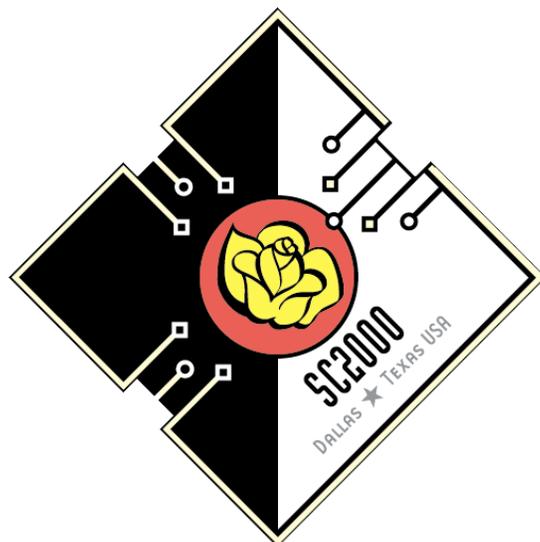




Welcome to the **VISTA**, or **VIS**ualization **T**ool for **A**lignments home page

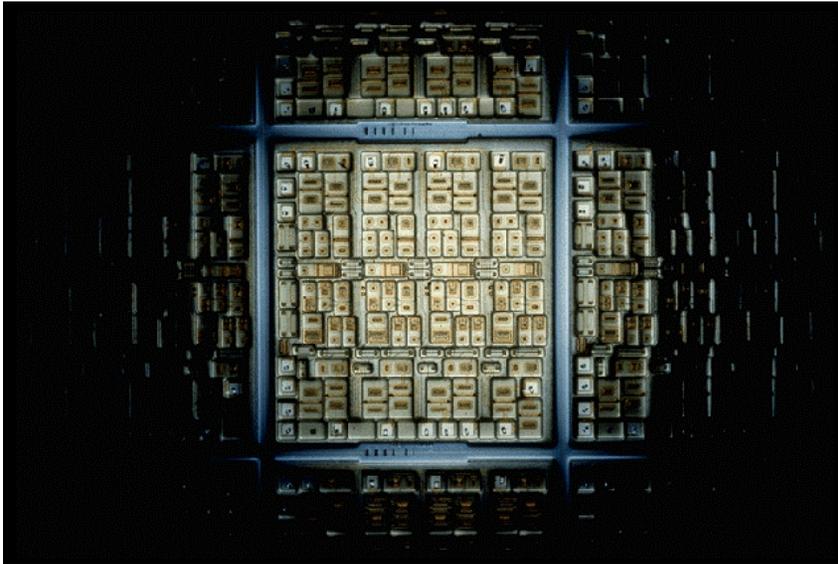
VISTA is an integrated system for global alignment and visualization, designed for comparative genomic analysis.

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.*

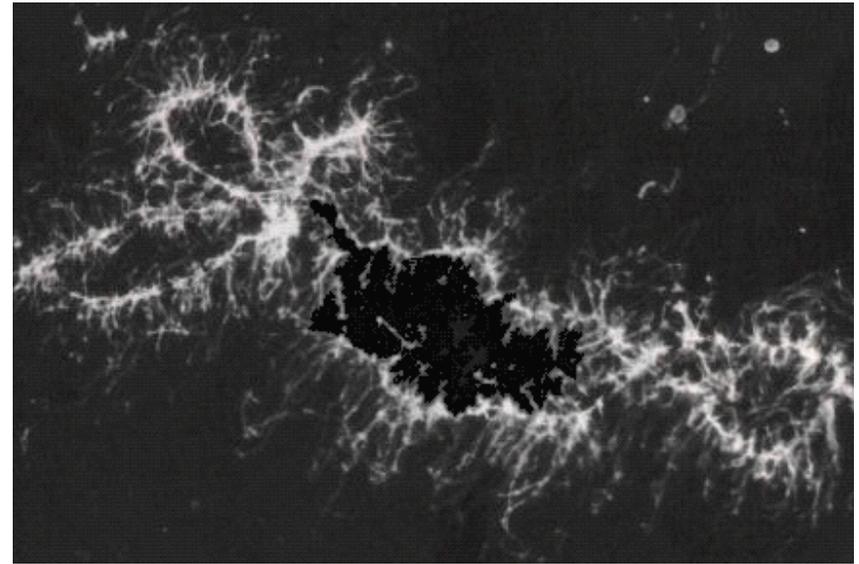


Gene Regulatory Networks and Cellular Processes

Adam Arkin
APArkin@lbl.gov
LBL



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

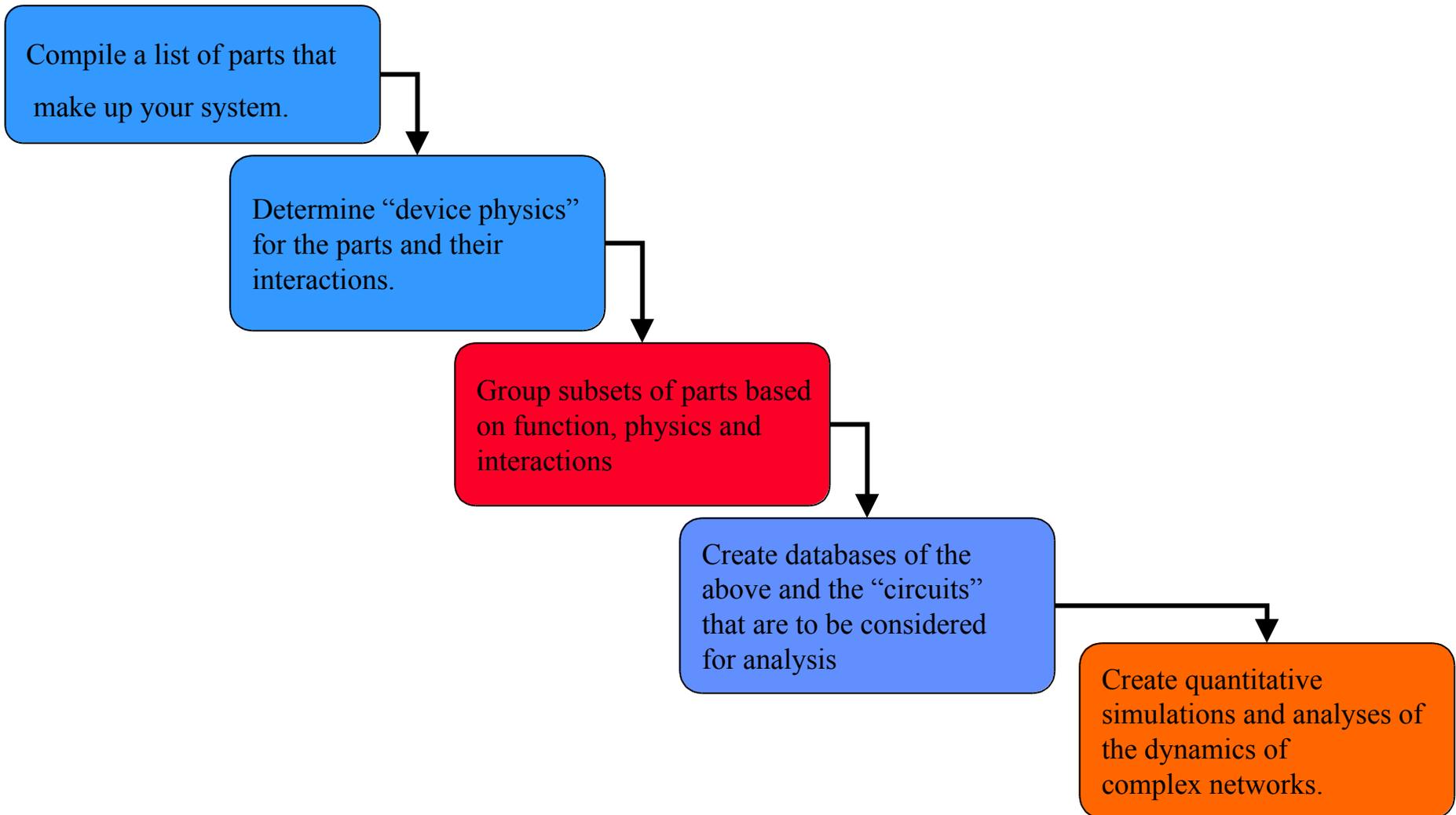
Partial knowledge of parts list
Partial knowledge of "device physics"
Partial knowledge of interactions

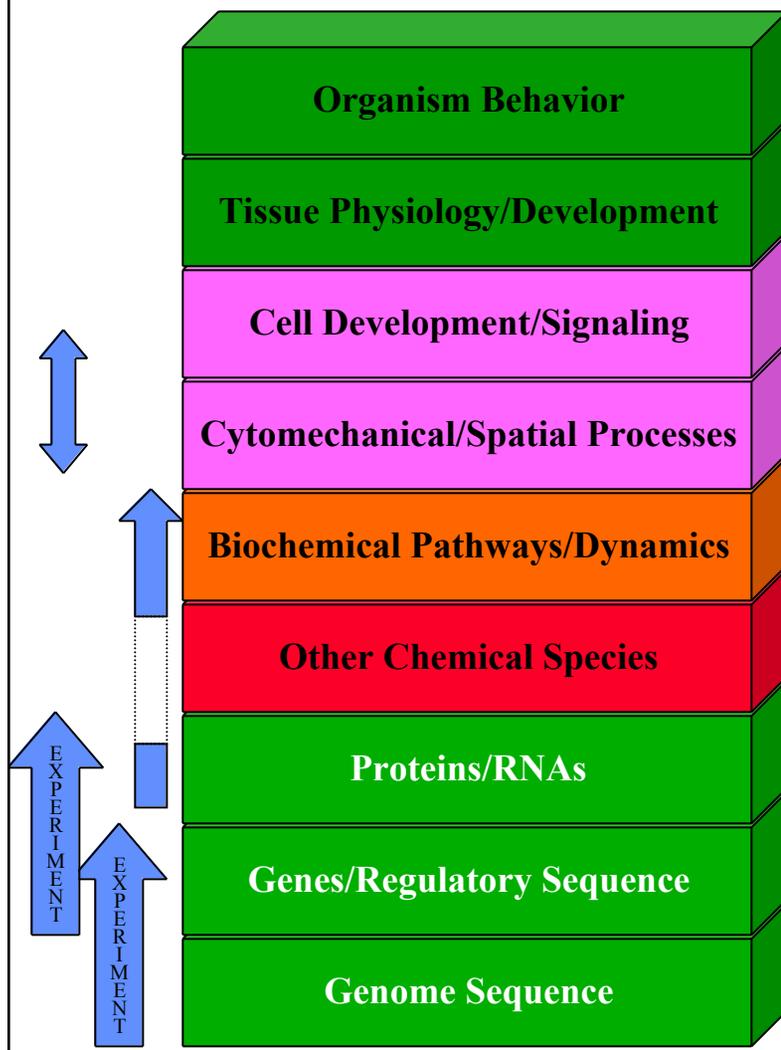
No one fully understands how this circuit works!!
Its just too complicated.

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:





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

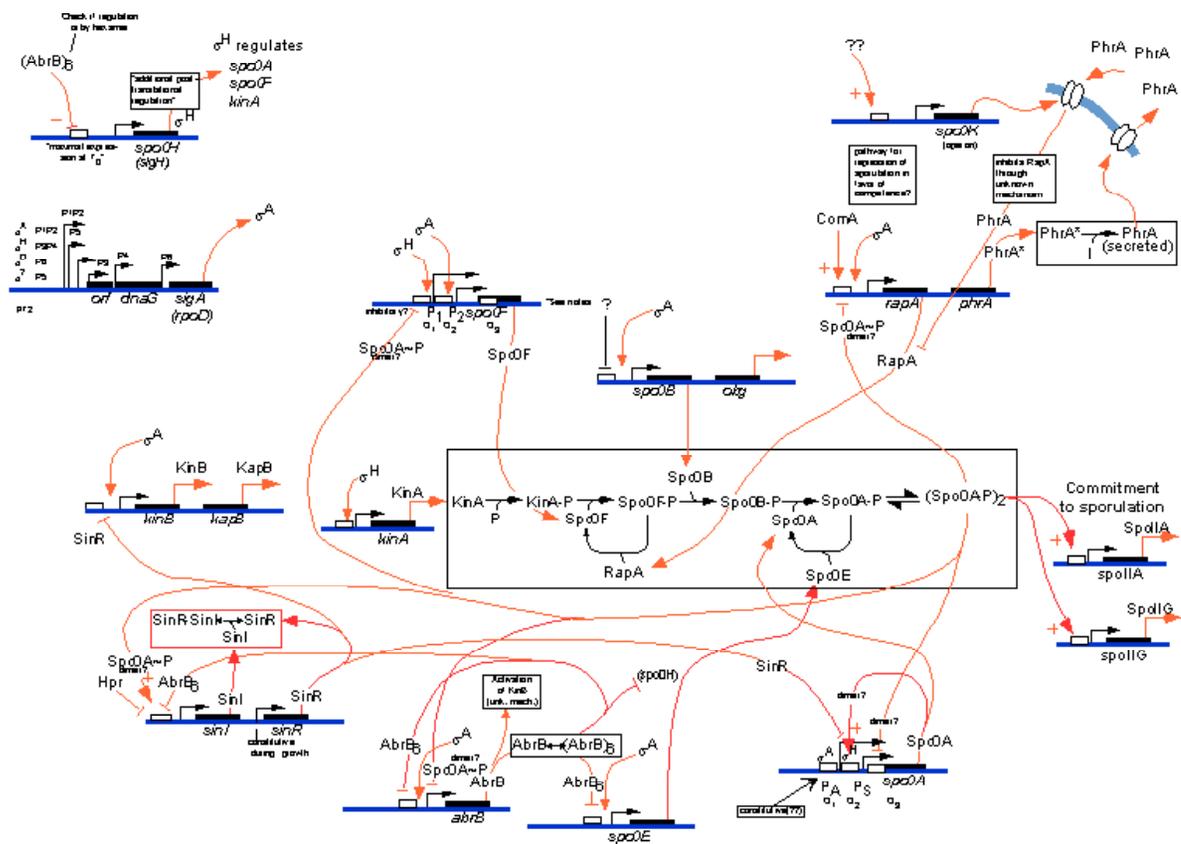
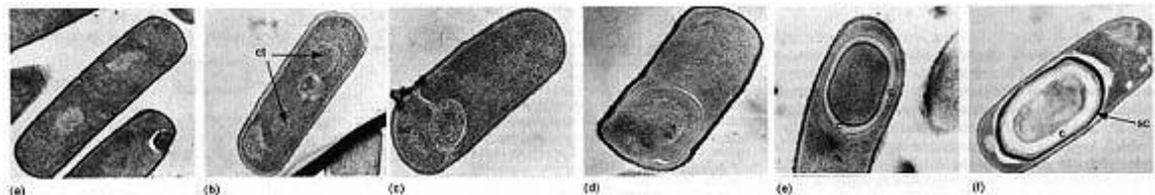
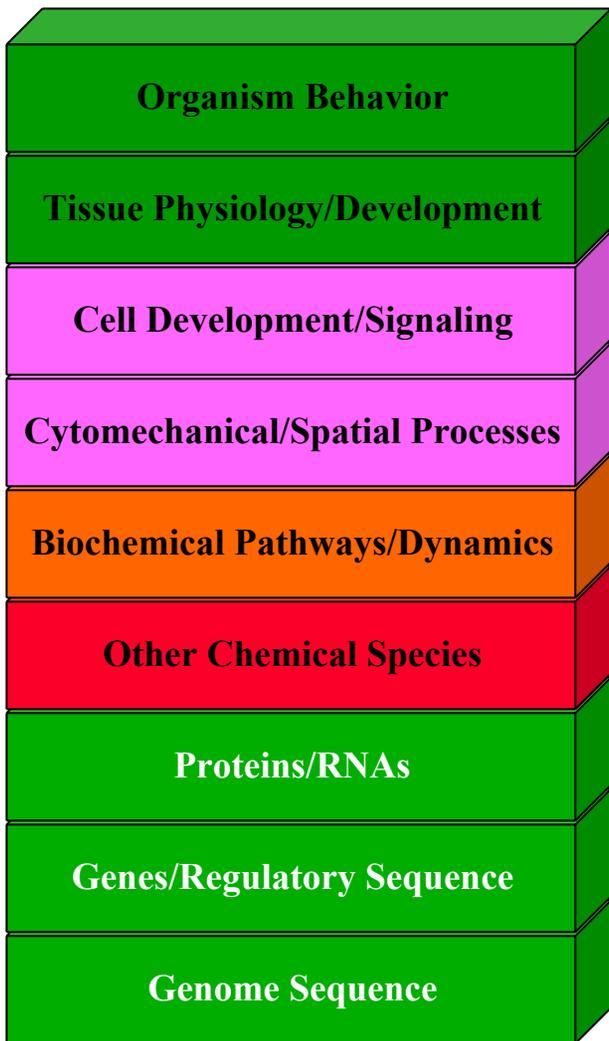
■ **There are challenges in:**

- **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**

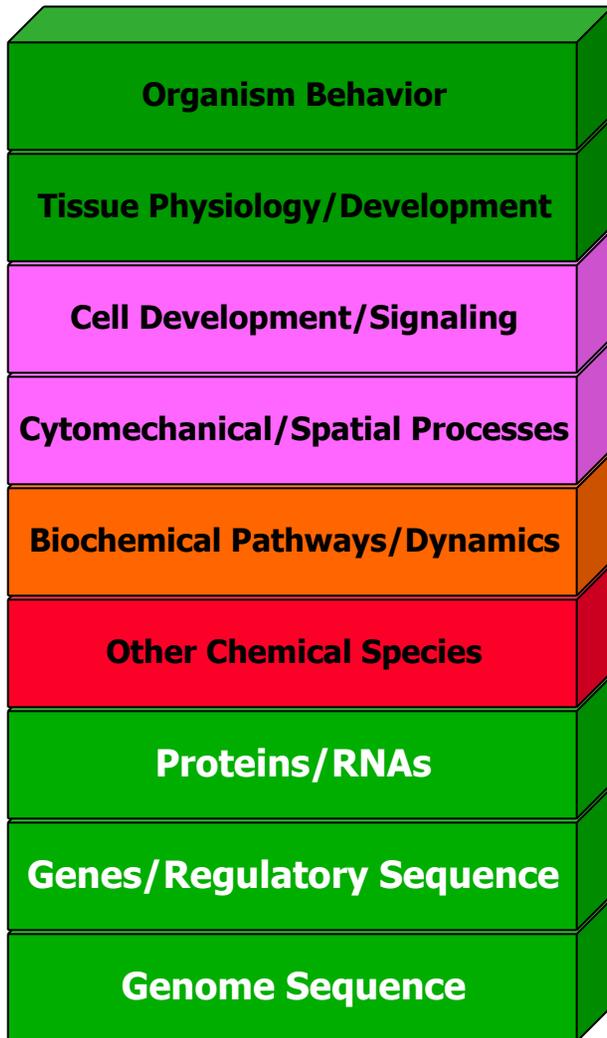
■ **Pay-off in**

- **Industrial bioengineering**
- **Rational pharmaceutical design**
- **Basic biological understanding**

Complexities of Cellular Function



Heterogeneity of Data

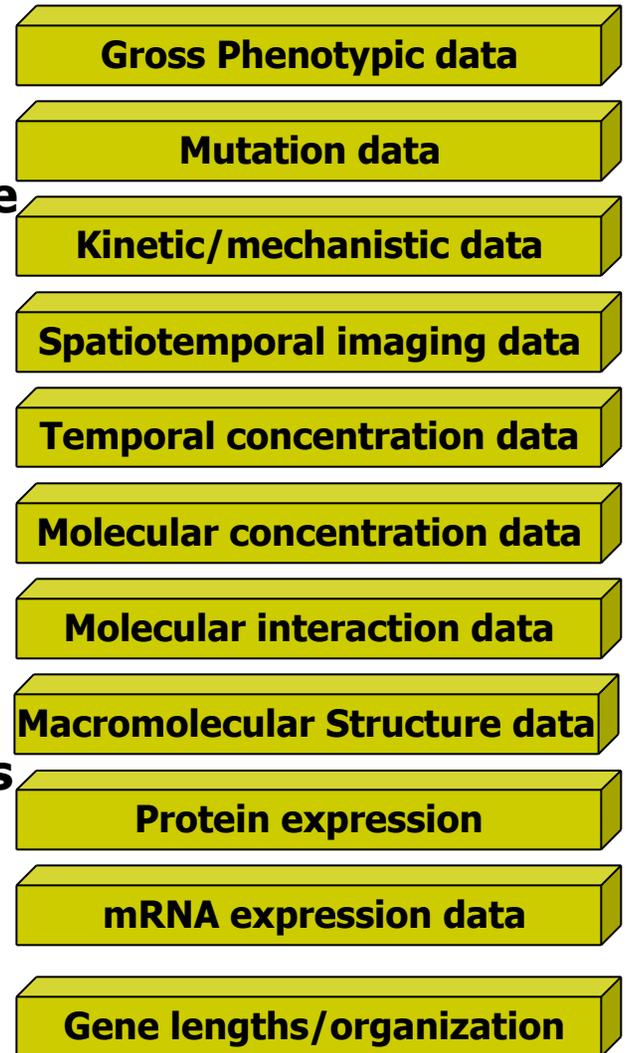


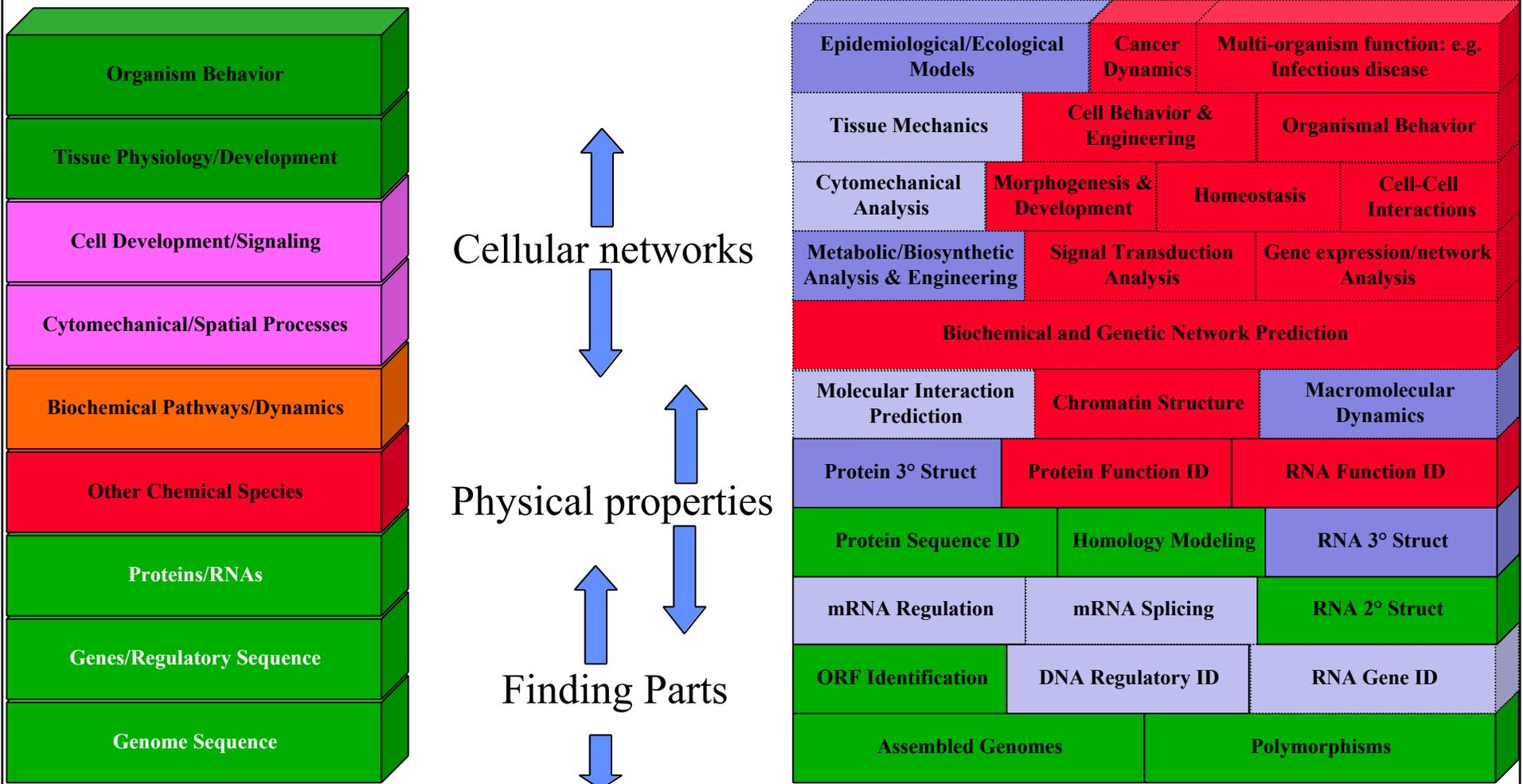
Data are:

- **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.





Why now?

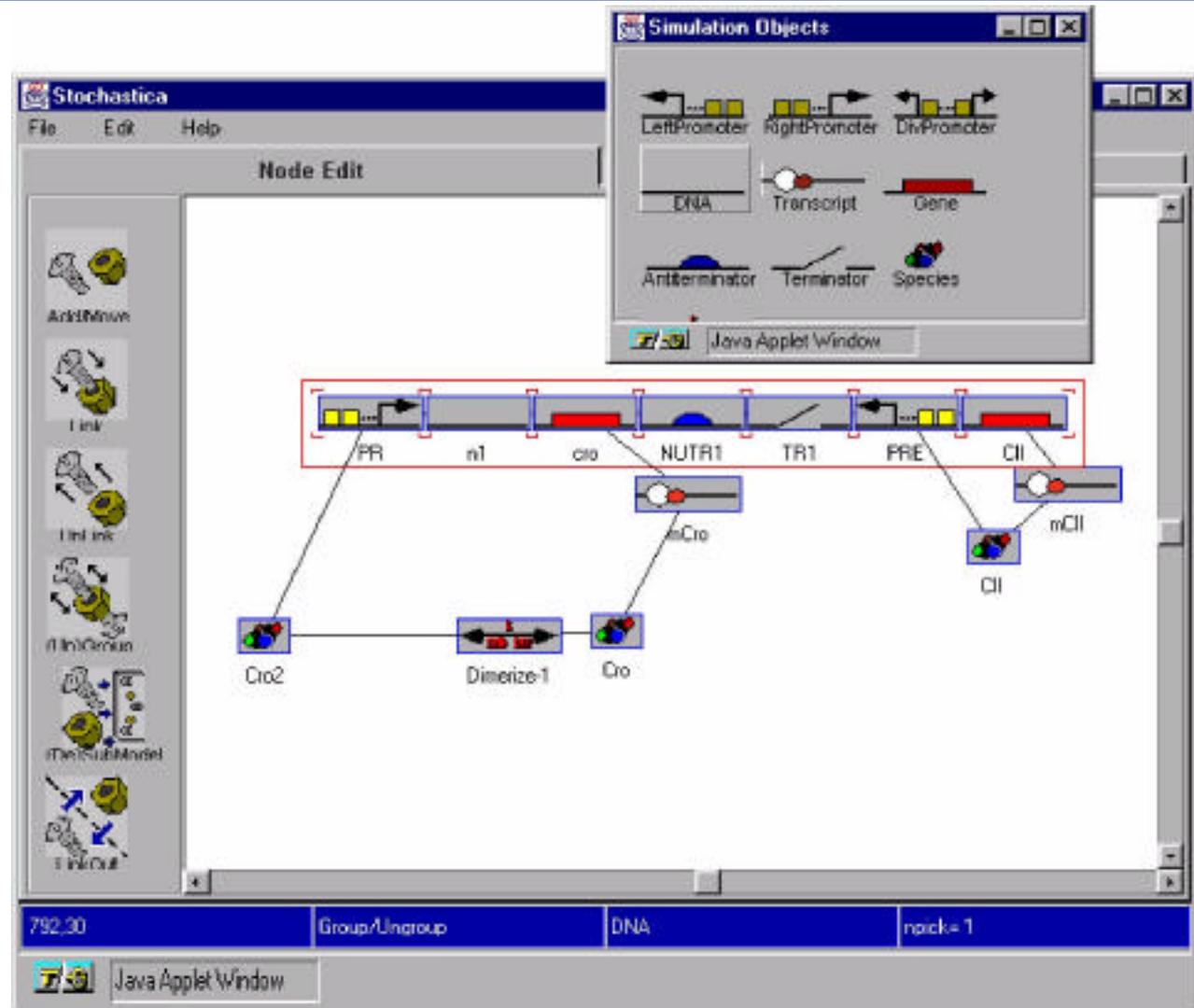
- **Genome projects are providing a large (but partial) list of parts**
- **New measurement technologies are helping to identify further components, their interactions, and timings**
 - ✓ **Gene microarrays**
 - ✓ **Two-Hybrid library screens**
 - ✓ **High-throughput capillary electrophoresis arrays 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**

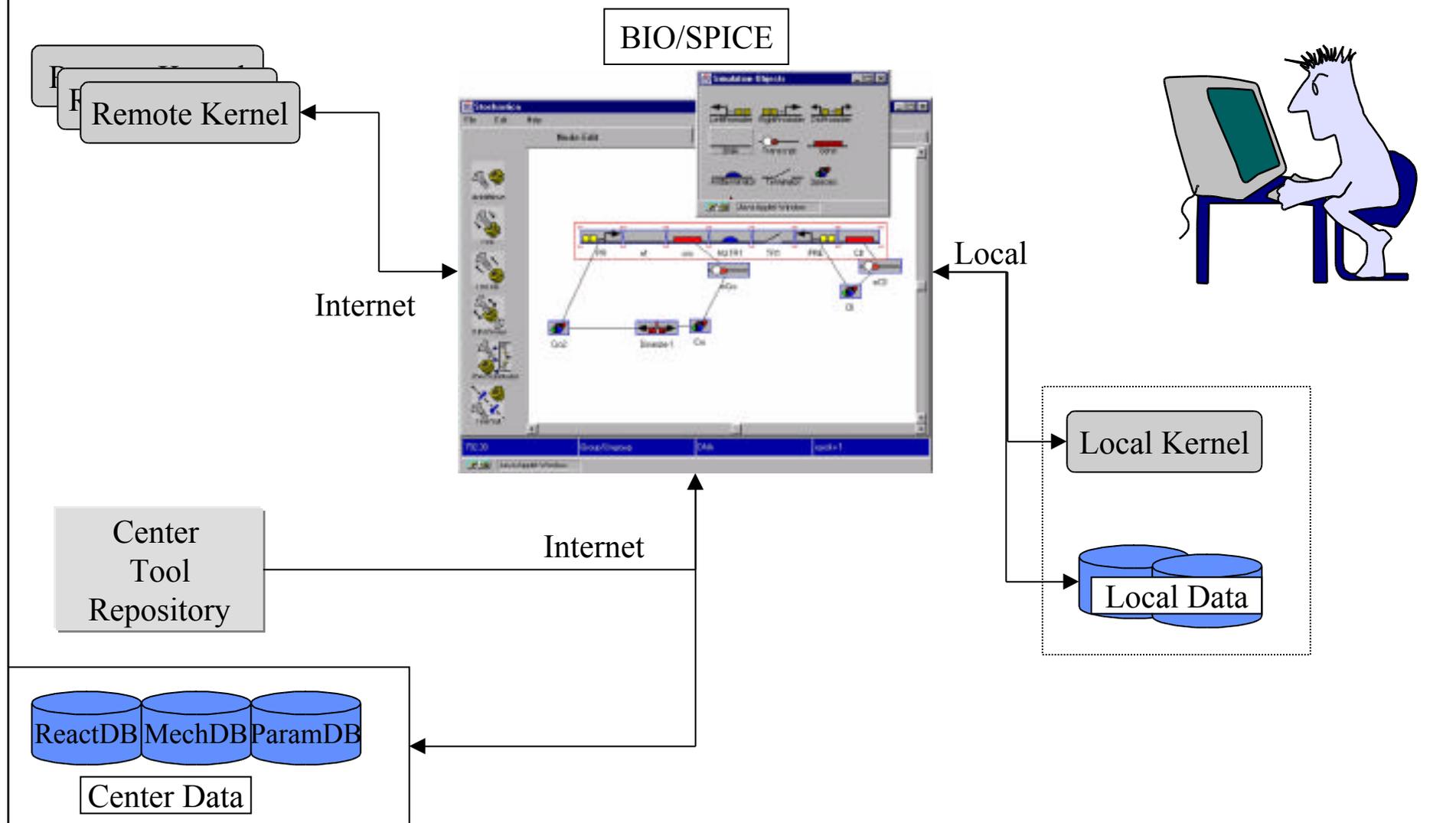
Bio/Spice: A Web-Servable, Biologist-Friendly, database, analysis and simulation interface was developed into a true beta product.

Interfaces to ReactDB, MechDB, and ParamDB.

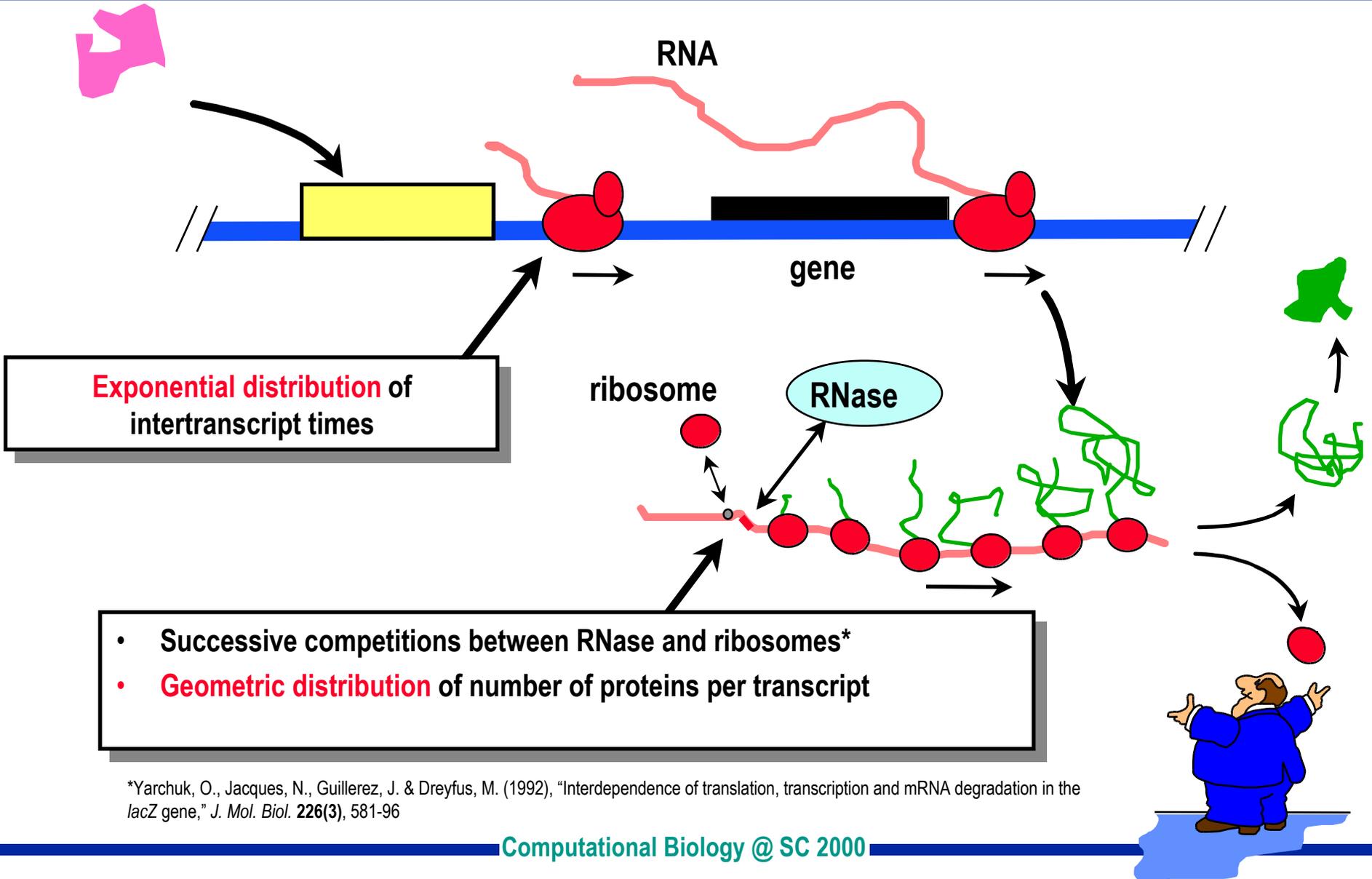
With Kernel, performs basic: flux-balance analysis, stochastic and deterministic kinetics, Scientific Visualization of results.

Notebook/Kernel design optimized for distributed computing.





Stochastic Mechanisms in Gene Expression



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 λ lysis/lysogeny decision**

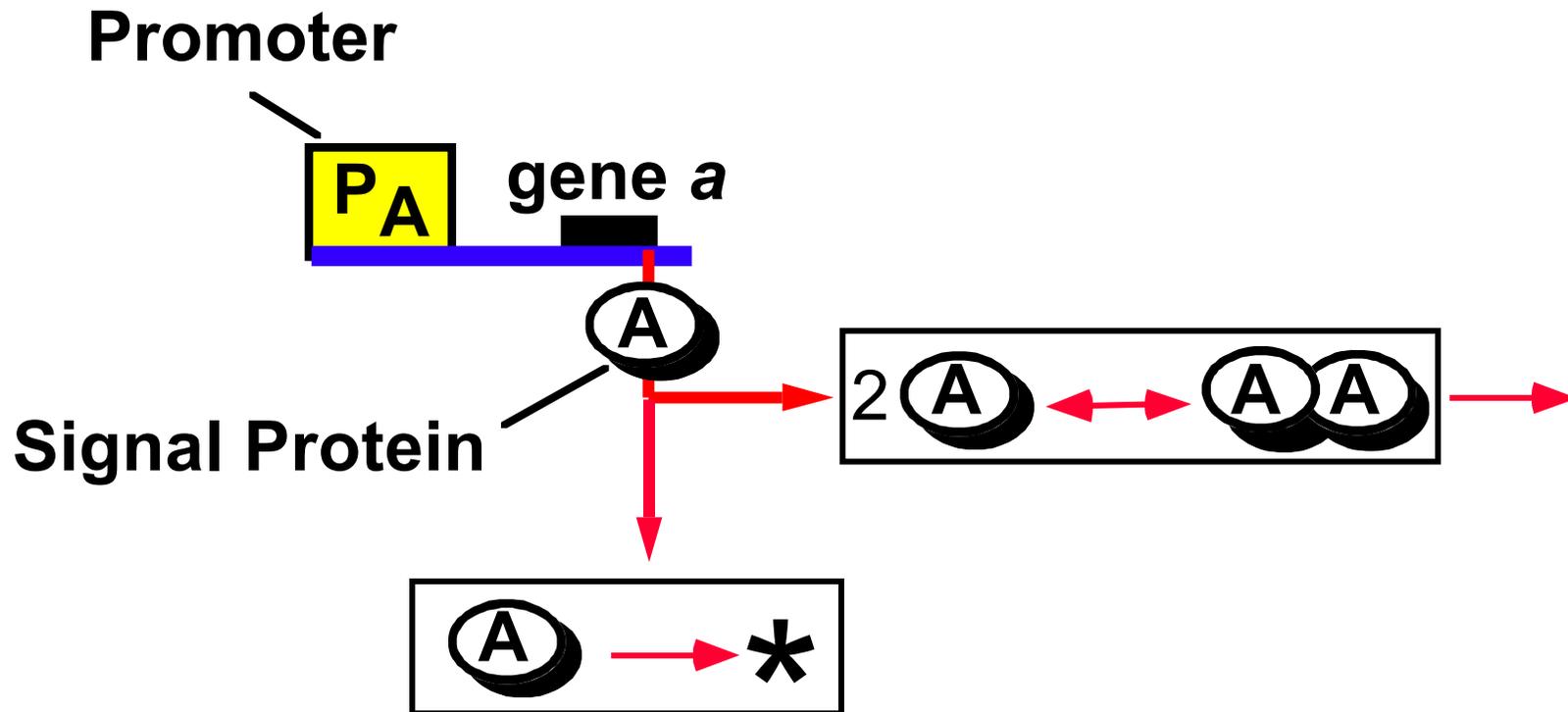
- **Random environmental influences**

- **Mutations**

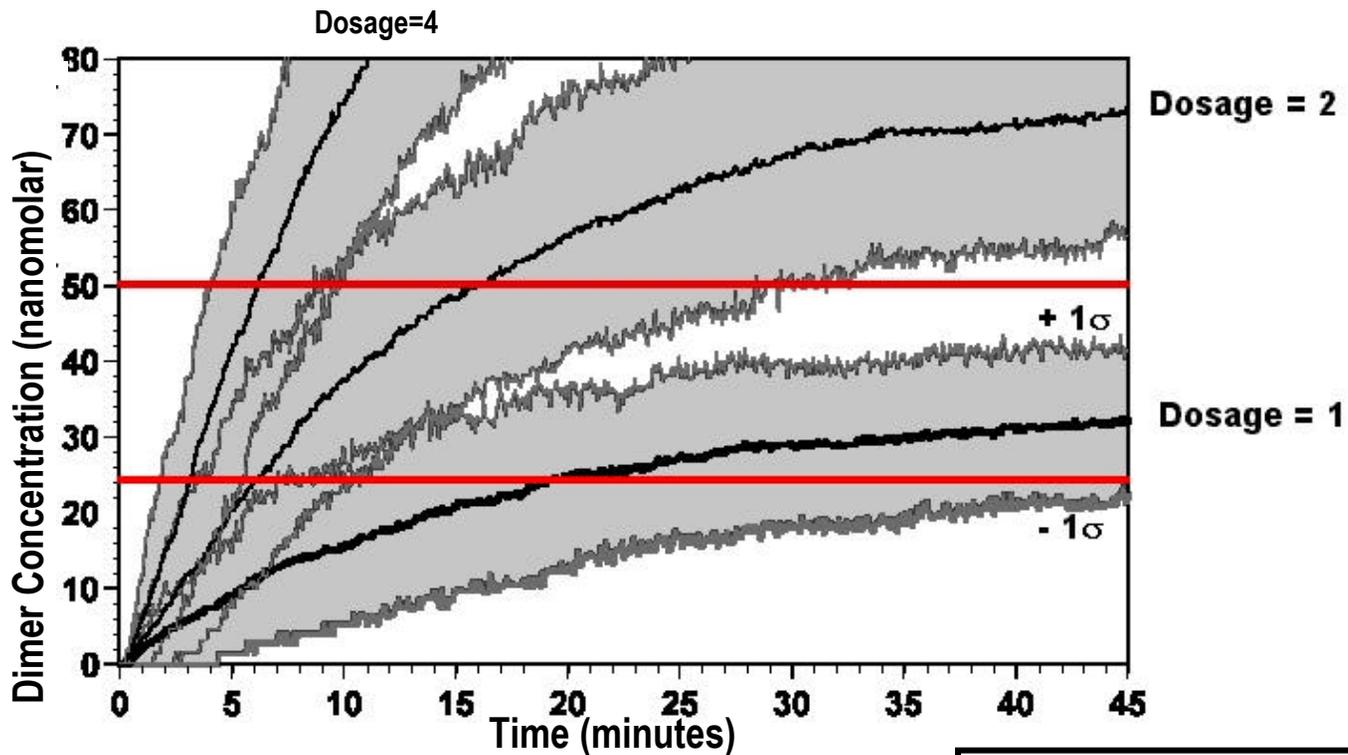
- **Asymmetric partitioning at cell division**

- **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



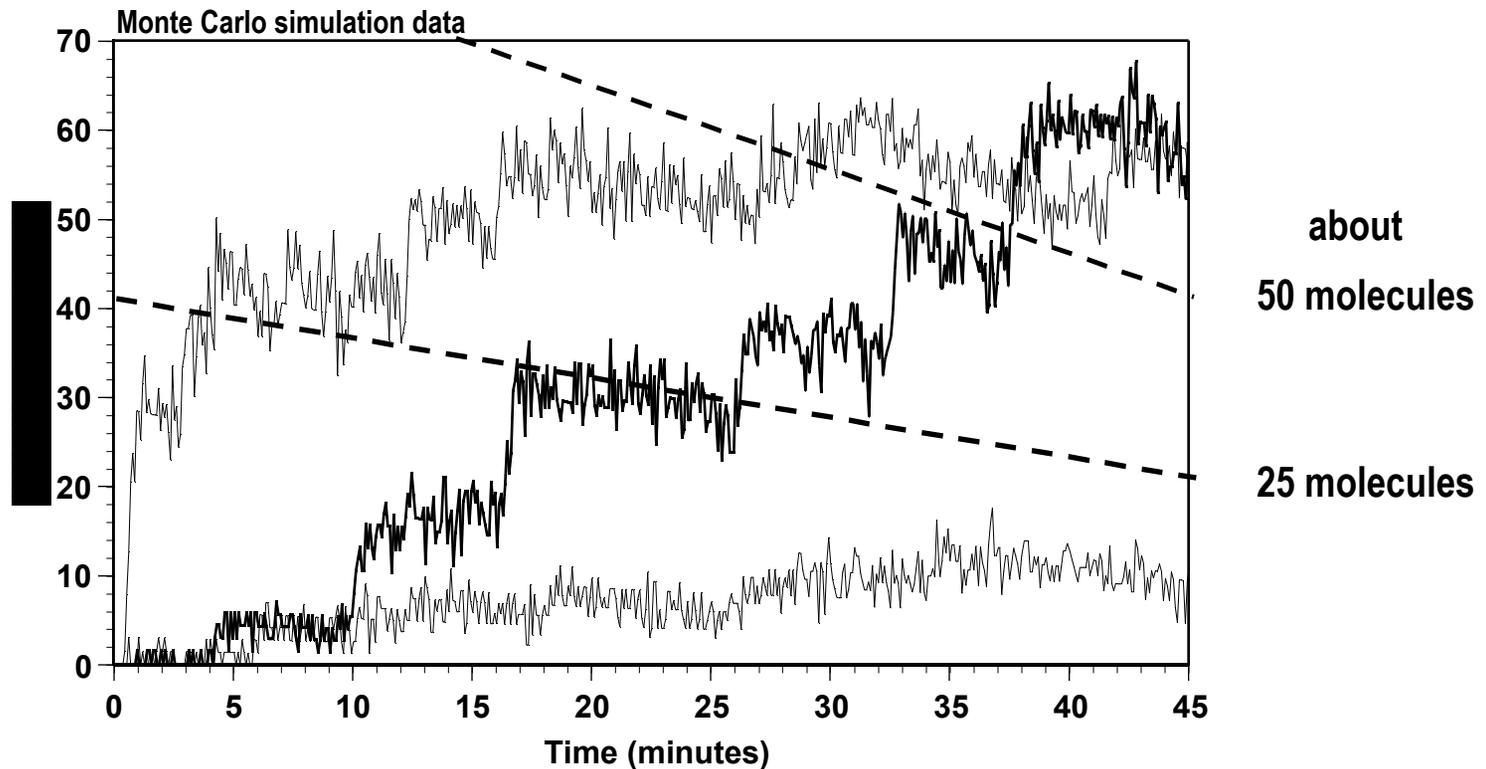
Time to Effectivity



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:

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

