CSE 527
Phylogeny & RNA: Pfold

Lectures 20-21

Autumn 2006

Phylogenies
(aka Evolutionary Trees)

"Nothing in biology makes sense, except in the light of evolution"

-- Dobzhansky

Modeling Sequence Evolution

Simple but useful models; assume:

- Independence of separate positions
- Independence of separate lineages
- Stationarity - e.g., nuc freqs aren't changing
- Markov property - nuc at a given position is independent of nuc there t_2 years ago *given* nuc there $t_1 < t_2$ years ago.

Simple Example: Jukes-Cantor

Rate matrix R =

	A	C	G	T
A	-3a	a	a	a
C	a	-3a	a	a
G	a	a	-3a	a
T	a	a	a	-3a

rate of C→T changes per unit time

diagonal s.t. row sums = 0

Consequences:

- equilibrium nuc freqs π_i all = 1/4
- all changes equally likely

Multiplicativity

Matrix $P^t[i,j]$: prob of change i → j in time t

$P^{s+t}[i,j] = \Sigma_k P^s[i,k] P^t[k,j]$

I.e.,

$P^{s+t} = P^s P^t$

Finding Change Probabilities

For small time ε, transition probabilities

$P^\varepsilon \approx I + \varepsilon R$

By multiplicativity

$P^{t+\varepsilon} = P^t P^\varepsilon \approx P^t (I + \varepsilon R)$

$(P^{t+\varepsilon} - P^t) / \varepsilon \approx P^t R$

I.e., solve system of diff eqns:

$$\frac{d}{dt} P^t = P^t R$$

Jukes-Cantor, cont.

Solving $\frac{d}{dt} P^t = P^t R$

Gives P^t =

r	s	s	s
s	r	s	s
s	s	r	s
s	s	s	r

where

r = (1+3 exp(-4at))/4

s = (1 - exp(-4at))/4

Other Models

Jukes-Cantor is simple, but inaccurate for some uses. E.g.,

- Many genomes deviate sharply from $\pi_i = 1/4$
- In fact, "transversions"
  - (purine {A,G} ↔ pyrimidine {C,T})
- less frequent than "transitions"
  - (pur ↔ pur or pyr ↔ pyr).

Various other models often used

General Reversible Model

Model is *reversible* if for all i, j

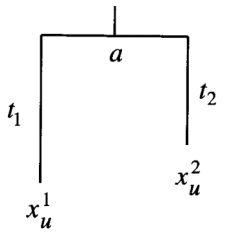
π_i P[i,j] = π_j P[j,i]

I.e., i→j and j→i changes are equally frequent; statistically, the past looks like the future

No closed form solution for $\frac{d}{dt}P^t = P^t R$ but numerically solvable using eigenvalues of rate matrix R

Evolutionary Models: Key points

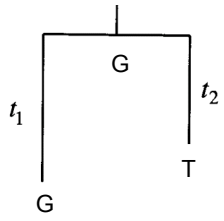
Given small number of parameters (e.g., 4 x 4 symmetric rate matrix, …), an evolutionary tree, and branch lengths, you can calculate probabilities of changes on the tree



Uses: Example 1

Probability of changes shown on this (given) tree:

$P(t_1, G \to G) * P(t_2, G \to T)$

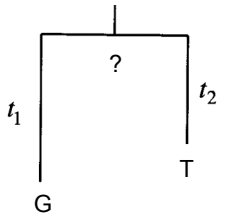


Uses: Example 2

What if ancestral state unknown?

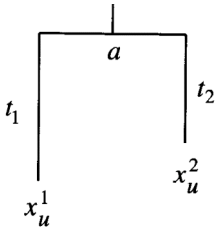
$\Sigma_a \pi_a P(t_1, a \to G) * P(t_2, a \to T)$

draw a at root from equilibrium distribution



Uses: Example 3

What if *sequences* at leaves and ancestral sequence unknown?

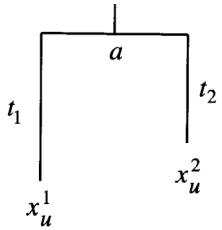


$$\prod_{u=1}^{n} \sum_{a_u} \pi_{a_u} P(t_1, a_u \to x_u^1) P(t_2, a_u \to x_u^2)$$

Uses: Example 4

What if branch lengths also unknown?

Can find MLE by numerical optimization of



$$\arg\max_{t_1,t_2} \prod_{u=1}^{n} \sum_{a_u} \pi_{a_u} P(t_1, a_u \to x_u^1) P(t_2, a_u \to x_u^2)$$

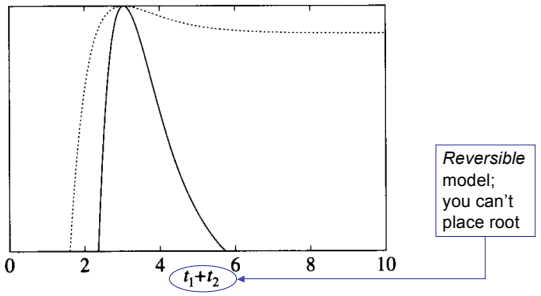
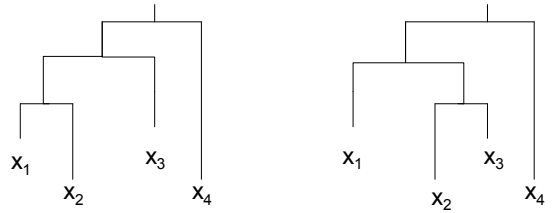


Figure 8.3 *The log likelihood* $P(x^1, x^2 | T, t_1, t_2)$ *given by (8.9), with* $n_1 = 100, n_2 = 250$, *and with* $n_1 = 1000, n_2 = 2500$. *The latter curve is sharper, as there are more data to define the maximum likelihood peak. The curves have been shifted so their peaks superimpose.*

Uses: Example 5

What if *Tree also unknown?*



Can try MLE of tree topology, too (>> parsimony)

A Complex Question:

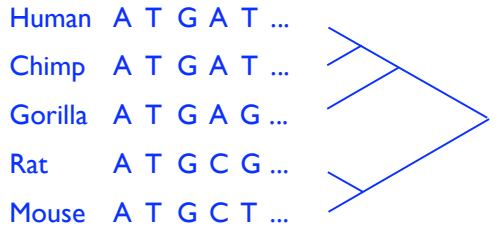
Given data (sequences, anatomy, ...) infer the phylogeny

A Simpler Question:

Given data and a phylogeny, evaluate "how much change" is needed to fit data to tree

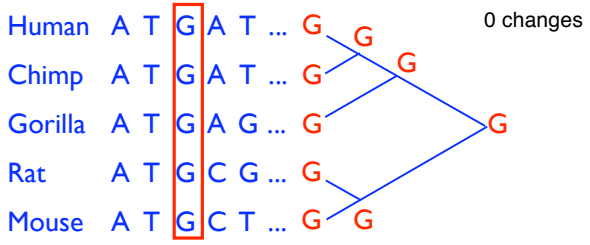
Parsimony

General idea ~ Occam's Razor: If change is rare, prefer explanations requiring few changes



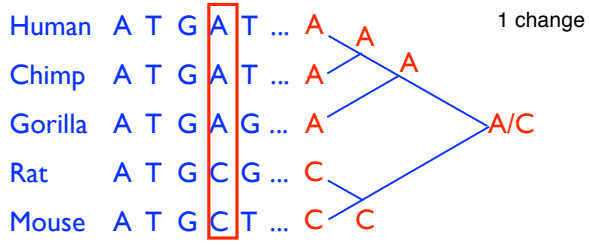
Parsimony

General idea ~ Occam's Razor: If change is rare, prefer explanations requiring few changes



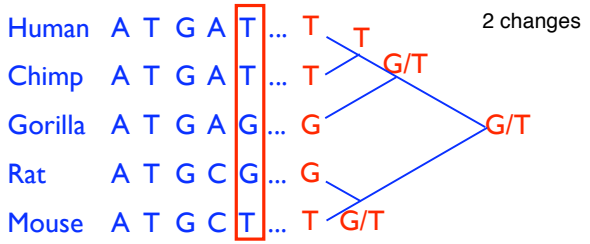
Parsimony

General idea ~ Occam's Razor: If change is rare, prefer explanations requiring few changes



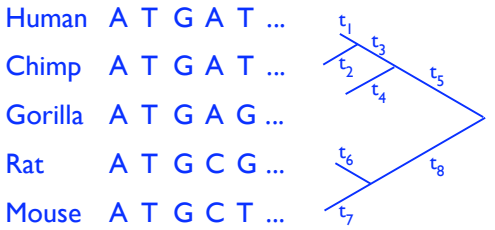
Parsimony

General idea ~ Occam's Razor: If change is rare, prefer explanations requiring few changes



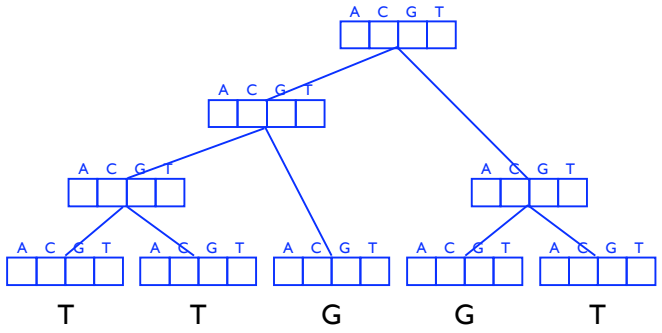
Likelihood

Given a statistical model of evolutionary change, prefer the explanation of *maximum likelihood*



Sankoff & Rousseau, '75

$P_u(s)$ = best parsimony score of subtree rooted at node u, assuming u is labeled by character s



Sankoff-Rousseau Recurrence

$P_u(s)$ = best parsimony score of subtree rooted at node u, assuming u is labeled by character s

For leaf u:

$$P_u(s) = \begin{cases} 0 & \text{if } u \text{ is a leaf labeled } s \\ \infty & \text{if } u \text{ is a leaf not labeled } s \end{cases}$$

For internal node u:

$$P_u(s) = \sum_{v \in \text{child}(u)} \min_{t \in \{A,C,G,T\}} \text{cost}(s,t) + P_v(t)$$

Time: O(alphabet^2 x tree size)

So, Parsimony easy; What about Likelihood?

Straightforward generalization of "simple" formula for 2-leaf tree

$$\prod_{u=1}^{n} \sum_{a_u} \pi_{a_u} P(t_1, a_u \rightarrow x_u^1) P(t_2, a_u \rightarrow x_u^2)$$

is infeasible, since you need to consider all (exponentially many) labelings of non-leaf nodes. Fortunately, there's a better way…

Felsenstein Recurrence

$L_u(s \mid \theta)$ = Likelihood of subtree rooted at node u, assuming u is labeled by character s, given θ

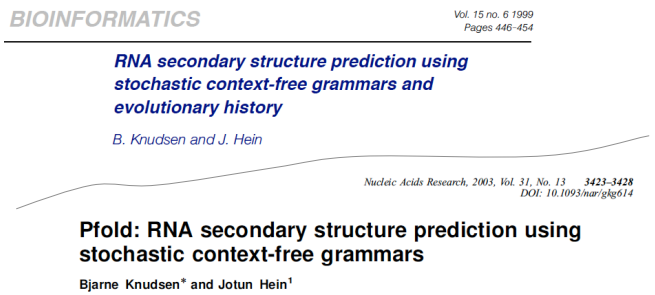
For Leaf u:

$$L_u(s \mid \theta) = \begin{cases} 1 & \text{if } u \text{ is a leaf labeled } s \\ 0 & \text{if } u \text{ is a leaf not labeled } s \end{cases}$$

For Internal node u:

$$L_u(s \mid \theta) = \prod_{v \in \text{child}(u)} \sum_{t \in \{A,C,G,T\}} P(s \rightarrow t \mid \text{length}(u,v), \theta) \cdot L_v(t \mid \theta)$$

Another Application: RNA folding



BIOINFORMATICS

Vol. 15 no. 6 1999
Pages 446–454

RNA secondary structure prediction using stochastic context-free grammars and evolutionary history

B. Knudsen and J. Hein

Nucleic Acids Research, 2003, Vol. 31, No. 13 **3423–3428**
DOI: 10.1093/nar/gkg614

Pfold: RNA secondary structure prediction using stochastic context-free grammars

Bjarne Knudsen* and Jotun Hein[1]

Using Evolution for RNA Folding

Assume you have

1. Training set of trusted RNA alignments
   - build evo model for unpaired columns
   - build evo model for paired columns
2. *Alignment* (& tree) for some RNAs presumed to have an (unknown) common structure
   - look at every col pair - better fit to paired model or 2 indp unpaired models? (Alternative to mutual information, using evo)

Training Data

Trusted alignments of 1968 tRNAs + 305 LSU rRNAs

Table 1. Base frequencies, showing nearly equal overall distribution of bases, with a slight underrepresentation of Cs. Stems have high GC/CG base pair frequencies, while loops have low content of Cs and Gs. The lowest row shows the distribution of bases between loops and stems

Stem		Loop		Overall	
AU/UA	35.6%	A	36.4%	A	26.8%
GC/CG	53.4%	C	15.1%	C	21.4%
UG/GU	9.8%	G	21.2%	G	26.7%
Other	1.2%	U	27.3%	U	25.1%
Total: 52.6%		Total:	47.4%		

Rate Matrix (Unpaired)

Table 2. The entries, r_{XY}, for the loop rate matrix. Transitions are more frequent than transversions

$X\backslash Y$	A	C	G	U
A	–0.75	0.16	0.32	0.26
C	0.40	–1.57	0.24	0.93
G	0.55	0.17	–0.96	0.24
U	0.35	0.51	0.19	–1.05

Rate Matrix (Paired)

Table 3. Some of the entries for the stem rate matrix. Only rates between the six most frequent base pairs are shown

$X\backslash Y$	AU	UA	GC	CG	UG	GU
AU	–1.16	0.18	0.50	0.12	0.02	0.27
UA	0.18	–1.16	0.12	0.50	0.27	0.02
GC	0.33	0.08	–0.82	0.13	0.02	0.23
CG	0.08	0.33	0.13	–0.82	0.23	0.02
UG	0.08	1.00	0.10	1.26	–2.56	0.04
GU	1.00	0.08	1.26	0.10	0.04	–2.56

What about Gaps?

option 1: evo model for them
- hard & slow

option 2: treat "-" as a 5th character
- they don't "evolve" quite like others

option 3: treat "-" as unknown
- ditto
- end up pairing? (drop if < 75%)
+ easy

```
Seq 1      CGAC-----AGCUGAGUGUGACUUUAGAAU
Seq 2      UGACGGUCUAGCUGACUGAUACUUCAGAGU
Seq 3      CGAC-----AGCUGAAUGAGACUUCAGAAU
Structure  ...(((((.....)))))............

Seq 1      CGAC-----AGCUGAGUGUGACUUUAGAAU
Seq 2      UGACGGUCUAGCUGACUGAUACUUCAGAGU
Seq 3      CGAC-----AGCUGAAUGAGACUUCAGAAU
Structure  ..........((((........))))...
```

Which Tree?

KH-99 : try to find MLE tree (using SCFG et al.)

good but slow

KH-03 : est tree without structure

average unpaired & (marginalized) paired rates

calc pairwise distances between seqs

tree topology from "neighbor joining"

MLE tree branch lengths

Synopsis of last lecture

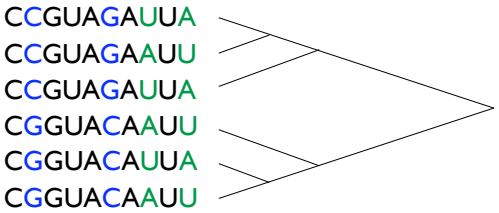
Based on simplifying assumptions (stationarity, independence, Markov, reversible), there are simple sequence-evolution models with a modest number of parameters giving, e.g., Pr(G→T | 1.5 m yr), …

It can model base-pairing in RNA, too

Felsenstein allows ML estimation of probabilities, branch lengths, even trees,… in this model. (Somewhat like "parsimony" algorithm, but better.)

Goal: Use all this for inference of RNA 2ary struct

Phylogeny vs Mutual Information



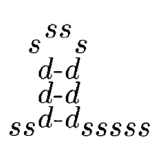
MI = 1 bit in both cases, but green pair is more compelling evidence of interaction: 3 events, not 1

The Glue: An SCFG

$$S \to LS \mid L$$
$$L \to s \mid dFd$$
$$F \to LS \mid dFd$$

a) $S \to LS \to LLLLLLLS \to LLLLLLLL$
$\to ssLssssS \to ssdFdssssS$
$\to ssdddFddddssssS$
$\to ssdddLSddddssssS$
$\to ssdddLLLLddddssssS$
$\to ssddddssssddddssssS$

b)



c) $F \to dFd \to ddFdd \to ddLSdd$
$\to ddLLdd \to ddLsdd \to dddFdsdd$

Full SCFG

$S \rightarrow LS$ (0.868534) | L (0.131466)

$L \rightarrow s$ (0.894603*p(s)) | dFd (0.105397*p(dd))

$F \rightarrow LS$ (0.212360) | dFd (0.787640*p(dd))

Where p(s) & p(dd) are the probabilities of the single/paired alignment columns s/dd as calculated by the Felsenstein algorithm based on the fixed evolutionary model and the given tree topology and branch lengths.

What SCFG Gives

"Prior" probabilities for

- fraction paired vs unpaired
- lengths of each
- frequency of bulges in stems
- etc., and

Inherits column probabilities from evo model

Cocke-Kasami-Younger for CFG

Suppose all rules of form A → BC or A → a
(by mechanical grammar transform, or use orig grammar & mechanically transform alg below…)

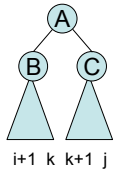
Given $x = x_1 \ldots x_n$, want $M_{i,j} = \{ A \mid A \rightarrow x_{i+1} \ldots x_j \}$

For j=2 to n

M[j-1,j] = {A | A → x_j is a rule}

for i = j-1 down to 1

M[i,j] = $\cup_{i<k<j}$ M[i,k] ⊗ M[k,j]



Where X ⊗ Y = {A | A → BC , B ∈ X, and C ∈ Y }

The "Inside" Algorithm for SCFG
(analogous to HMM "forward" alg)

Just like CKY, but instead of just recording possibility of A in M[i,j], record its *probability*:

For each A, do *sum* instead of *union,* over all possible k and all possible A → BC rules, of *products* of their respective probabilities.

Result: for each i, j, A, have Pr(A → $x_{i+1} \ldots x_j$)

(There's also an "outside" alg, analogous to backward…)

The "Viterbi" algorithm for SCFGs

Just like inside, but use max instead of sum.

So what's the structure?

The usual dynamic programming traceback:

Starting from S in upper right corner of matrix, find which k and which S → BC gave max probability, then (recursively) find where that B and that C came from…

(Really, you want to do it with the F → dFd grammar, and where those rules are used tells you where the base pairs are.)

Results & Validation

KH-99: 4 bacterial RNAse P, 340-380 nt

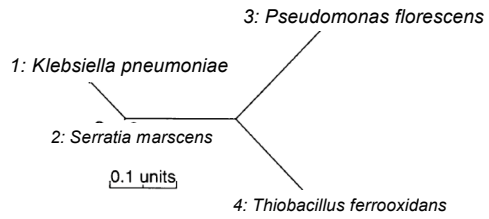
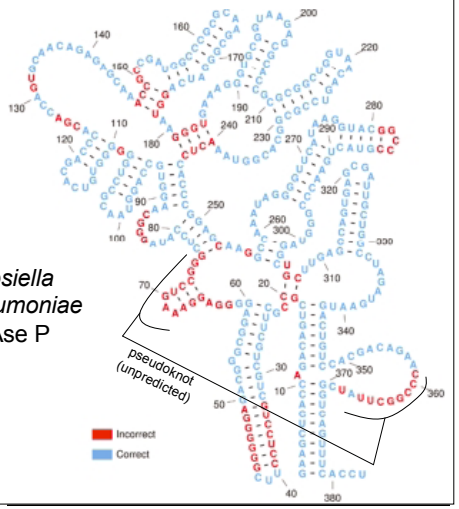


Fig. 2. The phylogenetic tree relating the four analysed sequences, as calculated using the ML estimation described above. The length units correspond to the rate matrices of the model.

Good overall structure prediction

Klebsiella pneumoniae RNAse P



Good Overall Structure Prediction

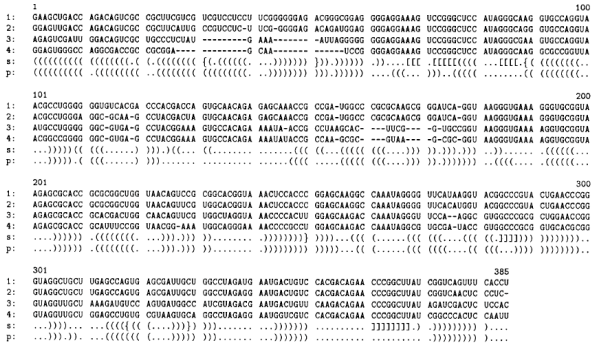


Fig. 3. The alignment of the four RNase P RNA sequences. The predicted structure, using all four sequences, is denoted p. The structure fro the database is denoted s, with square brackets denoting parts of pseudoknots. The square brackets used here match the structure descriptic in the database. The curly brackets denote positions where the structure differs: the sequences that have a non-standard pair in these positio have loop regions or bulges, the rest have pairs.

Not bad, even with only one seq

Table 7. Accuracy table, showing comparisons of single sequence predictions using the method described in this paper and MFOLD Version 3.0, by Zuker (1989) and Walter *et al.* (1994). Predictions of secondary structures were made on single sequences, which is the only possibility using MFOLD. The average results are comparable

Sequence	SCFG method	MFOLD
Seq 1	57.7%	67.1%
Seq 2	48.2%	54.0%
Seq 3	41.2%	35.6%
Seq 4	46.2%	50.3%
Average	48.3%	51.7%

More sequences help

So do phylogeny and a good alignment

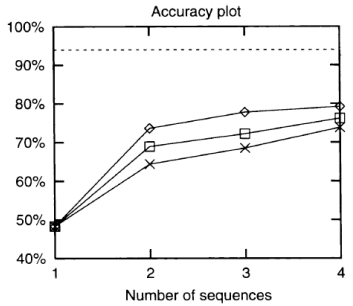


Fig. 4. A comparison of results with and without phylogeny. Diamonds (◇) denote the curve for predictions with phylogeny, while boxes (□) denote the one without. Crosses (×) denote results using CLUSTAL W alignments and phylogeny estimation. The dotted line at 94% represents the maximum possible prediction accuracy with regard to the pseudoknots.

	Structural alignment			
No. of sequences	1	2	3	4
Min result	41.2%	65.2%	73.9%	79.2%
Max result	57.7%	82.1%	79.6%	79.2%
Average	48.3%	73.6%	77.8%	79.2%
	CLUSTAL W alignment			
No. of sequences	1	2	3	4
Min result	41.2%	54.9%	60.1%	73.8%
Max result	57.7%	69.1%	76.9%	73.8%
Average	48.3%	64.4%	68.5%	73.8%
	Structural alignment, no phylogeny			
No. of sequences	1	2	3	4
Min result	41.2%	59.9%	67.7%	76.2%
Max result	57.7%	76.6%	76.6%	76.2%
Average	48.3%	68.9%	72.2%	76.2%

Results & Validation
KH-03

Test Set	Sequences
A: 9 tmRNAs (363.8)	act.act., hae.inf., kle.pne., pas.mul., sal.par., sal.typ., she.put., vib.cho., yer.pes.
B: 13 bacterial SRP RNAs (270.5)	bac.alc., bac.bre., bac.cer., bac.cir., bac.mac., bac.meg., bac.pol., bac.pum., bac.sph., bac.ste., bac.thu., bre.bre., clo.per.
C: 10 eukaryotic SRP RNAs (300.9)	ory.sat., tri.ae-a, tri.ae-b, zea.ma-a, zea.ma-b, zea.ma-c, zea.ma-d, zea.ma-e, zea.ma-f, zea.ma-h
D: 51 eukaryotic SRP RNAs (297.4)	ara.th-a, ara.th-b, cae.el-a, cae.el-b, cae.el-c, cae.el-d, can.spe., cin.hyb., dro.mel., fug.rub., hom.sa-a, hom.sa-b, hom.sa-c, hum.ja-a, hum.ja-b, hum.lu-a, hum.lu-b, hum.lu-c, hum.lu-d, lep.col., lyc.es-a, lyc.es-b, lyc.es-c, lyc.es-e, lyc.es-f, lyc.es-g, lyc.es-h, lyc.es-i, lyc.es-j, lyc.es-k, lyc.es-m, lyc.es-n, lyc.es-o, ory.sat., rat.rat., sch.pom., tet.ros., tet.the., tri.ae-a, tri.ae-b, try.br-a, try.br-b, xen.lae., yar.li-a, yar.li-b, zea.ma-a, zea.ma-b, zea.ma-c, zea.ma-d, zea.ma-e, zea.ma-f

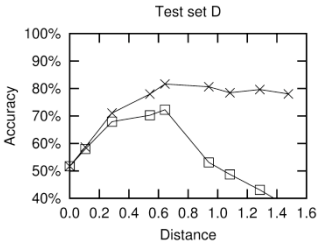


Figure 7. Accuracy as a function of pairwise distance between two sequences being analysed. As in Figure 6, crosses are from results using 'correct' alignments, while boxes are from ClustalW alignments. The pairs were grouped according to their Jukes–Cantor distances, in the intervals [0;0.2), [0.2;0.4), [0.4;0.6) etc. The points represent average results for 50 random sequence combinations from a specific range of distances. The *x*-value of a point is the average of the 50 distances.

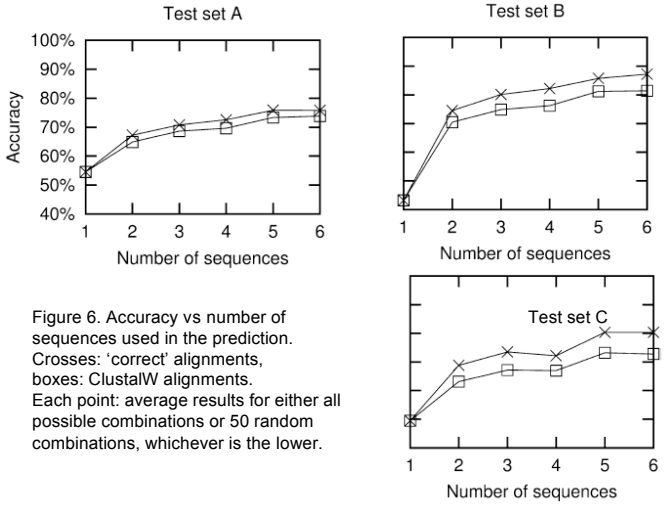


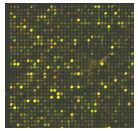
Figure 6. Accuracy vs number of sequences used in the prediction. Crosses: 'correct' alignments, boxes: ClustalW alignments. Each point: average results for either all possible combinations or 50 random combinations, whichever is the lower.

Course Wrap Up

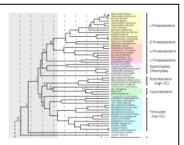
Course Project Presentations

Wednesday, 12/13, 1:00-2:30
CSE 674

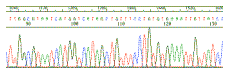
Everyone's invited



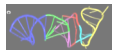
"High-Throughput BioTech"

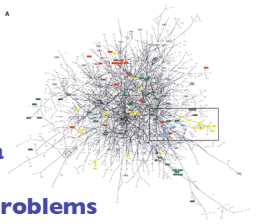


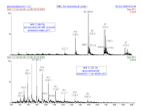
Sensors
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction



Controls
- Cloning
- Gene knock out/knock in
- RNAi







***Floods* of data**

"Grand Challenge" problems

CS/Math/Stats Points of Contact

Scientific visualization
- Gene expression patterns

Databases
- Integration of disparate, overlapping data sources
- Distributed genome annotation in face of shifting underlying coordinates

AI/NLP/Text Mining
- Information extraction from journal texts with inconsistent nomenclature, indirect interactions, incomplete/inaccurate models,…

Machine learning
- System level synthesis of cell behavior from low-level heterogeneous data (DNA sequence, gene expression, protein interaction, mass spec,

Algorithms

…

Frontiers & Opportunities

New data:
- Proteomics, SNP, arrays CGH, comparative sequence information, methylation, chromatin structure, ncRNA, interactome

New methods:
- graphical models? rigorous filtering?

Data integration
- many, complex, noisy sources

Systems Biology

Frontiers & Opportunities

Open Problems:

- splicing, alternative splicing
- multiple sequence alignment (genome scale, w/ RNA etc.)
- protein & RNA structure
- interaction modeling
- network models
- RNA trafficing
- ncRNA discovery
- …

Exciting Times

Lots to do

Various skills needed

I hope I've given you a taste of it

Thanks!