Multiple Sequence Alignment

Pradipta Ray

UT Dallas

BIOL6385 / BMEN6389

" One or two homologous sequences whisper . . . a full multiple alignment shouts out loud "

- Arthur M. Lesk

Penn State University

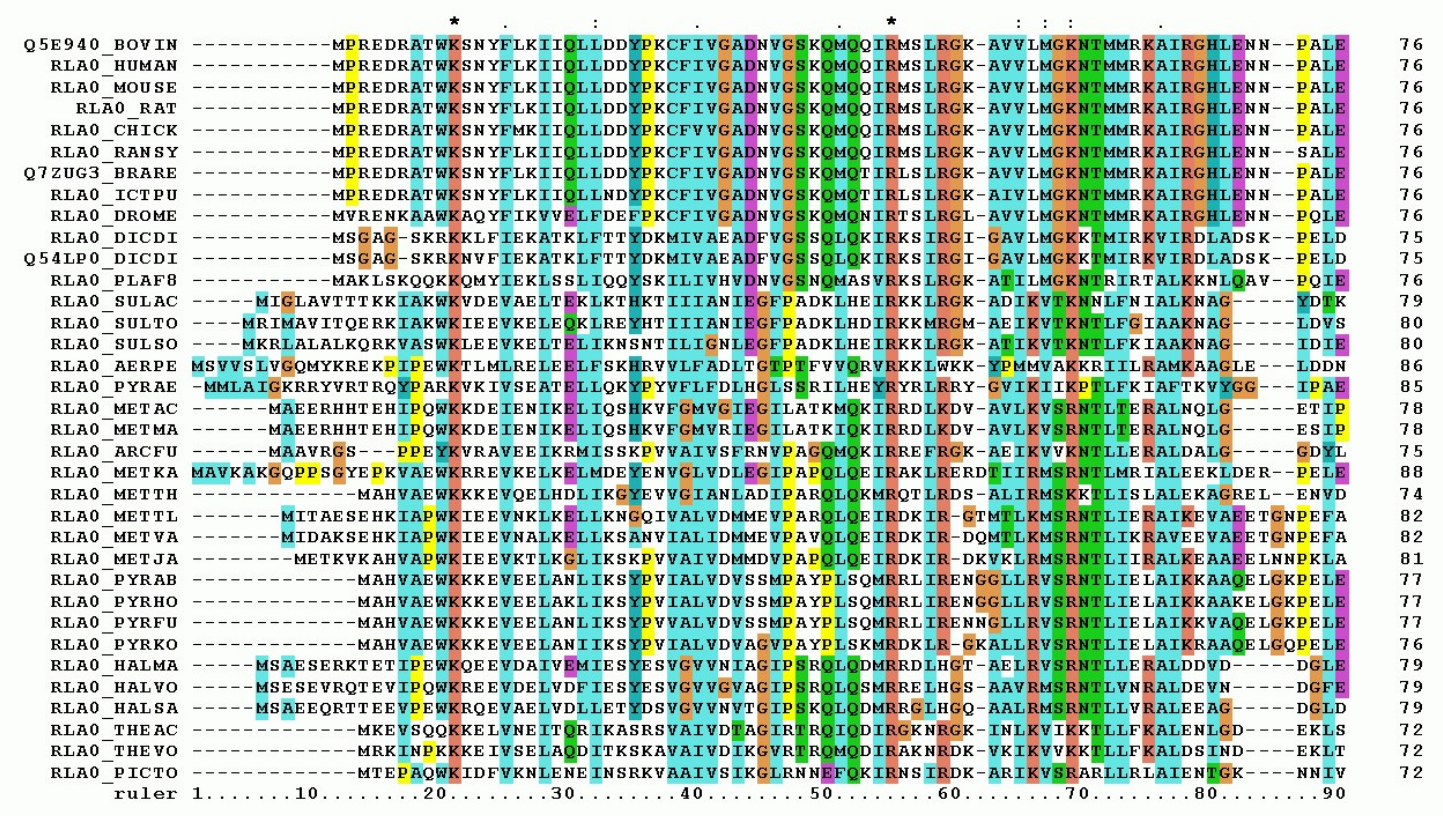
www.psu.edu

Outline

1. **What is Multiple Sequence Alignment (MSA) ?**
   - Early Chronology
   - Utilities
2. Challenges in MSA
3. Making MSA work
4. Limits of MSA
5. The future of alignment

Multiple sequence alignment : what

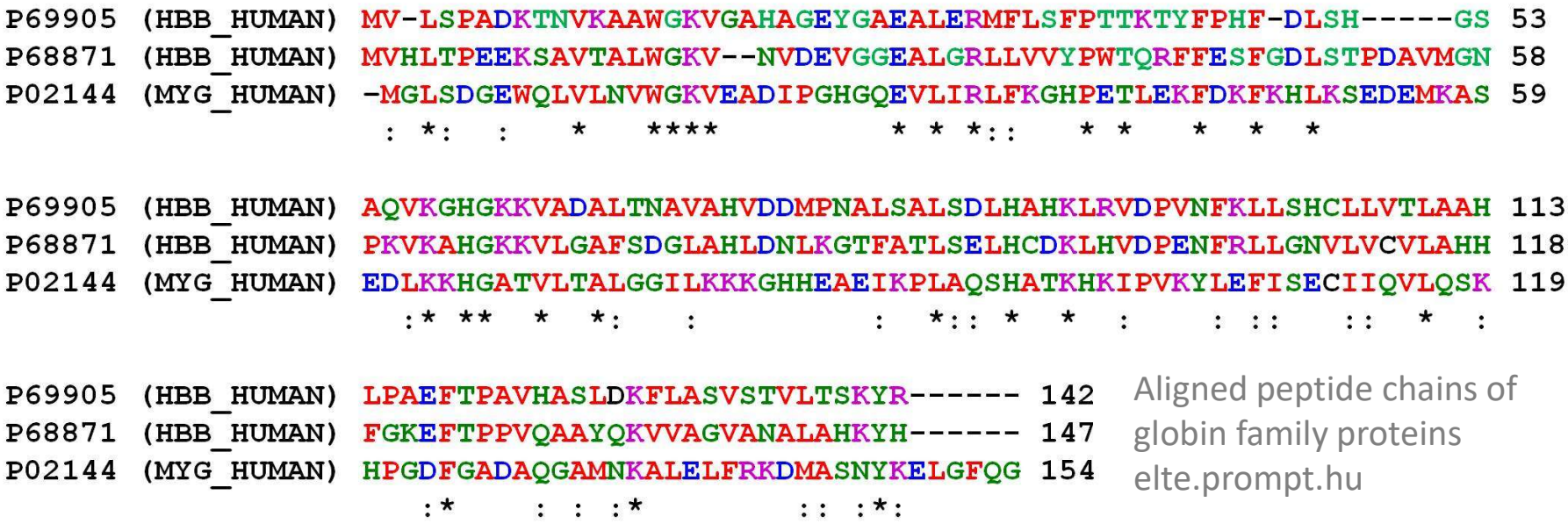
- Alignment : way of arranging sequences to identify “commonality” (homology ?)
- Proteins, RNA, DNA : a way to contrast regions



wikipedia

Homologous residues : meaning of MSA

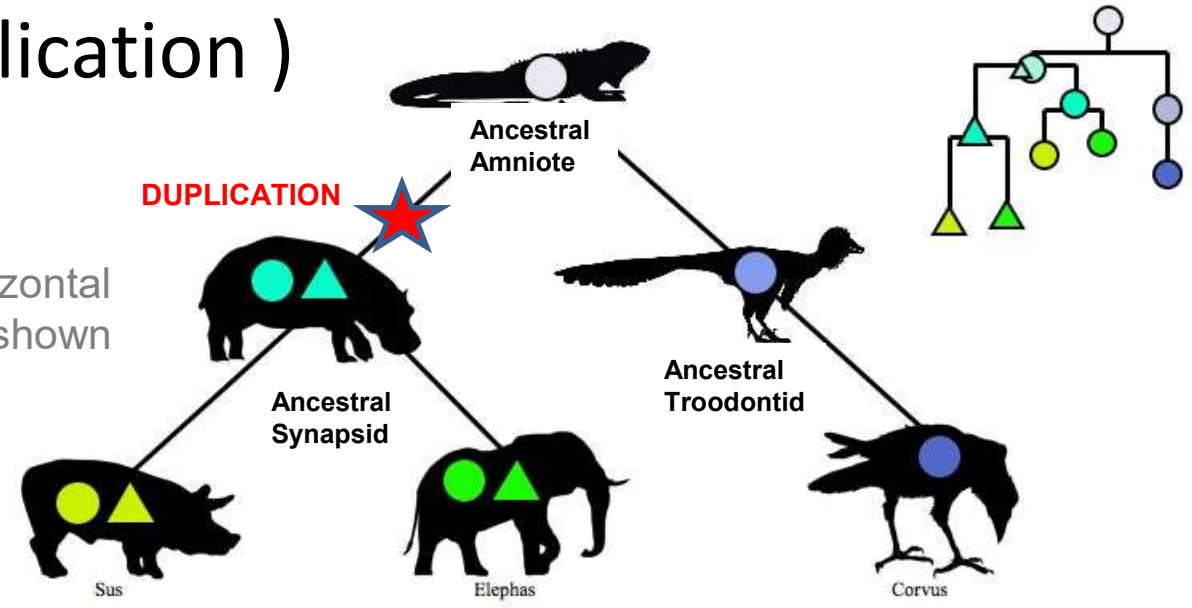
- Aligned residues : those present in a single column , typically assumed to be diverging from common ancestral residue
  - Could share common function and / or structure as a consequence of sharing a common evolutionary ancestor / having similar sequence pattern.



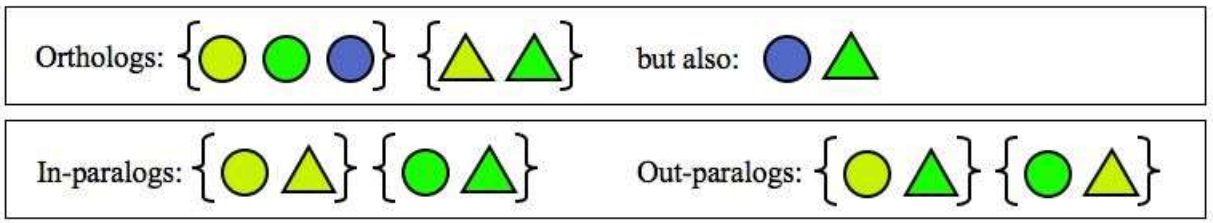
Aligned peptide chains of globin family proteins
elte.prompt.hu

Homologous : common ancestor by descent or duplication ?

- Homologous : orthologous (vertical : descent / speciation) or paralogous (horizontal : duplication)



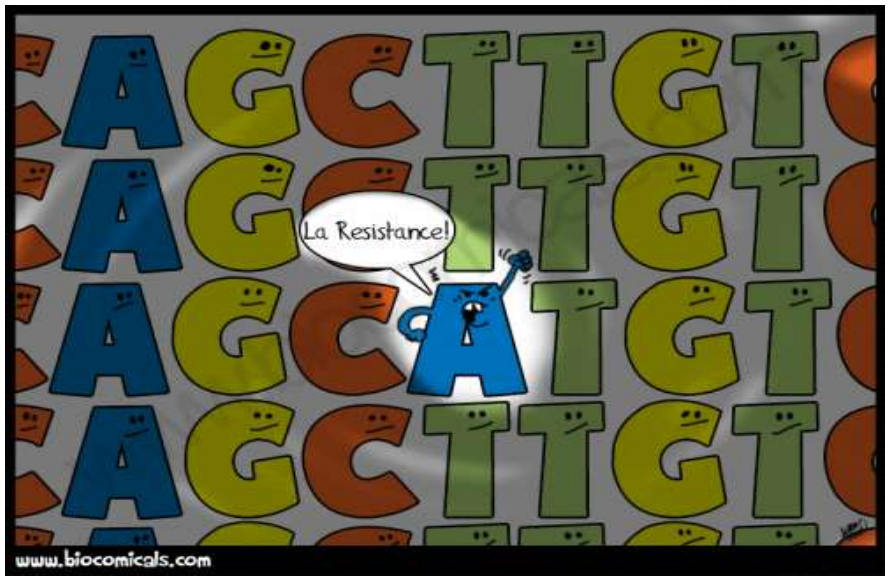
Xenology: horizontal transfer – not shown in diagram



Inferring a common ancestor

- By means of explicit evolutionary models : requires modelling how amino acids / nucleic acids evolve over time
  - focus on the nature and rate of **changes** (next class)
- By means of identifying potentially homologous sequences : identifying / aligning similar subsequences (may or may not use explicit evolutionary models)
  - focus on the location of **conserved regions**
- Approaches are **interdependent**

  - which approach to use depends on what you want to shine a light on



Outline

1. What is MSA ?
   - **Early Chronology**
   - Utilities
2. Challenges in MSA
3. Making MSA work
4. Limits of MSA
5. The future of alignment

Comparative -omics

- **Comparison of multiple sequences** to arrive at conclusions about ancestry, function, structure
- Groundbreaking Linus Pauling paper in 1963

ACTA CHEMICA SCANDINAVICA 17 (1963) S9–S16

Chemical Paleogenetics

Molecular "Restoration Studies" of Extinct Forms of Life

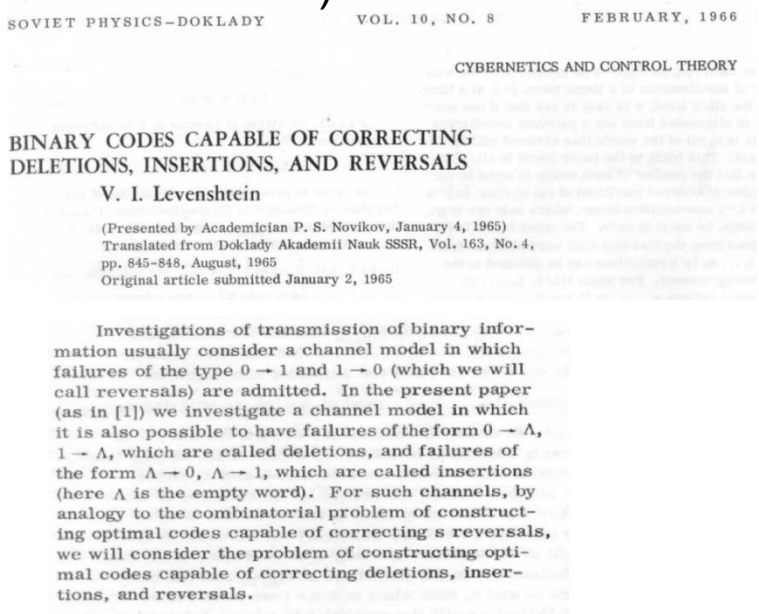
LINUS PAULING and EMILE ZUCKERKANDL*

*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA***

An unusual history

PAIRWISE ALIGNMENT THEORY

First started based on Levenshtein's paper in 1966 of detecting indels and "reversals" in a binary sequence - **communication theory** : nucleotide sequence modelling followed (Needleman – Wunsch 1970 , Smith – Waterman 1981)



SOVIET PHYSICS–DOKLADY VOL. 10, NO. 8 FEBRUARY, 1966

CYBERNETICS AND CONTROL THEORY

BINARY CODES CAPABLE OF CORRECTING DELETIONS, INSERTIONS, AND REVERSALS

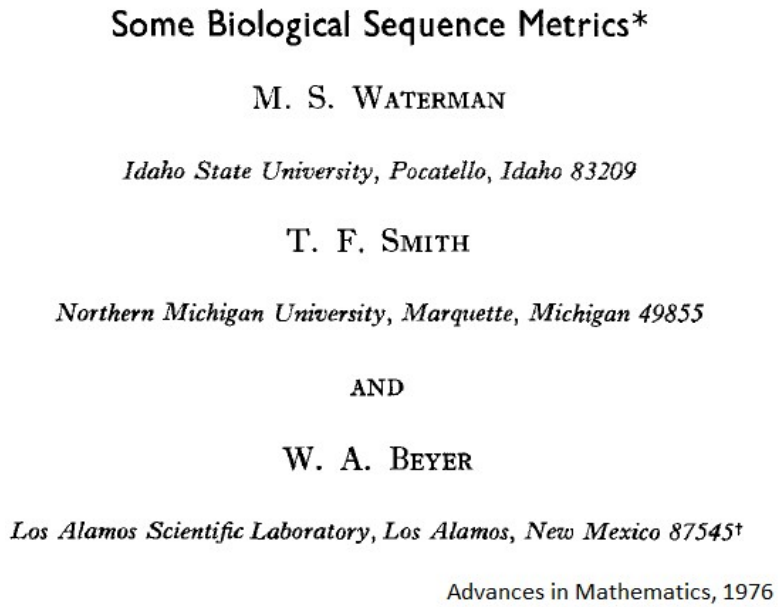
V. I. Levenshtein

(Presented by Academician P. S. Novikov, January 4, 1965)
Translated from Doklady Akademii Nauk SSSR, Vol. 163, No. 4, pp. 845–848, August, 1965
Original article submitted January 2, 1965

Investigations of transmission of binary information usually consider a channel model in which failures of the type $0 \rightarrow 1$ and $1 \rightarrow 0$ (which we will call reversals) are admitted. In the present paper (as in [1]) we investigate a channel model in which it is also possible to have failures of the form $0 \rightarrow \Lambda$, $1 \rightarrow \Lambda$, which are called deletions, and failures of the form $\Lambda \rightarrow 0$, $\Lambda \rightarrow 1$, which are called insertions (here Λ is the empty word). For such channels, by analogy to the combinatorial problem of constructing optimal codes capable of correcting s reversals, we will consider the problem of constructing optimal codes capable of correcting deletions, insertions, and reversals.

MULTIPLE ALIGNMENT THEORY

First started based on metrics to measure distance between multiple biological sequences in 1976. **Preceded** by early 3-sequence alignment studies (Dickerson 1971, Bewley, Dickson & Li 1972)



Some Biological Sequence Metrics*

M. S. Waterman

Idaho State University, Pocatello, Idaho 83209

T. F. Smith

Northern Michigan University, Marquette, Michigan 49855

AND

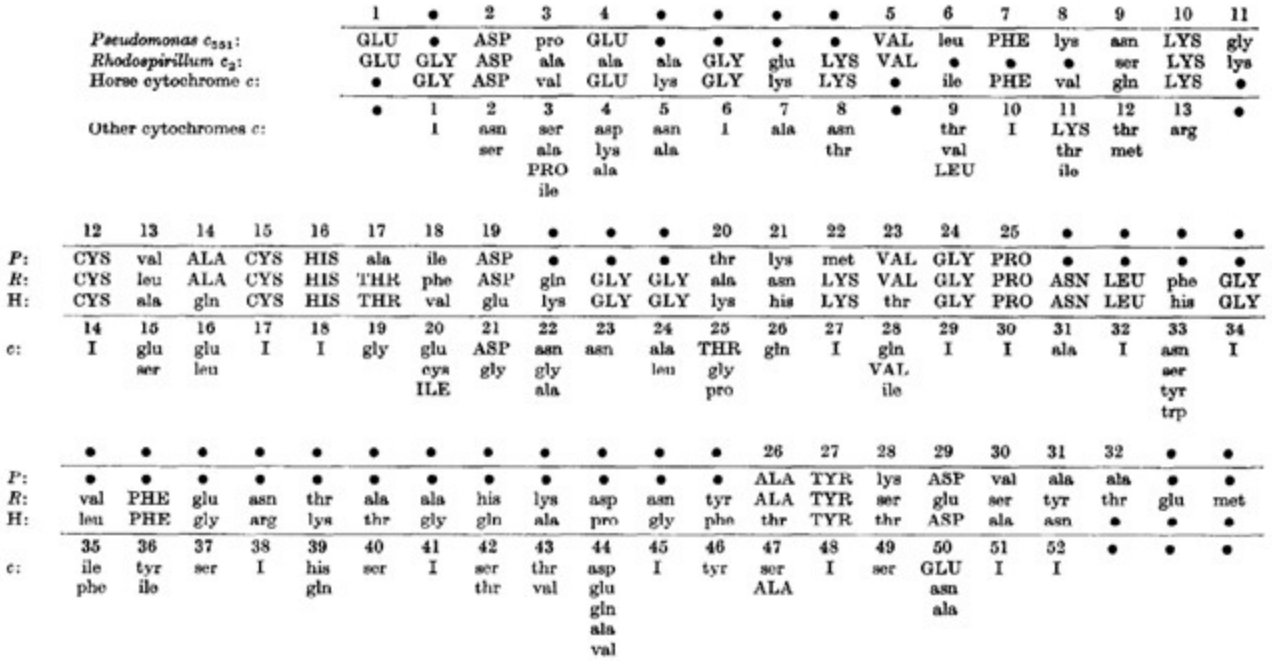
W. A. Beyer

Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545†

Advances in Mathematics, 1976

Early multiple sequence alignment

- Cue : pairwise codon alignment, theoretical framework for multiple sequence not developed yet
  - only works for partially diverged protein / coding DNA sequence



	1	•	2	3	4	•	•	•	•	5	6	7	8	9	10	11
Pseudomonas c_{551}:	GLU	•	ASP	pro	GLU	•	•	•	•	VAL	leu	PHE	lys	asn	LYS	gly
Rhodospirillum c_2:	GLU	GLY	ASP	ala	ala	ala	GLY	glu	LYS	VAL	•	•	•	ser	LYS	lys
Horse cytochrome *c*:	•	GLY	ASP	val	GLU	lys	GLY	lys	LYS	•	ile	PHE	val	gln	LYS	•

	•	1	2	3	4	5	6	7	8	•	9	10	11	12	13	•
Other cytochromes *c*:		1	asn	ser	asp	asn	1	ala	asn		thr	I	LYS	thr	arg	
			ser	ala	lys	ala			thr		val		thr	met		
				PRO	ala						LEU		ile			
				ile												

	12	13	14	15	16	17	18	19	•	•	•	20	21	22	23	24	25	•	•	•	•
P:	CYS	val	ALA	CYS	HIS	ala	ile	ASP	•	•	•	thr	lys	met	VAL	GLY	PRO	•	•	•	•
R:	CYS	leu	ALA	CYS	HIS	THR	phe	ASP	gln	GLY	GLY	ala	asn	LYS	VAL	GLY	PRO	ASN	LEU	phe	GLY
H:	CYS	ala	gln	CYS	HIS	THR	val	glu	lys	GLY	GLY	lys	his	LYS	thr	GLY	PRO	ASN	LEU	his	GLY
	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
c:	I	glu	glu	I	I	gly	glu	ASP	asn	asn	ala	THR	gln	I	gln	I	I	ala	I	asn	I
		ser	leu				cys	gly	gly		leu	gly			VAL					ser	
							ILE		ala			pro			ile					tyr	
																				trp	

	•	•	•	•	•	•	•	•	•	•	•	•	26	27	28	29	30	31	32	•	•
P:	•	•	•	•	•	•	•	•	•	•	•	•	ALA	TYR	lys	ASP	val	ala	ala	•	•
R:	val	PHE	glu	asn	thr	ala	ala	his	lys	asp	asn	tyr	ALA	TYR	ser	glu	ser	tyr	thr	glu	met
H:	leu	PHE	gly	arg	lys	thr	gly	gln	ala	pro	gly	phe	thr	TYR	thr	ASP	ala	asn	•	•	•
	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	•	•	•
c:	ile	tyr	ser	I	his	ser	I	ser	thr	asp	I	tyr	ser	I	ser	GLU	I	I			
	phe	ile			gln			thr	val	glu			ALA			asn					
										gln						ala					
										ala											
										val											

(Dickerson, J Mol Bio 1971)

Early challenge: homology or chance ?

- Billions of nucleic / amino acids : only 4 / 21 states (Doolittle, Science 1981)

Similar Amino Acid Sequences: Chance or Common Ancestry?

Russell F. Doolittle

Summary. The systematic comparison of every newly determined amino acid sequence with all other known sequences may allow a complete reconstruction of the evolutionary events leading to contemporary proteins. But sometimes the surviving similarities are so vague that even computer-based sequence comparison procedures are unable to validate relationships. In other cases similar sequences may appear in totally alien proteins as a result of mere chance or, occasionally, by the convergent evolution of sequences with special properties.

Science, 1981

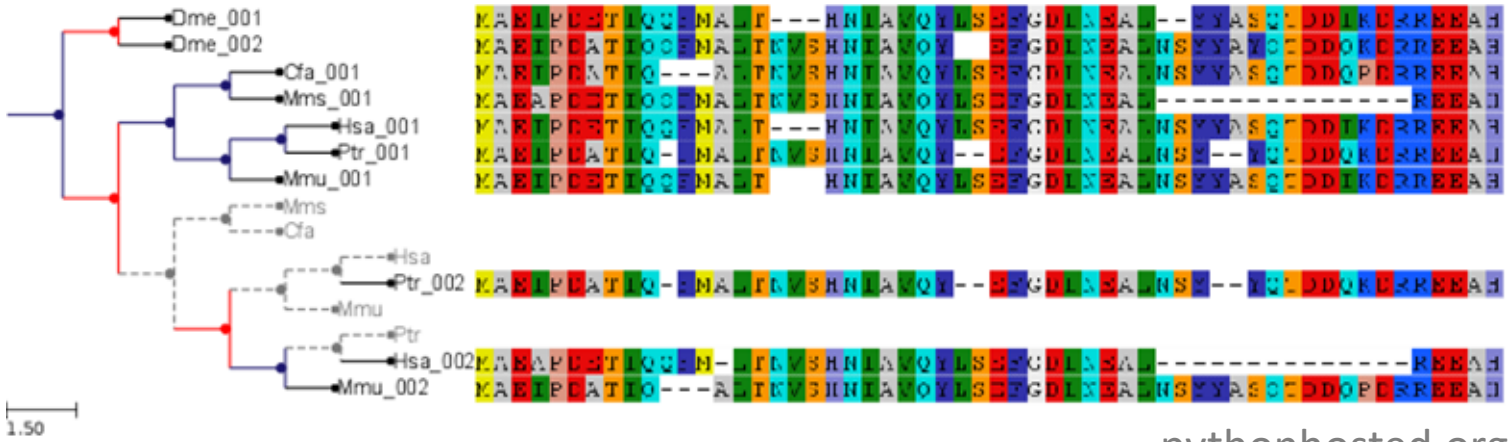
- **Context and contiguity**
  - identical residues next to a pair of homologous residues have high probability of being homologous
  - a stretch of identical residues have a greater probability of being homologous

Outline

1. What is MSA ?
   - Early Chronology
   - **Utilities**
2. Challenges in MSA
3. Making MSA work
4. Limits of MSA
5. The future of alignment

Why perform alignment ?

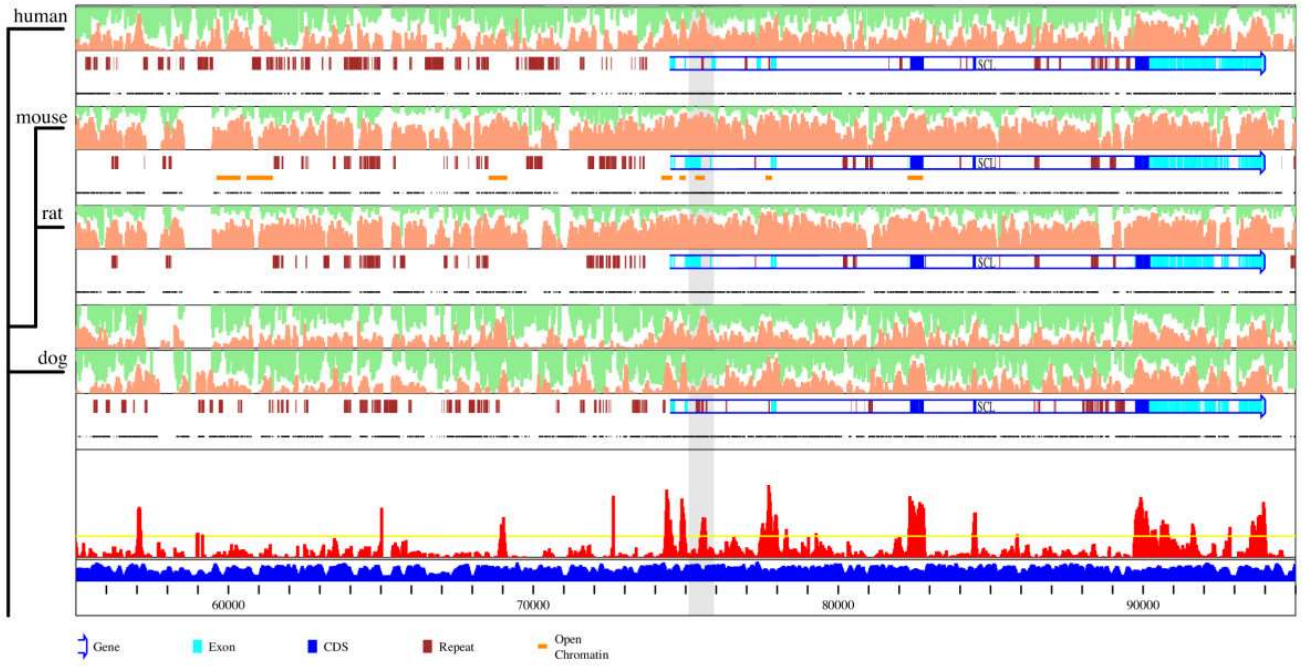
- Estimation of evolutionary relationships / time (**phylogenetics**) : by identifying homologs, we can say
  - how different they are (rate / time of evolution)
  - which subsets of sequences may share a more recent common ancestor (clades)



pythonhosted.org

Why perform alignment ?

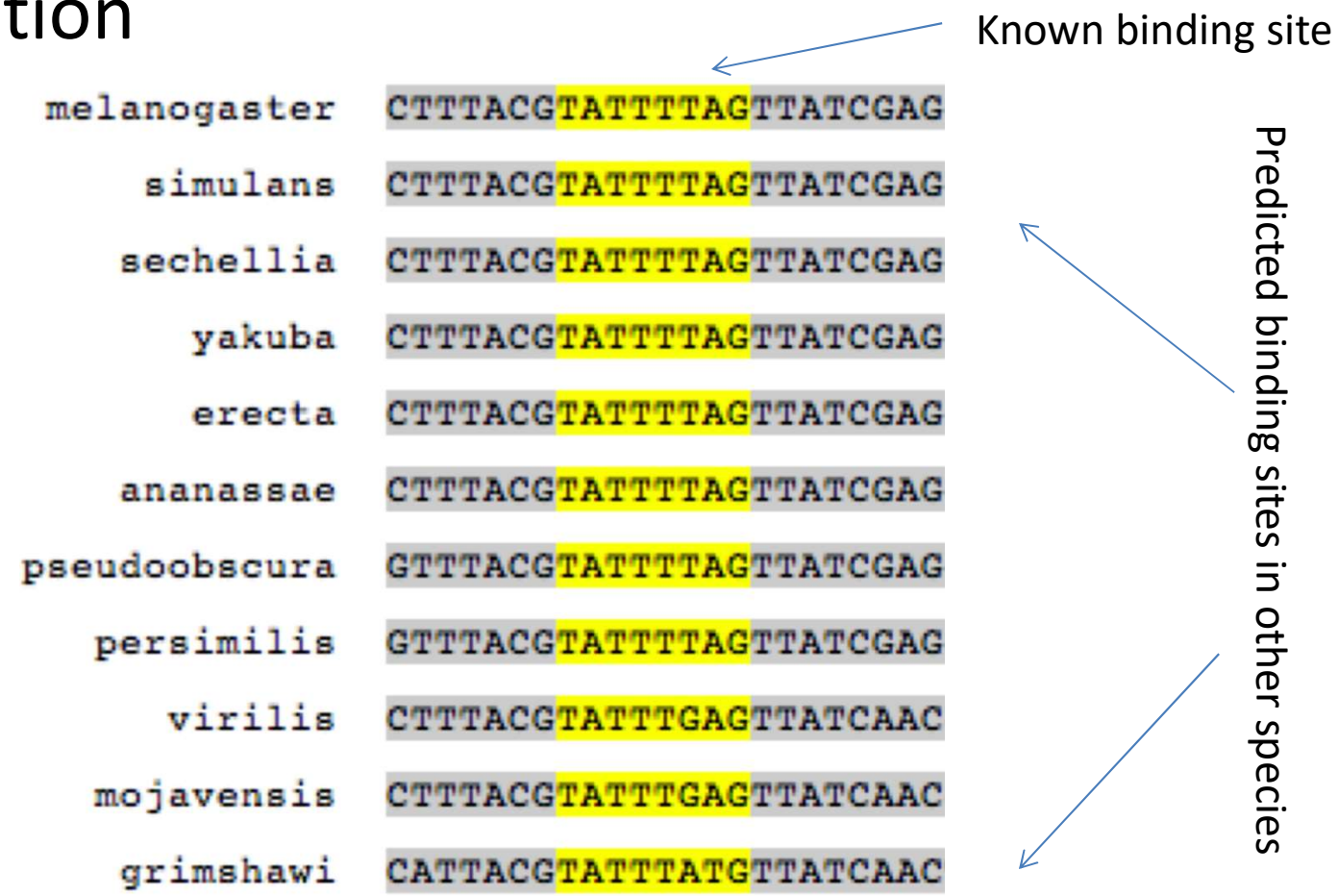
- "Shadowing" or "footprinting" studies : studying orthologs
  - lack of homology : what does it tell us ?



Wakefield et al, BMC Bioinformatics 2005

Why perform alignment ?

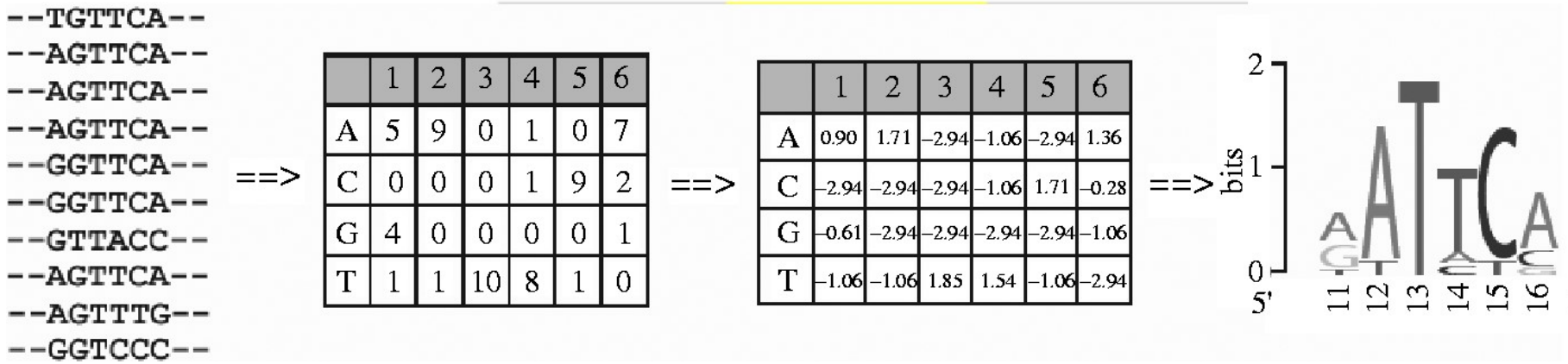
- Transfer learning : of functional or structural annotation



Ray et al, PLoS Comp Bio 2008

Why perform alignment ?

- Identify “motif”s : statistically over-represented patterns across sequences : sometimes we may be studying paralogs



	1	2	3	4	5	6
A	5	9	0	1	0	7
C	0	0	0	1	9	2
G	4	0	0	0	0	1
T	1	1	10	8	1	0

	1	2	3	4	5	6
A	0.90	1.71	–2.94	–1.06	–2.94	1.36
C	–2.94	–2.94	–2.94	–1.06	1.71	–0.28
G	–0.61	–2.94	–2.94	–2.94	–2.94	–1.06
T	–1.06	–1.06	1.85	1.54	–1.06	–2.94

Bi et al, IJICC 2008

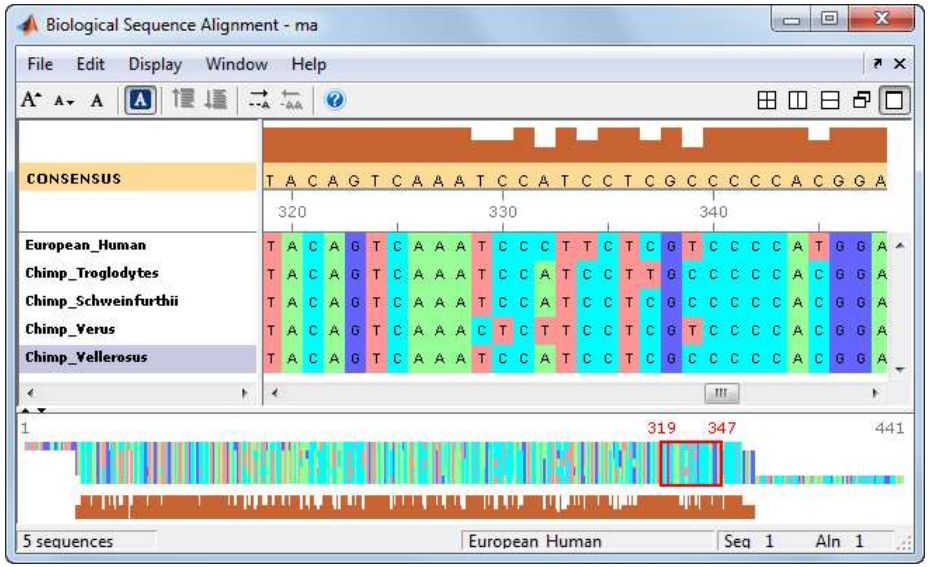
Outline

1. What is MSA ?
2. **Challenges in MSA**
   - Gold standard MSAs
   - MSA seeds
   - Scoring an MSA
   - Space of all MSAs
3. Making MSA work
4. Limits of MSA
5. The future of alignment

How to perform MSA

- Manual : Historically, biologists performed multiple sequence alignment by hand, guided by
  - ultra-conserved subsequences, functional cues (alignment of protein domains), biochemical cues (embedded hydrophobic residues), based on known structure of protein, using well-known patterns of insertions and deletions
- Very tedious ! Sources of bias : alignment in regions of high conservation are easy to spot

Mathworks



Most automated frameworks (including Matlab) still allow manual post-processing of alignments

How to perform MSA

- Automated : Algorithms to
  - search the space of all possible MSAs
  - score the MSAs : then choose one with best score

    Searching and scoring happens simultaneously in DP : efficient as “bad” subalignment scores are “forgotten” (only max retained in each cell)

    May not be simultaneous in non-DP settings

  Choose the MSA with the best score
- Two big challenges for both aspects : why ?

How to score a MSA ?

- The whole notion of “scoring” assumes the presence of a gold standard MSA, against which one can grade candidate MSAs.

- So, how can we get gold standard MSAs ?

Outline

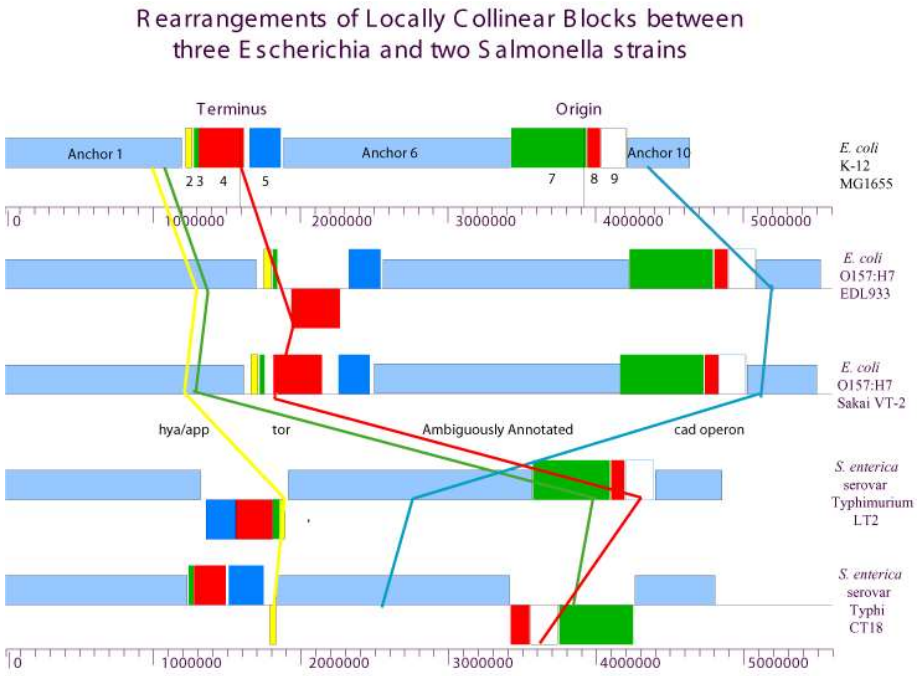
1. What is MSA ?
2. Challenges in MSA
   - **Gold standard MSAs**
   - MSA seeds
   - Scoring an MSA
   - Space of all MSAs
3. Making MSA work
4. Limits of MSA
5. The future of alignment

Scoring a MSA : probabilistically or otherwise

- Converting an expert's evaluation criteria into a scoring scheme : score based on the evidence and prior knowledge
  - essence of bayesian probabilistic modelling
  - typically requires a ground truth
- But what is the ground truth : in terms of evolutionary or structural homology ?
  - a single "correct" MSA can only be obtained only in trivial cases

Evolutionary homology

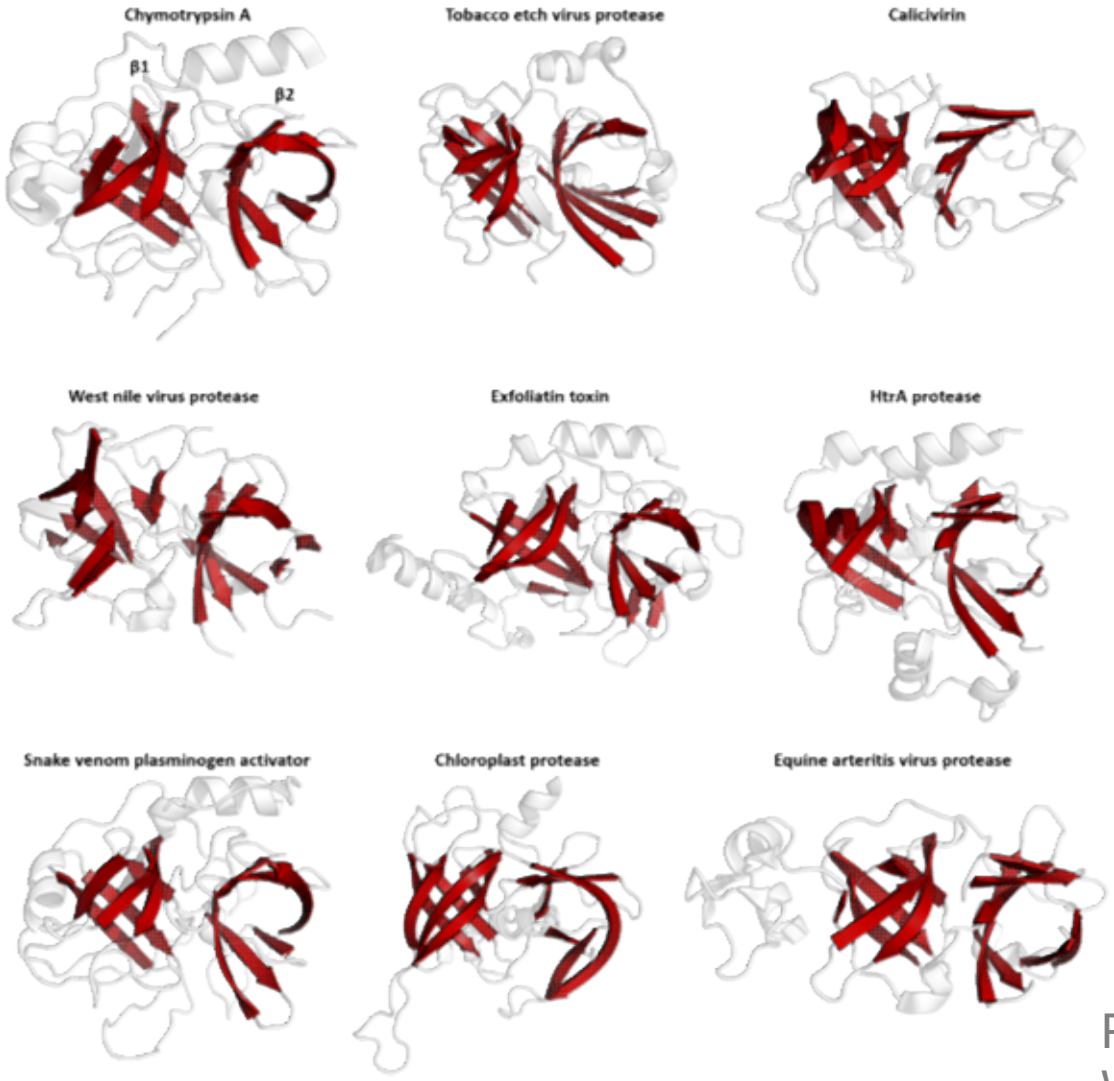
- The **language of MSA is insufficient** to capture all kinds of genomic evolutionary events, so this approach doesn't work !



Rearrangements, inversions, repeats

gel.ahabs.wisc.edu

Structural homology



PA superfamily, Wikipedia

Structural homology

- MSA test benches developed based on structural homology
- Still, many challenges :
  - Pair of divergent but homologous (30% identical) proteins have about 50% of residues not structurally superposable
  - Definition of structural superposition varies from expert to expert : **not ironclad**
- Globin family, used as “typical” example in MSA, is an exception : structure and sequence are strongly conserved throughout family

In the absence of ground truth …

- Bottomline about alignments : artificial constructs, hence no ground truth
- Use scoring schemes that score those alignments highly that look like **“meaningful” alignments**
- Meaningful alignments : informative ones with regard to the use you put the alignment to
  - means to an end

When is a MSA meaningful ?

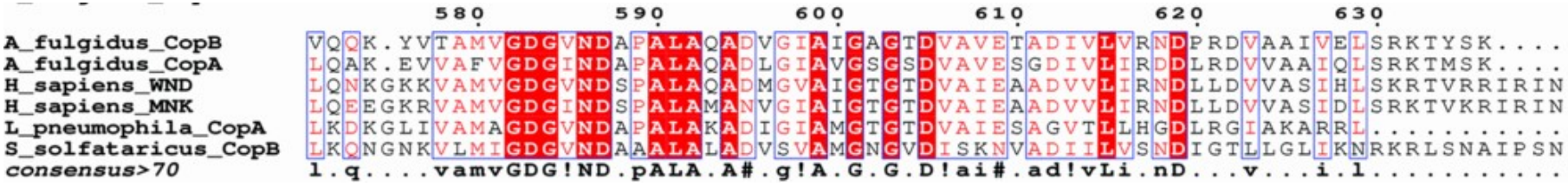
- **Degree of similarity matters** in alignment : why ?
- Ability to identify “correct” alignment depends on how closely related the sequences are
  - ultraconserved sequences : alignment unambiguous : not of interest
  - highly divergent sequences : not possible if degree of homology among sequences of interest ~ degree of homology among two randomly chosen sequences
  - **partially divergent** sequences : meaningful / informative but hard ! Where does the information come from ?

Outline

1. What is MSA ?
2. Challenges in MSA
   - Gold standard MSAs
   - **MSA seeds**
   - Scoring an MSA
   - Space of all MSAs
3. Making MSA work
4. Limits of MSA
5. The future of alignment

“Seeds” of a meaningful MSA

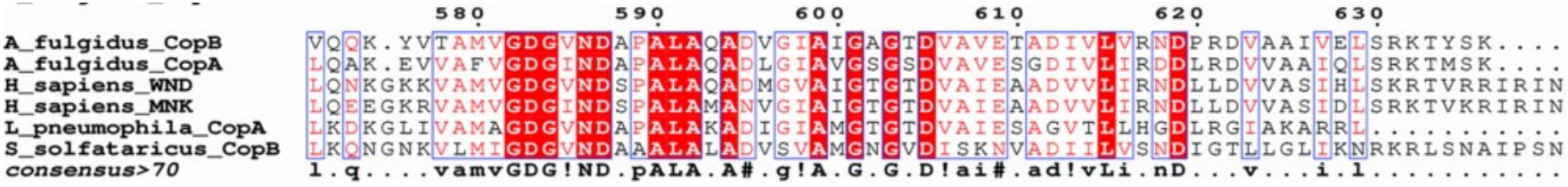
- Small contiguous sets of key residues which align unambiguously **irrespective of degree of total sequence divergence** : “seed”s of MSA
- Core structural (and functional) elements typically conserved – “negative selection”
- **Contrast between seeds and neighboring regions** : provide information / calibration for performing annotation, evolutionary studies



Jayakanthan et al, Biosc Rep 2012

“Seeds” of a meaningful MSA

- Identification of seeds : somewhat similar in spirit to **local sequence alignment**
  - more in whole genome alignment later
- Seeds : explain why only partially divergent sequences make meaningful alignments
  - ultraconserved sequences : everything conserved, no contrast (seq 3 & 4 below)
  - highly divergent sequences : seeds missing / hard to find



Jayakanthan et al, Biosc Rep 2012

Highly diverged sequences + seeds

- Seeds are short : prob significance of finding homologous “seeds” comes from sequence identity AND fact they are located in the same region of the genome
- Highly divergent sequences
  - multiple / major genome re-arrangement events : **loci of orthologs may be far apart** : dont show up as significant only on basis of sequence identity
  - chances of functional **seeds being deleted/mutated** beyond recognition are low, but increase over evolutionary distance, eg. duplication events may relieve negative selectional pressure from locus

Outline

1. What is MSA ?
2. Challenges in MSA
   - Gold standard MSAs
   - MSA seeds
   - **Scoring an MSA**
   - Space of all MSAs
3. Making MSA work
4. Limits of MSA
5. The future of alignment

Scoring framework

- Presuppose alignment exists, and score it
- Assumptions made to come up with a tractable scoring scheme
  - assumptions about columns
  - assumptions about rows

Assumption about columns

- **Independence of columns**
- Probabilistic models : total score of alignment = product of scores of each column, translates to a sum in log-likelihood framework
- Dynamic programming uses a monotonic sum to score pairwise alignments : similar notion for multiple alignment

$$S(m) = G + \sum_i S(m_i)$$

where S(m) : scoring scheme for whole alignment, $S(m_i)$: scoring function for each ungapped column, G : scoring function for gapped columns (possibly affine to help optimization problem)

- Validity of assumption :
  - Not independent , but in practice approximately Markovian (weaker assumption)

“Goodness” of a MSA column

- Similar in spirit to scoring schemes for pairwise alignment : common ancestor implies **homogeneity in column** (at least for short evolutionary distances)
  - What amount of homogeneity do we expect “**by chance alone**” ?
  - Related question : how are the taxa related ? i.e. what are the assumptions about the rows ?
  - Gaps should be grouped, horizontally AND vertically

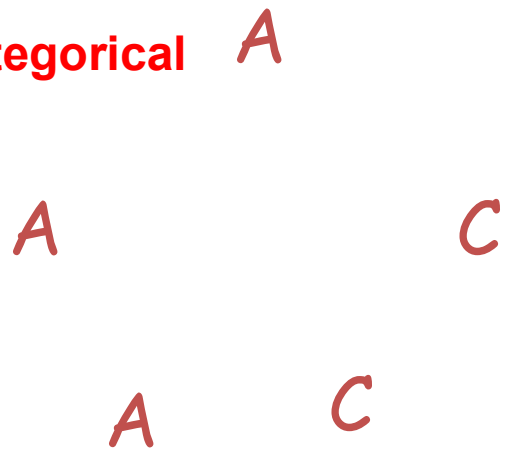
Assumptions about rows : relations between taxa

A
A
A
C
C

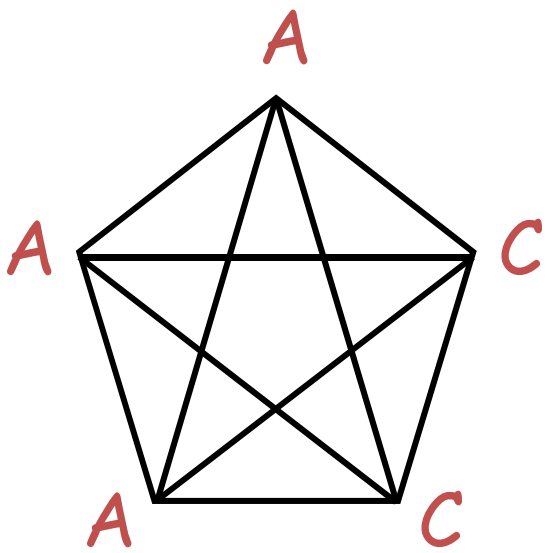
Scoring schemes derived based on the relations

Weighting sequences unequally for scoring is possible for any of these

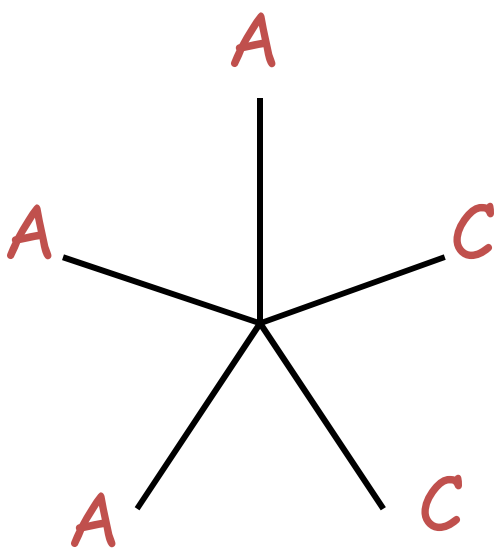
iid categorical



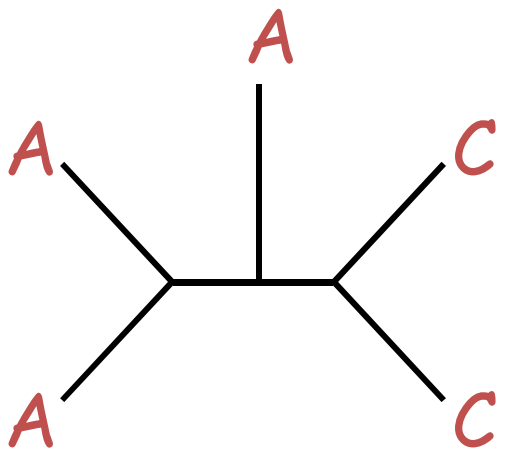
Sum-of-pairs



Star phylogeny

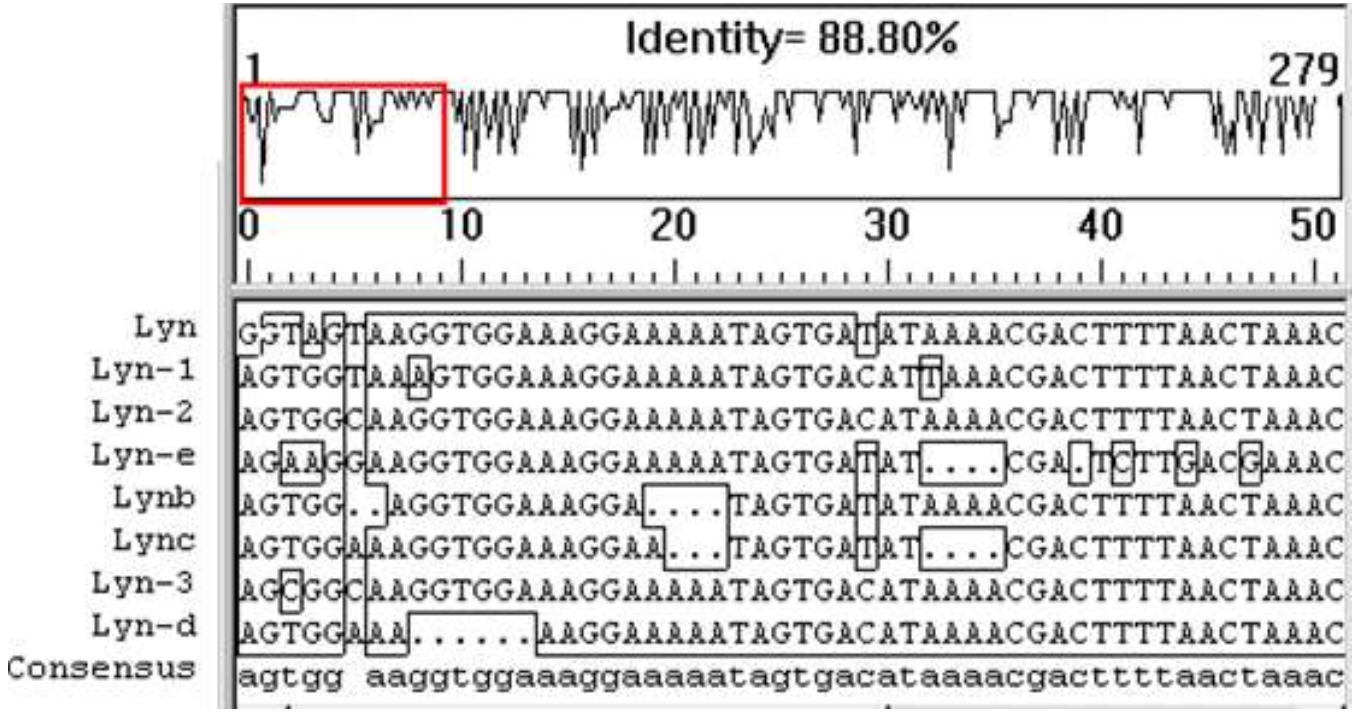


Binary phylogeny



A simple notion that doesn't work

- Degree of identity in each column
  - which nucleotide is the "reference" nucleotide ?
  - Are 2 As and 2 Ts better than 2 As, 1 T and 1 G ?

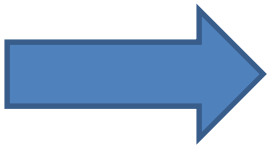


Information content / entropy

- Minimizing column-wise entropy

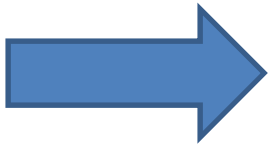
GGAGGT
GGCGGT
GGAGGT
GGCGGT
GGCGGT
GGCGGT
GGCGGT
GCATGT

Alignment



p(A) = 3/12
p(C) = 9/12
p(G) = 0/12
p(T) = 0/12

Multinomial estimated
for random variable
in one column



$$\text{Entropy} = -\sum_{i} P_i \log_2 P_i$$

= 0.81

Entropy of the R.V.

(log 0 = 0
for entropy
calculations !)

Information content / entropy

- Problem : Rows assumed to be iid draws (actually related by evolutionary tree)
  - Works well in practice for closely related sequences
  - If sequences are highly divergent, finds false homology
  - No way to prefer certain kind of changes over others (eg. transitions vs transversions)

Sum – of – pairs

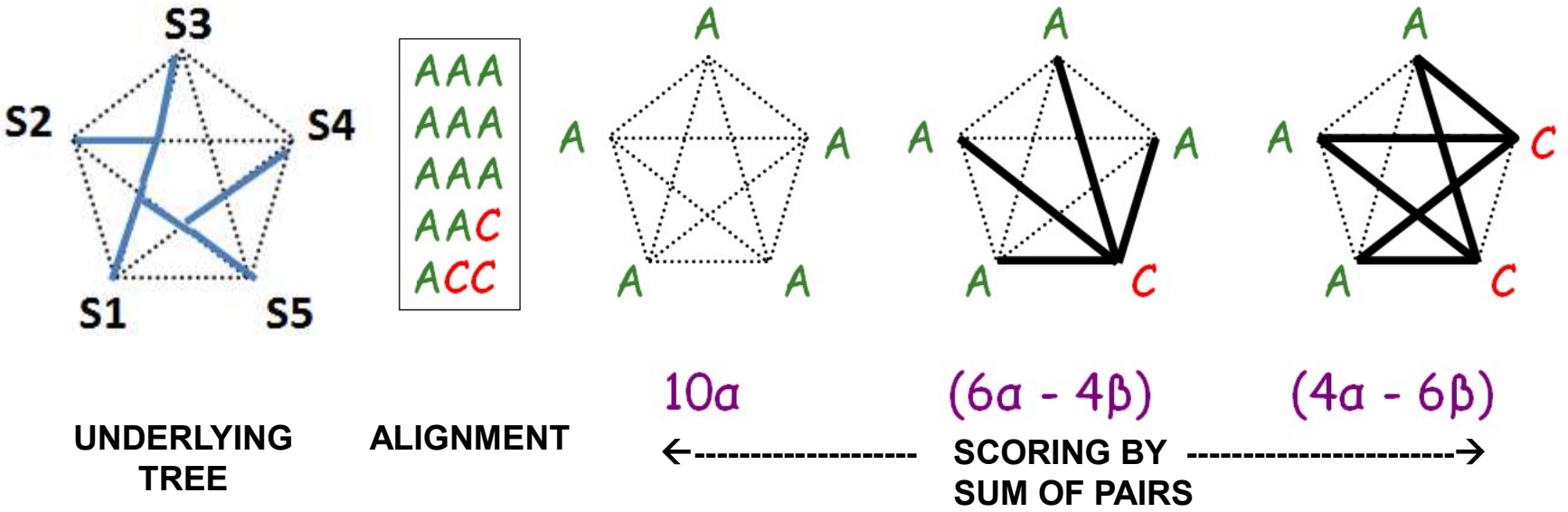
- Score = sum of all pairwise scores (since the pairwise scores prefer homogeneity, their sum should, too)

$$S(m_i) = \sum_{k<l} s(m_i^k, m_i^l)$$

- Sum all N (N – 1) / 2 pair of scores
- Not probabilistically correct extension of log odds score : $\log(p_{abc}/q_a q_b q_c)$ required, we use

$$\log(p_{ab}/q_a q_b) + \log(p_{bc}/q_b q_c) + \log(p_{ca}/q_c q_a)$$

How sum of pairs overcounts mutations



- Most likely phylogenetic history : there is one substitution in the 2nd case, and 1 / 2 substitutions in the 3rd case
  - mismatch penalties are thus disproportionate

Evolutionary score

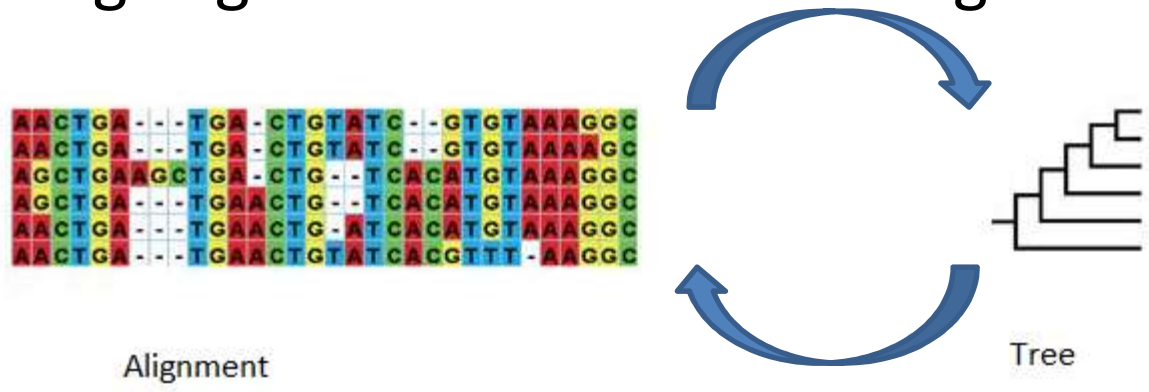
- Expected no of substitutions **counted on the tree** according to an evolutionary model : how much homogeneity is important, but which taxa are expected to be homogeneous is more important
- Sets of mutations or indels consistent with evolutionary tree are better

should not be scored the same

- We will learn more about such scores when we learn evolutionary models (next class)

But …

- **Chicken and egg** problem : so a “guide” tree may be built independently of the alignment
  - using small sequence of reliable alignment or using alignment free tree building methods



(requires evolutionary model for scoring, guide tree for progressive alignment)

(requires homology map for estimating phylogeny rates, and topology)

doi:10.3390/biology2010419

Outline

1. What is MSA ?
2. Challenges in MSA
   - Gold standard MSAs
   - MSA seeds
   - Scoring an MSA
   - **Space of all MSAs**
3. Making MSA work
4. Limits of MSA
5. The future of alignment

Enumerating the space of all MSAs

- No of alignments for 2 sequences of length n_1 and n_2 : Stanton – Cowan recursion

recursion on positive integers:

$$f(n_1, n_2) = f(n_1 - 1, n_2) + f(n_1 - 1, n_2 - 1) + f(n_1, n_2 - 1)$$

$$f(n_1, n_2) = 1 \text{ for } n_1 \text{ and/or } n_2 = 0$$

Stanton & Cowan, 1970
Laquer, 1981

$$f(n_1, n_2) = \sum_{i=0}^{n_1} \binom{n_1}{i} \binom{n_2 + i}{n_1}$$

- Multiple sequence alignment :

$$f(n_1, n_2, \ldots, n_m) = \sum_{N=\max n_j}^{\Sigma n_j} \sum_{i=0}^{N} (-1)^i \binom{N}{i} \prod_{j=1}^{m} \binom{N - i}{N - n_j - i}.$$

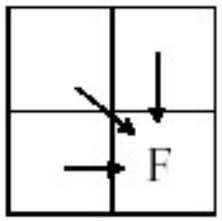
If every sequence is the same length, then the equation becomes

$$f(n, m) = \sum_{N=n}^{mn} \sum_{i=0}^{N} (-1)^i \binom{N}{i} \binom{N - i}{N - n - i}^m.$$

How many alignments are there for 5 DNA sequences of 5 nucleotides each?
A : 1.05 X10^{18} different alignments

J Slowinski, J of MOL PHYL AND EVOL Vol. 10, No. 2, 1994

Systematic traversal of MSA-space : Dynamic programming for MSA



2D

In the 3D case F gets fed from 7 possible cubes.

- Naïve expansion from 2 sequence to n sequence alignment
- How many inputs per cell for aligning n sequences ?

Dynamic programming for MSA

$$\alpha_{i_1,i_2,\ldots,i_N} = \max \begin{cases} \alpha_{i_1-1,i_2-1,\ldots,i_N-1} + S(x^1_{i_1}, x^2_{i_2}, \ldots, x^N_{i_N}), \\ \alpha_{i_1,i_2-1,\ldots,i_N-1} + S(-, x^2_{i_2}, \ldots, x^N_{i_N}), \\ \alpha_{i_1-1,i_2,i_3-1,\ldots,i_N-1} + S(x^1_{i_1}, -, \ldots, x^N_{i_N}), \\ \vdots \\ \alpha_{i_1-1,i_2-1,\ldots,i_N} + S(x^1_{i_1}, x^2_{i_2}, \ldots, -), \\ \alpha_{i_1,i_2,i_3-1,\ldots,i_N-1} + S(-, -, \ldots, x^N_{i_N}), \\ \vdots \\ \alpha_{i_1,i_2-1,\ldots,i_{N-1}-1,i_N} + S(-, x^2_{i_2}, \ldots, -), \\ \vdots \end{cases}$$

R. Durbin

$$\Delta_i \cdot x = \begin{cases} x & \text{if } \Delta_i = 1 \\ - & \text{if } \Delta_i = 0 \end{cases}$$

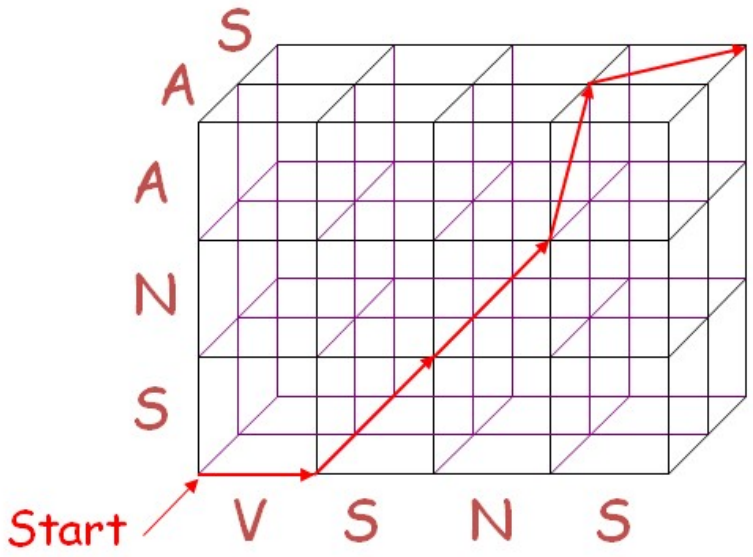
$$\alpha_{i_1,i_2,\ldots,i_N} = \max_{\Delta_1+\ldots+\Delta_N>0} \left\{ \alpha_{i_1-\Delta,i_2-\Delta,\ldots,i_N-\Delta} + S\left(\Delta_1 \cdot x^1_{i_1}, \Delta_2 \cdot x^2_{i_2}, \ldots, \Delta_N \cdot x^N_{i_N}\right) \right\}$$

Dynamic programming for MSA

- Extension of the DP for pairwise alignment
- Assumptions (like pairwise alignment)
  - The columns of an alignment are statistically independent
  - The gaps are scored with affine gap cost
  - Score for an alignment can be calculated as a sum of the scores for each column.
- $(2^k\text{-}1)\ n^k$ comparisons performed by DP for k sequences of length n

An example

- An alignment : a path through the n-dimensional hypercube (n = no of sequences)



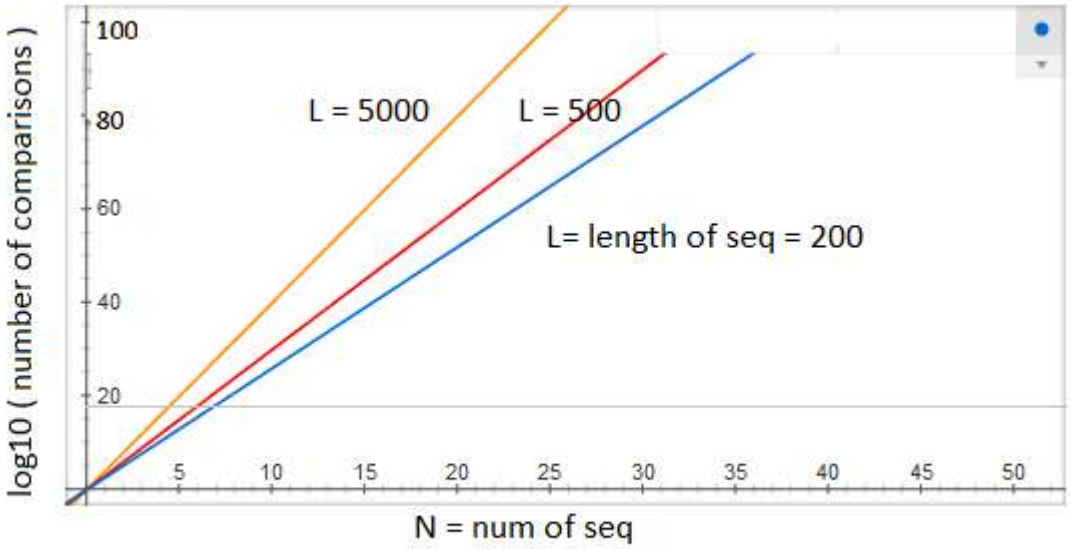
V S N – S
– S N A –
– – – A S

http://www.cs.iastate.edu/~cs544/

Curse of dimensionality

- Many computational problems face “the curse of dimensionality”
- Heuristic solution : only explore a subspace of the space where alignments live – restricted MSA
- Heuristic, practical approaches required
  - Build a MSA from pairwise alignments
  - Add one sequence at a time into the MSA
  - Doesn’t guarantee optimal alignment, but will get you a good one

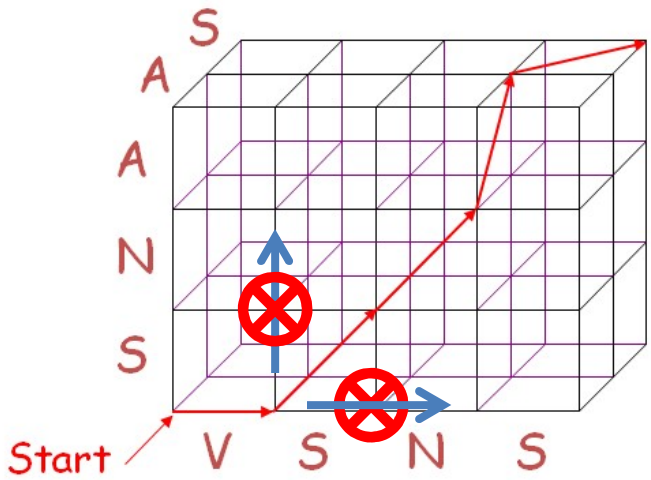
Can we use DP ?



- No of nanoseconds in a decade = 3.15e+17 (grey line)
- Parallelization of algorithm : convert DP hypercube to partially ordered set (based on fillup order) : still challenging

Heuristics

- Simplest heuristic : **branch and bound** : do not further explore unpromising sub-alignments in DP hypercube
  - Trade off time for optimality guarantee
  - Choosing which subalignments to throw away is not easy : a badly scoring indel block may lead to a perfectly aligned block : design of heuristic critical



V S N — S
— S N A —
— — — A S

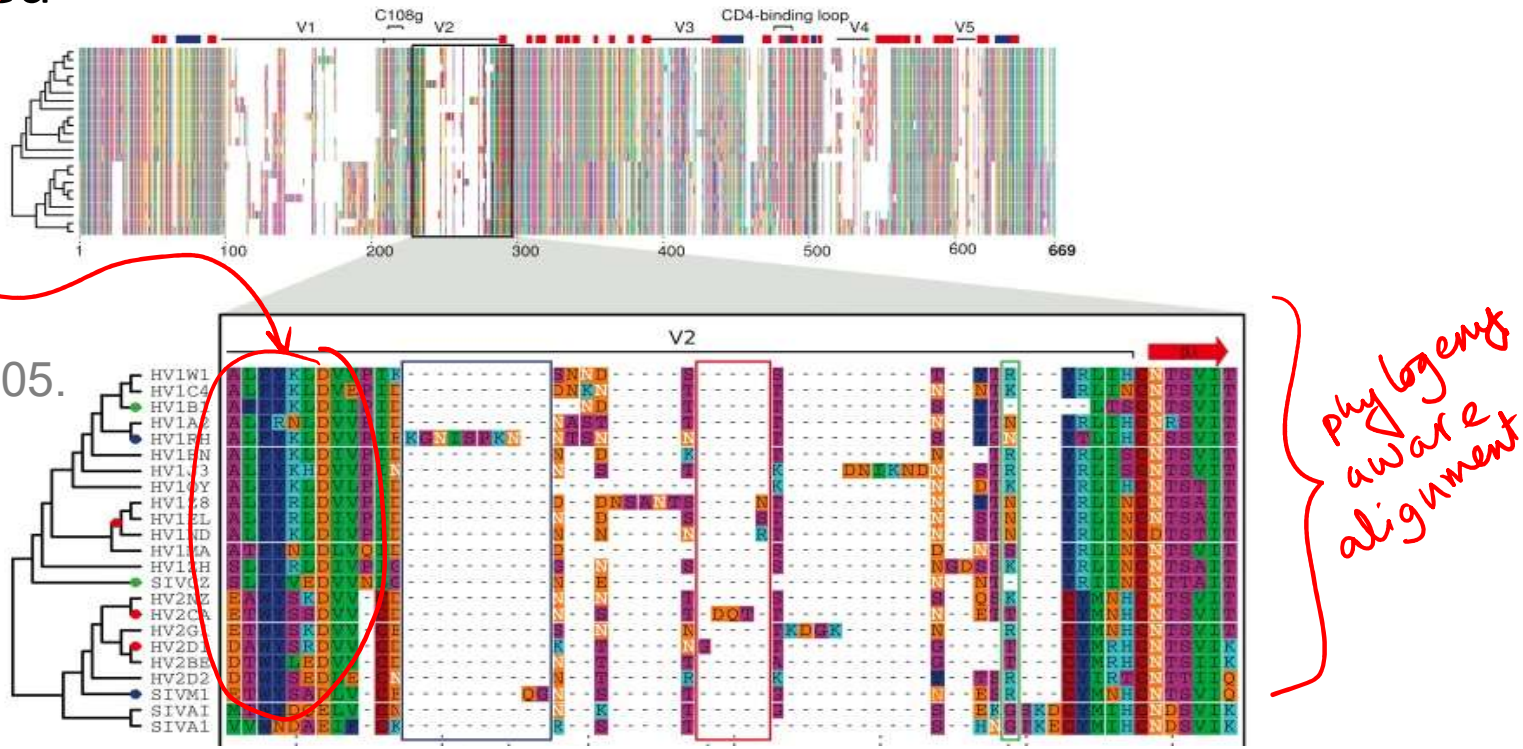
http://www.cs.iastate.edu/~cs544/

Outline

1. What is MSA ?
2. Challenges in MSA
3. **Making MSA work**
   - Progressive alignments
   - Profile HMMs
   - Simultaneous tree + alignment estimation
   - Whole genome alignment
4. Limits of MSA
5. The future of alignment

Real-world MSA algorithms

- Leverage features of “good” MSAs
  - Some sites are more conserved than others, conserved sites occur in “blocks” (sites ~ Markovian model)
  - Sequences related by phylogeny, not independent : reliable phylogeny reqd, sites with patterns consistent w/ phylogeny preferred



Loytynoja A, Goldman N. 2005. PNAS

phylogeny aware alignment

Real-world MSA algorithms

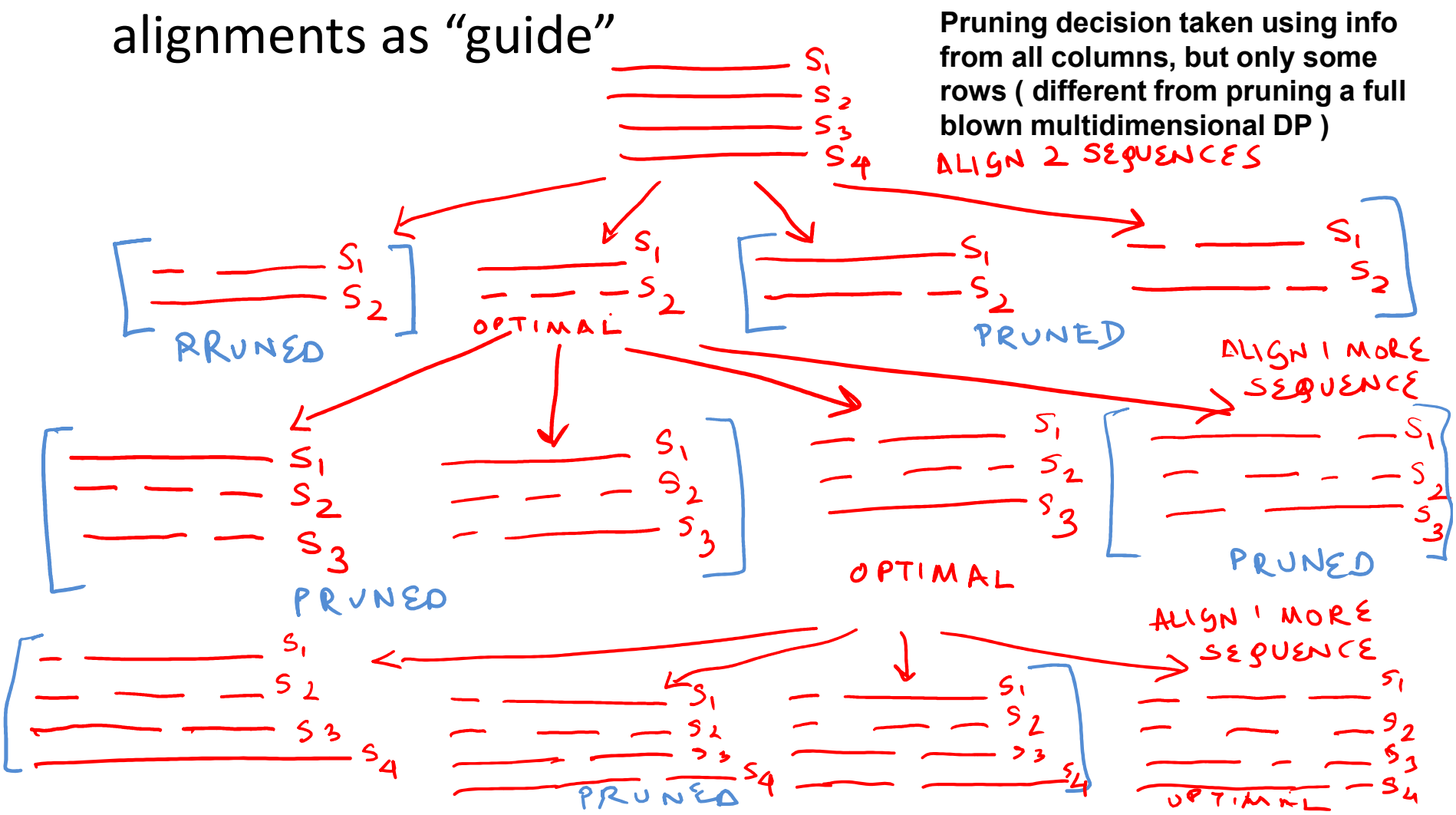
- Deals with inherent issues of building MSAs
  - Curse of dimensionality : use heuristics to prune the space of all alignments
  - Intractable for large sequence sizes : use clever indexing and divide-&-conquer for whole – genome alignments
  - Declining alignment quality for large no of sequences : only align as many sequences as you need
  - False homology for large evolutionary distances : can “intermediate” sequences be found and used ?

Outline

1. What is MSA ?
2. Challenges in MSA
3. Making MSA work
   - **Progressive alignments**
   - Profile HMMs
   - Simultaneous tree + alignment estimation
   - Whole – genome alignment
4. Limits of MSA
5. The future of alignment

Progressive alignment strategies

- Way to prune search space : uses optimal pairwise alignments as “guide”



Progressive alignment

- Consider each sequence as an alignment of 1 sequence
- Choose the two most “similar” alignments and align these alignments
  - Which alignments are most similar ?
  - How to align alignments ?
- Repeat until only a single MSA remains

Feng Doolittle algorithm

- How to rank similarity of sequences ?
  - Choose a distance metric between pairs of sequences
  - Perform hierarchical clustering
    - Historically uses Fitch-Margoliash method, but we will use an algorithm called UPGMA (Unweighted Pair Means Algorithm)

$d(S_i, S_j)$

METRIC :

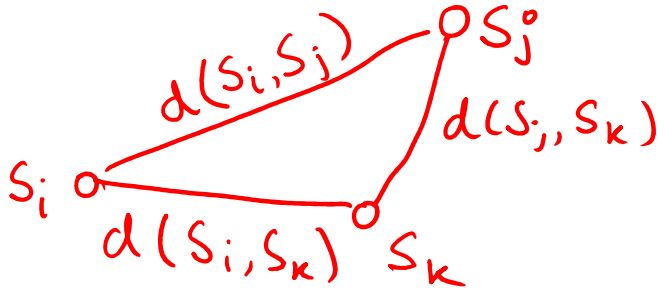
$d(S_i, S_j) \geq 0$

$[\, = 0 \text{ iff } i = j \,]$

$d(S_i, S_j) = d(S_j, S_i)$

$d(S_i, S_j) \geq d(S_i, S_k) + d(S_k, S_j)$

TRIANGLE INEQUALITY



Feng Doolittle

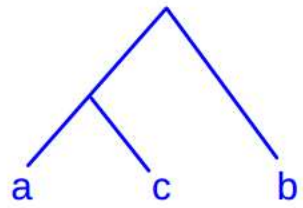
- 3 sequences: $a = ACCAT$, $b = ACGGAT$, $c = AACCAT$ and score: $s(x,y) = \begin{cases} 1 & \text{if } x = y \\ -1 & \text{else} \end{cases}$

- pairwise alignments (similarities !):

$a \leftrightarrow b = 2$	A C – C A T A C G G A T
$b \leftrightarrow c = 0$	A C G G A T A A C C A T
$a \leftrightarrow c = 4$	– A C C A T A A C C A T

$$D = -\log S_{eff} = -\log \frac{S_{obs} - S_{rand}}{S_{\max} - S_{rand}}$$

guide tree



http://www.bioinf.uni-freiburg.de/Lehre/

Feng Doolittle

- start with $a \leftrightarrow c = 4$ and replace gap by X

$$\text{group 1:} \begin{bmatrix} X & A & C & C & A & T \\ A & A & C & C & A & T \end{bmatrix} \quad a' \downarrow$$

- join $b \Rightarrow$ generate all pairwise alignments from b against group 1

$a' \leftrightarrow b = 2$	X	A	C	–	C	A	T
	–	A	C	G	G	A	T
$b \leftrightarrow c = 0$	A	C	G	G	A	T	
	A	A	C	C	A	T	

- use best alignment $a' \leftrightarrow b$ to determine alignment to group

$$\text{group2:} \begin{bmatrix} X & A & C & – & C & A & T \\ A & A & C & – & C & A & T \\ – & A & C & G & G & A & T \end{bmatrix} \rightarrow \begin{bmatrix} X & A & C & X & C & A & T \\ A & A & C & X & C & A & T \\ X & A & C & G & G & A & T \end{bmatrix}$$

http://www.bioinf.uni-freiburg.de/Lehre/

Once a gap, always a gap

- –After an alignment is completed, gap symbols are replaced with a neutral X character.
- –This rule allows pairwise sequenc alignments to be used to guide the alignment of sequences to groups or groups to groups; otherwise, any given pairwise sequence alignment would not necessarily be consistent with the pre-existing alignment of a group.
- –Desirable side effect:encouraging gaps to occur in the same columns in subsequent pairwise alignments.

Problem with Feng - Doolittle

- A problem with the Feng-Doolittle approach - all alignments are determined by pairwise sequence alignments.
- It is advantageous to use position-specific information from the group's multiple alignment to align a new sequence to it. (e.g. degree of sequence conservation)
- •Many progressive alignment methods use pairwise alignment of *sequences to profiles or of profiles to profiles as a subroutine which is used many times in the process.*

CLUSTAL-W

- Profile-based progressive multiple alignment
- Works in much the same way as the Feng-Doolitle method except for its carefully tuned use of profile alignment methods.
- Uses various heuristics

CLUSTAL-W

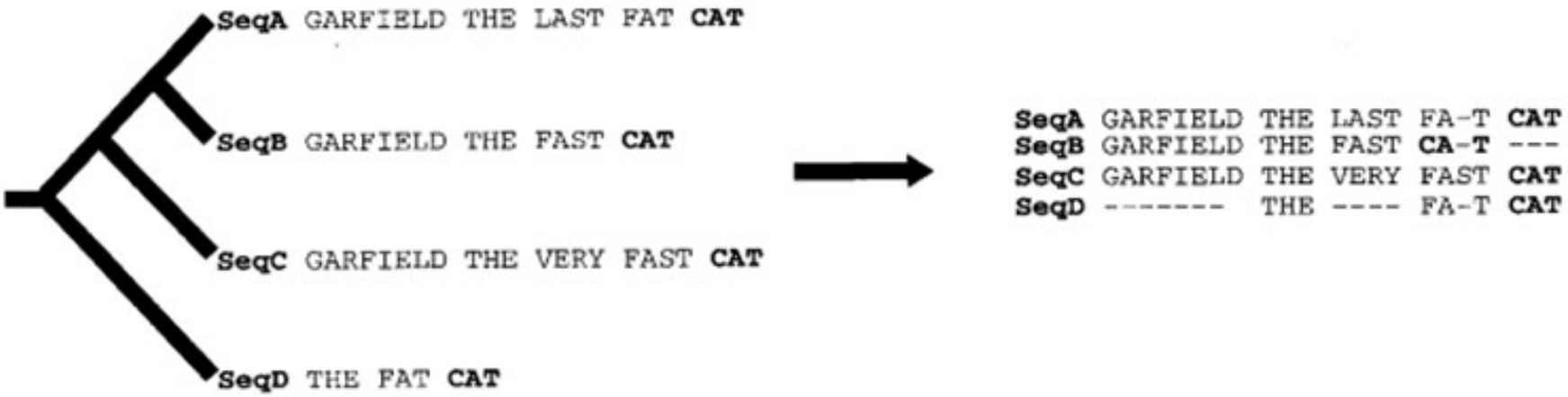
- Construct a distance matrix of all N(N-1)/2 pairs by pairwise dynamic programming.
- Construct a guide tree by clustering (neighbour-joining) .
- Progressively align at nodes in order of decreasing similarity, using sequence-sequence, sequence-profile, and profile-profile alignment.
  - Scoring is basically SP.

CLUSTAL-W

- Heuristics used
- Sequences are weighted to compensate for biased representation in large subfamilies.
- The substitution matrix is chosen on the basis of the similarity expected of the alignment.
- Position-specific gap-open penalties are used.
- Gap penalties are increased if there are no gaps in a column but gaps occur nearby in the alignment.

Pitfalls of sequential alignment

- Mistakes made early on cannot be corrected later



Notredame et al, JMB, 2000

Barton-Sternberg multiple alignment

- Find the two sequences with the highest pairwise similarity and align them using standard pairwise DP alignment.
- Find the sequence that is most similar to a profile of the alignment of the first two, and align it to the first two by profile-sequence alignment. Repeat until all sequences have been included in the multiple aligment.
- Remove sequence x1 and realign it to a profile of the other aligned sequences x2,… xN by profile-sequence alignment. Repeat for sequences x2…xN.
- Repeat the previous realignment step a fixed number of times, or until the alignment score converges.

Distance and similarity function

- Models evolutionary forces
- Order of alignment : affects MSA hugely
- Evolutionary model : make or break progressive alignment methods

Outline

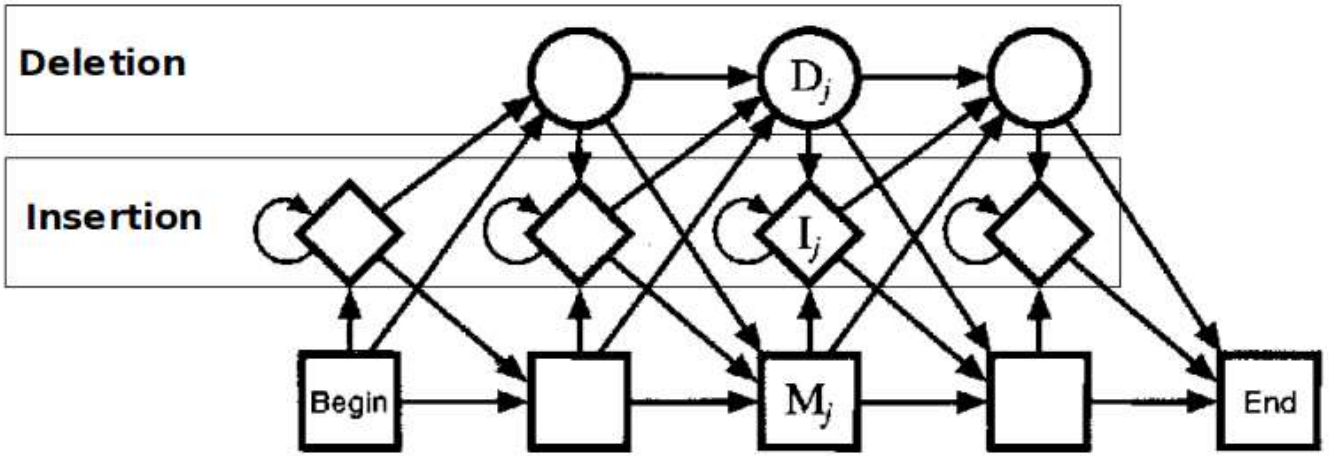
1. What is MSA ?
2. Challenges in MSA
3. Making MSA work
   - Progressive alignments
   - **Profile HMMs**
   - Simultaneous tree + alignment estimation
   - Whole – genome alignment
4. Limits of MSA
5. The future of alignment

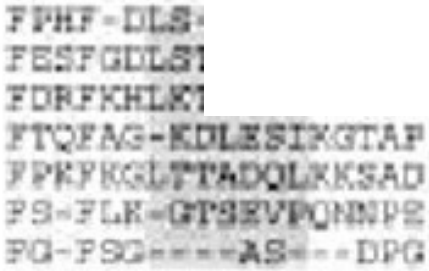
Multiple alignment by profile HMM training

- “Profiles” : sequence template as a sequence of multinomials - profile HMMs.
- Profile HMMs could simply be used in place of standard profiles in progressive or iterative alignment methods.
- Ad hoc SP scoring scheme can be replaced by more explicit profile HMM assumption.
- Trained from initially unaligned sequences : Baum-Welch : EM + Viterbi

Profile HMM

- Start from an initial profile, and sequentially add sequences
  - how to obtain initial profile ?





http://codecereal.blogspot.com/

Baum Welch

- No ground truth
- Viterbi + Expectation Maximization

- Local maxima
  - search stochastically
    - simulated annealing and other approaches

Development in the 2000s

Review : Cedric Notredame, PLoS CompBio, 2000

Method	Score	Templates	Validation Values		Server
			PreFab	HOMSTRAD	
ClustalW [14]	Matrix	—	61.80 [12]	—	http://www.ebi.ac.uk/clustalw/
Kalign	Matrix	—	63.00 [18]	—	http://msa.cgb.ki.se/
MUSCLE [6]	Matrix	—	68.00 [16]	45.0 [9]	http://www.drive5.com/muscle/
T-Coffee [10]	Consistency	—	69.97 [12]	44.0 [9]	http://www.tcoffee.org/
ProbCons [7]	Consistency	—	70.54 [12]	—	http://probcons.stanford.edu/
MAFFT [8]	Consistency	—	72.20 [12]	—	http://align.genome.jp/mafft/
M-Coffee [12]	Consistency	—	72.91 [12]	—	http://www.tcoffee.org/
MUMMALS [16]	Consistency	—	73.10 [16]	—	http://prodata.swmed.edu/mummals/
DbClustal [24]	Profiles	—	—	—	http://bips.u-strasbg.fr/PipeAlign/
PRALINE [9]	Matrix	Profiles	—	50.2 [9]	http://zeus.cs.vu.nl/programs/pralinewww/
PROMALS [16]	Consistency	Profiles	79.00 [16]	—	http://prodata.swmed.edu/promals/
SPEM [28]	Matrix	Profiles	77.00 [28]	—	http://sparks.informatics.iupui.edu/Softwares-Services_files/spem.htm
Expresso [13]	Consistency	Structures	—	71.9 [11][a]	http://www.tcoffee.org/
T-Lara [29]	Consistency	Structures	—	—	https://www.mi.fu-berlin.de/w/LiSA/

Validation values were compiled from several sources, and selected for comparability. PreFab validations were made using PreFab version 3. HOMSTRAD validations were made on datasets having less than 30% identity. The source of each value is indicated by the accompanying reference citation.

[a]The Expresso value comes from a slightly more demanding subset of HOMSTRAD (HOM39) made of sequences less than 25% identical.

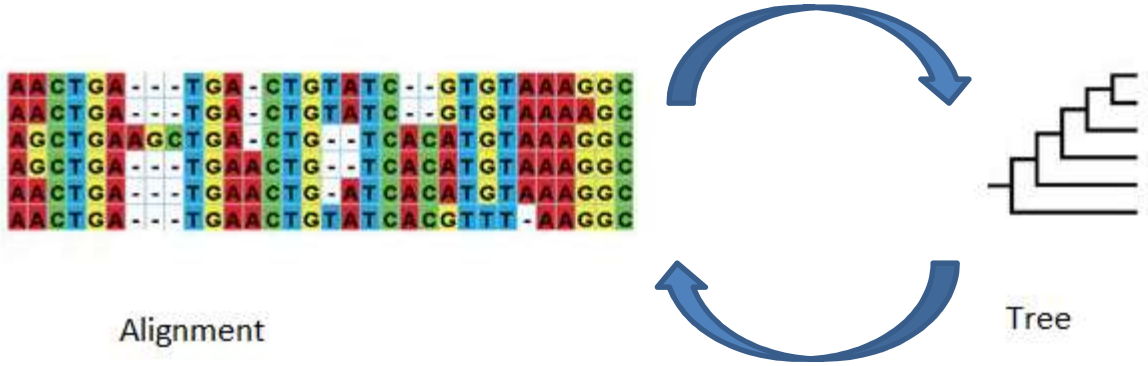
doi:10.1371/journal.pcbi.0030123.t001

Outline

1. What is MSA ?
2. Challenges in MSA
3. Making MSA work
   - Progressive alignments
   - Profile HMMs
   - **Simultaneous tree + alignment estimation**
   - Whole – genome alignment
4. Limits of MSA
5. The future of alignment

Wait, didn't you say …

- MSA s are used to calculate evolutionary models and trees !
- How can "guide trees" be used to calculate MSAs ?
- Chicken and egg problem : can iterate until convergence (Tandy Warnow lab, UT Austin)



PASTA : Simultaneous alignment and tree construction

Step 1 Decompose the input set S into subsets $S_1 \ldots S_m$ of size at most k.
Step 2 Compute a spanning tree T^* to connect the subsets $S_1 \ldots S_m$.
Step 3 Align each subset using the subset alignment technique.
Step 4 Merge the two alignments on endpoints of each edge in T^*.
Step 5 Use successive applications of transitive closure to merge the overlapping and compatible alignments obtained in Step 4.
Step 6 Compute a maximum likelihood (ML) tree on the full MSA using FastTree-2 [13].

- Repeat until convergence

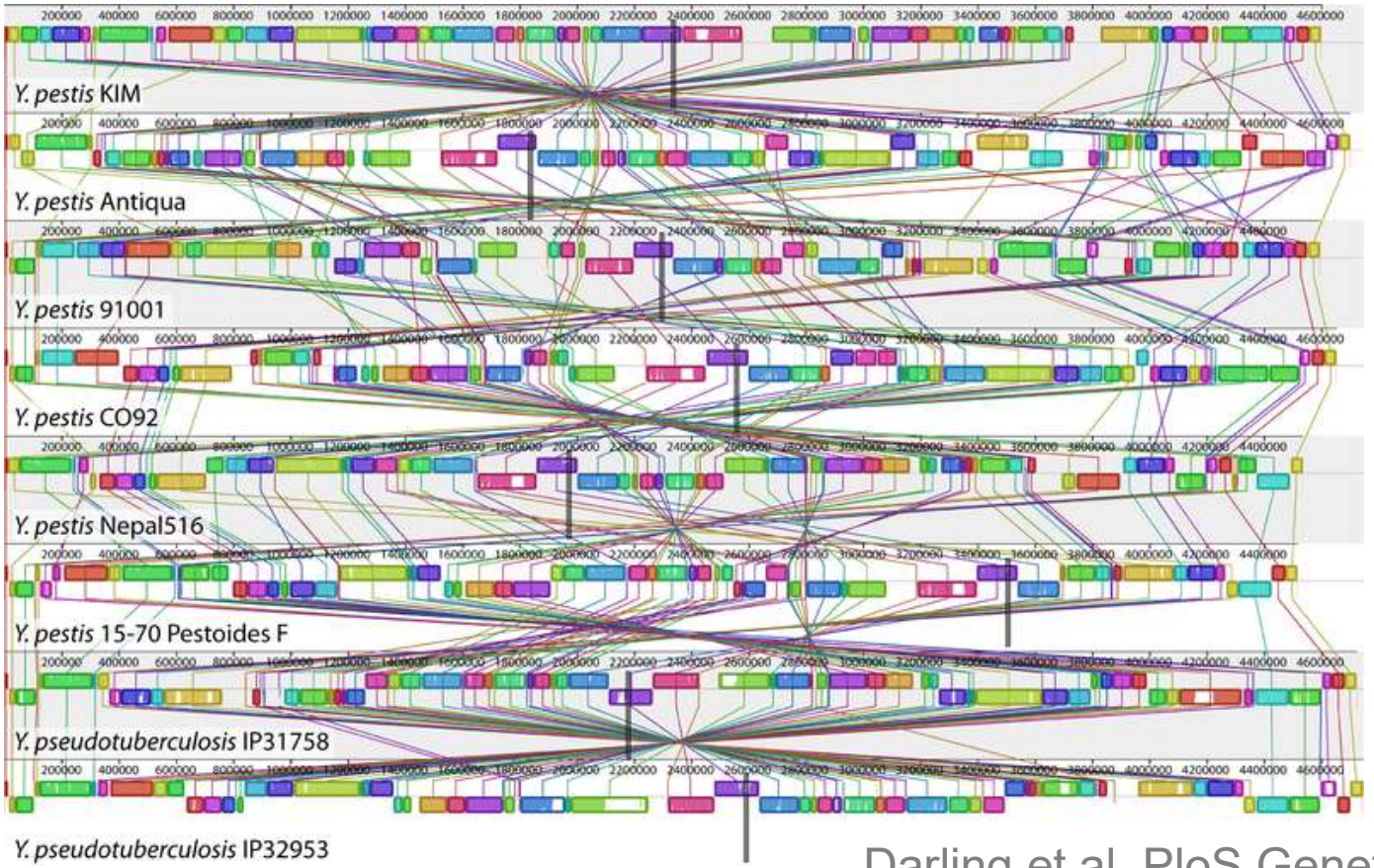
Mirarab et al, RECOMB 2014

Outline

1. What is MSA ?
2. Challenges in MSA
3. Making MSA work
   - Progressive alignments
   - Profile HMMs
   - Simultaneous tree + alignment estimation
   - **Whole – genome alignment**
4. Limits of MSA
5. The future of alignment

Whole genome alignment

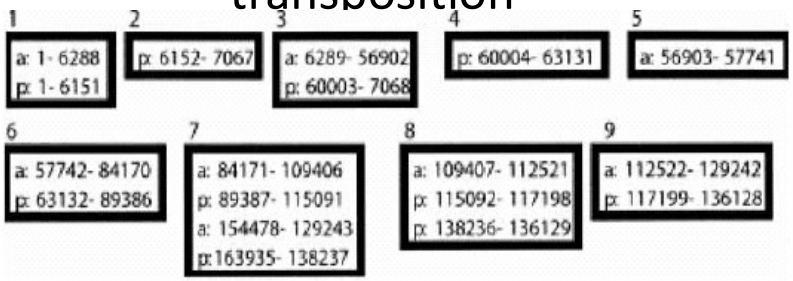
- Identify “collinear” (orthologous) regions or blocks and perform piecewise alignment



Darling et al, PloS Genetics, 2008

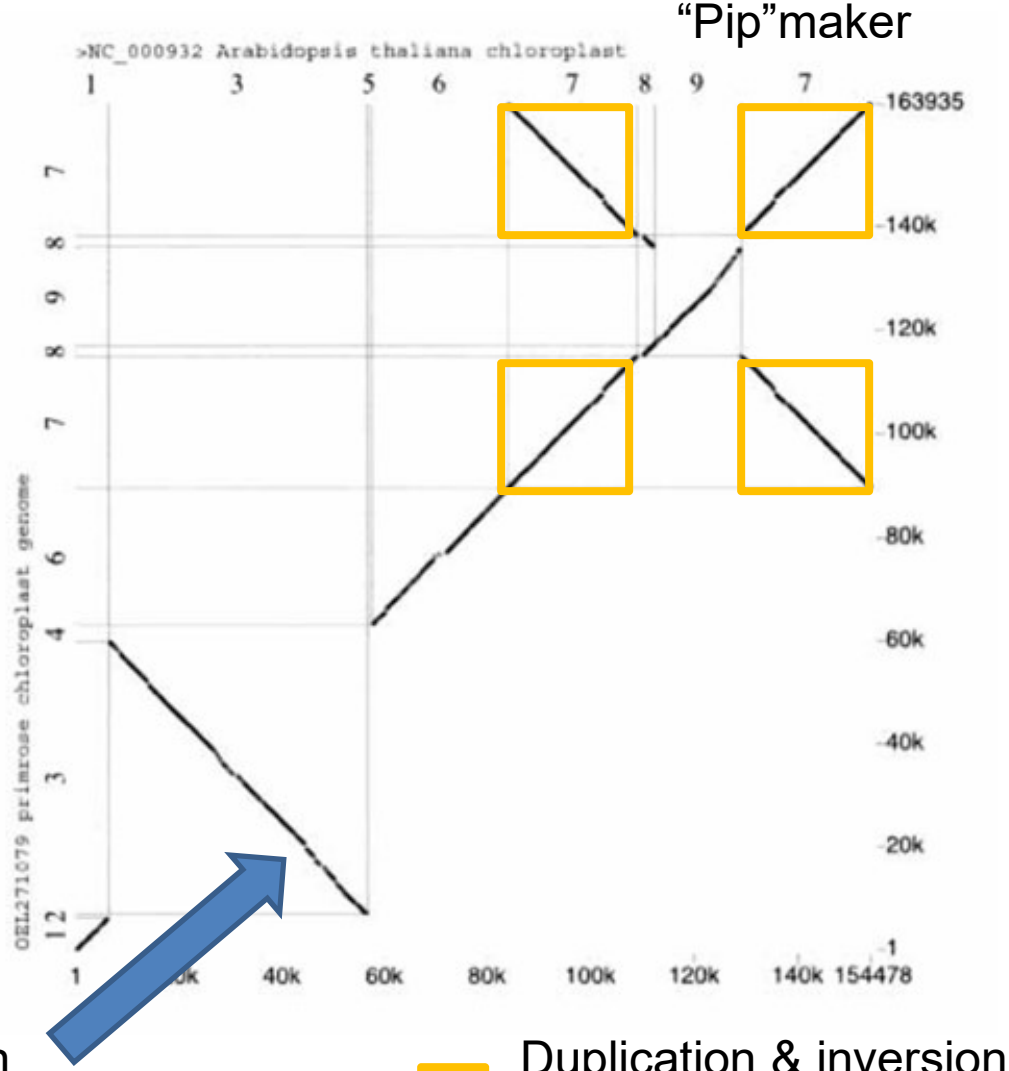
TBA : Threaded Blockset Aligner

- Threaded blockset : **generalization of MSA**
  - Input : Set of sequences
  - Output : Set of “block”s (MSAs) without duplication, inversion, transposition



- Partially order blocks (how to find the blocks ?)

“Pip”maker



Inversion after common ancestor

Duplication & inversion before common ancestor

Blanchette et al, Genome Research, 2004

Outline

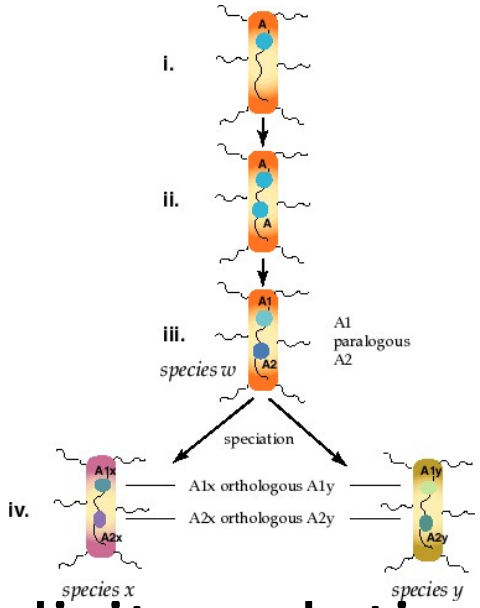
1. What is MSA ?
2. Challenges in MSA
3. Making MSA work
4. Limits of MSA
   - Progressive “fracturing” or false homologs
   - Limited “dynamic range”
   - Limited powers of inference
5. The future of alignment

Things in the real world arent always simple

- Homologous columns don’t behave identically

Things in the real world arent always simple

- More complicated homology



stdgen.northwestern.edu

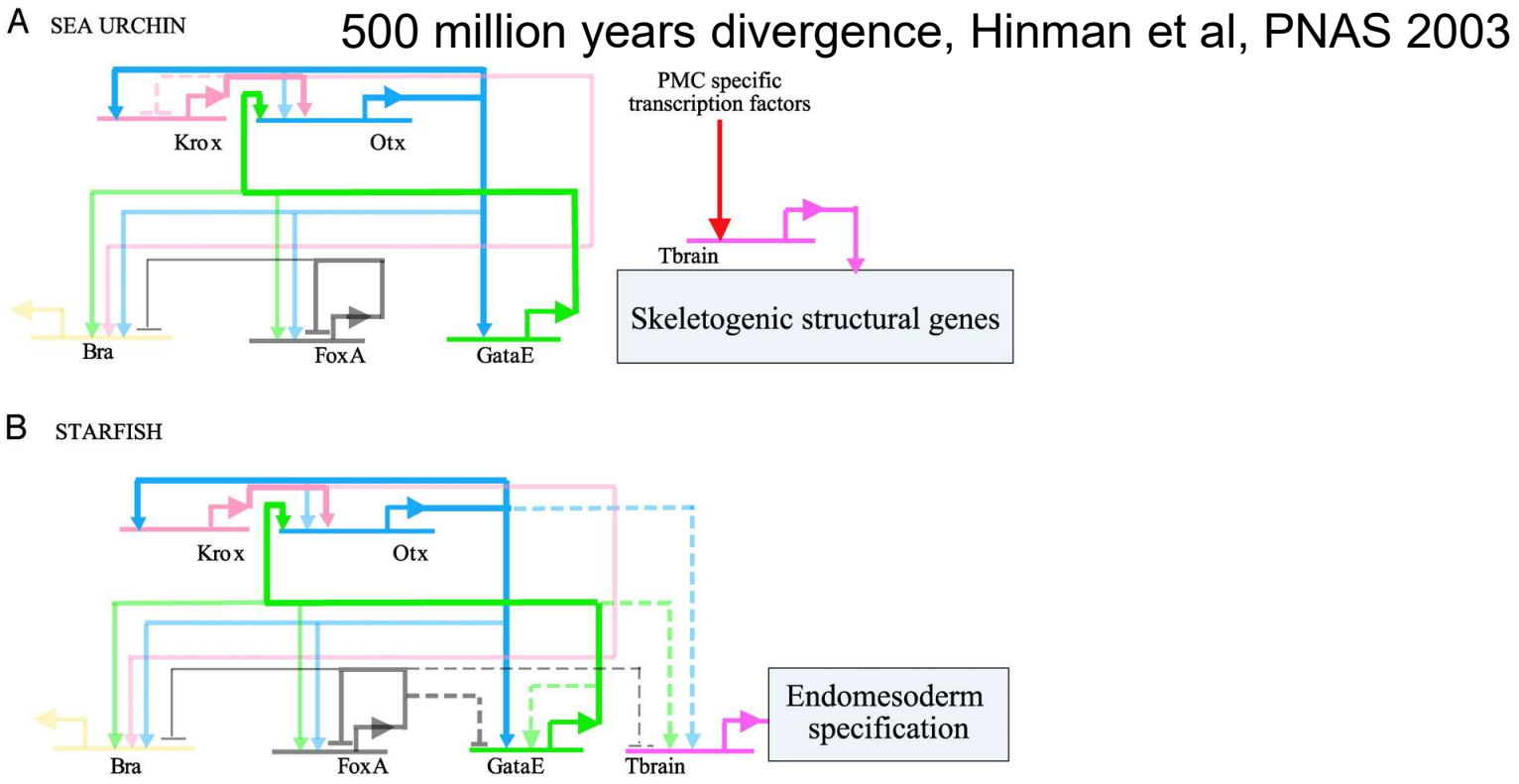
- Requires explicit evolutionary modelling

Outline

1. What is MSA ?
2. Challenges in MSA
3. Making MSA work
4. Limits of MSA
5. **The future of alignment**

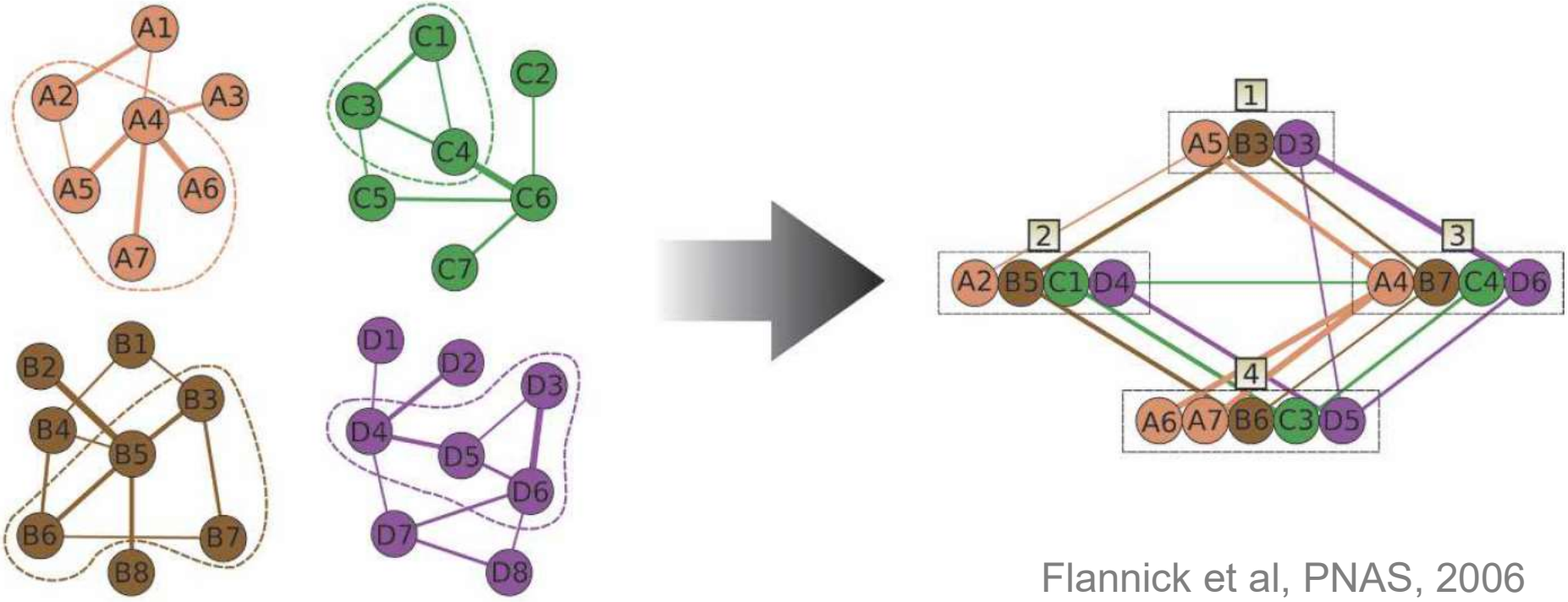
Network alignment

- For large timescales, gene regulatory network may be conserved even though sequence may not be conserved



Network alignment algorithms

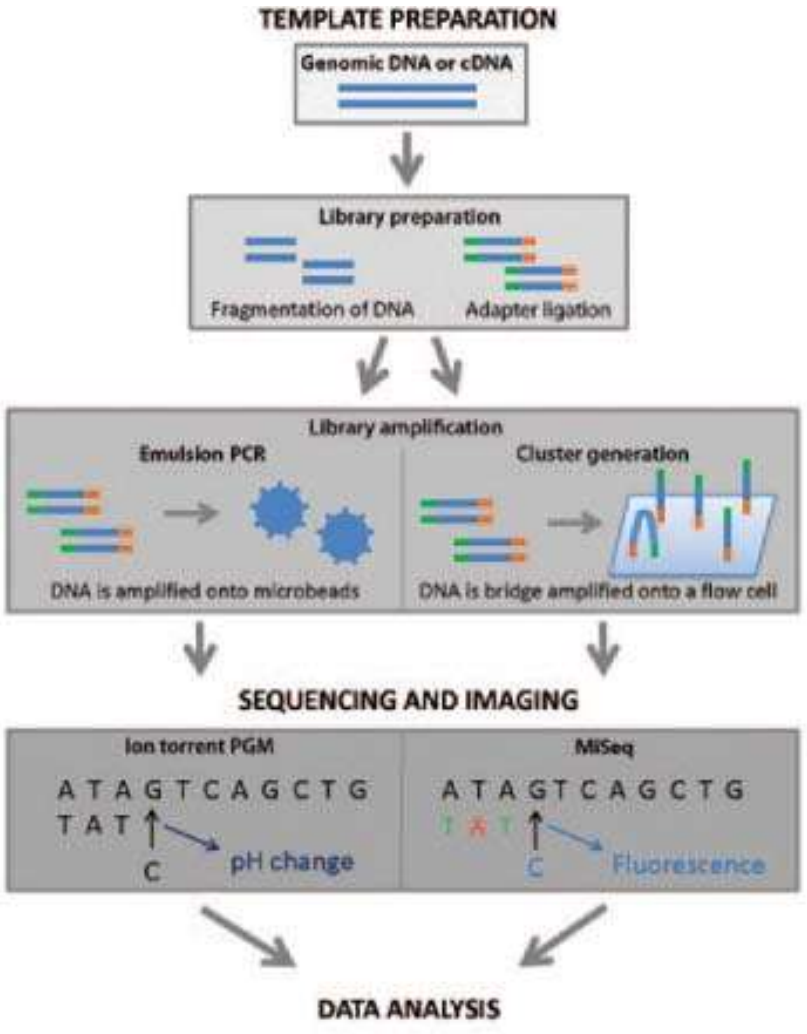
- Identifying network motifs (Qnet 2007, TOPAC 2012)
- Performing multiscale network alignment (GRAEMLIN 2006, BiNa 2009)



Flannick et al, PNAS, 2006

Next generation sequencing (NGS)

- "Microscope of 21st century"
  - Many important problems reduced to NGS : reference sequence generation, sequence variant detection, protein – DNA binding, transcriptome quantification, chromatin structure, DNA / RNA epigenetics
  - Necessary first step : sequenced reads to be assembled / aligned



Grada & Weinbecht, J Inv Dermat, 2013

NGS mapping to reference genome

- Local alignment of millions of small read to whole genome / transcriptome : “mapping”

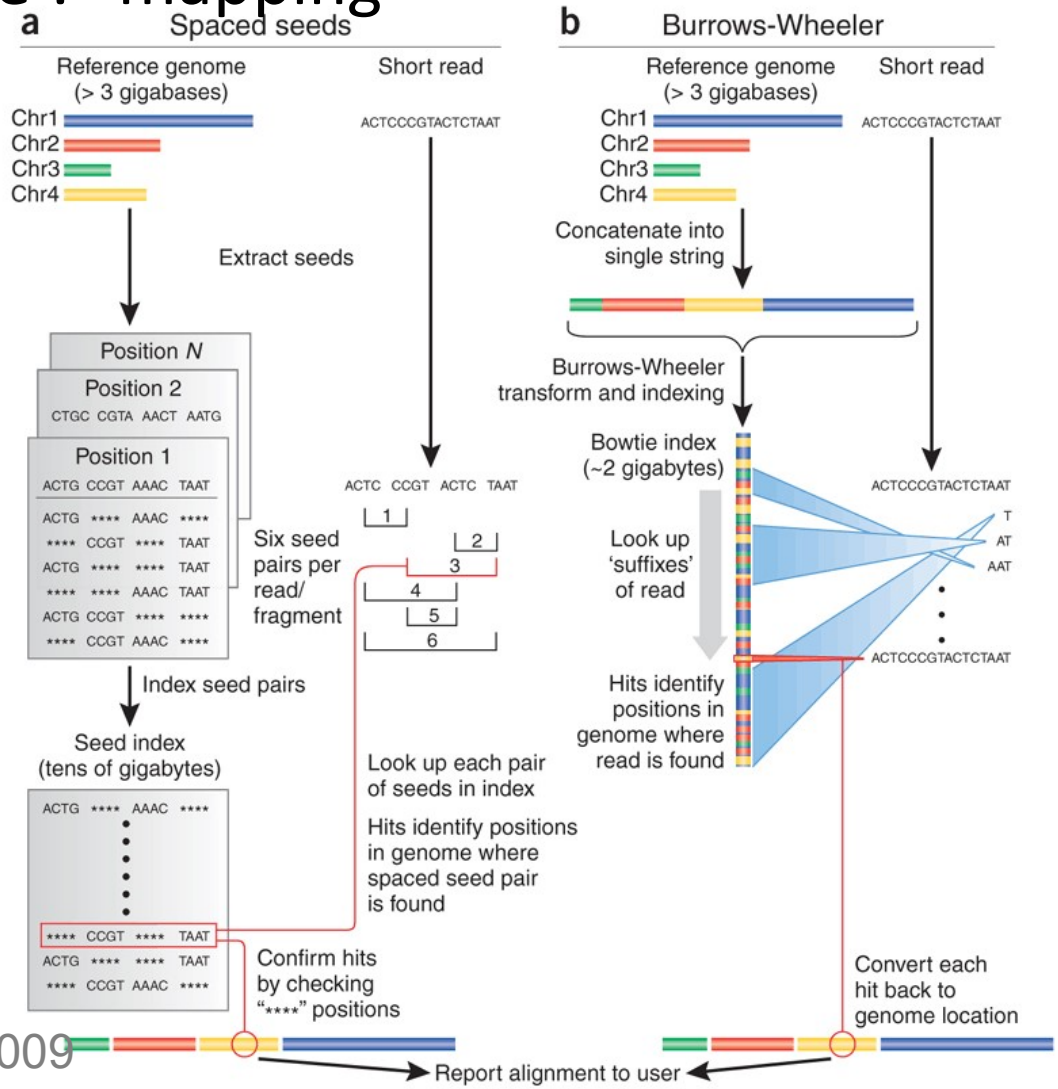
“Search” for perfect / near-perfect match

Traditional local alignment : too slow

Suffix tree (MPScan)

Seed-and-extend / exact hash tables (BLAST, FASTA, MAQ, BLAT, RMAP)

Burrows – Wheeler Transform + suffix lookup (Bowtie)



Trapnell & Salzburg, Nature Biotech 2009

NGS mapping + assembly w/o reference

- De novo assembly
  - determine the assembly (similar to shotgun sequencing – DeBruijn traversals and variants)
  - can it be assembled / uniquely assembled ?
  - if it can, where do the reads map ?

Ambiguous assembly

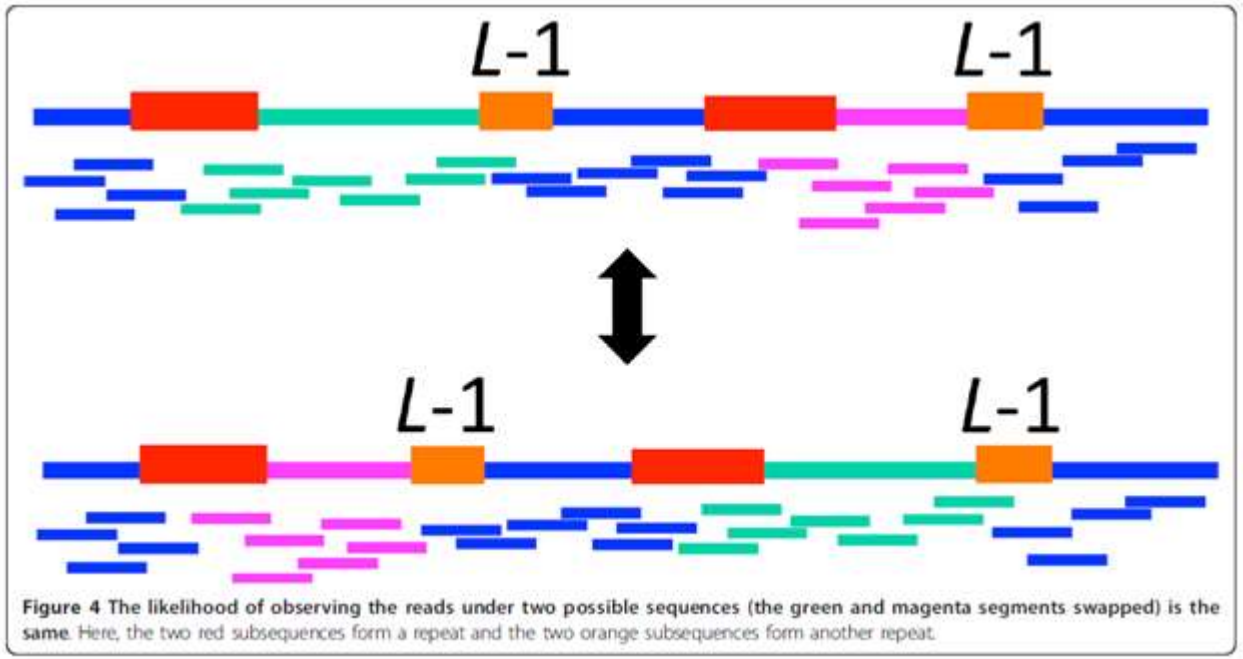


Figure 4 The likelihood of observing the reads under two possible sequences (the green and magenta segments swapped) is the same. Here, the two red subsequences form a repeat and the two orange subsequences form another repeat.

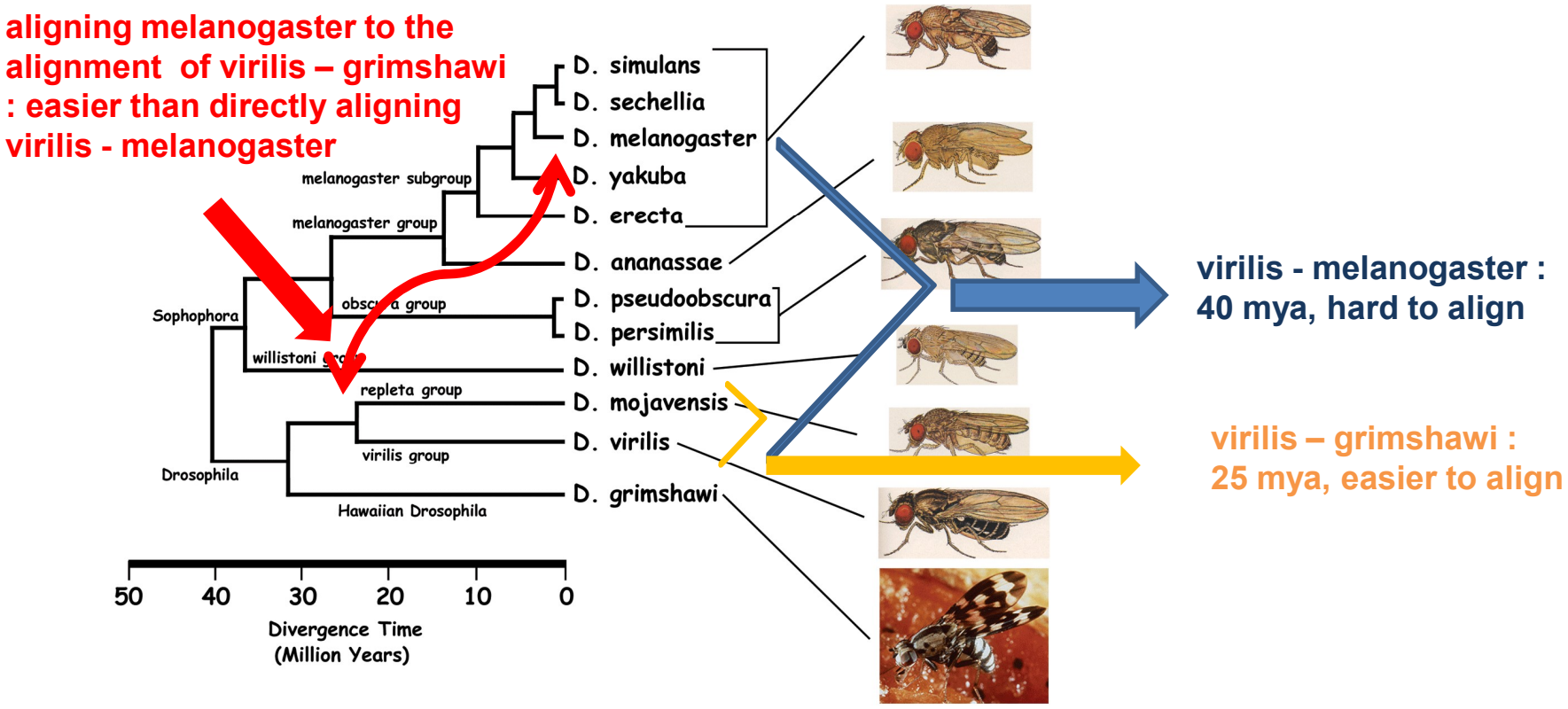
Notion of "bridging reads"

proaktive.co.uk

Bresler et al, RECOMB – Seq 2013

Sequencing intermediate species

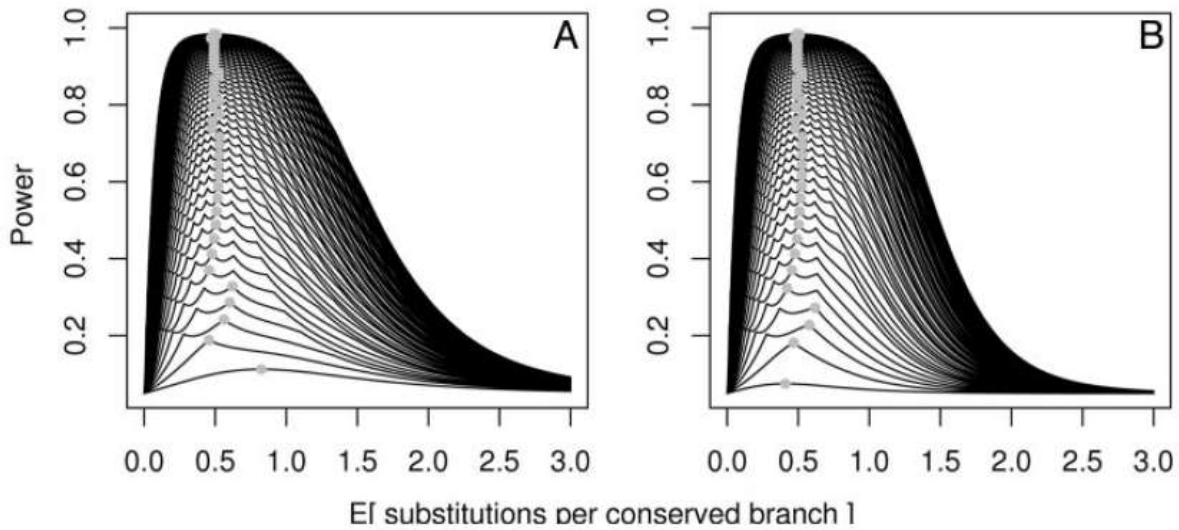
- Helps in progressive alignments, evolutionary analysis



arthropods.eugenes.org

Sequencing intermediate species

- A white paper for choosing the next white paper to write (which species to sequence next ?)



Power to detect conservation as a function of common branch length for the fully observed (*A*) and hidden-ancestor (*B*) SSTs,

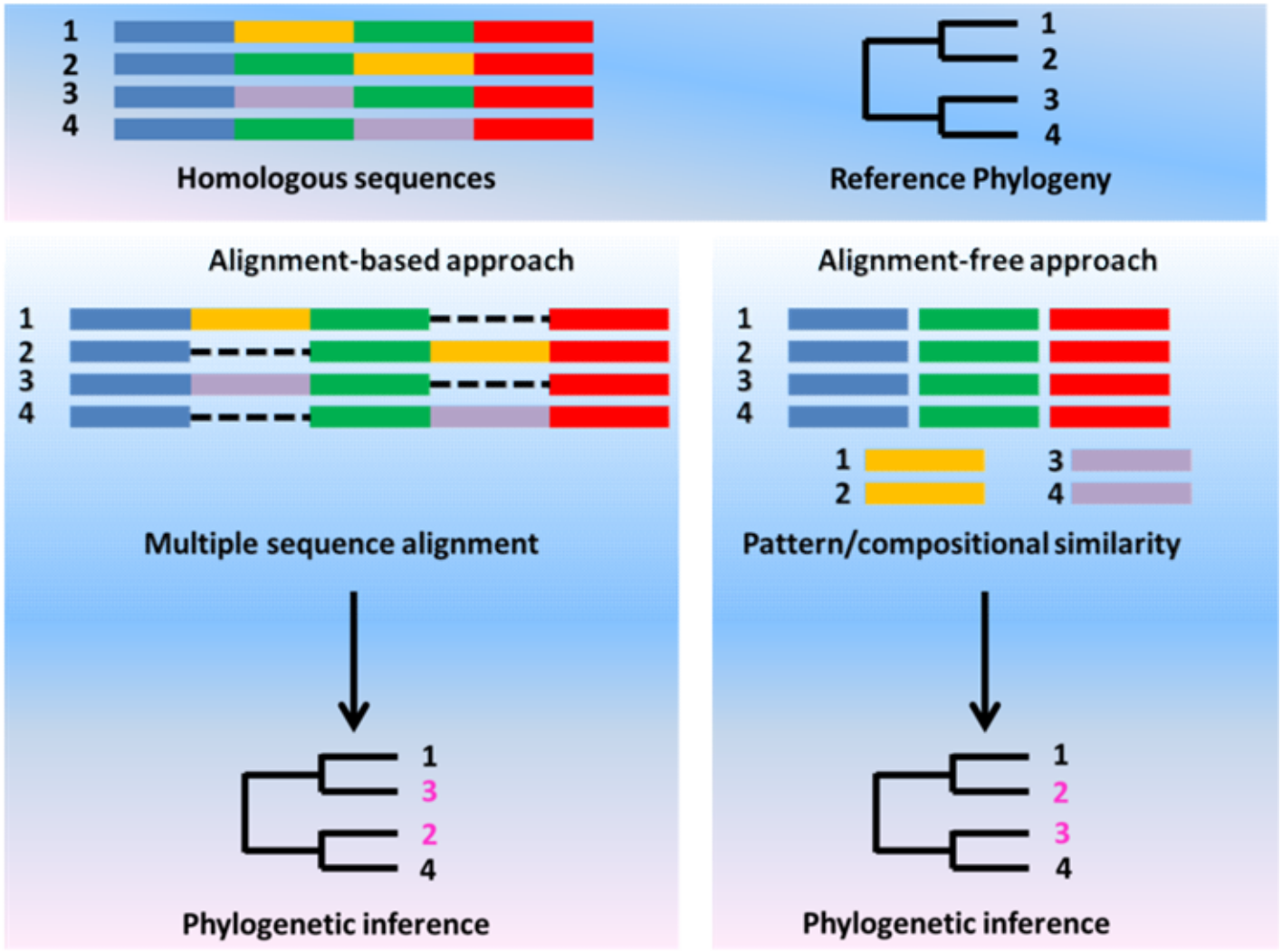
Subtree power analysis and species selection for comparative genomics

Jon D. McAuliffe[†], Michael I. Jordan[†‡], and Lior Pachter[§¶]

Departments of [†]Statistics and [§]Mathematics and [‡]Division of Computer Science, University of California, Berkeley, CA 94720

Communicated by Peter J. Bickel, University of California, Berkeley, CA, April 6, 2005 (received for review December 13, 2004)**PNAS**

A world without alignments



Wikipedia

Further reading

- Review papers
  - **A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives ,** Thompson et al, PLoS One 2011
  - **Profile hidden Markov models**, SR Eddy, Bioinformatics, 1998
  - **Recent progress in multiple sequence alignment: a survey**, C Notredame, Pharmacogenomics, 2002