

**Genome 559:
Introduction to Statistical and
Computational Genomics**

**Lecture 15a
Multiple Sequence Alignment
Larry Ruzzo**

Multiple Alignment: Motivations

Common structure, function, or origin may be only weakly reflected in sequence; multiple comparisons may highlight weak signal

Major uses

- represent protein, RNA families

- represent & identify conserved seq features

- “whole genome” alignments

Ribosomal Protein L10E

Accession	Species	Sequence	Position
Q5E940	BOVIN	-----MPREDRATWKS [*] SNYFLKIIQLLDDYPKCFIVGADNVGS [:] SKOMQOIRMSLRGK-AVVL [*] LMGKNTMMRKAIRGHLENN--PALE	76
RLA0	HUMAN	-----MPREDRATWKS [*] SNYFLKIIQLLDDYPKCFIVGADNVGS [:] SKOMQOIRMSLRGK-AVVL [*] LMGKNTMMRKAIRGHLENN--PALE	76
RLA0	MOUSE	-----MPREDRATWKS [*] SNYFLKIIQLLDDYPKCFIVGADNVGS [:] SKOMQOIRMSLRGK-AVVL [*] LMGKNTMMRKAIRGHLENN--PALE	76
RLA0	RAT	-----MPREDRATWKS [*] SNYFLKIIQLLDDYPKCFIVGADNVGS [:] SKOMQOIRMSLRGK-AVVL [*] LMGKNTMMRKAIRGHLENN--PALE	76
RLA0	CHICK	-----MPREDRATWKS [*] SNYFMKIIQLLDDYPKCFVVGADNVGS [:] SKOMQOIRMSLRGK-AVVL [*] LMGKNTMMRKAIRGHLENN--PALE	76
RLA0	RANSY	-----MPREDRATWKS [*] SNYFLKIIQLLDDYPKCFIVGADNVGS [:] SKOMQOIRMSLRGK-AVVL [*] LMGKNTMMRKAIRGHLENN--SALE	76
Q7ZUG3	BRARE	-----MPREDRATWKS [*] SNYFLKIIQLLDDYPKCFIVGADNVGS [:] SKOMQOIRMSLRGK-AVVL [*] LMGKNTMMRKAIRGHLENN--PALE	76
RLA0	ICTPU	-----MPREDRATWKS [*] SNYFLKIIQLLNDYPKCFIVGADNVGS [:] SKOMQOIRMSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0	DROME	-----MVRENKAAWKAQYFIKVV [*] ELFDEFPKCFIVGADNVGS [:] SKOMQOIRMSLRGL-AVVL [*] LMGKNTMMRKAIRGHLENN--POLE	76
RLA0	DICDI	-----MSGAG-SKRKKLFIEKATKLF [:] TTYDKMIVAEAD [:] FVGS [:] SQLOKIRKSI [:] IRGI-GAVLMGKKT [:] MIRKVI [:] RD [:] LADSK--PELD	75
Q54LP0	DICDI	-----MSGAG-SKRKNVFIEKATKLF [:] TTYDKMIVAEAD [:] FVGS [:] SQLOKIRKSI [:] IRGI-GAVLMGKKT [:] MIRKVI [:] RD [:] LADSK--PELD	75
RLA0	PLAF8	-----MAKLSKQKKQMYIEKLS [:] SLIQQYSKILIVHVDNVGS [:] SNQMASVRKSLRGK-ATILMGKNT [:] IR [:] TALK [:] KNLQAV--POIE	76
RLA0	SULAC	-----MIGLAVTTTKKIAKWKVDEVAELTEK [:] LK [:] TH [:] KTI [:] IANIEGFPADKLHEIRK [:] LRGK-ADIKVT [:] KN [:] NLF [:] NI [:] ALK [:] NAG----YDTK	79
RLA0	SULTO	-----MRIMAVITQERKIAKWKIEEVKELEK [:] LR [:] EYHT [:] II [:] ANIEGFPADKLHDIRK [:] MRGM-AEIKVT [:] KN [:] TLF [:] GIAAKNAG----LDVS	80
RLA0	SULSO	-----MKRLALALKQRKVASWKLEVEKELTE [:] LK [:] SN [:] TIL [:] IGNLEGFADKLHEIRK [:] LRGK-ATIKVT [:] KN [:] TLF [:] KIAAKNAG----IDIE	80
RLA0	AERPE	MSVVSIVGQMYKREKPIPEWKTLMLELEEL [:] FSK [:] HR [:] VVLFADLTGTPTFVVQ [:] RVR [:] KKLWKK-Y [:] PMM [:] VAK [:] KRIILRAMKAAGLE---LDDN	86
RLA0	PYRAE	-MMLAIGKRRYVRTRQYPARKVKIVSEATE [:] LLQ [:] KY [:] PV [:] FL [:] DL [:] HGL [:] SSRILHE [:] YR [:] RLRRY-G [:] VIK [:] IK [:] P [:] TLF [:] KIAFT [:] KVYGG---IPAE	85
RLA0	METAC	-----MAEERHTEHIPQWKKDEIENIKELIQ [:] SH [:] KV [:] FG [:] MV [:] GIEGILATKMKIRRD [:] LKDV-AVL [:] KV [:] SR [:] N [:] TL [:] TERALNQLG----ETIP	78
RLA0	METMA	-----MAEERHTEHIPQWKKDEIENIKELIQ [:] SH [:] KV [:] FG [:] MV [:] RIE [:] GILATKMKIRRD [:] LKDV-AVL [:] KV [:] SR [:] N [:] TL [:] TERALNQLG----ESIP	78
RLA0	ARCFU	-----MAAVRGS--PPEYKVRAVEEIKRMIS [:] SK [:] PV [:] VAV [:] SFR [:] NVPAGOMKIRRE [:] FRGK-AEIK [:] V [:] KN [:] TLLERALD [:] DALG----GDYL	75
RLA0	METKA	MAVKAKGQPPSGYE [*] PKVAEWKRREVKELKELMDE [:] YENV [:] GL [:] VD [:] LEGIPAPQLQEIRAKLRERD [:] TIIRMSRNTLMRIA [:] LEEK [:] LDER--PELE	88
RLA0	METHH	-----MAHVAEWKKKEVQELHDLIKGYE [:] VV [:] GIANLADIPARQLQKMRQT [:] LRDS-ALIRMSK [:] KT [:] LISL [:] ALEK [:] AGREL--ENVD	74
RLA0	METTL	-----MITAESEHKIAPWKIEEVNKLKEL [:] LK [:] NG [:] Q [:] IVAL [:] VDM [:] MEVPARQLQEIRDKIR-GT [:] M [:] TL [:] KMS [:] RNT [:] LIE [:] RAI [:] KEVAEETGNPEFA	82
RLA0	METVA	-----MIDAKSEHKIAPWKIEEVNKLKEL [:] LK [:] NS [:] VIAL [:] IDM [:] MEVPARQLQEIRDKIR-DQ [:] M [:] TL [:] KMS [:] RNT [:] LIE [:] RAVEEVAEETGNPEFA	82
RLA0	METJA	-----METKVKAHVAPWKIEEVKTLKGLIK [:] SK [:] PV [:] VAV [:] VD [:] MMDVPAPQLQEIRDKIR-DK [:] V [:] KL [:] RS [:] RNT [:] LIE [:] RA [:] KEAAEELNNPKLA	81
RLA0	PYRAB	-----MAHVAEWKKKEVEELANLIKSY [:] PVIAL [:] VD [:] VSSMPAYPLSQMRR [:] LIRE [:] NG [:] LL [:] RV [:] SRNT [:] LIE [:] LAI [:] KKAAQELGKPELE	77
RLA0	PYRHO	-----MAHVAEWKKKEVEELAKLIKSY [:] PVIAL [:] VD [:] VSSMPAYPLSQMRR [:] LIRE [:] NG [:] LL [:] RV [:] SRNT [:] LIE [:] LAI [:] KKAAKELGKPELE	77
RLA0	PYRFU	-----MAHVAEWKKKEVEELANLIKSY [:] PVVAL [:] VD [:] VSSMPAYPLSQMRR [:] LIRE [:] NN [:] GL [:] RV [:] SRNT [:] LIE [:] LAI [:] KKVAQELGKPELE	77
RLA0	PYRKO	-----MAHVAEWKKKEVEELANIKSY [:] PVIAL [:] VD [:] VAGVPAYPLSKMRDK [:] LR-GK [:] ALL [:] RV [:] SRNT [:] LIE [:] LAI [:] KRAAQELGQPELE	76
RLA0	HALMA	-----MSAESERKTETIPEWQQEEVD [:] AIV [:] EMIESY [:] ESV [:] GV [:] VNIAGIPSRQLQDMRRD [:] LHGT-AEL [:] RV [:] SRNT [:] LLE [:] RALDDVD----DGLE	79
RLA0	HALVO	-----MSESEVRQTEVIPQWKREEVDEL [:] VDFIESY [:] ESV [:] GV [:] VGVAGIPSRQLQSMRRE [:] LHGS-AAV [:] RS [:] RNT [:] LVN [:] RALDEVN----DGFE	79
RLA0	HALSA	-----MSAEEQRTTEEVPEWKRQEV [:] AEL [:] V [:] DL [:] LETY [:] DSV [:] GV [:] VNV [:] TGIPSKQLQDMRRGLH [:] GQ-AAL [:] RS [:] RNT [:] LL [:] V [:] RALEEAG----DGLD	79
RLA0	THEAC	-----MKEVSQKKKELVNEIT [:] ORIKAS [:] RSVAIV [:] DTAGIR [:] TROI [:] QDIR [:] GKNR [:] GK-INL [:] KV [:] IK [:] K [:] TL [:] LF [:] KAL [:] ENLGD----EKLS	72
RLA0	THEVO	-----MRKINPKKKEIVSELAQDIT [:] SKAV [:] AV [:] VDIK [:] GV [:] RTROMQDIRAKNRDK-VK [:] IK [:] V [:] KK [:] TL [:] LF [:] KAL [:] DSIND----EKLT	72
RLA0	PICTO	-----MTEPAQWKIDFVKNLENEINSR [:] KVA [:] AV [:] IVS [:] IKGLR [:] NN [:] EFQ [:] KIRNS [:] IRDK-ARI [:] KV [:] SRARLLRLAIENTGK----NNIV	72
ruler		1.....10.....20.....30.....40.....50.....60.....70.....80.....90	

First 90 Residues, Human to Archaea

Alignment of 7 globins.

```

Helix          AAAAAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBBBCCCCCCCCCCC
HBA_HUMAN     -----VLSPADKTNVKAAWGKVG--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN     -----VHLTPEEKSAVTALWGKV---NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA     -----VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP    -----LSADQISTVQASFDKVKG-----DPVGILYAVFKADPSIMAKFTQF
GLB5_PETMA    PIVDTGSVAPLSAAEKTIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKF
LGB2_LUPLU    -----GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI    -----GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-F
Consensus     Ls.... v a W kv . . g . L.. f . P . F F

```

```

Helix          DDDDDDDDEEEEEEEEEEEEEEEEEEEEEEE   FFFFFFFFFFFFFF
HBA_HUMAN     -DLS-----HGSAQVKGHGKKVADALTNVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN     GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFFATLSELHCDKL-
MYG_PHYCA     KHLKTEAEMKASEDLKKGVTVLTALGAILKK---K-GHHEAELKPLAQSHATKH-
GLB3_CHITP    AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA    KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU    LK-GTSEVPQNNPELQAHAGKVFKLVEAAIQLVQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI    SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAQVVRHKGYGN
Consensus     . t . . . v..Hg kv. a a...l d . a l. l H .

```

```

Helix          FGGGGGGGGGGGGGGGGGGGGGG   HHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN     -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR-----
HBB_HUMAN     -HVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAAYQKVAVAGVANALAHKYH-----
MYG_PHYCA     -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP    --VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-----
GLB5_PETMA    -QVDPQYFKVLAAVIADTVAAG-----DAGFEKLSMICILLRSAY-----
LGB2_LUPLU    --VADAHFPVVKAILKTIKEVVGAKWSEELNSAWTIAAYDELAIVIKKEMNDAA---
GLB1_GLYDI    KHIKAQYFEPLGASLLSAMEHRIGGKMNAAKDAWAAAYADISGALISGLQS-----
Consensus     v. f l . . . . . f . aa. k. . l sky

```

Human, whale, midge, lamprey, lupin, bloodworm.

A-H mark 8 alpha helices. Consensus line: upper case = 6/7, lower = 4/7, dot=3/7.

Multiple Alignments: Key Issues

Scoring:

How to evaluate a proposed alignment

Computational demands:

How to do it in reasonable time

Multiple Alignment Scoring

A Key Issue

Varying goals, methods (& controversy)

Ideal is perhaps phylogenetic, position specific, but typically too slow, too many parameters

Most methods assume independence between columns, so you can score them separately

(Very inappropriate for RNA alignments, e.g.)

Multiple Alignment Scoring within one column

Two common ways:

1. Min Entropy – if you assume a star phylogeny with long branches, positions in one column are independent and a proper probabilistic model reduces to per-column entropy (akin to last week). Intuitively sensible; favors alignments with less in-column variability

2. SP score: Sum of Pairs

E.g., use BLOSUM62 score
between all pairs of sequences

$$\begin{array}{l} abcde \\ ac-de \\ xccxd \end{array} \triangleright \sum_{i < j} D(S_i, S_j)$$

It is *not* theoretically justifiable, but is easy, not terrible

Optimal SP Alignment via DP

k sequences of length n

$(n+1) \times (n+1) \times \dots \times (n+1)$ k-dim array

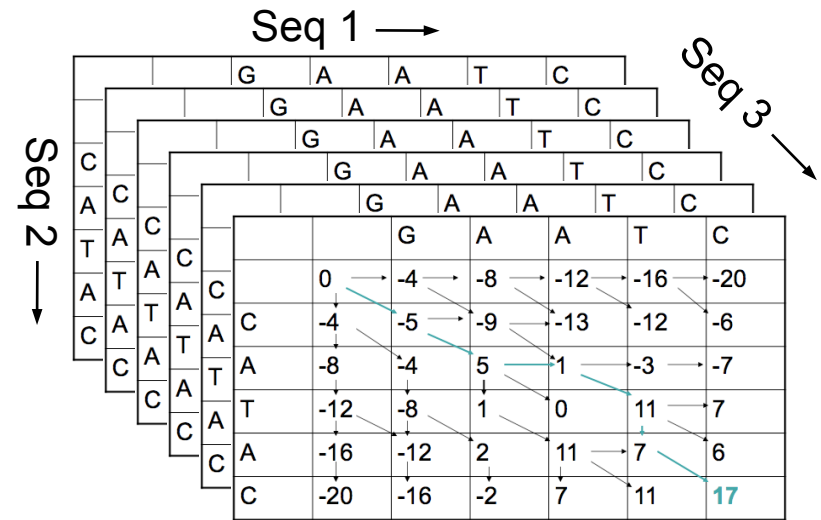
Max of $2^k - 1$ neighbors per cell; $(n+1)^k$ cells

Time: at least $(2n)^k$

Want n, k 10's to 100's

Unlikely to do dramatically better –
it's “NP-hard”

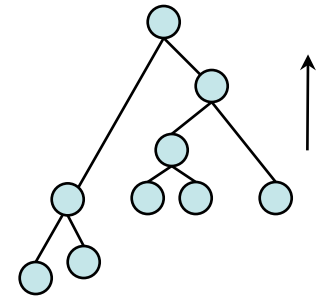
Wang & Jiang, '94



E.g., n = 100
 10^6 ops/sec

k	Time
2	40 ms
3	8 sec
4	.5 hr
5	100 hrs
6	2 years

Common Heuristic: Progressive Alignment



Pick a “guide tree”

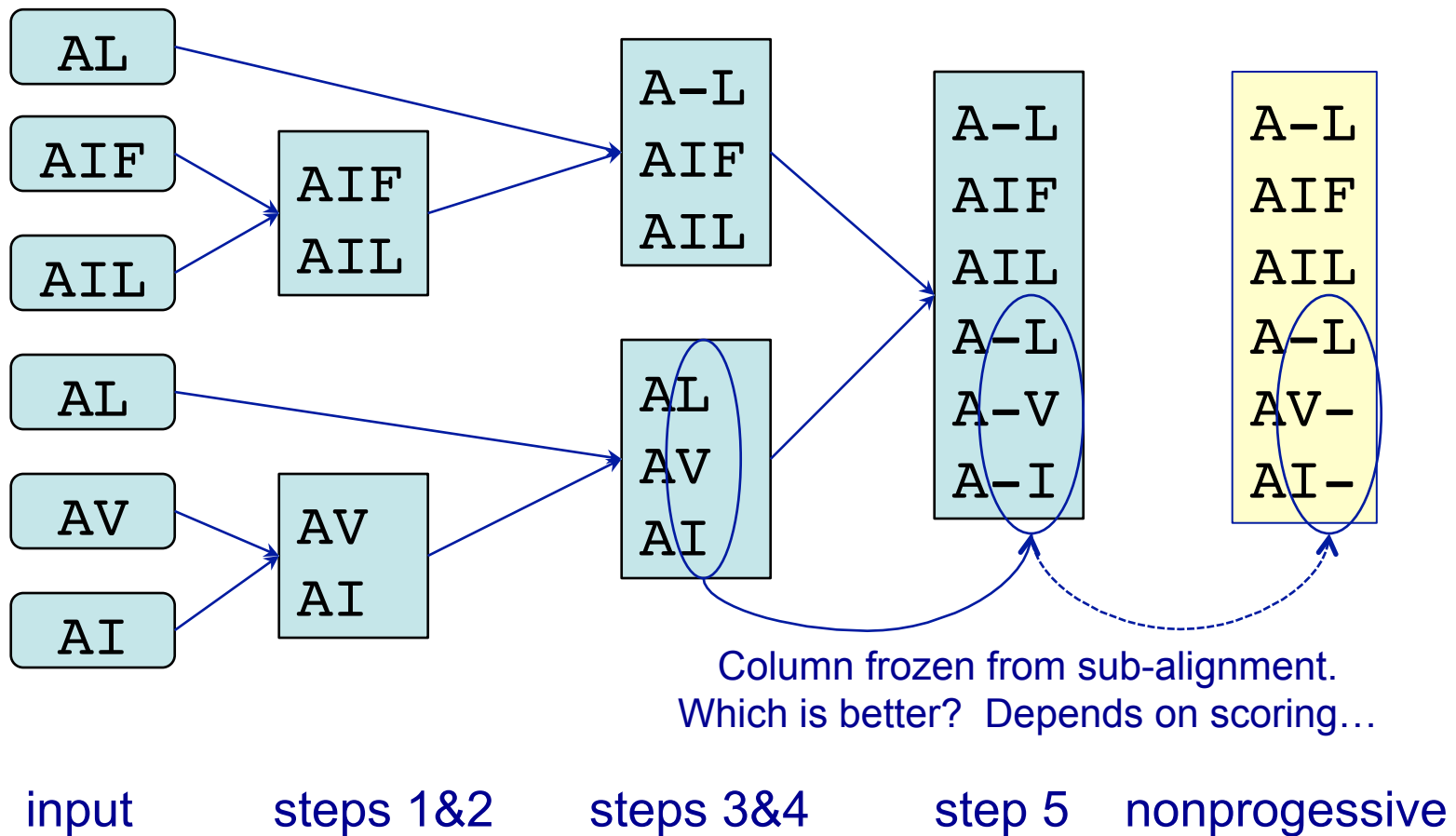
phylogeny is ideal, but expensive

quicker alternative: get pairwise alignment scores,
convert to distances, use, e.g., “neighbor joining”

Work up tree, leaves to root, doing pairwise
alignments

(Many implementations, many variants, e.g. ClustalW)

Progressive Alignment



BLOSUM 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-2	-1	1	0	-3	-2	0	
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Summary

Very important problem

Scoring is very difficult to get right

Fast, exact solutions appear impossible (even with simple scoring schemes)

Many heuristics have been tried

Useful methods like ClustalW are available

Still an open field

e.g., “genome scale” and RNA especially challenging

Iterative Pairwise Alignment (More Detail)

align some pair

while not done

Pick an unaligned string “near” some aligned one(s)

Align with the **profile** of the previously aligned group

Resulting new spaces inserted in all

Many variants

Summarizing a Multiple Alignment

A *profile* of a multiple alignment gives letter frequencies per column

a b a
a b -
- b a
c a -

	col 1	col 2	col 3
a	50%	25%	50%
b	0%	75%	0%
c	25%	0%	0%
-	25%	0%	50%

Alternatively, use log likelihood ratios

$p_i(a)$ = fraction of a's in col i

$p(a)$ = fraction of a's overall

$\log p_i(a)/p(a)$

Aligning to a Phylogenetic Tree

Given a tree with a sequence at each leaf,
assign labels to internal nodes so as to

minimize $\sum_{\text{edges } (i,j)} D(S_i, S_j)$

[Note: NOT SP score]

Also NP-Complete

Poly time approximation within 2 x possible;
better with more time (PTAS)

Multiple Sequence Alignment

Defn: An *alignment* of S_1, S_2, \dots, S_k ,
is a set of strings S'_1, S'_2, \dots, S'_k , (with spaces) s.t.

(1) $|S'_1| = |S'_2| = \dots = |S'_k|$, and

(2) removing all spaces leaves S_1, S_2, \dots, S_k

a c b c d b

c a d b d

a c a b c d

a c - - b c d b

- c a d b - d -

a c a - b c d -

Multiple Alignment Scoring

Varying goals

Varying methods (& controversy)

3 examples:

Consensus string;
sum distances to it

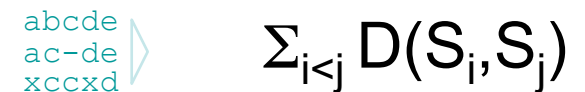


Align to (evolutionary) tree;
sum edges



SP score:

Sum of Pairs



NP-Complete Problems

A problem X is *NP-Complete* if

(1) it's in NP, and

(2) a poly time algorithm for X would give a poly time algorithm for *all* problems in NP

Thousands known; superficially very different
- algebra, geometry, cs, bio, ...

Smart Money betting against $P = NP$