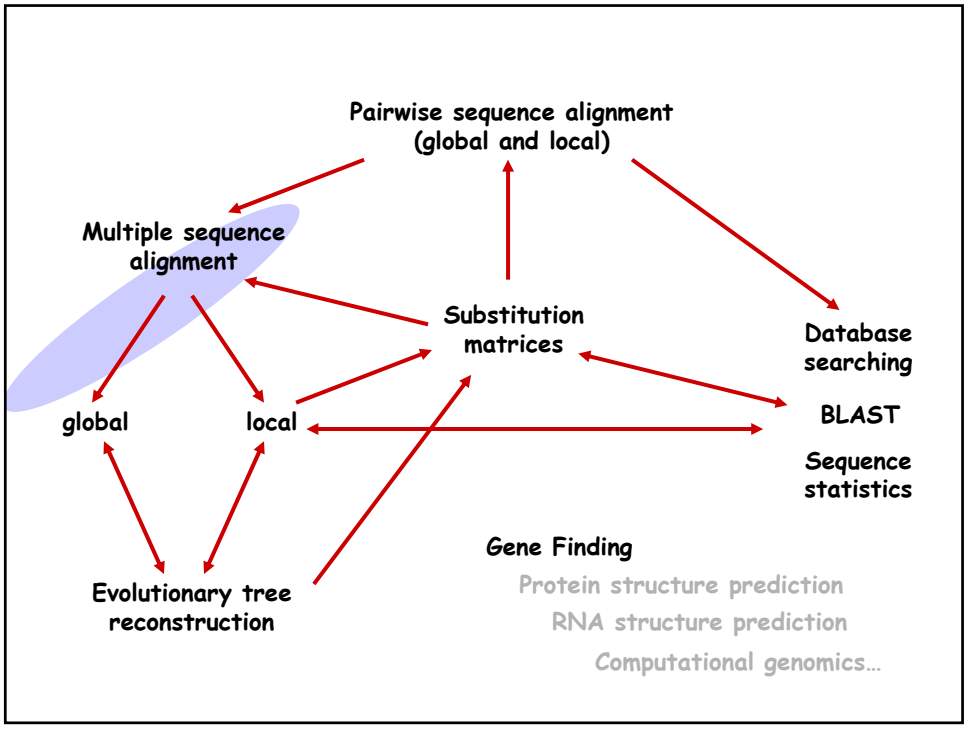


Global Multiple Sequence Alignment

```
HUMAN   MKWVTFISLL FLFSSAYSRG V..FRRDA.H KSEVAHRFKD LGEENFKALV
RABIT   MKWVTFISLL FLFSSAYSRG V..FRREA.H KSEIAHRFND VGEEHFIGLV
PIG     ~~WVTFISLL FLFSSAYSRG V..FRRDT.Y KSEIAHRFKD LGEQYFKGLV
CHICK   MKWVTLISFI FLFSSATSRN LQRFARDAEH KSEIAHRYND LKEETFKAVA
```

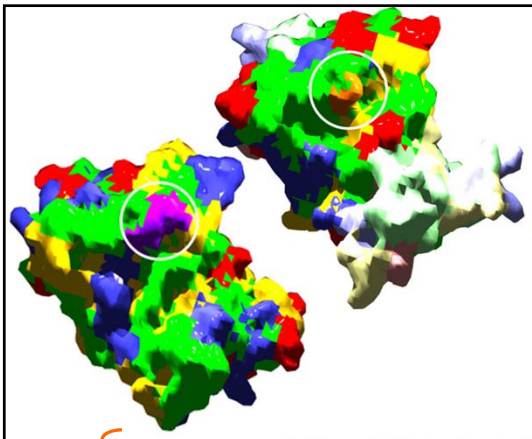
Align *k* sequences, so that residues in each column share a property of interest:

- a common ancestor
- a structural or functional role

Applications
Global Multiple Alignment

```
HUMAN   MKWVTFISLL FLFSSAYSRG V..FRRDA.H KSEVAHRFKD LGEENFKALV
RABIT   MKWVTFISLL FLFSSAYSRG V..FRREA.H KSEIAHRFND VGEEHFIGLV
PIG     ~~WVTFISLL FLFSSAYSRG V..FRRDT.Y KSEIAHRFKD LGEQYFKGLV
CHICK   MKWVTLISFI FLFSSATSRN LQRFARDAEH KSEIAHRYND LKEETFKAVA
```

- Protein structure and function
- RNA structure
- Evolutionary tree reconstruction



...replacement of the central arginine of the "arginine finger"… confirmed a local alteration in shape and charge of the conserved GTPase interaction interface, while the overall conformation of the molecule appears preserved.

```
OstaFH1     ESLRTHGVEL GTKLFAEAMV NKSHGARALE EGEIFIACDG
OsluFH1     GSLKHRGMEL GAELFTRASE ADALSDKAVS EGETYLAV-G
Phypa7      MLFLFAAIEV -EGLFSIPIP LEDVLEVKKC FDSCDPLETK
Selmo2a/b   ATLQIHADKL -DKLFCAPVP VHEVLRLRSL LIQGHGQQEL
GgGRAF      HAVETRGINE -QGLYRIVGV NSRVQKLLSI LMDPETEICA
HsRacGAP    REIEARGLKS -EGLYRVSGF TEHIEDVKMA F-DRDGEKAD
Hs p85      EAIEKKGLEC -STLYRTQSS SN-LAELRQL L-DCDTPSVD
Hs p50      AYLQAHALTT -EGIFRRSAN TQVVREVQQK Y-NMGLPVDF
```

Grunt *et al. BMC Evolutionary Biology* 2008 **8**:115

A MULTIPLE SEQUENCE ALIGNMENT EXAMPLE

Ribosome: an RNA/protein complex

rpS14: a ribosomal protein in yeast

Goal: Determine residues responsible for binding rpS14 to ribosomal RNA

Known:

- Sequence of rpS14
- Structure of homolog in bacteria
- Sequences in many species

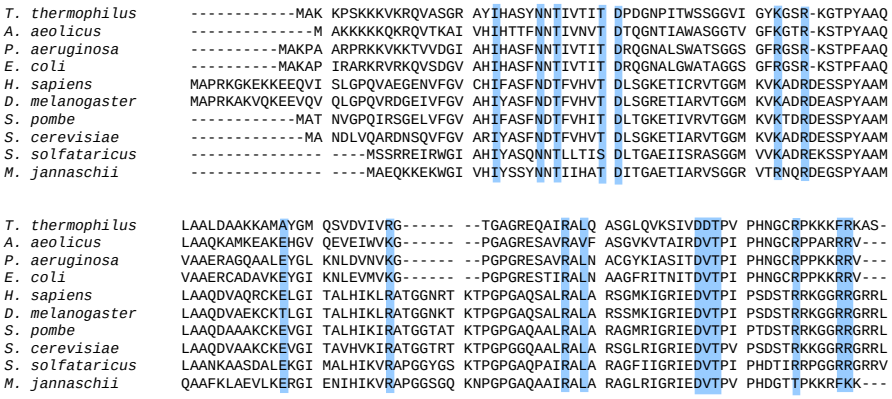
Pam Bush, PhD CMU, 02

A MULTIPLE SEQUENCE ALIGNMENT EXAMPLE

Strategy: alanine scan

- Find "likely" candidate amino acids
- Replace candidates with alanine
- Hope that
  - Alanine preserves structure
  - Alanine will destroy binding

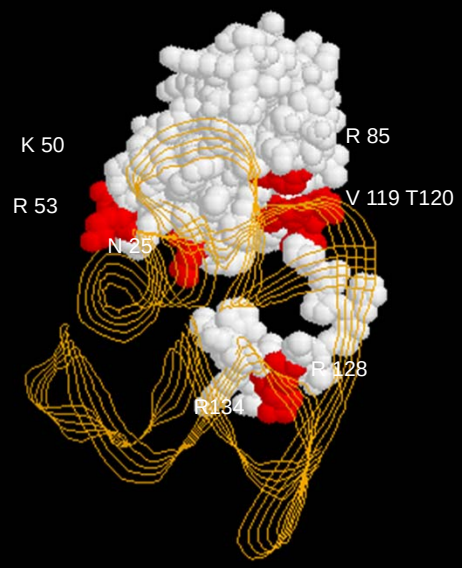
Seventeen residues are conserved in rpS14



Criteria for selecting conserved amino acids:

- Conserved among all three phylogenetic groups
- Conserved in at least 90% sequences analyzed, allowing for conservative substitution K↔R
- No residue could be an alanine, proline or glycine

Eight conserved residues are on the surface of the bacterial protein interacting with the rRNA



…and these are distributed throughout the protein sequence

```
T. thermophilus      ------------MAK KPSKKKVKRQVASGR AYIHASYNNTIVTIT DPDGNPITWSSGGVI GYKGSR-KGTPYAAQ
A. aeolicus          -------------M AKKKKKQKRQVTKAI VHIHTTFNNTIVNVT DTQGNTIAWASGGTV GFKGTR-KSTPYAAQ
P. aeruginosa        ----------MAKPA ARPRKKVKKTVVDGI AHIHASFNNTIVTIT DRQGNALSWATSGGS GFRGSR-KSTPFAAQ
E. coli              ----------MAKAP IRARKRVRKQVSDGV AHIHASFNNTIVTIT DRQGNALGWATAGGS GFRGSR-KSTPFAAQ
H. sapiens          MAPRKGKEKKEEQVI SLGPQVAEGENVFGV CHIFASFNDTFVHVT DLSGKETICRVTGGM KVKADRDESSPYAAM
D. melanogaster     MAPRKAKVQKEEVQV QLGPQVRDGEIVFGV AHIYASFNDTFVHVT DLSGRETIARVTGGM KVKADRDEASPYAAM
S. pombe             ------------MAT NVGPQIRSGELVFGV AHIFASFNDTFVHIT DLTGKETIVRVTGGM KVKTDRDESSPYAAM
S. cerevisiae        -------------MA NDLVQARDNSQVFGV ARIYASFNDTFVHVT DLSGKETIARVTGGM KVKADRDESSPYAAM
S. solfataricus      --------------- ----MSSRREIRWGI AHIYASQNNTLLTIS DLTGAEIISRASGGM VVKADREKSSPYAAM
M. jannaschii        --------------- ----MAEQKKEKWGI VHIYSSYNNTIIHAT DITGAETIARVSGGR VTRNQRDEGSPYAAM

T. thermophilus     LAALDAAKKAMAYGM QSVDVIVRG------ --TGAGREQAIRALQ ASGLQVKSIVDDTPV PHNGCRPKKKFRKAS-
A. aeolicus         LAAQKAMKEAKEHGV QEVEIWVKG------ --PGAGRESAVRAVF ASGVKVTAIRDVTPI PHNGCRPPARRRV---
P. aeruginosa       VAAERAGQAALEYGL KNLDVNVKG------ --PGPGRESAVRALN ACGYKIASITDVTPI PHNGCRPPKKRRV---
E. coli             VAAERCADAVKEYGI KNLEVMVKG------ --PGPGRESTIRALN AAGFRITNITDVTPI PHNGCRPPKKRRV---
H. sapiens          LAAQDVAQRCKELGI TALHIKLRATGGNRT KTPGPGAQSALRALA RSGMKIGRIEDVTPI PSDSTRRKGGRRGRRL
D. melanogaster     LAAQDVAEKCKTLGI TALHIKLRATGGNKT KTPGPGAQSALRALA RSSMKIGRIEDVTPI PSDSTRRKGGRRGRRL
S. pombe            LAAQDAAAKCKEVGI TALHIKIRATGGTAT KTPGPGAQAALRALA RAGMRIGRIEDVTPI PTDSTRRKGGRRGRRL
S. cerevisiae       LAAQDVAAKCKEVGI TAVHVKIRATGGTRT KTPGPGGQAALRALA RSGLRIGRIEDVTPV PSDSTRKKGGRRGRRL
S. solfataricus     LAANKAASDALEKGI MALHIKVRAPGGYGS KTPGPGAQPAIRALA RAGFIIGRIEDVTPI PHDTIRRPGGRRGRRV
M. jannaschii       QAAFKLAEVLKERGI ENIHIKVRAPGGSGQ KNPGPGAQAAIRALA RAGLRIGRIEDVTPV PHDGTTPKKRFKK---
```

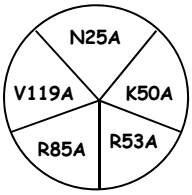
About the changes to alanine of the conserved residues:

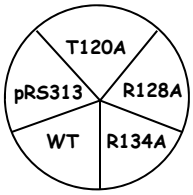
N25, K50, R53, R85, V119, T120, R128, R134

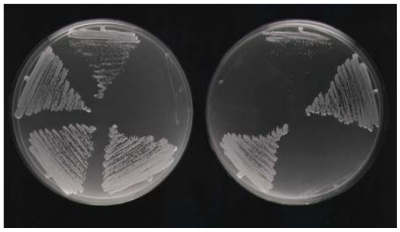
- Like most ribosomal proteins, high percentage of basic residues
- Change to alanine is a big charge difference
- Alanines on surface usually do not have a big structural change

The predicted secondary structure of rpS14 does not change significantly when the conserved residues are changed to alanine

Functional assays suggest that K50A and R134 play a role in rpS14/RNA binding



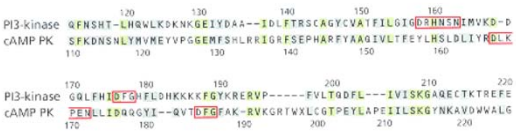




Alignment of the PI3 and cAMP dependent kinases

Local:

Global:



Multiple alignment with other kinases:

From Zvelebil and Baum, "Understanding Bioinformatics"

Global Multiple Sequence Alignment

Given sequences $s_1 \ldots s_k$ of lengths $n_1 \ldots n_k$

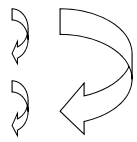
seek $s'_1 \ldots s'_k$ of length $l \geq max\{n_i\}$ such that

- Obtain s_i from s'_i by removing gaps
- No column contains all gaps
- The score of the alignment is optimal

Scoring function: Sum-of-Pairs

$$Score = \sum^{k}_{a=1} \sum^{k}_{a=1} \sum_{b > a} p(s'_a[i], s'_b[i])$$

(1) **A_TT**
(2) **A_T_**
(3) **ACAT**



$p[\_,\_] = 0$

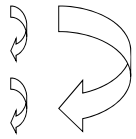
$$Score = p[s_1,s_2] + p[s_1,s_3] + p[s_2,s_3]$$
$$= 0 + g + g = 2g$$

Note: this example uses a similarity function. We can also use Sum-of-Pairs with distance scoring.

Scoring function: Sum-of-Pairs

$$Score = \sum^{k}_{a=1} \sum^{k}_{a=1} \sum_{b > a} p(s'_a[i], s'_b[i])$$

(1) **A_TT**
(2) **A_T_**
(3) **ACAT**



$p[\_,\_] = 0$

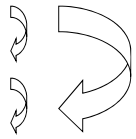
$$Score = p[s_1,s_2] + p[s_1,s_3] + p[s_2,s_3]$$
$$= M + m + m = 2m + M$$

Note: this example uses a similarity function. We can also use Sum-of-Pairs with distance scoring.

Scoring function: Sum-of-Pairs

$$Score = \sum^{k}_{a=1} \sum^{k}_{a=1} \sum_{b > a} p(s'_a[i], s'_b[i])$$

(1) **A_TT**
(2) **A_T_**
(3) **ACAT**

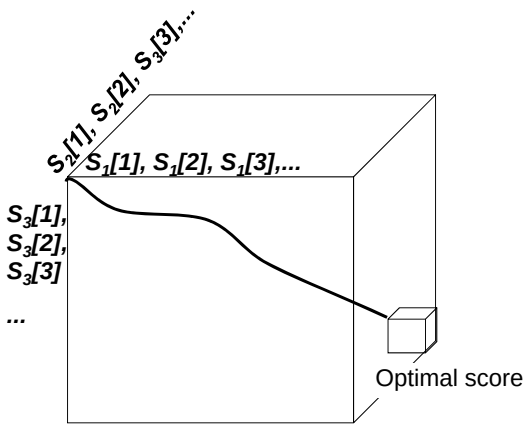


$p[\_,\_] = 0$

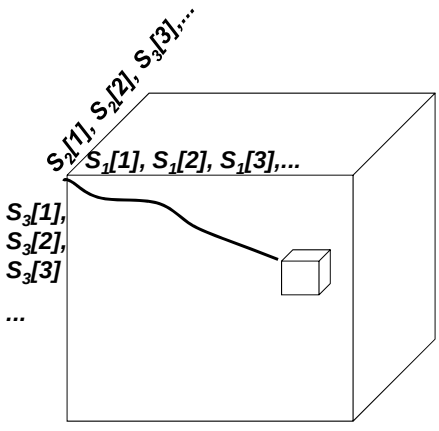
$$Score = p[s_1,s_2] + p[s_1,s_3] + p[s_2,s_3]$$
$$= g + M + g = 2g + M$$

Note: this example uses a similarity function. We can also use Sum-of-Pairs with distance scoring.

Dynamic Programming for Multiple Alignment



Dynamic Programming for Multiple Alignment



Limits:

~ k = 8 - 10 sequences

~ n = 500 residues

Each cell has $O(2^k)$ neighboring cells

Calculating the sum-of-pairs score for each neighbor is $O(k^2)$

Number of cells in matrix: $O(n^k)$

Total computational complexity:

$O(n^k 2^k k^2)$

MSA is NP-complete for Sum-of-Pairs scoring

Observations

1. A multiple alignment induces pairwise alignments

2. A column in the induced pairwise alignment may contain all gaps, even though no column in the MSA contains all gaps.

```
(1)  AG_CT
(2)  AG_CT
(3)  ACT_T
```

3. The pairwise alignments induced by the *optimal multiple alignment* are *not* the same as the *optimal pairwise alignments.*

Optimal Pairwise Alignments

```
(1)  ACT
(2)  AGT
```

1 substitution

Optimal Multiple Alignment

```
(1)  AC_T
(2)  A_GT
(3)  ACGT
```

2 indels

Although this costs more, it may be a biologically more realistic alignment

Since exact methods for MSA have exponential time complexity, heuristic approaches are used. Progressive alignment is the most commonly used.

Basic progressive alignment strategy:

- Compute *D,* a matrix of distances between all pairs of sequences
- From *D, c*onstruct a "guide tree" *T*
- Construct MSA by pairwise alignment of partial alignments ("profiles") guided by T
- Improve alignment by postprocessing steps.

Optimal Pairwise Alignments

(1) **ACTCAT**
(2) **AGTCAT**
(3) **ACGTCCT**

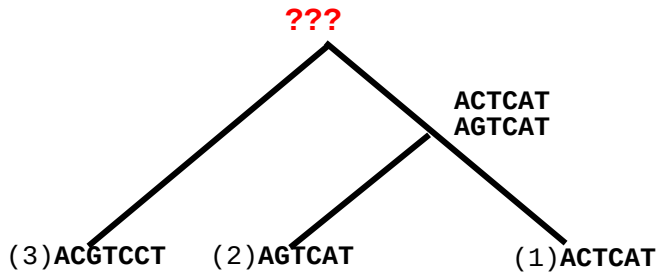
d(x,y) = 3
d(x,"_") = 2

(1) **ACTCAT**
(2) **AGTCAT** 3

(2) **A_GTCAT**
(3) **ACGTCCT** 5

(1) **AC_TCAT**
(3) **ACGTCCT** 5

Progressive Alignment



- Use *profile alignment* to merge sequences according to a guide tree.
- Typically, most closely related sequences are merged first.

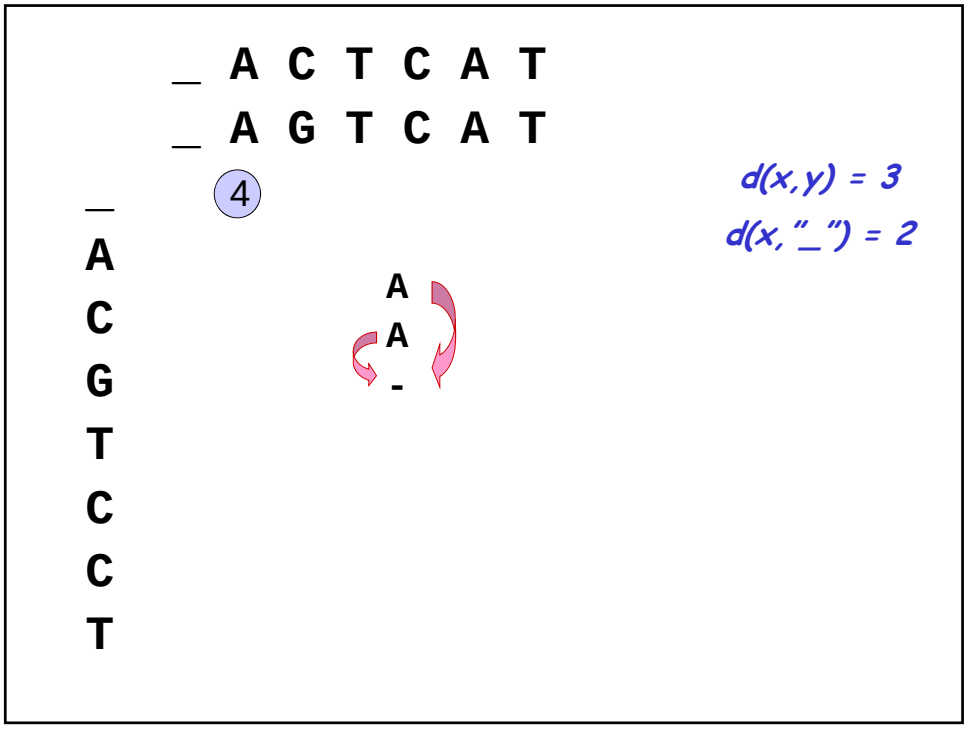
Merging strategy:

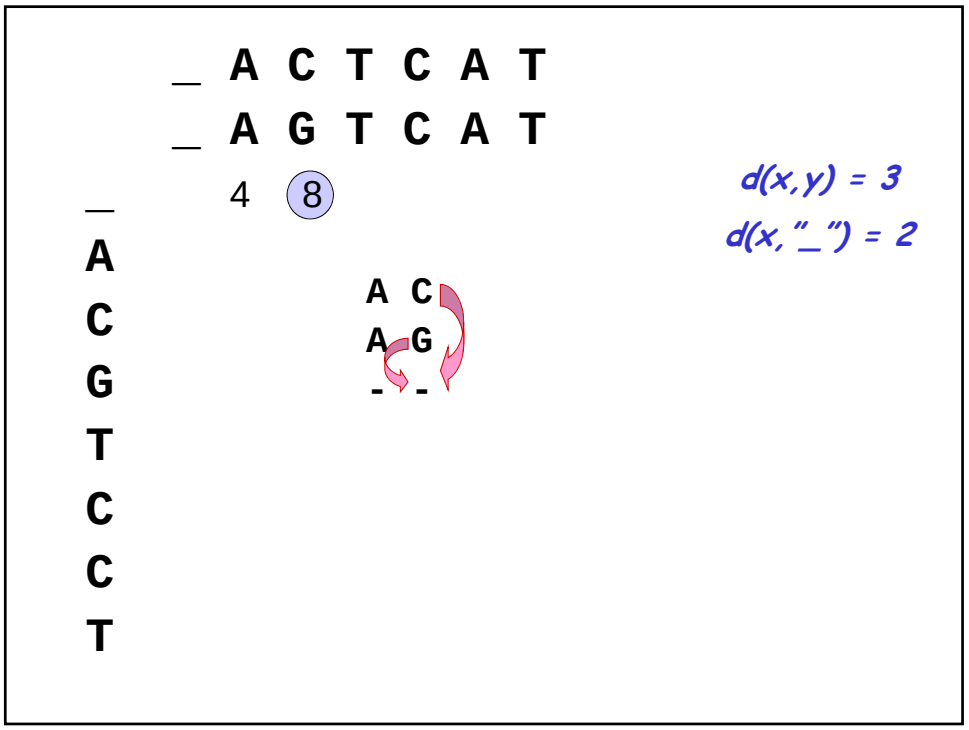
Align the profile (1,2) with sequence (3)

(1) **ACTCAT**
(2) **AGTCAT**
(3) **ACGTCCT**

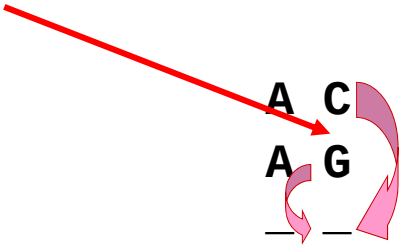
d(x,y) = 3
d(x,"_") = 2

(1)	**ACTCAT**	3
(2)	**AGTCAT**	
(2)	**A_GTCAT**	5
(3)	**ACGTCCT**	
(1)	**AC_TCAT**	5
(3)	**ACGTCCT**	


_ A C T C A T
_ A G T C A T
4
d(x,y) = 3
d(x,"_") = 2
_
A
C
G
T
C
C
T
A
A
-


_ A C T C A T
_ A G T C A T
4 8
d(x,y) = 3
d(x,"_") = 2
_
A
C
G
T
C
C
T
A C
A G
- -

Note: no penalty for mutations in the profile.
We paid for those in a previous step

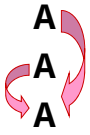


4+4

		_	A	C	T	C	A	T
		_	A	G	T	C	A	T
			4	8	12	16	20	24
_								
A	4							
C	8							
G	12							
T	16							
C	…							
C								
T								

d(x,y) = 3
d(x,"_") = 2

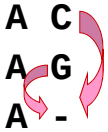
	_	A	C	T	C	A	T
	_	A	G	T	C	A	T
_		4	8	12	16	20	24
A	4	0					
C	8						
G	12						
T	16						
C	…						
C							
T							



d(x,y) = 3

d(x, "_") = 2

	_	A	C	T	C	A	T
	_	A	G	T	C	A	T
_		4	8	12	16	20	24
A	4	0	4				
C	8						
G	12						
T	16						
C	…						
C							
T							



d(x,y) = 3

d(x, "_") = 2

		_	A	C	T	C	A	T
		_	A	G	T	C	A	T
_		4	8	12	16	20	24	
A	4	0	4	8				
C	8							
G	12							
T	16							
C	…							
C								
T								

$d(x,y) = 3$

$d(x, "\_") = 2$

		_	A	C	T	C	A	T
		_	A	G	T	C	A	T
_		4	8	12	16	20	24	
A	4	0	4	8	12			
C	8							
G	12							
T	16							
C	…							
C								
T								

$d(x,y) = 3$

$d(x, "\_") = 2$

		_	A	C	T	C	A	T
		_	A	G	T	C	A	T
_			4	8	12	16	20	24
A	4	0	4	8	12	16		
C	8							
G	12							
T	16							
C	…							
C								
T								

$d(x,y) = 3$

$d(x, "\_") = 2$

		_	A	C	T	C	A	T
		_	A	G	T	C	A	T
_			4	8	12	16	20	24
A	4	0	4	8	12	16	20	
C	8	4						
G	12							
T	16							
C	…							
C								
T								

A _
A _
A C

$d(x,y) = 3$

$d(x, "\_") = 2$

	_	A	C	T	C	A	T
	_	A	G	T	C	A	T
_		4	8	12	16	20	24
A	4	0	4	8	12	16	20
C	8	4	3				
G	12						
T	16						
C	…						
C							
T							

A C
A G
A C

d(x,y) = 3
d(x,"_") = 2

	_	A	C	T	C	A	T
	_	A	G	T	C	A	T
_		4	8	12	16	20	24
A	4	**0**	4	8	12	16	20
C	8	**4**	**3**	7	8	12	16
G	12	8	**7**	9	12	14	18
T	16	12	11	**7**	11	15	14
C	20	16	15	11	**7**	11	15
C	24	20	19	15	11	**13**	17
T	28	24	23	19	15	17	**13**

d(x,y) = 3
d(x,"_") = 2

Optimal Pairwise Alignments

(1) ACTCAT
(2) AGTCAT

(2) A_GTCAT
(3) ACGTCCT

(1) AC_TCAT
(3) ACGTCCT

Progressive alignment

(1,2) + (3)

(3) ACGTCCT
(1) AC_TCAT
(2) AG_TCAT

4m+2g

An alternate alignment

(1) AC_TCAT
(2) A_GTCAT
(3) ACGTCCT

2m+4g

Optimal Pairwise Alignments

(1) ACTCAT
(2) AGTCAT

(2) A_GTCAT
(3) ACGTCCT

(1) AC_TCAT
(3) ACGTCCT

Progressive alignment

(1,2) + (3)

(3) ACGTCCT
(1) AC_TCAT
(2) AG_TCAT

16

An alternate alignment

(1) AC_TCAT
(2) A_GTCAT
(3) ACGTCCT

14

Progressive alignment

- "Once a gap, always a gap"
  - You can't go back and correct a bad decision at an earlier step.
- Progressive alignment is not guaranteed to give the optimal alignment.
- But it does have better complexity…

Complexity of progressive alignment

- Distance matrix
  - Each pairwise alignment $O(n^2)$
  - Number of pairwise alignments $O(k^2)$
- Iterative construction of MSA
  - Number of merge steps $O(k)$
  - Each pairwise alignment $O(k^2n^2)$

Entire method $O(k^2n^2)$

Summary: Progressive alignment heuristics

- Not guaranteed to give the optimal MSA
- Bad choice of gaps propagates
- Complexity
  - Progressive: $O(k^2n^2)$
  - versus DP: $O(n^k 2^k k^2)$
- Typically, merge the most closely related sequences first.