

Bioinformatics 1: lecture 4

Followup of lecture 3?

Molecular evolution

Global, semi-global and local

Affine gap penalty

How sequences evolve

- point mutations (single base changes)
- deletion (loss of residues within the sequence)
- insertion (gain of residue within the sequence)
- truncation (loss of either end)
- extension (gain of residues at either end)

Mechanisms of insertion or extension:

- duplication of whole gene or domain
- polymerase "stutter"
- transposable element
- more??

How evolution is scored

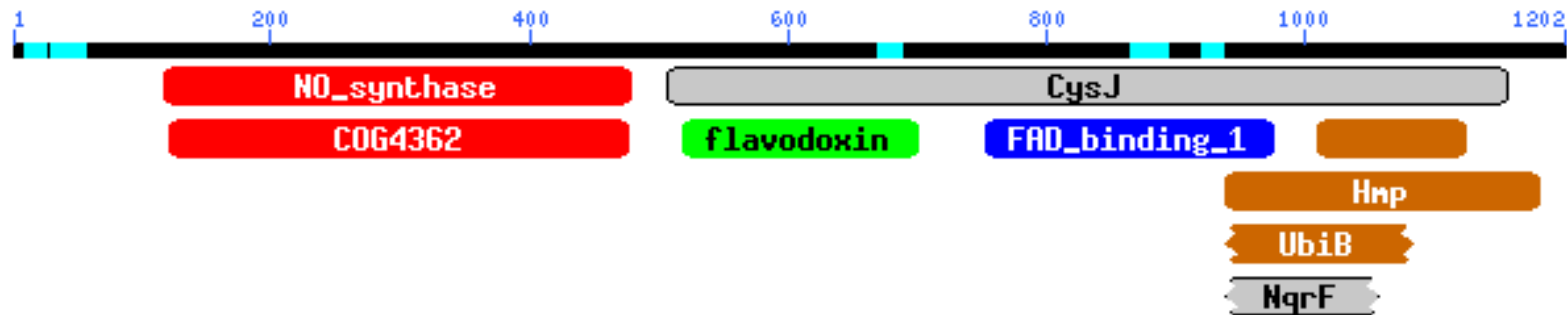
- point mutations substitution matrix
- deletion gap penalty
- insertion gap penalty
- truncation end gap penalty
- extension end gap penalty

Yes, an **alignment algorithm** is really
A Model for Sequence Evolution!

•✂• *That means the way we do alignment
should be closely aligned to what we
know about how things evolve.*

- point mutations relatively frequent, usually bad
- deletion infrequent, always bad, location dependent
- insertion infrequent, always bad, location dependent
- truncation frequent, not so bad
- extension frequent, not so bad

Extension/truncation, domains : end gaps

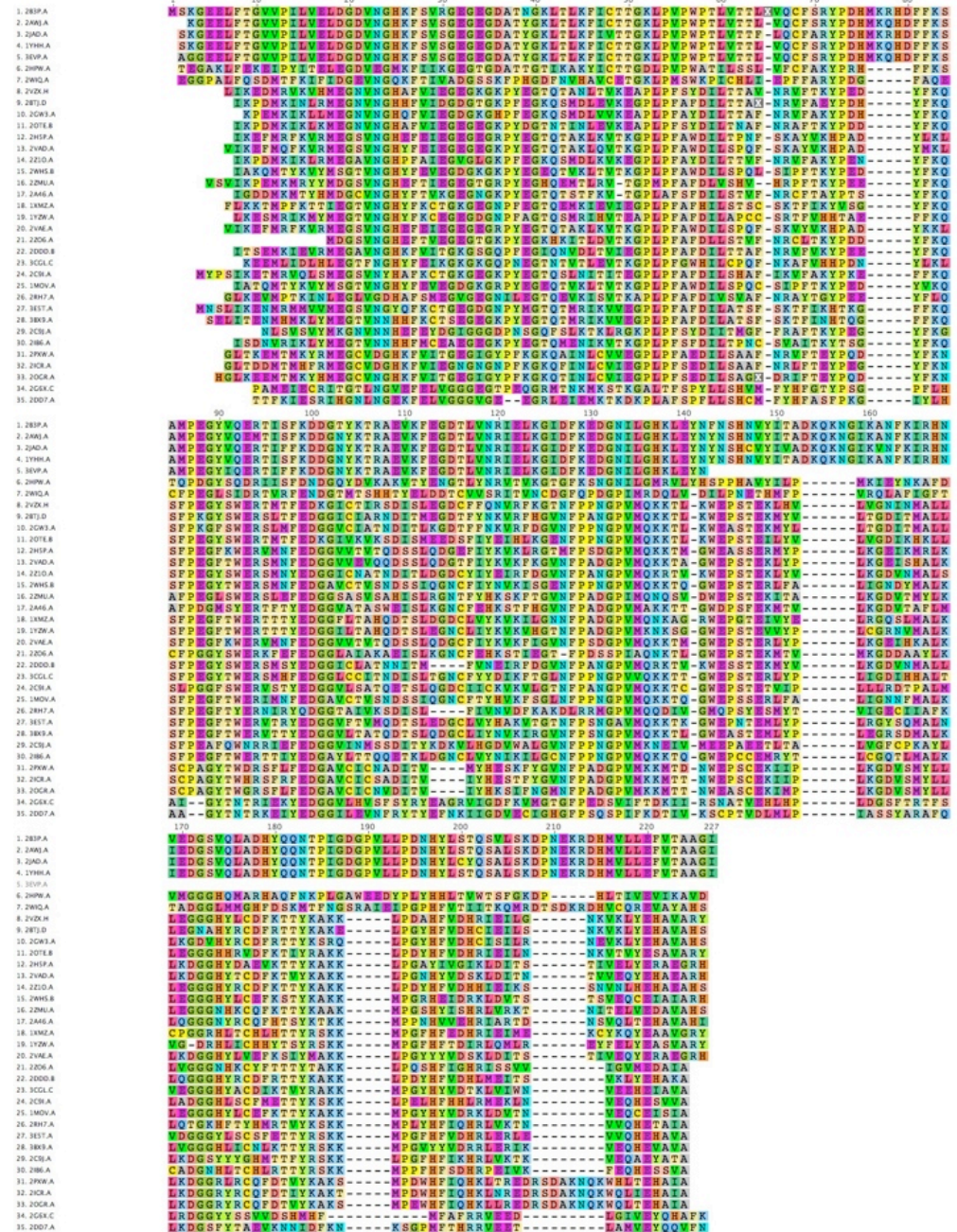


Example: here is an alignment of mouse nitric oxide synthase (thick black line). It has multiple domains which are homologous to several shorter proteins. If we penalize end gaps, what happens to the score of the true alignment? Did "end gaps" evolve the same way as internal gaps? (no!)

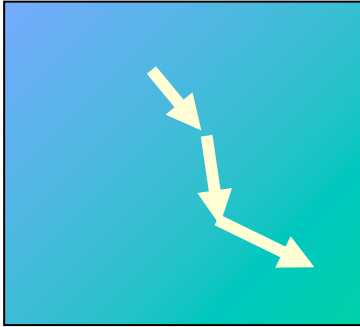
Unless the two proteins are known to be single domains, it makes more sense **NOT** to penalize end gaps.

A structure-based multiple sequence alignment of fluorescent proteins.

Note the ragged N-terminal edge, small number of indels, large number of point substitutions distributed unevenly over the sequence (conserved regions shown hot spots)



Local Alignment



A local alignment can start and end anywhere in the alignment matrix.

start

	A	T	S	F	M
P					
G					
T					
S					
F					
E					
P					

A purple arrow points from the top-left cell (row 2, column 1) to the cell containing 'S' (row 4, column 3). A black arrow points from the 'S' cell to the 'end' cell (row 5, column 4). The 'end' cell is highlighted with a red border.

$$A(i,j) = \text{MAX}$$

TSF
TSF

$$A(i-1,j-1) + S(i,j)$$

$$A(i,j-1) + \text{gap}$$

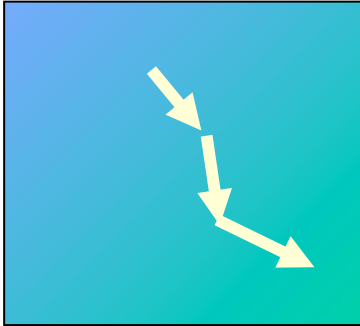
$$A(i-1,j) + \text{gap}$$

$$0 + \text{match score}$$



start
arrow

end is the maximum score anywhere in the *matrix*.



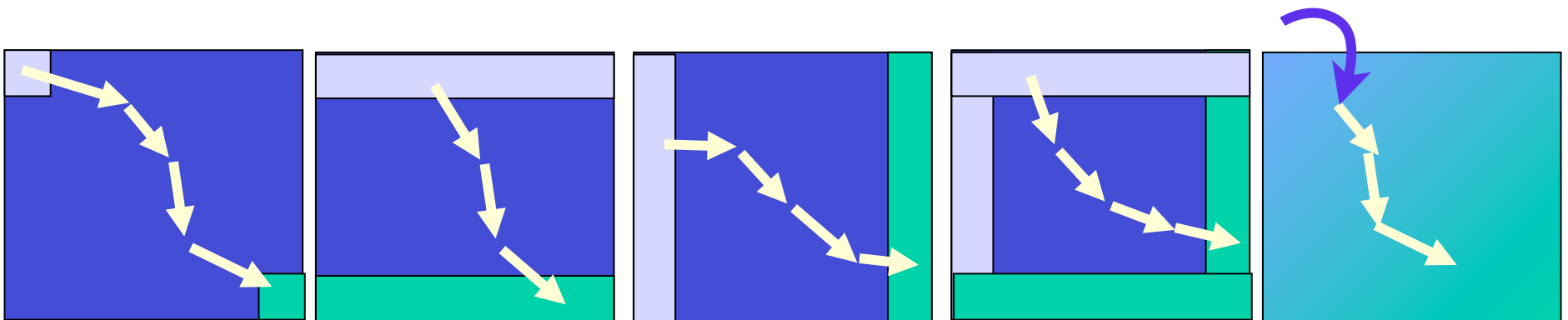
Local Alignment

- Asks for largest domain, sub-domain, or set of contiguous domains that are in common between two sequences.
- Worst score is always zero (0) for “no alignment”
- More appropriate than global or semi-global when there are no assumptions about the sequence relationship.
- Used for database searches.
- Required to obtain e-values

Global, semi-global, and local alignment

The choice of alignment method makes a statement about how the sequences are related. Was one sequence inserted into the other?

- **Global alignment** (end gaps) requires that all 4 termini are counted. In general, the two sequences are about the same length.
- **Semi-global** (no end gaps in 1 or both seqs) requires that one of the two sequences be completely contained in the other or that 2 or the 4 the termini be included.
- **Local alignment** finds subsequences in both. Does not require that the termini be included in the alignment.



The **optimal** alignment may be **no alignment**

If the maximum score in the alignment matrix is < 0 .,
then the optimal local alignment has score = 0 and
looks like this:

ATSFM~~~~~

~~~~~PGTSFEP

In class exercise: gaps

- In a browser, goto to NCBI. Search Protein database
- for **1DRF**. Save as FASTA file
- for **2DRC**. Do the same.
- Open both in Ugene. Select. Save as Alignment. Select. Align using Kalign. Gap open=12, Gap extend=3, Global with free end gaps.
- Count the number of gaps in the resulting alignment (initiations, not characters)
- Re-align (right click. Align, Kalign). Do the experiments on the next page.

Ugene align worksheet

gap extension
penalty

gap opening
penalty

	0	1	3	10
0	86, 28.8			>50, 24.7
1				
3				
12			6 , 26.2	
20				
50	0, 0.0			0, 0.0

Record: # of gaps , % Identity

Structure-based alignments are the "gold standard"

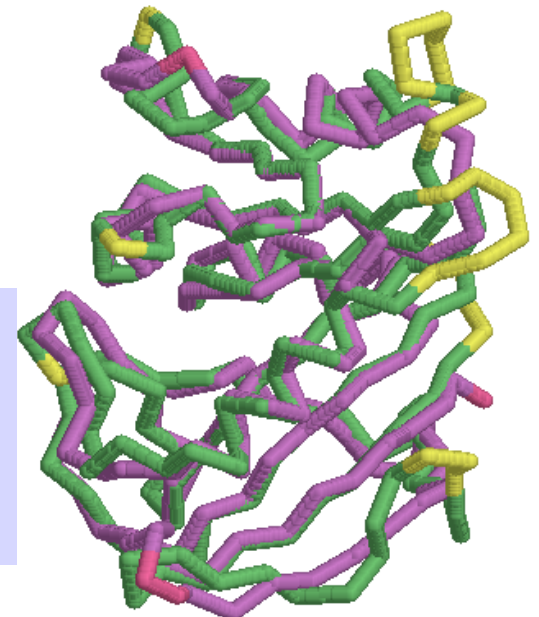
A structure-based alignment is a sequence alignment that comes from a protein structure superposition.

2DRC:A	1/2	MISLIAALAVDRVIGMENAM-PFNLPADLAWFKRNTL-----DKPVIMGRHTWESIG-
1DRF:_	3/4	SLNCIVAVSQNMGIGKNGDL P WPPLRNEFRYQRM T T S S V E G K QNLVIMGKKTWFSI E
2DRC:A	52/53	--RPLPGRKNIILSSQP--GTDDRVTWVKSVDIAAAG-----DVPEIMVIGGGRVYE
1DRF:_	63/64	K N R P L K G R I N L V L S R E L K E P P Q G A H F L S R S L D D A L K L T E Q P E L A N K V D M V I V G S S V Y K
2DRC:A	102/103	QFLPK--AQKLYLTHIDAEVEGDTHFPDYEPDDWESVF-----SEFHDA D A Q N S H S Y C F
1DRF:_	123/124	EAMNH P G H L K L F V T R I M Q D F E S D T F F P E I D L E K Y K L L P E Y P G V L S D V Q E E ---KGIKYKF
2DRC:A	154/155	EILERR
1DRF:_	180/181	EVYEKN

Look carefully. What do you see? Lots of mismatches (id=38%), few gaps (8), gaps are long (1-7).

Two similar structures may be superimposed. The parts that overlay well are the matches (purple and green), and the parts that do not overlay well are the insertions (yellow and red).

Aligned positions have similar chemical 3D environment



BAlibase

- A database of curated multiple sequence alignments derived from structure-based alignments. The Gold Standard for multiple sequence alignment!
- <http://www-bio3d-igbmc.u-strasbg.fr/balibase/>

Affine gap penalty-- theory

- Each gap represents an evolutionary event (duplication, polymerase stutter, deletion/ligation, etc.)
- If the alignment has "**evolutionary distance**" meaning, then the gap penalty score should be proportional to the number of gaps.

Are long gaps proportionally less likely?

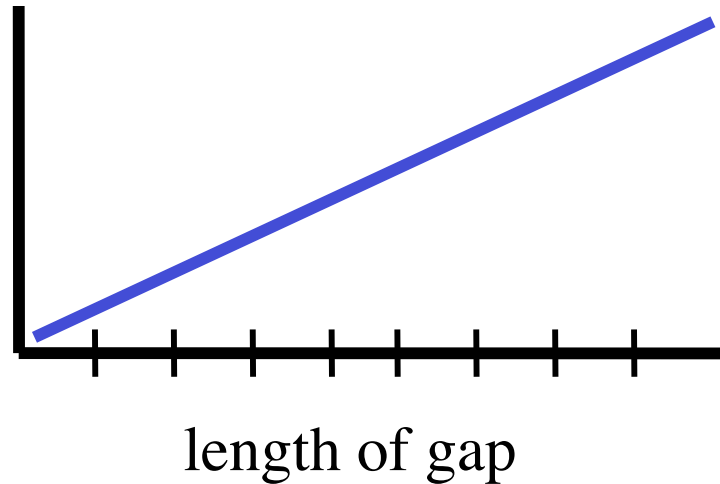
Which alignment is intuitively better?

AGGCTACT~T~TCA
GGCTACTATATCA

AGGCTACTTT~~CA
GGCTACTATATCA

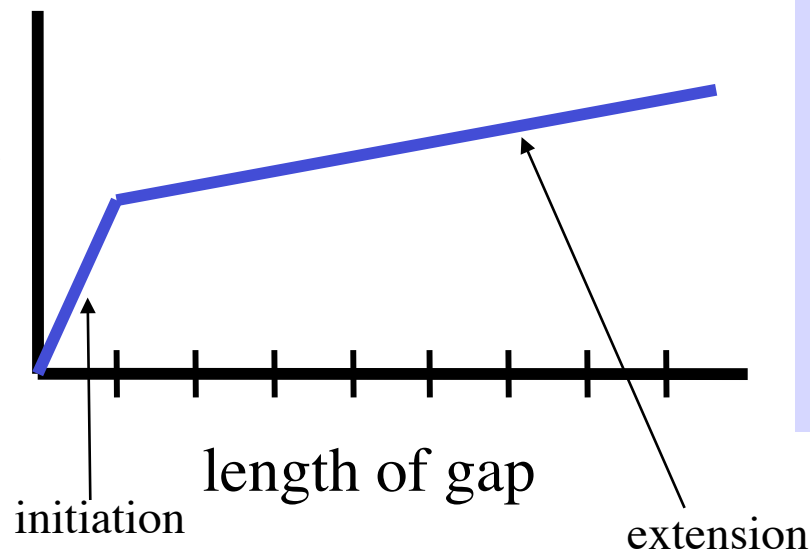
Linear versus Affine gap penalty

linear gap
penalty



Gap penalty for the whole
sequence is just the total
number of gap characters
times a constant.

affine gap
penalty



Gap penalty for the whole
sequence is the function.
 $N * (\text{gap initiation penalty}) +$
 $E * (\text{gap extension penalty})$

where N is the number of gap
initiation characters, E is the number
of gap extension characters

Example: affine gap

gap *initiation* = -5 gap *extension* = -1

-5 -5

AGGCTACT	T	~	T	~	T	CA
GGCTACT	T	A	T	A	T	CA

-10

-5 -1

AGGCTACT	TTT	~	~	CA		
GGCTACT	T	A	T	A	T	CA

-6

Affine Gap DP

• You can have **5 types of arrows**, instead of just three.

(1) Match

(2) Open a gap in first sequence.

(3) Open a gap in second sequence.

(4) Extend a gap in first sequence.

(5) Extend a gap in second sequence.

---or---

• You can have variable length arrows.

Affine gap DP algorithm using variable length arrows

	A	D	P	Q	F	G
A						
K						
L						
K						
L						
D						
O						
F						
G						
P						

$$S_{i,j} = \max_n \left\{ \begin{aligned} &S_{i-1,j-1} + s(i,j), \\ &S_{i-1-n,j-1} + s(i,j) - g_{\text{init}} - (n-1) g_{\text{ext}}, \\ &S_{i-1,j-1-n} + s(i,j) - g_{\text{init}} - (n-1) g_{\text{ext}} \end{aligned} \right\}$$

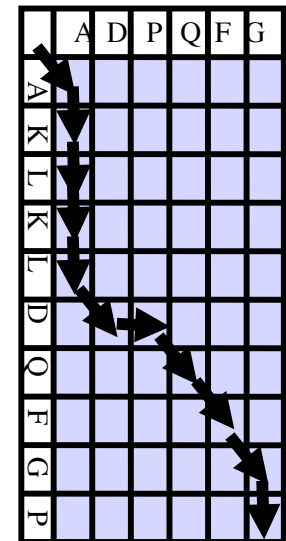
...where $s(i,j)$ is the substitution score, n is the length of the gap, g_{init} is the gap initiation penalty, and g_{ext} is the gap extension penalty.

Notes: All arrows end in match. Gap-to-gap not possible. Local or semi-global only. End-gaps not scored. Arrows still translate to an alignment. Still optimal.

Traceback for linear gap penalty

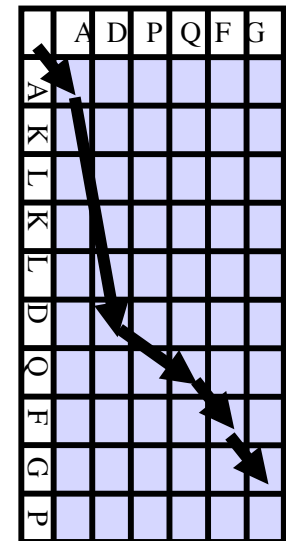
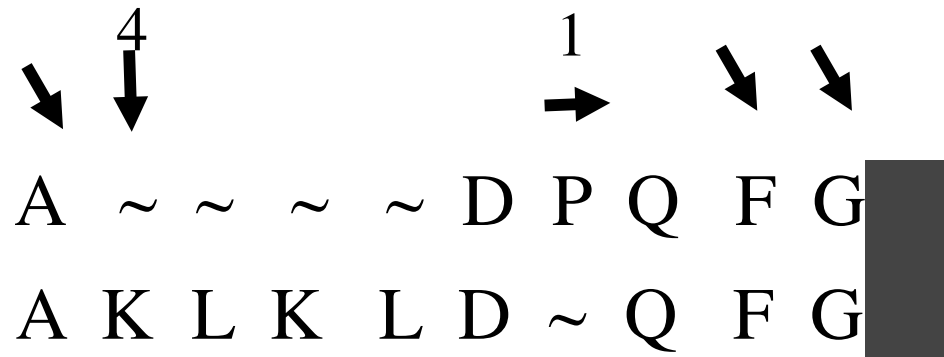
Each arrow advances either one sequence or both, by 1. Each column has one arrow.

↘	↓	↓	↓	↓	↘	→	↘	↘	↘	↓
A	~	~	~	~	D	P	Q	F	G	~
A	K	L	K	L	D	~	Q	F	G	P



Traceback for affine gap DP

Each arrow advances one sequence by 1, the other sequence by n . Output of one arrow is n columns. Last of n columns is a match. Number of arrows is \leq number of columns.



Does gap to gap make sense???

Special rules may apply for going from I to D and D to I.

AGGCTACT~TATCA
GGCTACTA~ATCA

If you think this alignment does not make sense, then D to I and I to D can simply be **disallowed** in the DP algorithm. Most programs do this.

[Exception: For a global alignment, D-to-I or I-to-D arrows are allowed at the ends of alignments because there is no other way to complete the matrix.]

Machine learning the gap penalty

- Create a database of sequence alignments (BAliBase)
 - Training set.
 - must be non-redundant
 - must be representative
- Define an objective function
 - function of all alignments
 - converges on a maximum as alignments converge on Training set
- Explore parameter space
 - may be exhaustive search, or something smarter
- Cross-validate.
 - Report the accuracy on a Test set (non-redundant, representative, no overlap with Training set)

Different substitution matrices give different alignments when sequence similarity is in the “Twilight zone”

Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions

*Yao-ming Huang and *Christopher Bystroff*

Center for Bioinformatics, Dept of Biology, Rensselaer Polytechnic Institute, Troy, New York 12180 USA

Objective function: count number of True Matches (Matches found in bAliBase alignment).

Fig. 6-a

	1R69	01	SISSRVKSKRIQLGLNQAELAQKVGTTQQSIE-Q-LENGKTKRPRFLPELASALGVSDWLLNGT
BLOSUM40	1NEQ	17	-----GLKKRKL SLSALS RQFGYAPTTLANA-
SDM	1NEQ	31	-----QFGYAPTTLANALERHWPKEG-QIIANALETKPEVI-----
HMMSUM-D ₃	1NEQ	13	DVIAGLKKRKL S L-----SALS RQFGYAPTTLANA-LE RHWPKEQII---ANALETKPEVIWPSR
HMMSUM-D _{3+NS}	1NEQ	13	DVIAGLKKRKL SLSALS RQFGYAPTTLANALE-R-----HWPKEQIIANALETKPEVIWPSR
HMMSUM-D	1NEQ	14	-----VIAGLKKRKL SLSALS RQFGYAPTTLA-N-ALERHWPKEQIIANALETKPEVIWPSR--
HMMSUM-D _{NS}	1NEQ	11	--RADVIAGLKKRKL SLSALS RQFGYAPTTLA-N-ALERHWPKEQII---IANALETKPEVIWPSR
BAl iBASE	1NEQ	09	WHRADVIAGLKKRKL SLSALS RQFGYAPTTLA-N-ALER--HWPKEQIIANALETKPEVIWPSR

Points to remember

- Scoring should reflect true evolutionary distance
- Semi-global alignment is good for finding domain boundaries
- Local alignment is used for database searches
- Affine gap penalty is better than linear
- Structure-based alignments are the gold standard