Lecture 2, 5/12/2001:

- Local alignment
  the Smith-Waterman algorithm

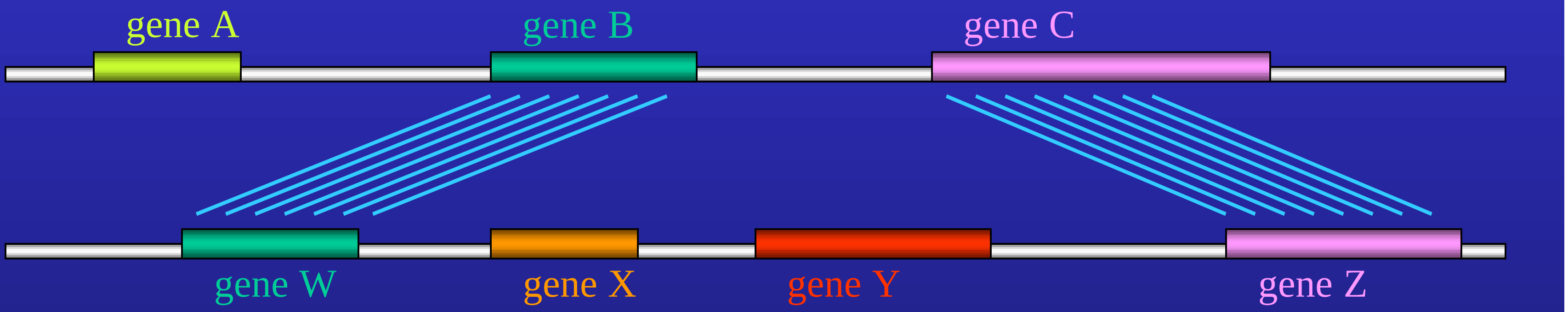
- Alignment scoring schemes and theory:
  substitution matrices and gap models

Local sequence alignments

Local sequence alignments are necessary for cases of:

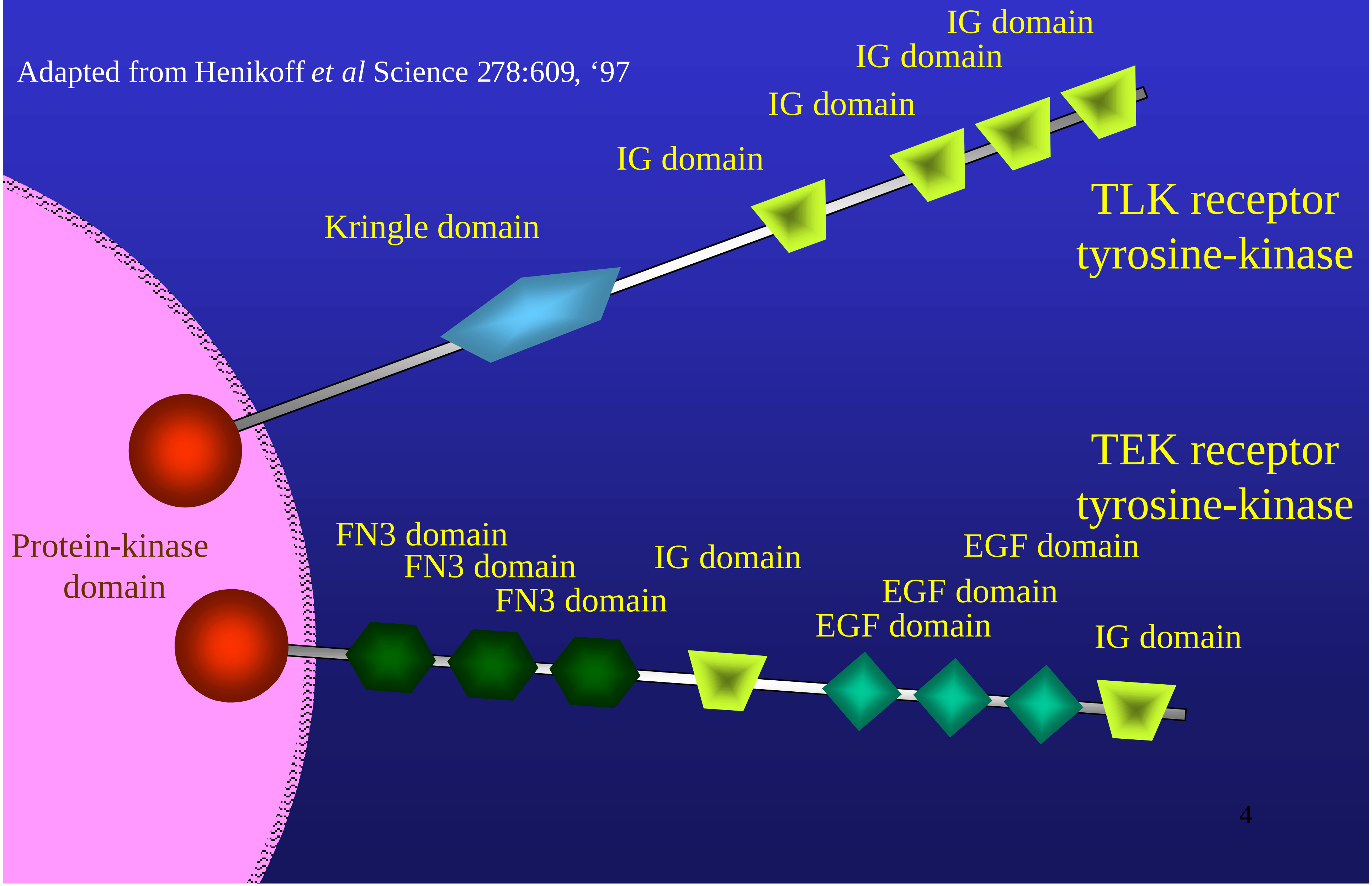
- Modular organization of genes and proteins (exons, domains, etc.)
- Repeats
- Sequences diverged so that similarity was retained, or can be detected, just in some sub-regions

Modular organization of genes



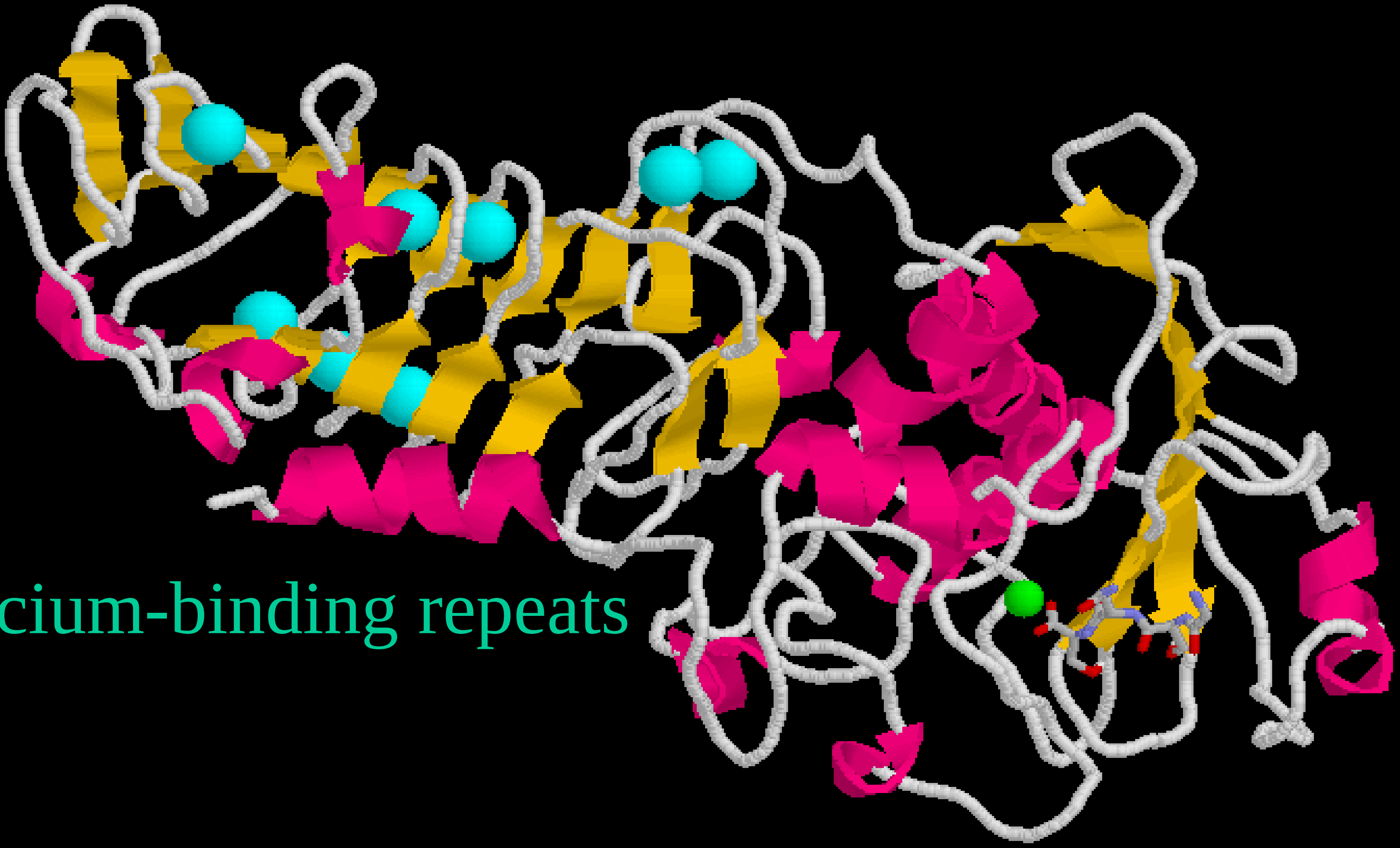
Modular protein organization

Adapted from Henikoff *et al* Science 278:609, ‘97



Modular protein organization

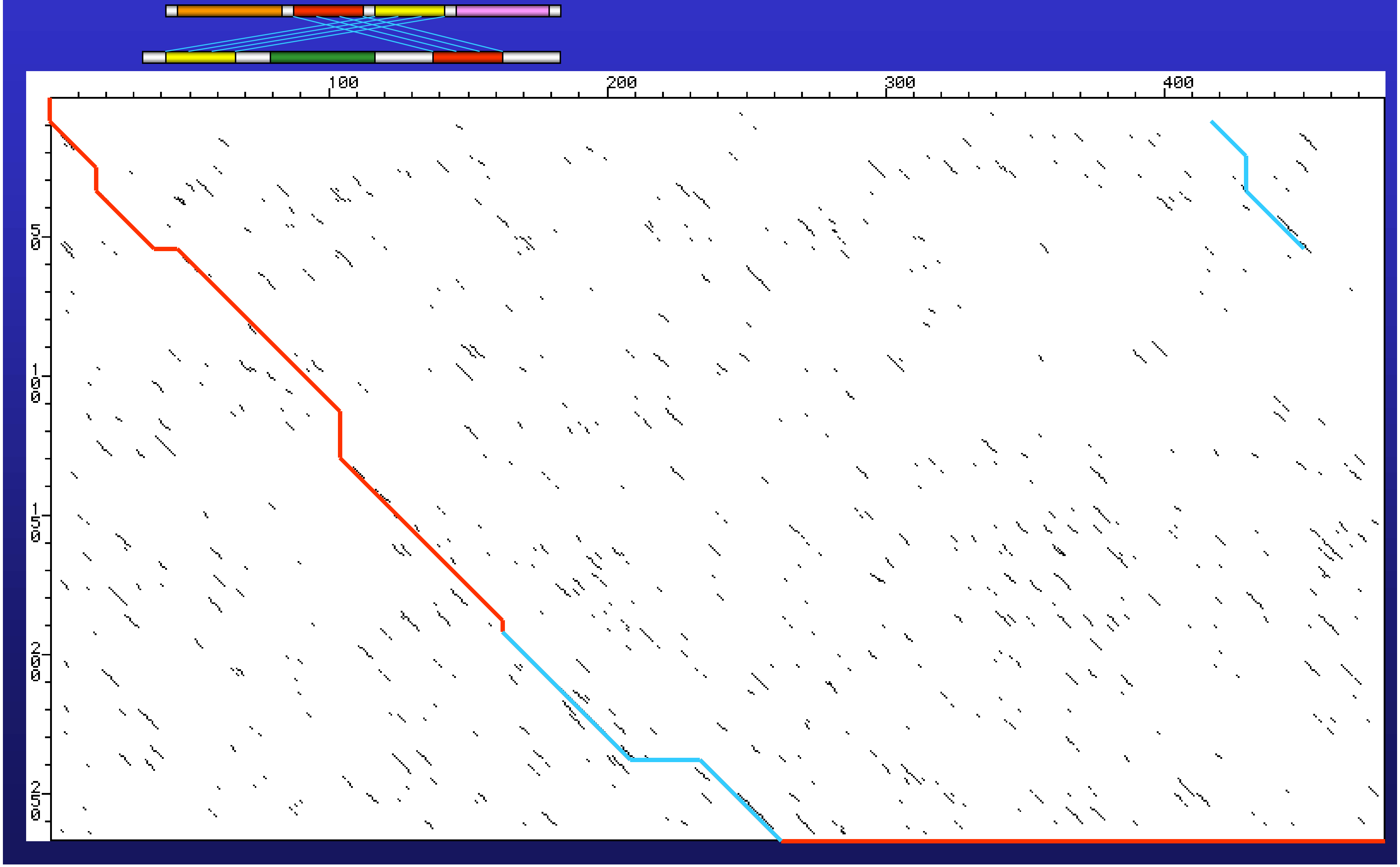
1KAP secreted calcium-binding alkaline-protease



Calcium-binding repeats

Protease domain

Local sequence alignment



Local sequence alignment

For local sequence alignment we wish to find what *regions* (sub-sequences) in the compared pair of sequences will give the best alignment scores with the parameters we supply (substitution matrix, gap penalty and gap scoring model.

The aligned regions may be anywhere along the sequences. More then one region might be aligned with a score above the threshold.

Local sequence alignment
Smith-Waterman algorithm

$\sigma[{}^{a}{}_{b}]$: score of aligning a pair of residues a and b

$-q$: gap penalty

$S'(i,j)$: optimal score of an alignment ending at residues i,j

best : highest score in the scores-matrix (S)

Local sequence alignment
Smith-Waterman algorithm

```
$best \Leftarrow 0$
    for $j \Leftarrow 1$ to $N$ do
        $S'(0,j) \Leftarrow 0$
    for $i \Leftarrow 1$ to $M$ do
    {   $S'(i,0) \Leftarrow 0$
        for $j \Leftarrow 1$ to $N$ do
            $S'(i,j) \Leftarrow \max \big( S'(i\text{-}1, j\text{-}1) + \sigma[^{ai}{}_{bj}],$
                $\max \{ S'(0, j) \ldots S(i\text{-}1, j) \}$ -q,
                $\max \{ S'(i, 0) \ldots S(i, j\text{-}1) \}$ -q,
                $0 \big)$
            $best \Leftarrow \max \big( S'(i, j)$ , $best \big)$
    }
```

Pearson & Miller
Meth Enz 210:575, ‘92

Local sequence alignment
Smith-Waterman algorithm
Finding the optimal alignment

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

Gap penalty -2

		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	1	0	0
T	0	0	1	0	0	0	0	0	2	0
C	0	0	0	2	0	0	0	0	0	3
A	0	1	0	0	3	1	1	1	1	1
G	0	0	0	0	1	4	2	2	2	2
T	0	0	1	0	1	2	3	1	3	1
C	0	0	0	2	1	2	1	2	1	4
A	0	1	0	0	3	2	3	1	1	2

The optimal local alignment is:

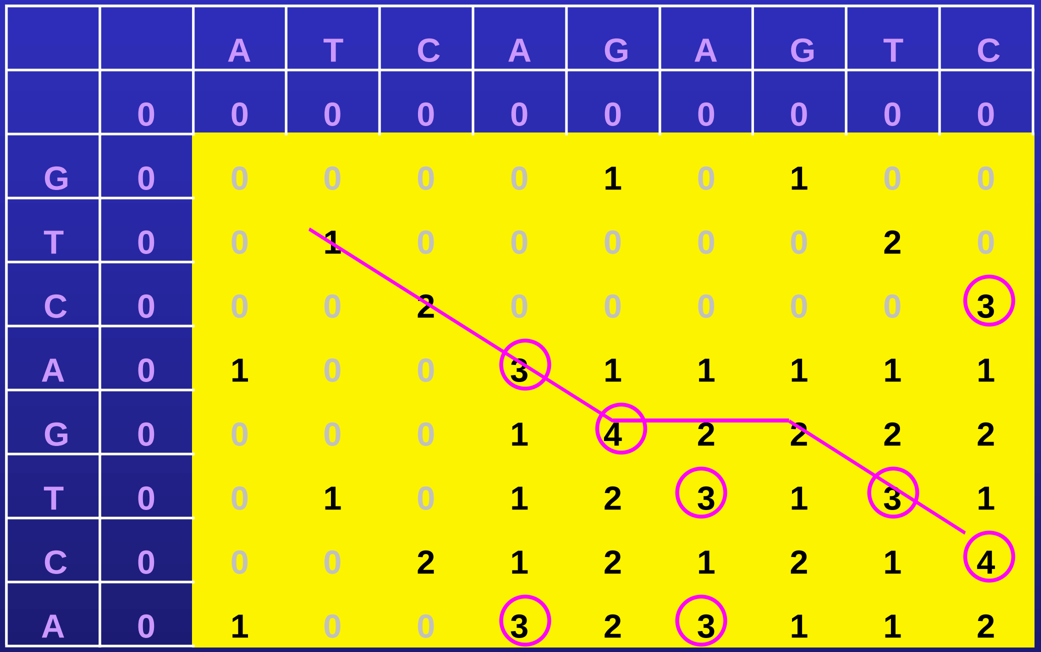
```
ATCAGAGTC
GTCAG--TCA
 ++++^^++
```

: 1+1+1+1-2+1+1=4

Local sequence alignment
Smith-Waterman algorithm
Finding the sub-optimal alignment

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

Gap penalty -2



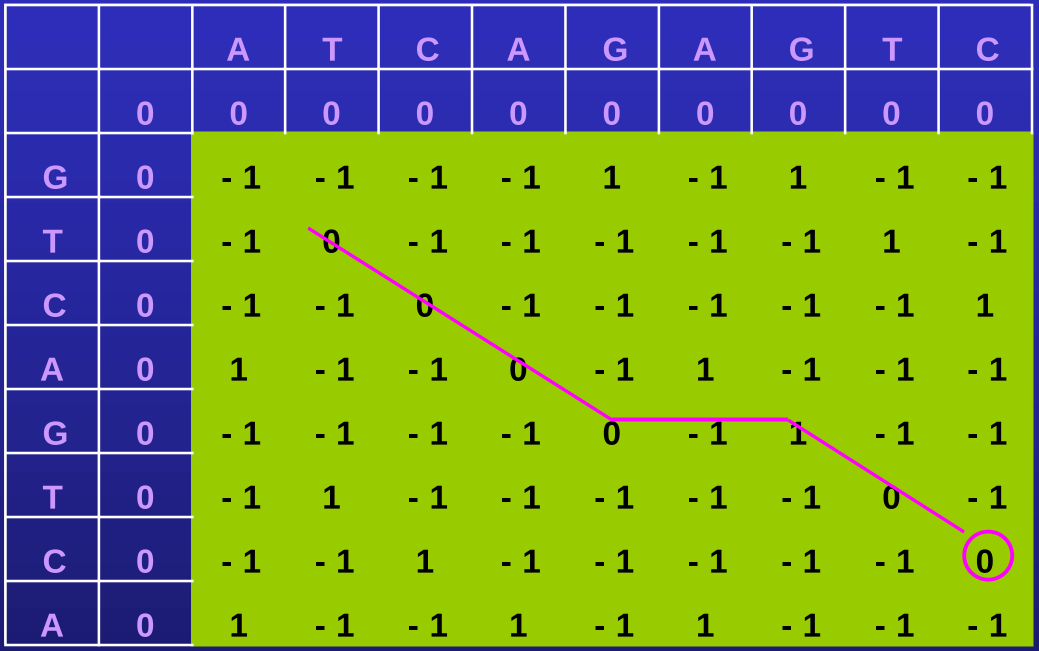
		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	1	0	0
T	0	0	1	0	0	0	0	0	2	0
C	0	0	0	2	0	0	0	0	0	3
A	0	1	0	0	3	1	1	1	1	1
G	0	0	0	0	1	4	2	2	2	2
T	0	0	1	0	1	2	3	1	3	1
C	0	0	0	2	1	2	1	2	1	4
A	0	1	0	0	3	2	3	1	1	2

Score threshold 3

Local sequence alignment
Smith-Waterman algorithm
Finding the sub-optimal alignment

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

Gap penalty -2



		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	-1	-1	-1	-1	1	-1	1	-1	-1
T	0	-1	0	-1	-1	-1	-1	-1	1	-1
C	0	-1	-1	0	-1	-1	-1	-1	-1	1
A	0	1	-1	-1	0	-1	1	-1	-1	-1
G	0	-1	-1	-1	-1	0	-1	1	-1	-1
T	0	-1	1	-1	-1	-1	-1	-1	0	-1
C	0	-1	-1	1	-1	-1	-1	-1	-1	0
A	0	1	-1	-1	1	-1	1	-1	-1	-1

Remove scores of the current optimal alignment and then recalculate the matrix to find the next best alignment /s

```
ATCAGAGTC
GTCAG--TCA
```

Local sequence alignment
Smith-Waterman algorithm
Finding the sub-optimal alignment

	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

Gap penalty -2

		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	1	0	1	0	0
T	0	0	0	0	0	0	0	0	2	0
C	0	0	0	0	0	0	0	0	0	3
A	0	1	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	0	2	0	0
T	0	0	1	0	0	0	0	0	0	0
C	0	0	0	2	0	0	0	0	0	0
A	0	1	0	0	3	1	1	1	1	1

Score threshold 3

```
   ATCAGAGTC
GTCAGTCA
     +++   : 1+1+1 =3
```

Local sequence alignment
Smith-Waterman algorithm

In order for the algorithm to identify local alignments the score for aligning unrelated sequence segments should typically be negative. Otherwise true optimal local alignments will be extended beyond their correct ends or have lower scores then longer alignments between unrelated regions.

Alignment scores are determined by substitution matrix and by the gap penalties and gap scoring model.

Alignment scoring schemes: gap models

Gap scoring by a constant relation to the gap length:

$\sigma \Leftarrow -q\,g$ (g is the number of gapped residues)

```
ATCACA
T---CA
```

$\sigma \Leftarrow -3q$

Gap scoring by a constant relation to the gap length:

$\sigma \Leftarrow -q$

```
ATCACA
T---CA
```

$\sigma \Leftarrow -q$

Affine gap scoring (opening [d] and extending gap penalties [e]):

$\sigma \Leftarrow -(d + e\,(g-1))$

```
ATCACA
T---CA
```

$\sigma \Leftarrow -(d + 2e)$

Local sequence alignment
Smith-Waterman algorithm

If alignment scores of unrelated sequences are mainly or solely determined by the substitution scores then such alignments would have negative scores if the sum of expected substitution scores would be negative:

$\sum_{i,j} p_i \, p_j \, s_{ij} < 0$

i & j	- residues,
p_i	- frequency of residue i
s_{ij}	- score of aligning residues i and j

Local sequence alignment
Smith-Waterman algorithm

We can easily identify substitution matrices that will not give positive scores to random alignments. However, we have no analytical way for finding which gap scores will satisfy the demand for random alignment scores to be less or equal to zero and produce local sequence alignments.

Nevertheless, certain sets of scoring schemes (substitution matrix and gap scores) were found to give satisfactory local alignments.

Sequence alignment
DNA substitution matrix

	A	C	G	T
A	5	-4	-4	-4
C	-4	5	-4	-4
G	-4	-4	5	-4
T	-4	-4	-4	5

Typical gap penalties for local alignment algorithms of DNA sequences are
16 for opening a gap & 4 for extending it

Alignment scoring schemes: substitution matrices

Unitary substitution matrix -

two scores are used, one for matches and one mismatches.

Practical usage of such matrices is for nucleotide alphabets.

In protein sequence alignments there are 20 types of residues (amino acids - aa) with complex relations by size, charge, genetic code, and chemistry. Unitary aa substitution matrices are outperformed by matrices that can have different scores for the 210 aa pairs. These matrices are calculated by scoring the relation between different of aa according to some of their features and/or which substitutions occur in correct alignments and what is the probability of having them by chance.

Alignment scoring schemes: substitution matrices

Altschul
JMB 219:555, '91

Every substitution matrix is either *explicitly* calculated from target frequencies of aligned residues (q_{ij}) and the frequencies of the residues (p_i), or these target and observed frequencies are *implicit* and can be back-calculated from the substitution scores.

The ratio of a target frequency to the frequencies it will occur by chance compares the probability an event will occur under two alternative hypotheses - $q_{ij}/(p_i\ p_j)$. This is called a likelihood, or odds, ratio.

Such probabilities should be multiplied to get the probability of their independent occurrence, or their log can be added. Log-odds score -
$s_{ij} = \left(\ln q_{ij}/(p_i\ p_j)\right) / \lambda$ (λ determines the base of the logarithm)

Sequence alignment
amino acids substitution matrix

BLOSUM62 in 1/2 Bit Units

The expected score per aligned position $(\sum_{i,j} p_i \, p_j \, s_{ij})$ is -0.52. Thus, this matrix is suitable for finding local sequence alignments.

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

Sequence alignment
amino acids substitution matrix

BLOSUM62 in 1/2 Bit Units

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

	P_i	P_j	Q_{ij}	Q_{ij}/P_iP_j	$2\log_2(Q_{ij}/P_iP_j)$
A:A	0.074	0.074	0.0215	3.926	3.946
R:R	0.074	0.052	0.0023	0.598	-1.485

See ftp://ncbi.nlm.nih.gov/repository/blocks/unix/blosum/README
and ftp://ncbi.nlm.nih.gov/repository/blocks/unix/blosum/BLOSUM/

Alignment scoring schemes: substitution matrices

Altschul
JMB 219:555, '91

Substitution matrices are characterized by their average score per residue pair

$$H = \Sigma_{i,j}\, q_{ij}\, s_{ij} = \Sigma_{i,j}\, q_{ij} \log_2 (q_{ij}/p_i p_j)$$

H is the information, in bit units, per aligned residue pair. It depends on the target frequencies (q_{ij}) - calculated from what we think are correct alignments - and on the alignments that would occur by chance ($p_i p_j$). It is termed the relative entropy of the matrix.

H measures the information provided by the matrix to distinguish correct alignments from chance ones. Matrices with lower values will identify more distant sequence relationships that produce weaker alignments.

Alignment scoring schemes: substitution matrices

Substitution matrices differ by the models and data used for their calculation. Each is suitable for identifying alignments of sequences with different evolutionary distances. Nevertheless, longer alignments are needed to identify the relationship between more distant sequences.

The scale of the substitution matrix (base of the log) is arbitrary. However, matrices must be in the same scale to be compared to each other, and gap penalties are specific to the matrix and scale used. Typical penalties for local alignment with the BLOSUM62 matrix in half-bit units are 12 for opening a gap and 2 for extending it.

More details, sources and things to do for next lecture

Sources: Pearson & Miller "Dynamic programming algorithms for biological sequence comparison." *Methods in Enz.* , **210**:575-601 (1992),

Altschul “Amino acid substitution matrices from an information theoretic perspective” *J Mol Biol* **219**:555-565 (1991),

Henikoff “Scores for sequence searches and alignments” Curr Opin Struct Biol **6**:353-360 (1996).

Assignment:

Read the source articles for this lecture. They have more details on the material we covered and introduce topics for next lectures.

Calculate the q_{ij} target frequencies of the DNA substitution matrix shown in class for equal nucleotide frequencies, and for $p_A = p_T = 0.3$ & $p_G = p_C = 0.2$.

More details, sources and things to do for next lecture

For those who are no acquainted with information theory or want to be certain they know the basics of it:

An information theory primer for molecular biologists-
http://www.lecb.ncifcrf.gov/~toms/paper/primer