Scoring matrix: amino acids

Different criteria can be considered when devising a scoring matrix for amino acid sequence alignments

Most common ones are based on observed physical/chemical similarity and observed substitution frequencies

E.g. pairing two amino acids that both have aromatic functional groups might receive a good positive score,

pairing an amino acid that has a nonpolar functional group with one that has a charged functional group might result in a scoring penalty.

Scoring matrices have been derived based on residue hydrophobicity, charge and size

Another option is based on genetic code: minimum number of nucleotide substitutions are necessary to convert a codon from one residue to other

Scoring matrix: amino acids

A common method for deriving scoring matrices is to observe the actual substitution rates among various amino acid residues in nature.

If substitution between two amino acid residues is observed frequently, then positions in which these residues are aligned favorably.

Likewise, alignments between residues that are not observed to interchange frequently in natural evolution is penalized.

One commonly used scoring matrix based on observed substitution rates is the point accepted mutation (PAM) matrix.

The scores in a PAM matrix are computed by observing the substitutions that occur in alignments between similar sequences.

Development of PAM matrix

1. Alignment is constructed with very high sequence identity (usually >85%).

2.The relative mutatbility, m_j, for each amino acid is computed. It is the number of times the amino acid was substituted by any other amino acids. E.g. Ala to others

3. Pair of amino acids, A_{ij}, the number of times amino acid j was replaced by amino acid i, tallied for each amino acid pairs i and j. E.g. A_{cm} is the number of time Met is replaced with cysteine.

4. The substitution tallies are divided by relative mutability.

5. Normalize with the frequency of occurrence of each amino acid

6. Take log of each resulting entries in the PAM-1 matrix (PAM-1 means 1 substitution per 100 residues or 1 PAM unit) . This matrix is also called log odds matrix, since the entries are based on the log of the substitution probability for each amino acid.

PAM-1 matrix is appropriate to compare sequences are closely related. PAM-1000 matrix might be used to compare sequences with distant relationships. Usually PAM-250 is used for sequence alignment.

Calculation of a PAM matrix

PAM matrix is a 20x20 matrix for all pairs

Assumption:

Substitutions are equal is both directions (A to G and G to A)

E.g.: Element GA

Frequency of pairs, $F_{G,A} = 3$

Relative mutability, $m_A = 4$

Normalizing factor = number of mutations in the entire tree times 2, times relative frequency of A residues multiplied by 100 (1 substitution per 100 residues)

i.e., 6 x 2 x (10/63) x 100 = 190.4762

Hence, normalized relative mutability,

$m_A = 4/190.4762 = 0.021$

Consider a multiple sequence alignment

```
1.ACGCTAFKI
2.GCGCTAFKI  (1: A→G)
3.ACGCTAFKL  (1: I→L
4.GCGCTGFKI  (2: A→G)
5.GCGCTLFKI  (2: A→L)
6.ASGCTAFKL  (3: C→S)
7.ACACTAFKL  (3: G→A)
```

Construct tree

Calculation of a PAM matrix

$m_A = 4/190.4762 = 0.021$

Mutation probability, $M_{ij} = m_j F_{ij}/\Sigma f_{ij}$

$M_{G,A} = 0.021 \times 3 / 4$

$= 0.0157$

Σf_{ij}, total number of substitutions involving A

$R_{ij} = \log(M_{ij}/f_i) = \log(M_{GA}/f_G)$

$f_G = 10/63 = 0.1587$

$R_{GA} = \log(0.0157/0.1587) = \log(0.0989)$

$R_{GA} = -1.005$

Repeat for all off-diagonal elements.

For diagonal elements: $M_{jj} = 1 - m_j$

Calculate R_{jj}

Consider a multiple sequence alignment

```
1.ACGCTAFKI
2.GCGCTAFKI  (1: A→G)
3.ACGCTAFKL  (1: I→L
4.GCGCTGFKI  (2: A→G)
5.GCGCTLFKI  (2: A→L)
6.ASGCTAFKL  (3: C→S)
7.ACACTAFKL  (3: G→A)
```

Calculate the element R_{AA}

Calculate the element R_{IL}

PAM-120 mutation matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	3	-3	-1	0	-3	-1	0	1	-3	-1	-3	-2	-2	-4	1	1	1	-7	-4	0	0	-1	-1	-8
R	-3	6	-1	-3	-4	1	-3	-4	1	-2	-4	2	-1	-5	-1	-1	-2	1	-5	-3	-2	-1	-2	-8
N	-1	-1	4	2	-5	0	1	0	2	-2	-4	1	-3	-4	-2	1	0	-4	-2	-3	3	0	-1	-8
D	0	-3	2	5	-7	1	3	0	0	-3	-5	-1	-4	-7	-3	0	-1	-8	-5	-3	4	3	-2	-8
C	-3	-4	-5	-7	9	-7	-7	-4	-4	-3	-7	-7	-6	-6	-4	0	-3	-8	-1	-3	-6	-7	-4	-8
Q	-1	1	0	1	-7	6	2	-3	3	-3	-2	0	-1	-6	0	-2	-2	-6	-5	-3	0	4	-1	-8
E	0	-3	1	3	-7	2	5	-1	-1	-3	-4	-1	-3	-7	-2	-1	-2	-8	-5	-3	3	4	-1	-8
G	1	-4	0	0	-4	-3	-1	5	-4	-4	-5	-3	-4	-5	-2	1	-1	-8	-6	-2	0	-2	-2	-8
H	-3	1	2	0	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-3	-1	-3	1	1	-2	-8
I	-1	-2	-2	-3	-3	-3	-3	-4	-4	6	1	-3	1	0	-3	-2	0	-6	-2	3	-3	-3	-1	-8
L	-3	-4	-4	-5	-7	-2	-4	-5	-3	1	5	-4	3	0	-3	-4	-3	-3	-2	1	-4	-3	-2	-8
K	-2	2	1	-1	-7	0	-1	-3	-2	-3	-4	5	0	-7	-2	-1	-1	-5	-5	-4	0	-1	-2	-8
M	-2	-1	-3	-4	-6	-1	-3	-4	-4	1	3	0	8	-1	-3	-2	-1	-6	-4	1	-4	-2	-2	-8
F	-4	-5	-4	-7	-6	-6	-7	-5	-3	0	0	-7	-1	8	-5	-3	-4	-1	4	-3	-5	-6	-3	-8
P	1	-1	-2	-3	-4	0	-2	-2	-1	-3	-3	-2	-3	-5	6	1	-1	-7	-6	-2	-2	-1	-2	-8
S	1	-1	1	0	0	-2	-1	1	-2	-2	-4	-1	-2	-3	1	3	2	-2	-3	-2	0	-1	-1	-8
T	1	-2	0	-1	-3	-2	-2	-1	-3	0	-3	-1	-1	-4	-1	2	4	-6	-3	0	0	-2	-1	-8
W	-7	1	-4	-8	-8	-6	-8	-8	-3	-6	-3	-5	-6	-1	-7	-2	-6	12	-2	-8	-6	-7	-5	-8
Y	-4	-5	-2	-5	-1	-5	-5	-6	-1	-2	-2	-5	-4	4	-6	-3	-3	-2	8	-3	-3	-5	-3	-8
V	0	-3	-3	-3	-3	-3	-3	-2	-3	3	1	-4	1	-3	-2	-2	0	-8	-3	5	-3	-3	-1	-8
B	0	-2	3	4	-6	0	3	0	1	-3	-4	0	-4	-5	-2	0	0	-6	-3	-3	4	2	-1	-8
Z	-1	-1	0	3	-7	4	4	-2	1	-3	-3	-1	-2	-6	-1	-1	-2	-7	-5	-3	2	4	-1	-8
X	-1	-2	-1	-2	-4	-1	-1	-2	-2	-1	-2	-2	-2	-3	-2	-1	-1	-5	-3	-1	-1	-1	-2	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

PAM 250 mutation matrix

	Cys	Gly	Pro	Ser	Ala	Thr	Asp	Glu	Asn	Gln	His	Lys	Arg	Val	Met	Ile	Leu	Phe	Tyr	Trp
Cys	12																			
Gly	-3	5																		
Pro	-3	-1	6																	
Ser	0	1	1	1																
Ala	-2	1	1	1	2															
Thr	-2	0	0	1	1	3														
Asp	-5	1	-1	0	0	0	4													
Glu	-5	0	-1	0	0	0	3	4												
Asn	-4	0	-1	1	0	0	2	1	2											
Gln	-5	-1	0	-1	0	-1	2	2	1	4										
His	-3	-2	0	-1	-1	-1	1	1	2	3	6									
Lys	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
Arg	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
Val	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
Met	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
Ile	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
Leu	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
Phe	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Tyr	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
Trp	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17

BLOSUM matrix

BLOSUM (Blocks Substitution Matrix) is another popular scoring matrix obtained with statistical clustering techniques.

Clustering approach helps to avoid some statistical problems that can occur when the observed substitution rate is very low for a particular pair of amino acids.

BLOSUM considers mainly conserved regions.

BLOSUM matrices can also be derived for alignments with different sequence identities.

Lower numbered PAM matrices are appropriate for comparing closely related sequences.

Lower numbered BLOSUM matrices are appropriate for comparing distantly related sequences.

E.g. BLOSUM-62 matrix is appropriate for comparing sequences of approximately 62% sequence similarity.

BLOSUM-62 matrix

	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	−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

BLAST

Basic Local Alignment Search Tool

BLAST: Process the Query Sequence and Database

Divide the query sequence into all "words" of length K=2 (default 3 for proteins)

Step 1: Hash table for sequence A

Query	Database
1234567	123456
QLNFSAGW	NLNYTPW
QL	NL
LN	LN
NF	NY
FS	YT
SA	TP
AG	PW
GW	

BLAST

Basic Local Alignment Search Tool

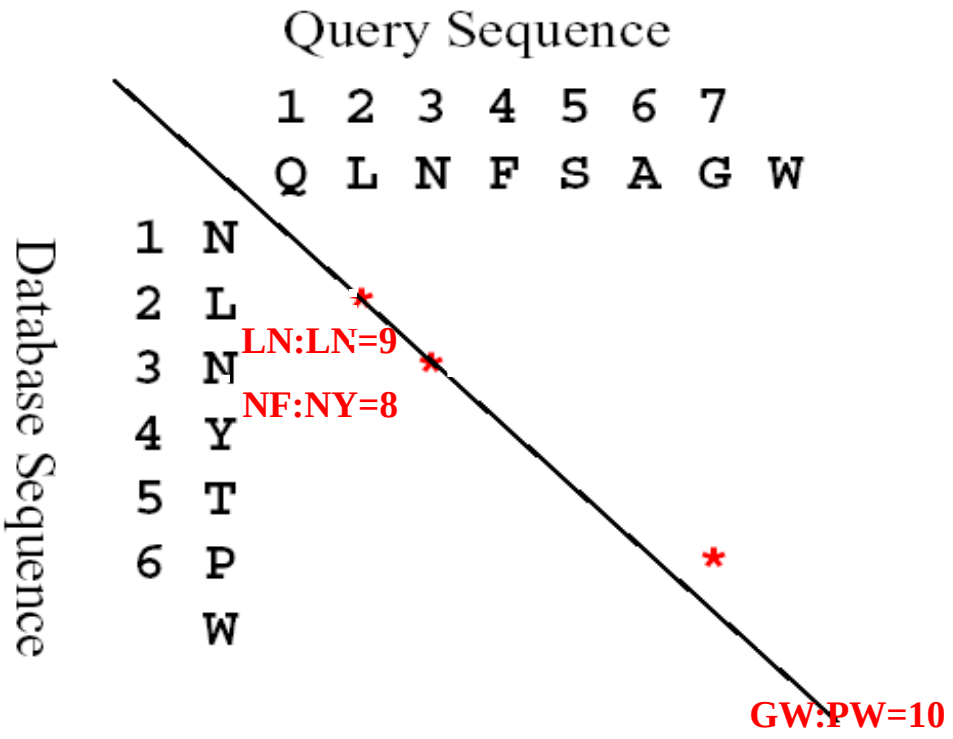
Identify Word Matches

Step 2:

Use all of the 2-letter words in query sequence to scan against database sequence and mark those with score > 8

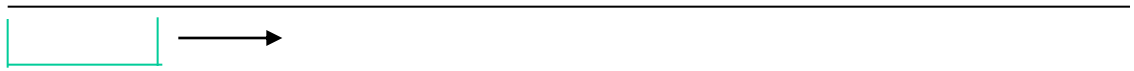
Note:

Marked points can be on the diagonal and off-diagonal



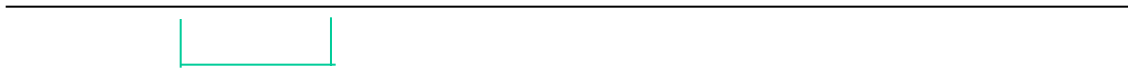
BLAST

Step2: Scan sequence b for hits.

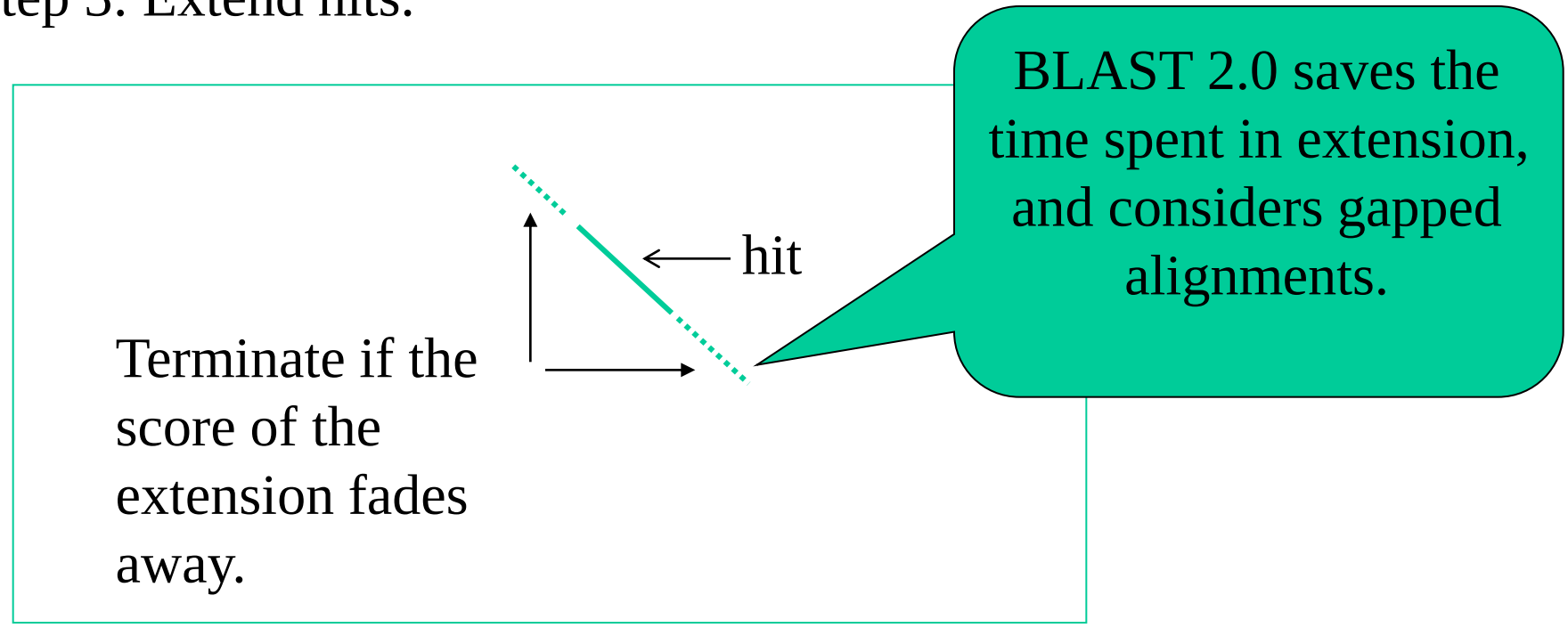


BLAST

Step2: Scan sequence b for hits.



Step 3: Extend hits.



Dynamic programming

Once a method for scoring alignments is selected, an algorithm to find the best alignment between two sequences can be aligned

The most obvious one is exhaustive searching and it is not possible.

E.g. two sequences with 100 and 95 residues, all possible alignments are ~55 million.

It is necessary to develop a smart algorithm

Method of breaking a problem apart into reasonably sized sub problems and using these partial results to compute the final answer: Dynamic programming

```
                First position      Score  Remaining sequence
                      C               +1          ACGA
                      C                           GA
                ----------------------------------------
CACGA           C                     -1          ACGA
CGA             -                                 CGA
                ----------------------------------------
                -                     -1          CACGA
                C                                 GA
                ----------------------------------------
```

Gene and Protein Sequence Alignment

Example:

Sequence a: ATTCTTGC

Sequence b: ATCCTATTCTAGC

Best Alignment:

```
        ATTCTTGC
ATCCTATTCTAGC
   /|\
   gap
```

Bad Alignment:

```
  AT   TCTT    GC
ATCCTATTCTAGC
/|\       /|\
   gap       gap
```

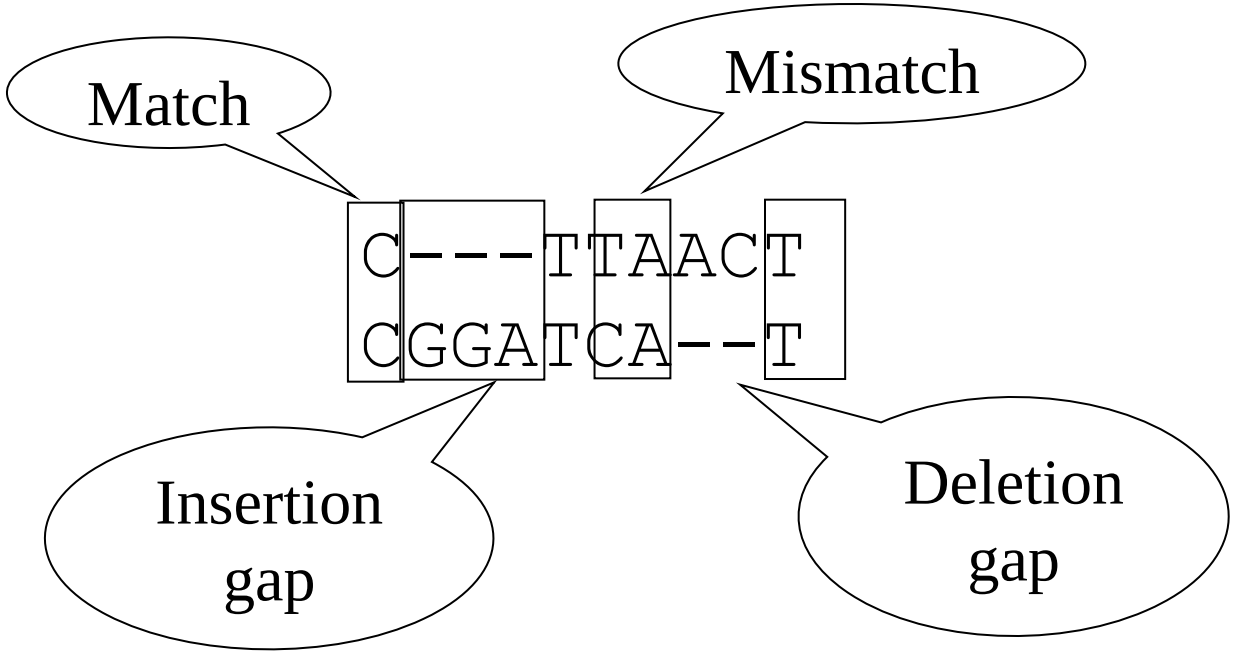
What is a good alignment?

Pairwise Alignment

Sequence a: CTTAACT

Sequence b: CGGATCAT

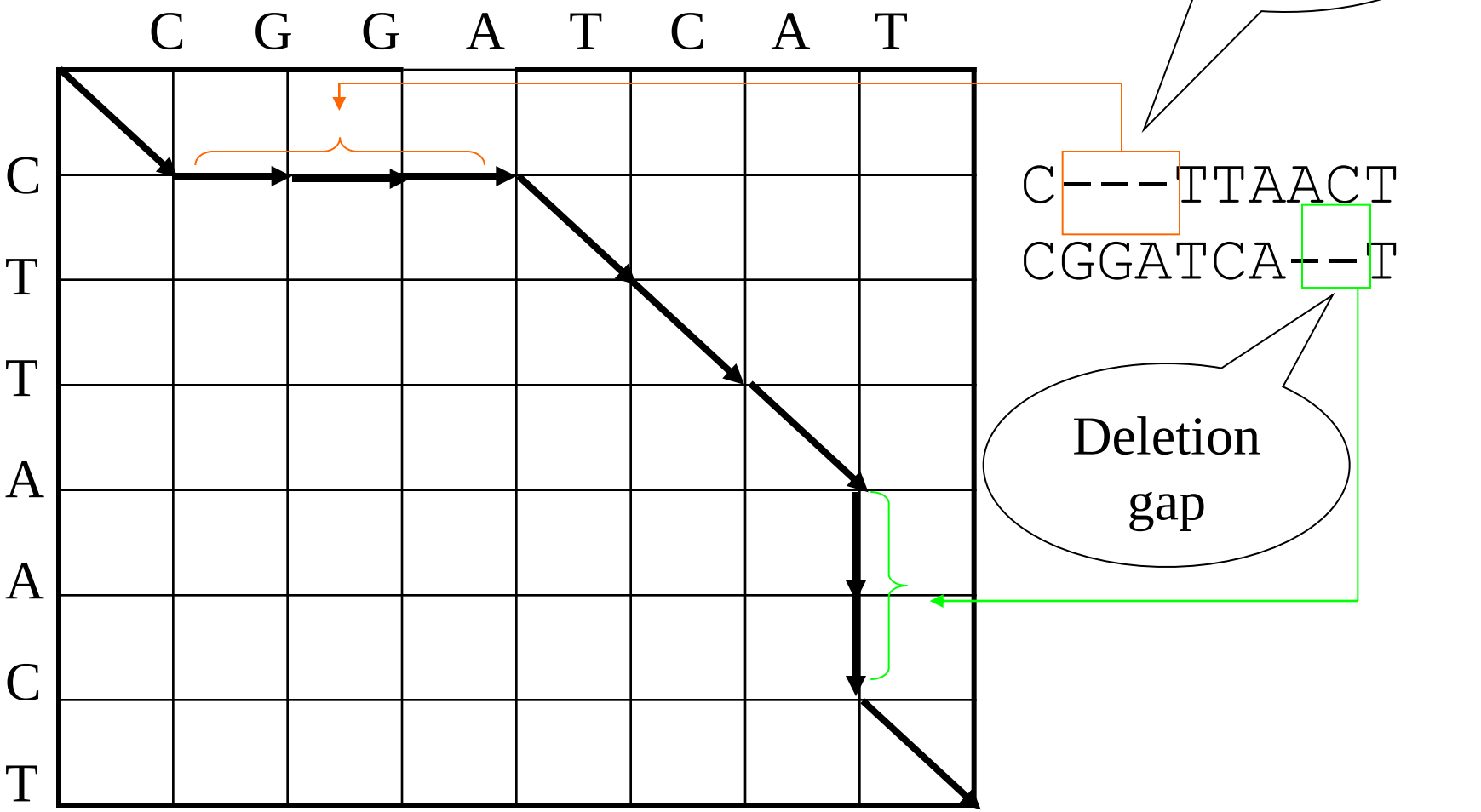
An alignment of a and b:



Alignment Graph

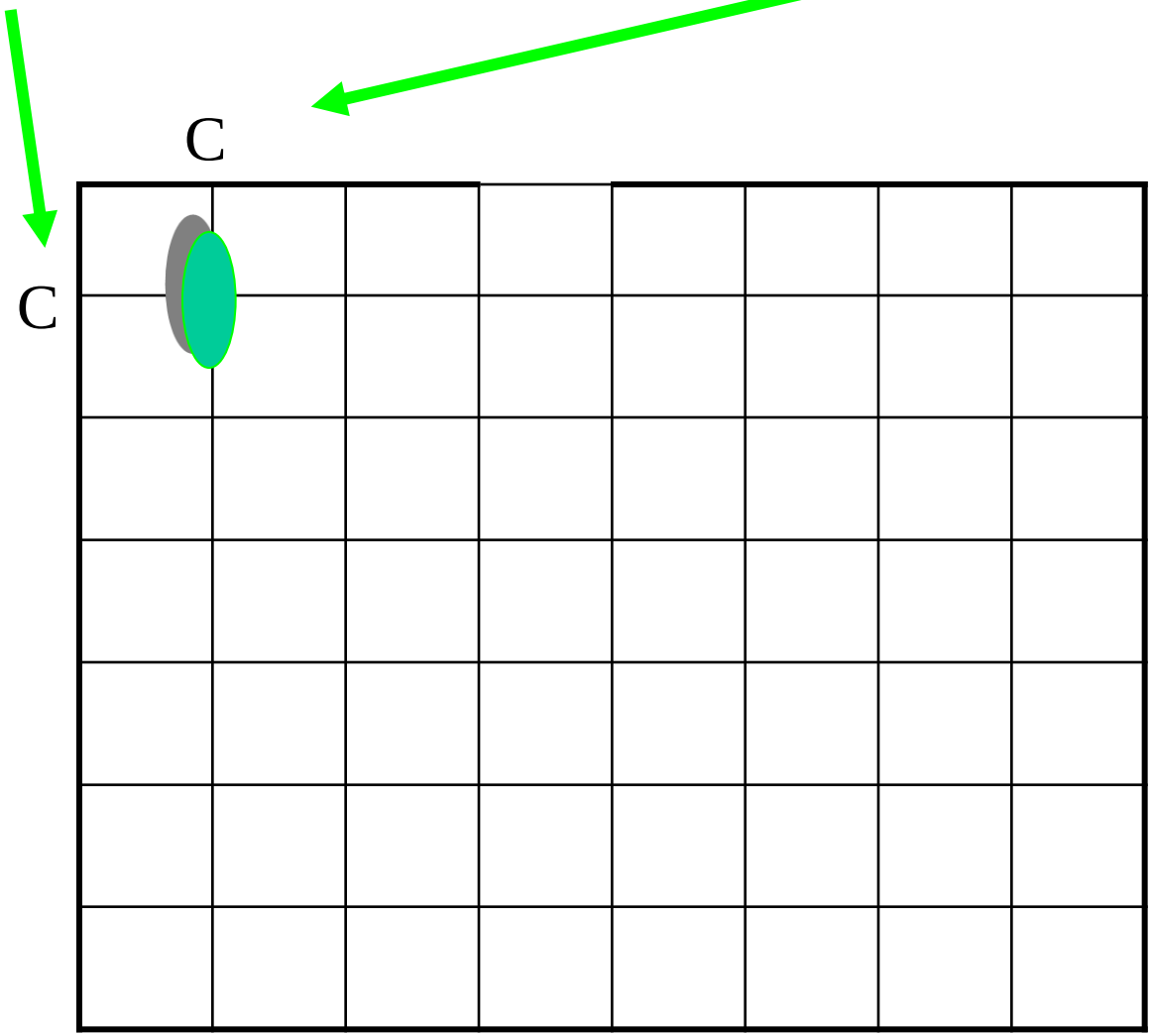
Sequence a: CTTAACT

Sequence b: CGGATCAT



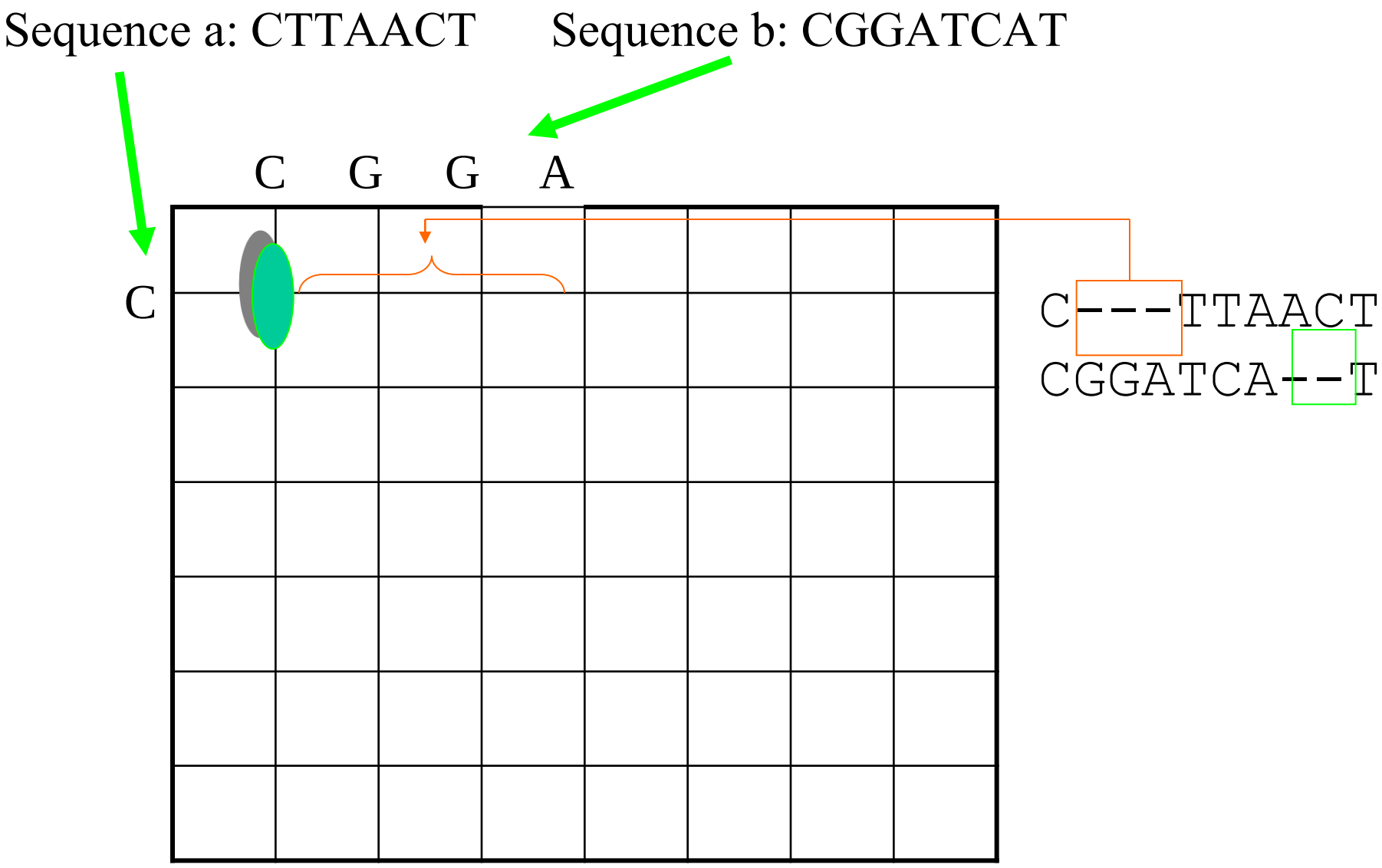
Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT



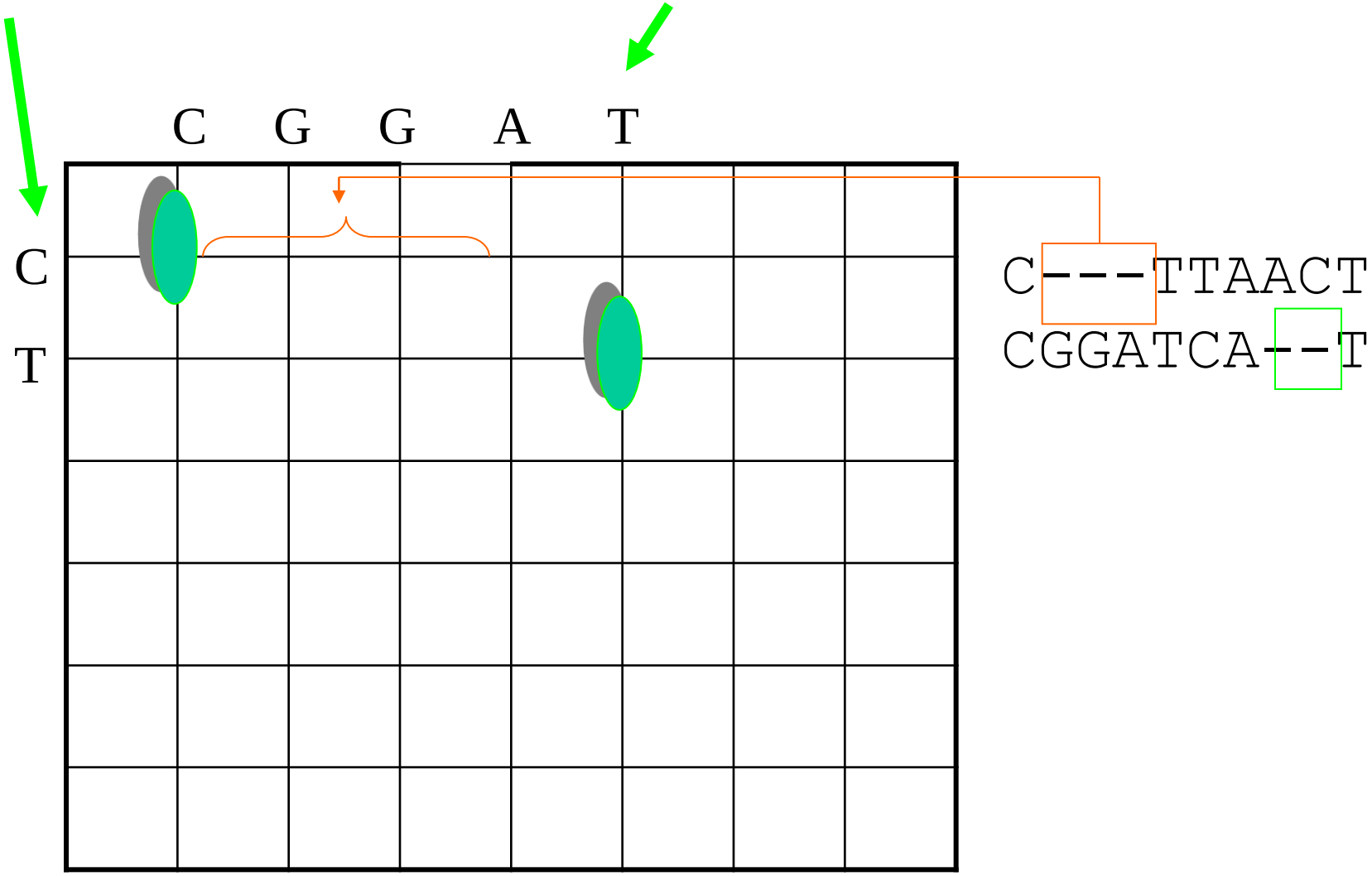
C---TTAACT
CGGATCA--T

Graphic representation of an alignment



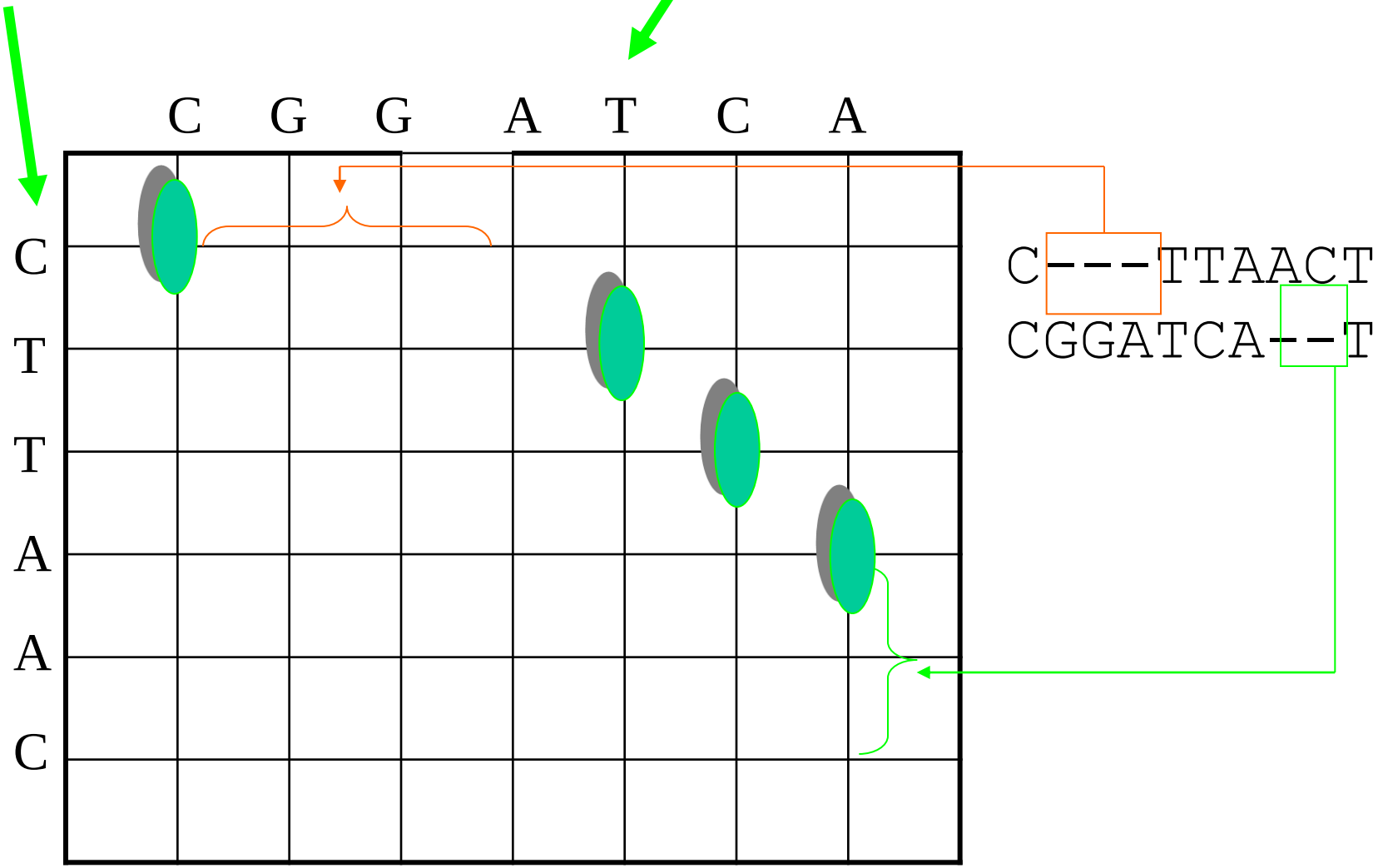
Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT

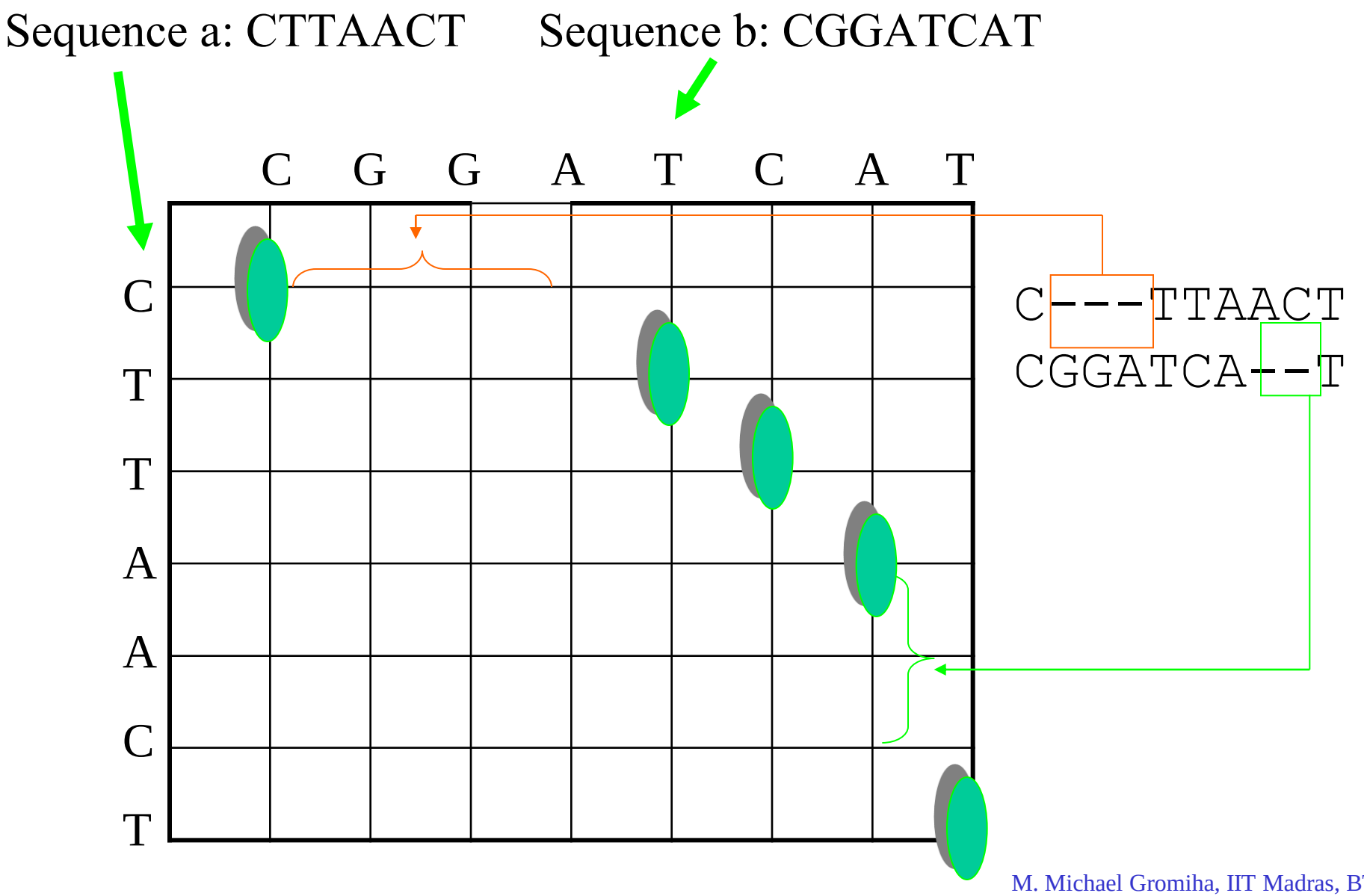


Graphic representation of an alignment

Sequence a: CTTAACT Sequence b: CGGATCAT



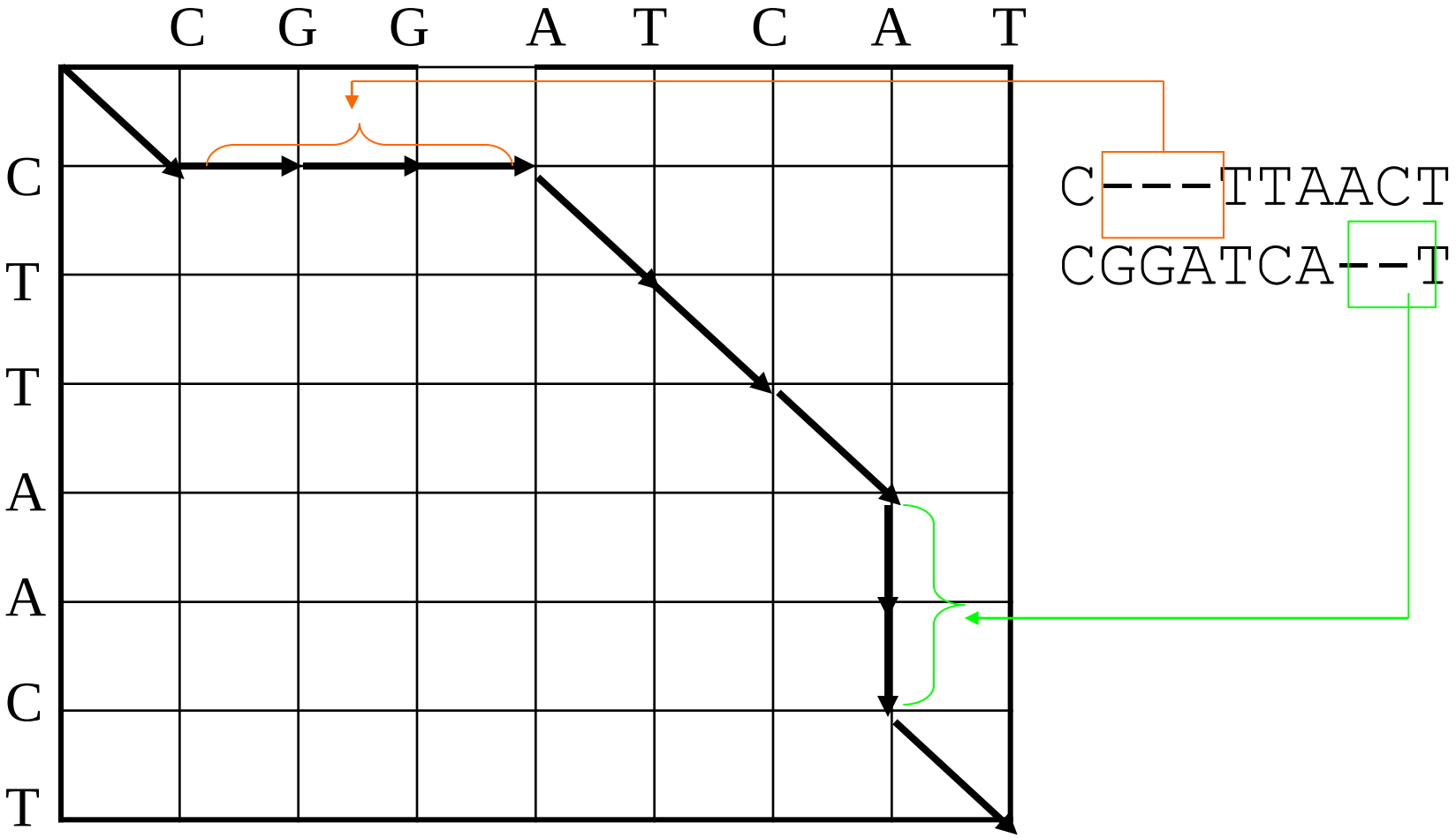
Graphic representation of an alignment



Pathway of an alignment

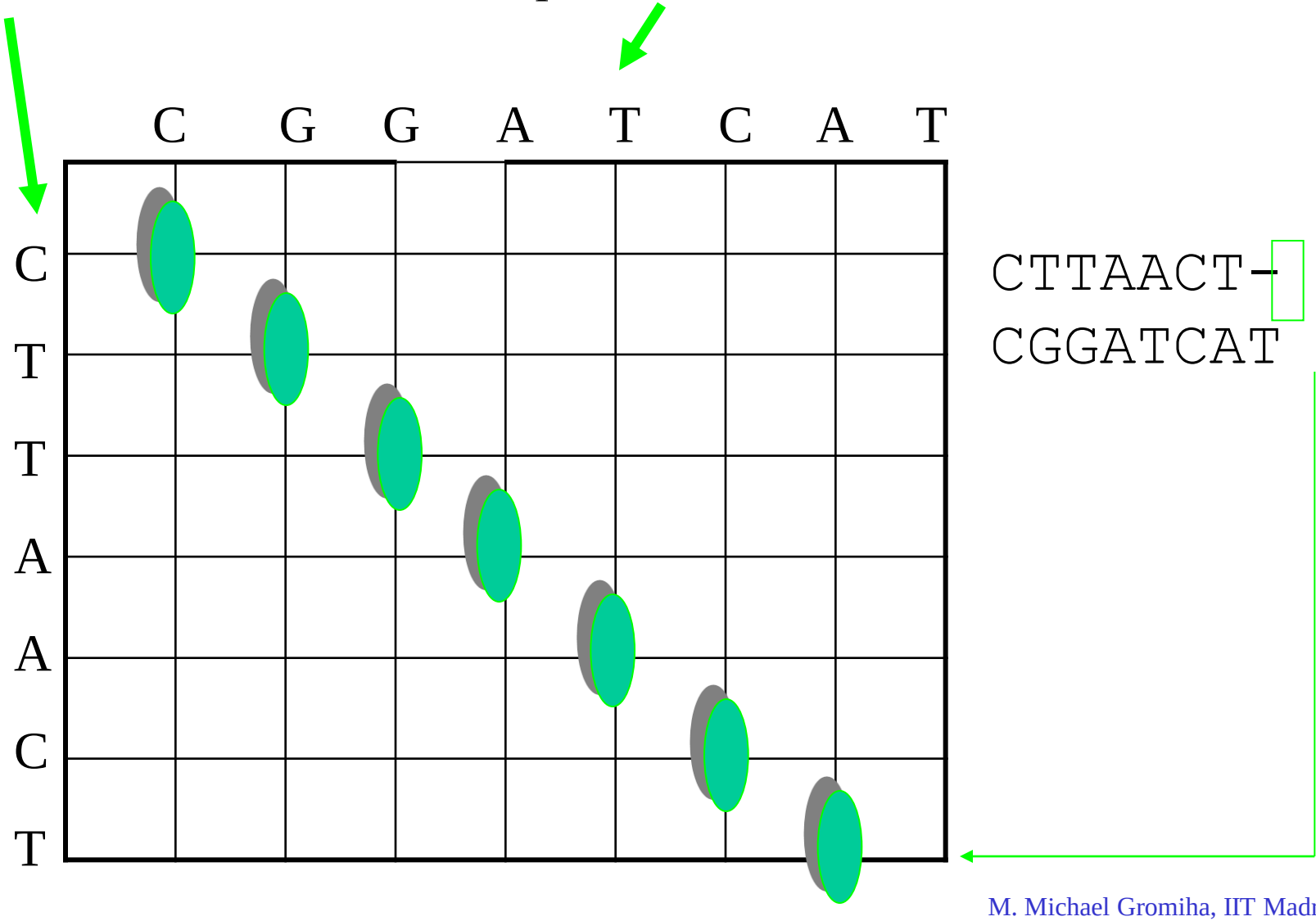
Sequence a: CTTAACT

Sequence b: CGGATCAT



Graphic representation of an alignment

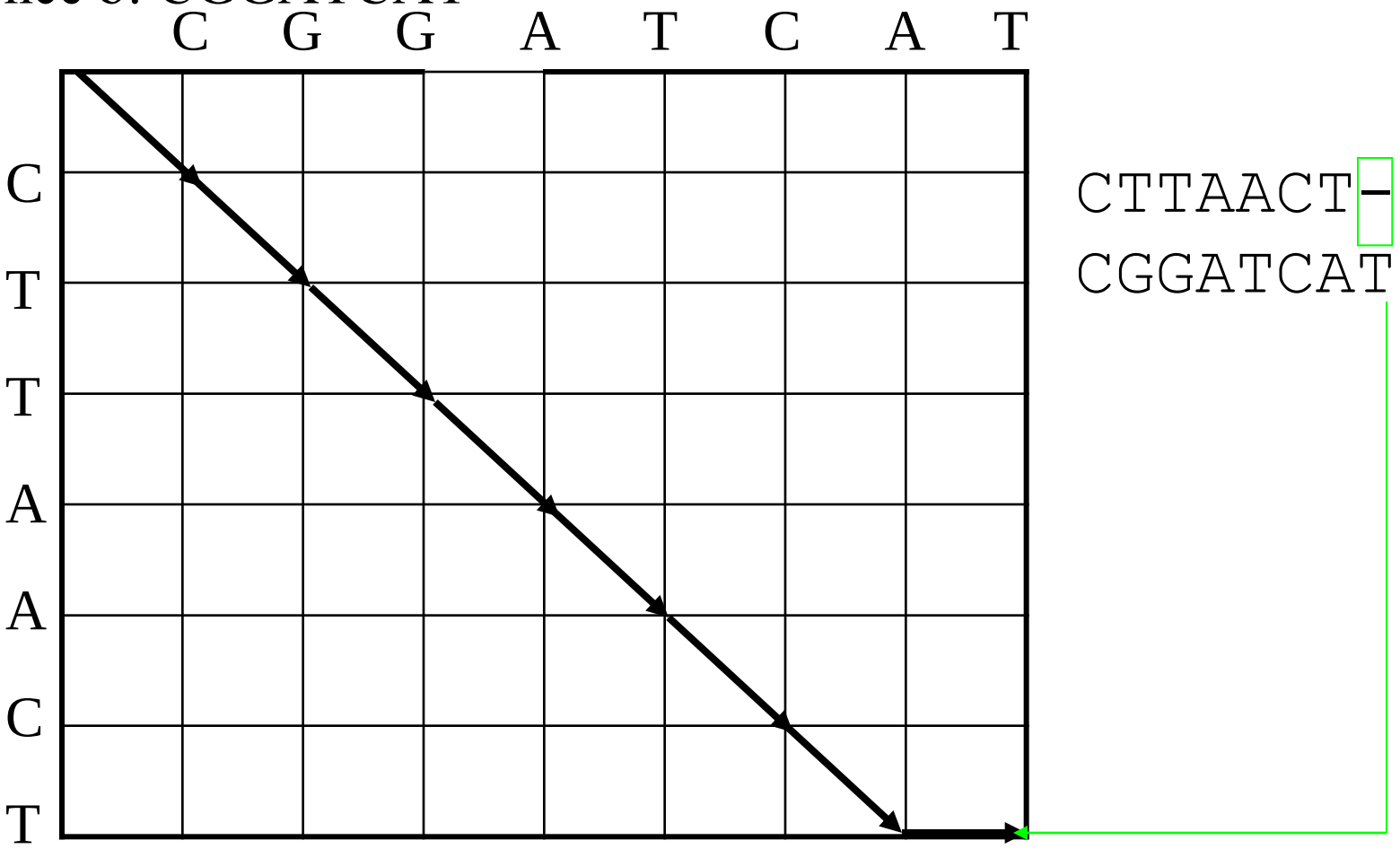
Sequence a: CTTAACT Sequence b: CGGATCAT



Pathway of an alignment

Sequence a: CTTAACT

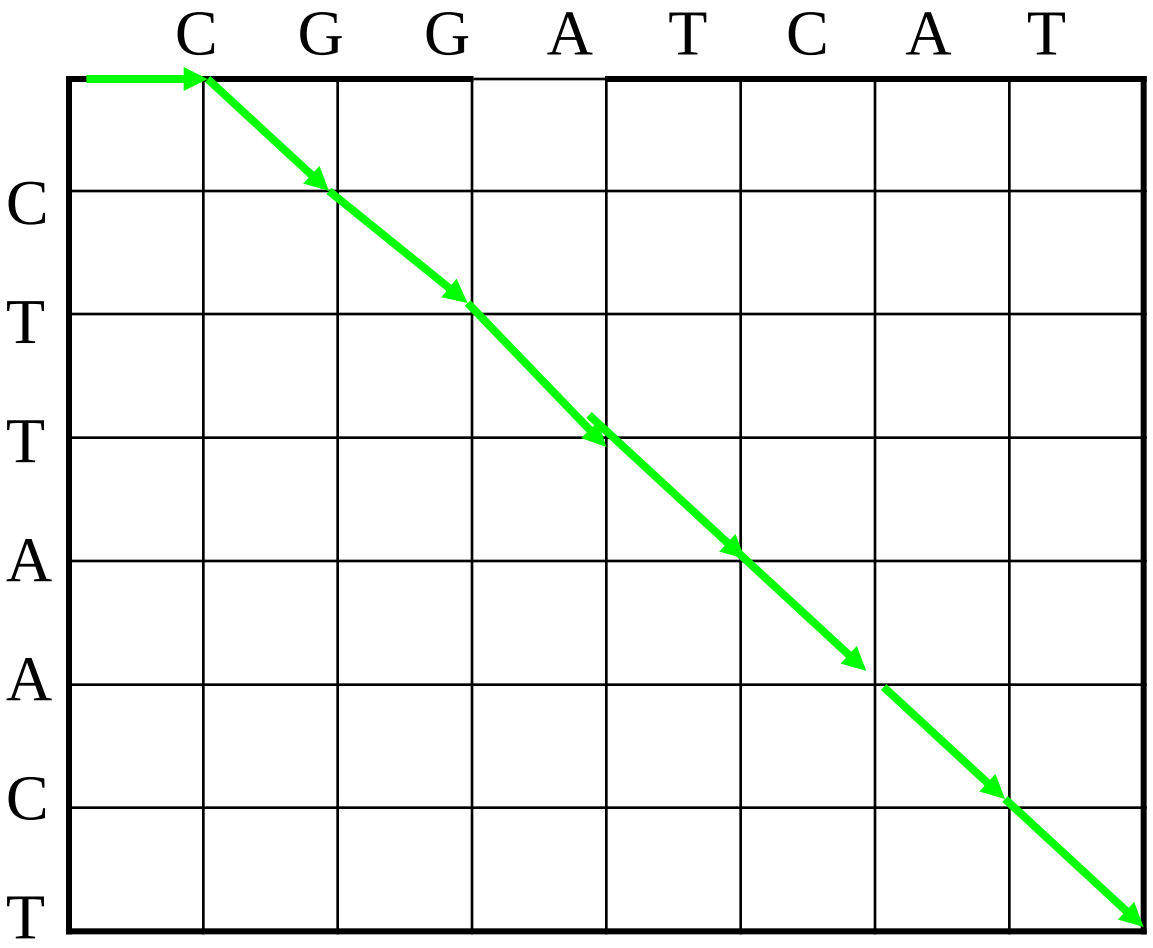
Sequence b: CGGATCAT



Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT

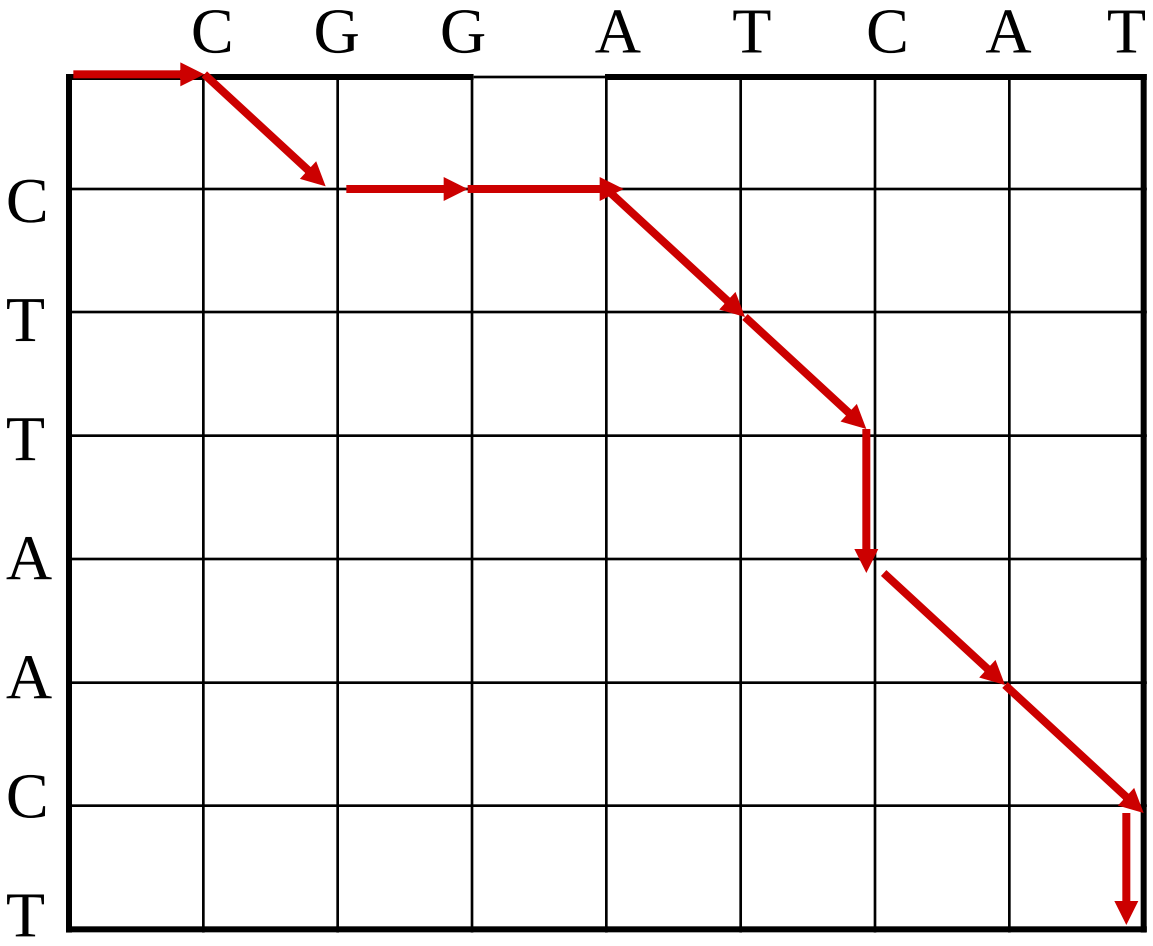


- CTTAACT
CGGATCAT

Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT

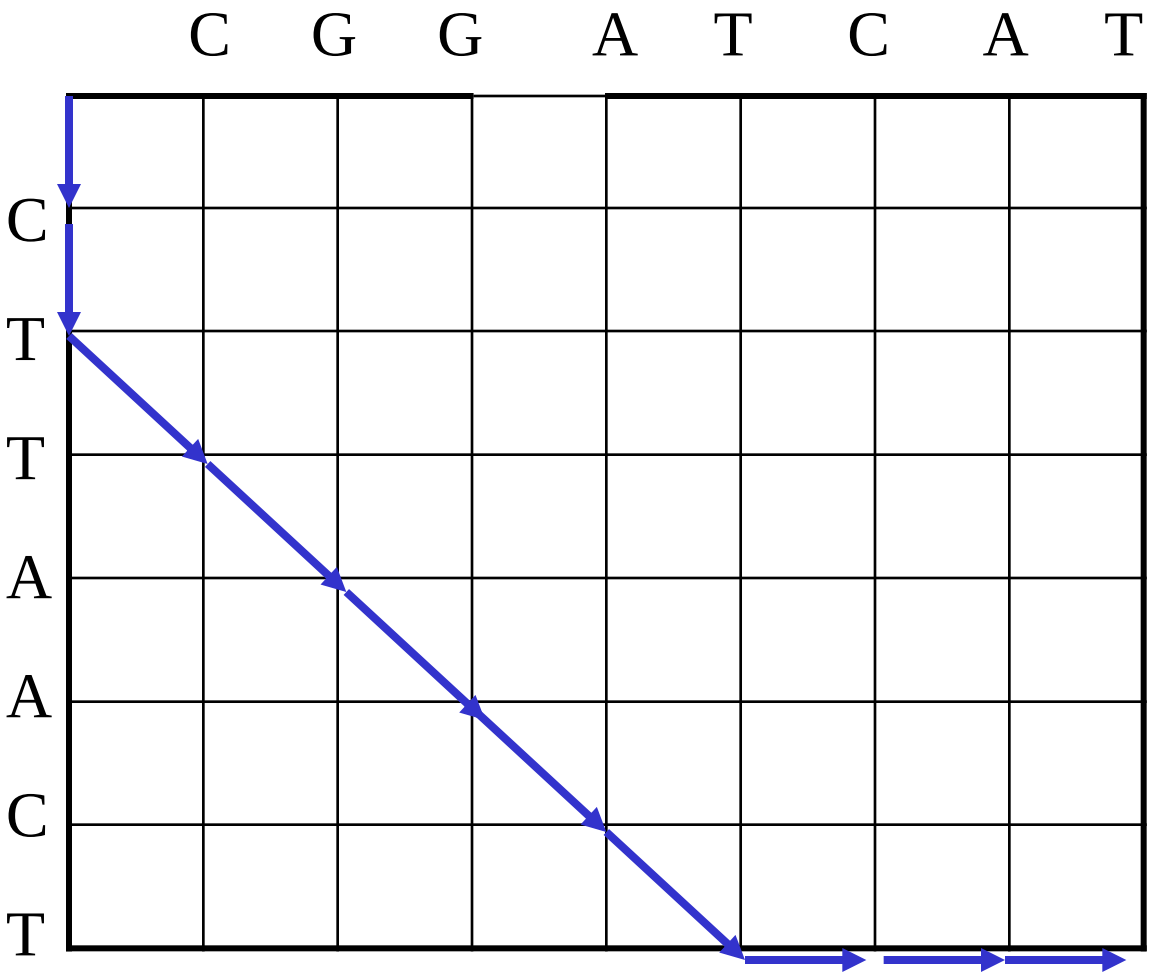


- C - - TTAACT
CGGATC - AT -

Use of graph to generate alignments

Sequence a: CTTAACT

Sequence b: CGGATCAT



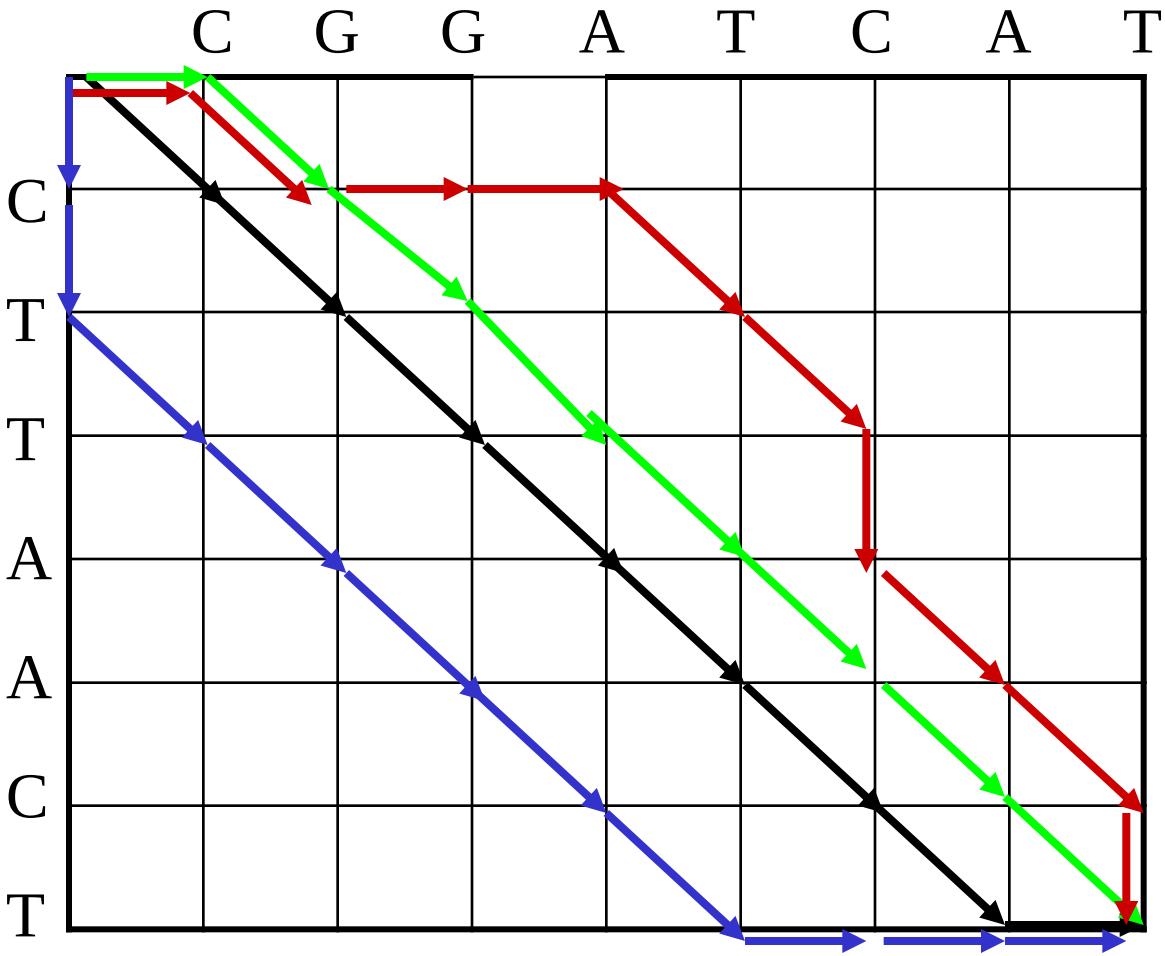
CTTAACT - - -

- - CGGATCAT

Which pathway is better?

Sequence a: CTTAACT

Sequence b: CGGATCAT



Multiple pathways

Each with a unique scoring function

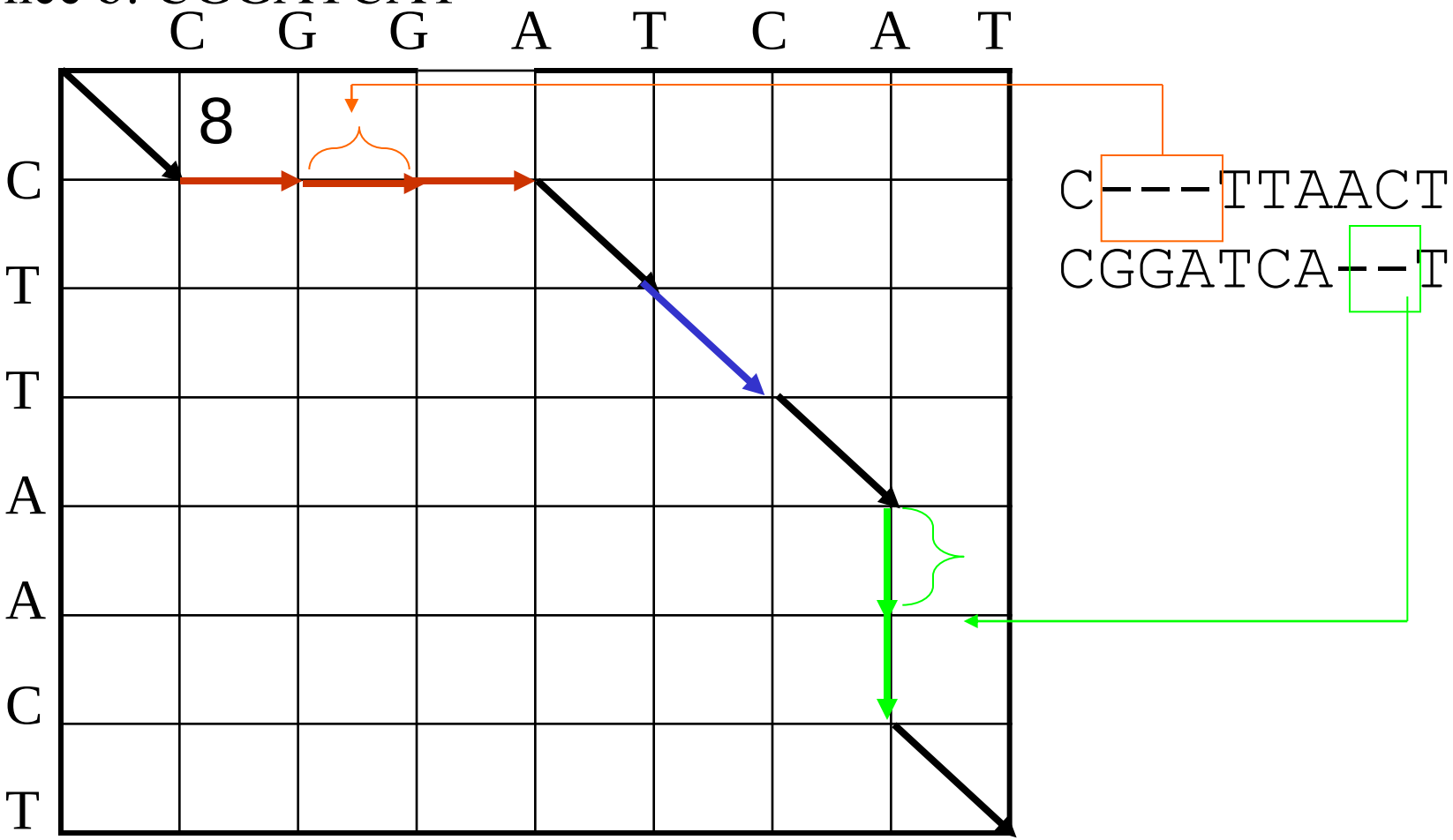
How to rate an alignment?

- Match: +8 ($w(x, y) = 8$, if $x = y$)
- Mismatch: -5 ($w(x, y) = -5$, if $x \neq y$)
- Each gap symbol: -3 ($w(-,x)=w(x,-)=-3$)

Alignment Score

Sequence a: CTTAACT

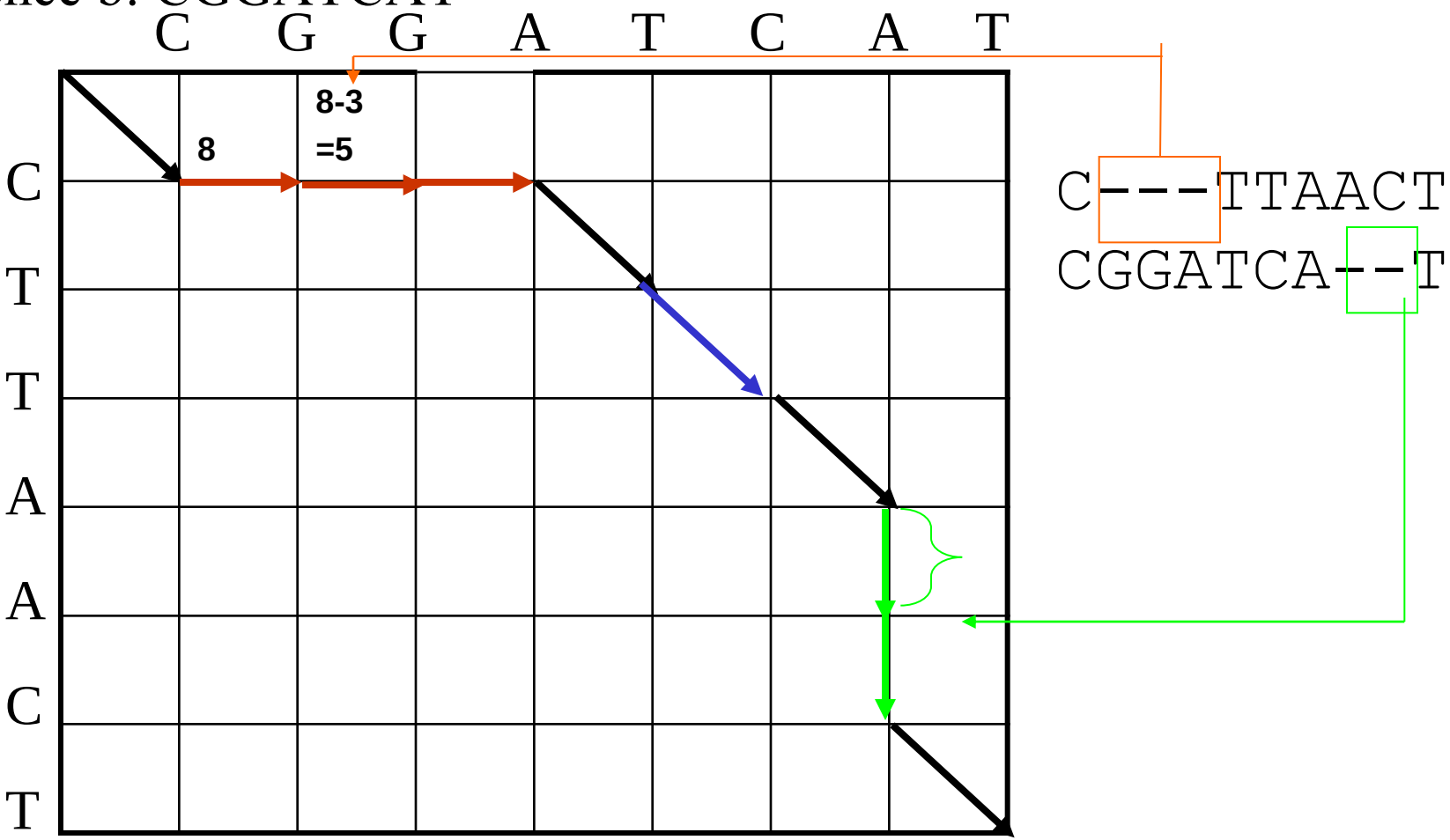
Sequence b: CGGATCAT



Alignment Score

Sequence a: CTTAACT

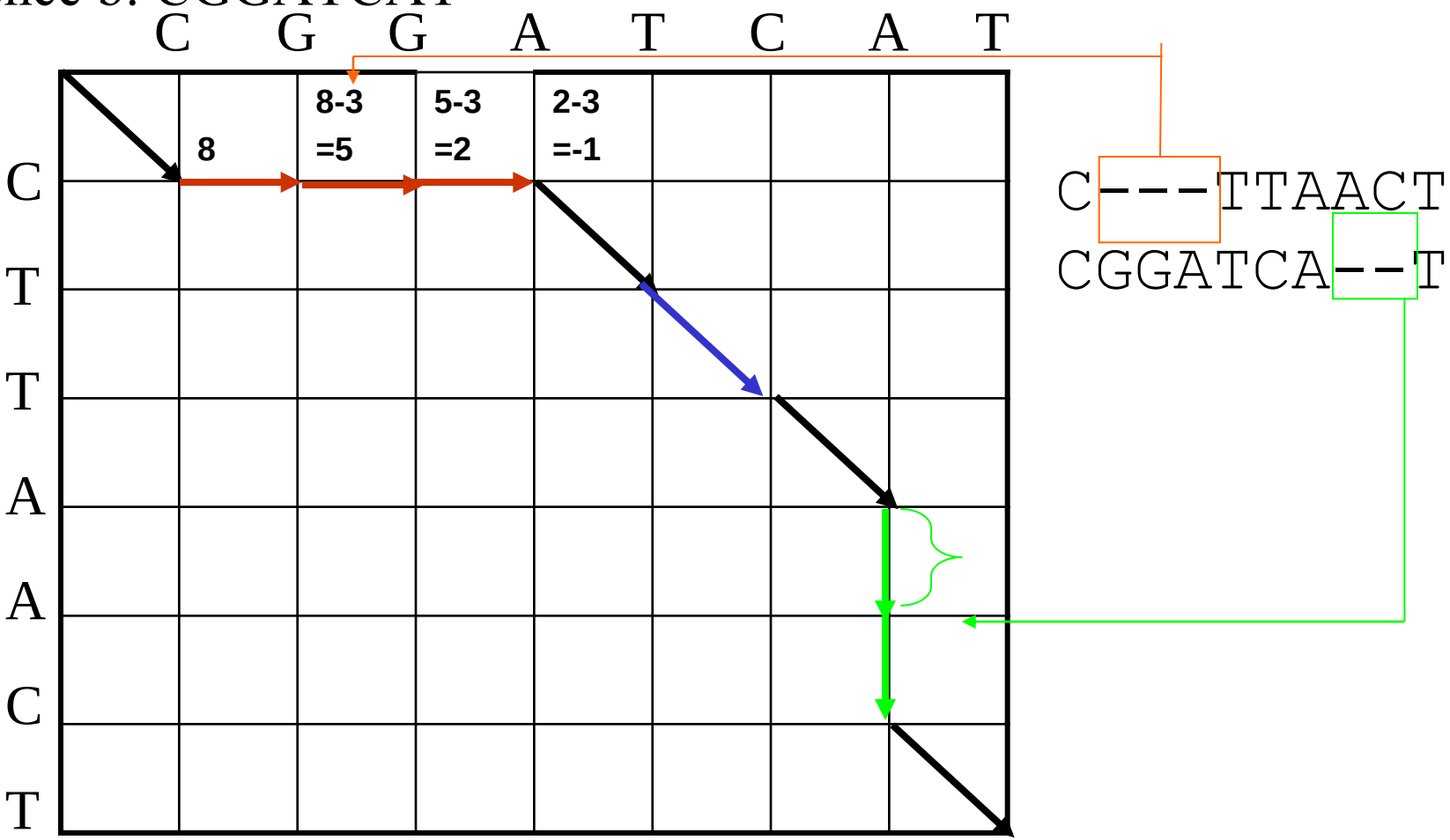
Sequence b: CGGATCAT



Alignment Score

Sequence a: CTTAACT

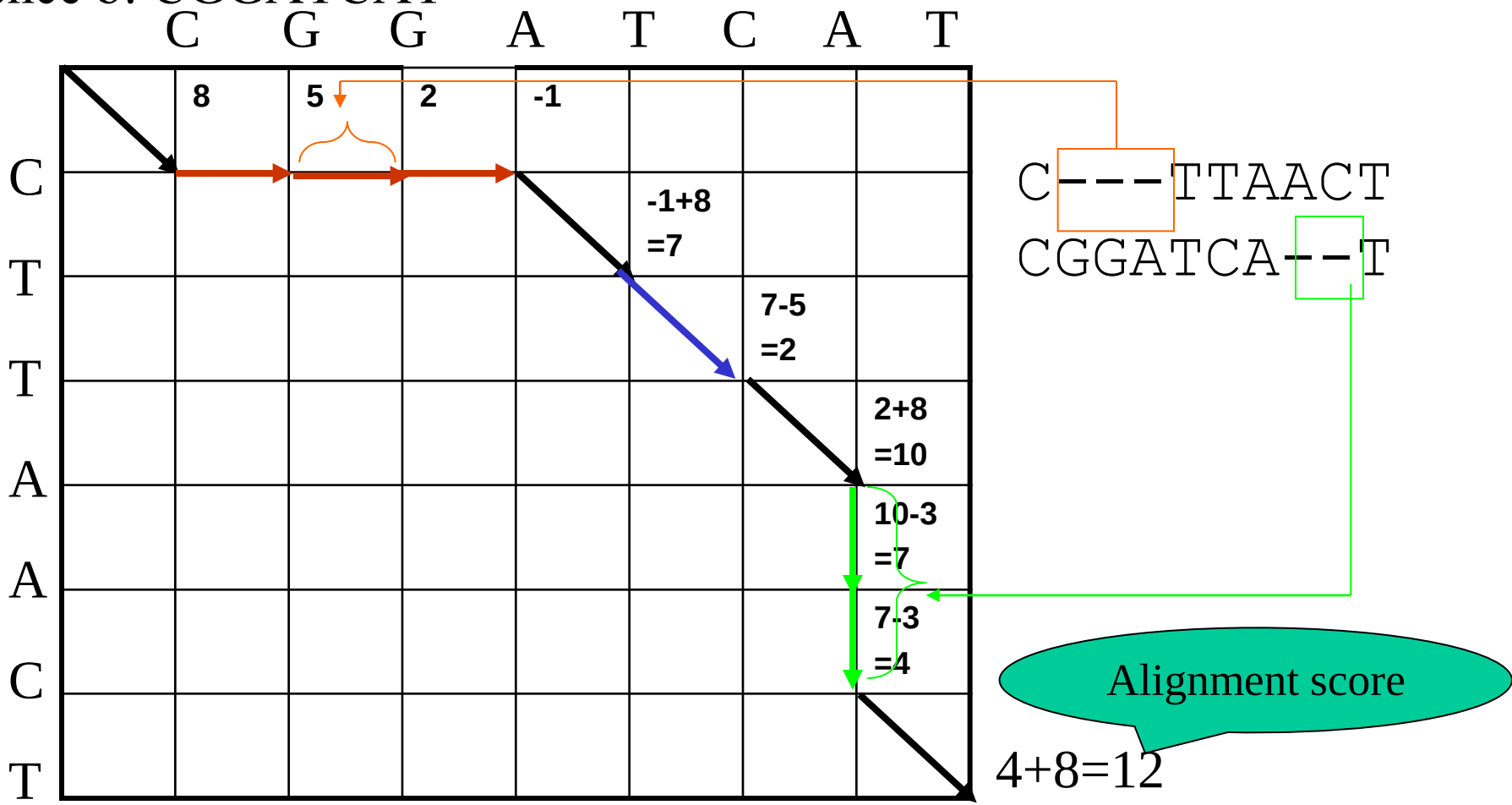
Sequence b: CGGATCAT



Alignment Score

Sequence a: CTTAACT

Sequence b: CGGATCAT



An optimal alignment

-- the alignment of maximum score

Needleman and Wunsch Algorithm

Let A=$a_1a_2...a_m$ and $B=b_1b_2...b_n$.

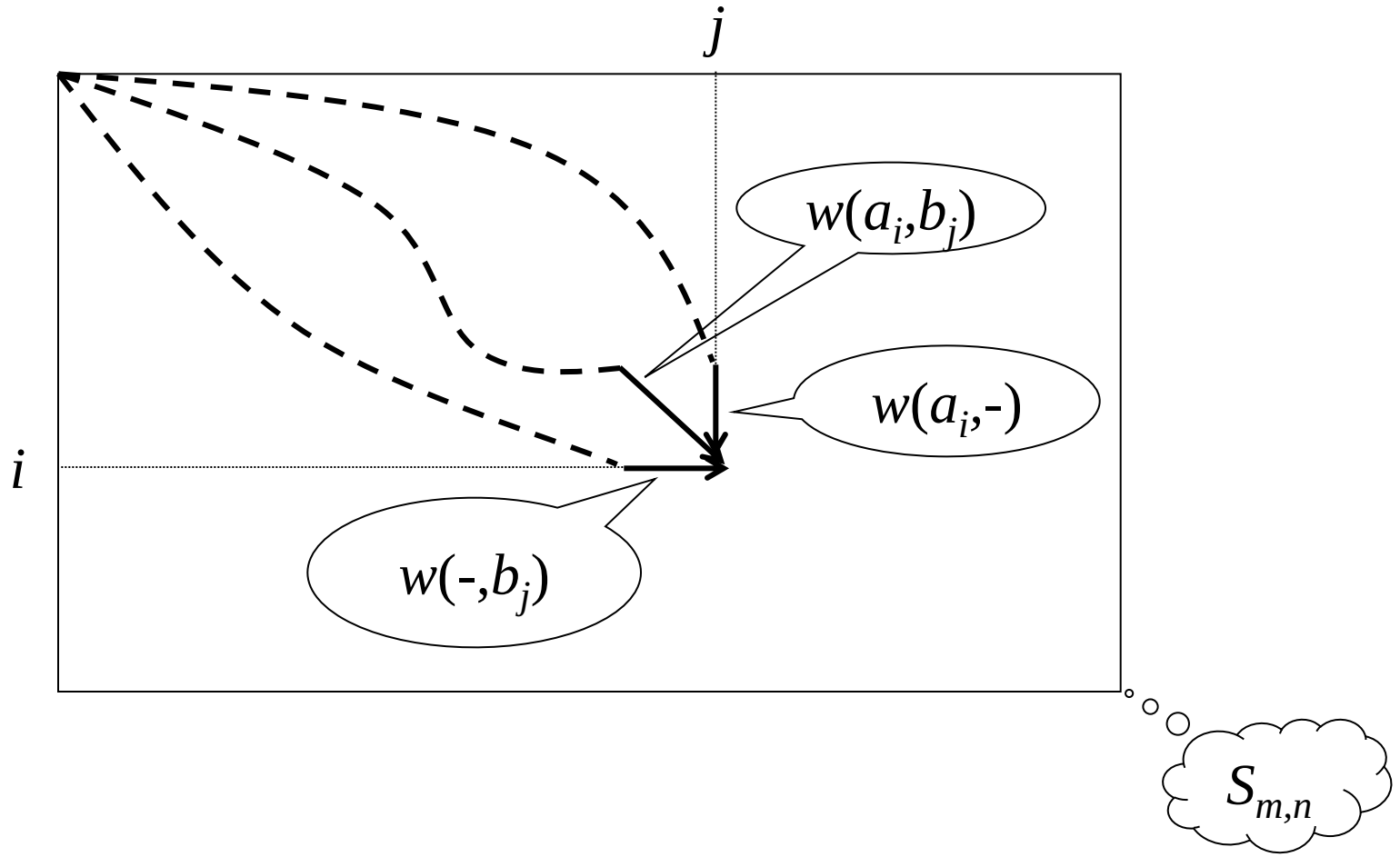
$S_{i,j}$: the score of an optimal alignment between
$a_1a_2...a_i$ and $b_1b_2...b_j$

With proper initializations, $S_{i,j}$ can be computed as follows.

$$s_{i,j} = \max\begin{cases} s_{i-1,j} + w(a_i,-) \\ s_{i,j-1} + w(-,b_j) \\ s_{i-1,j-1} + w(a_i,b_j) \end{cases}$$

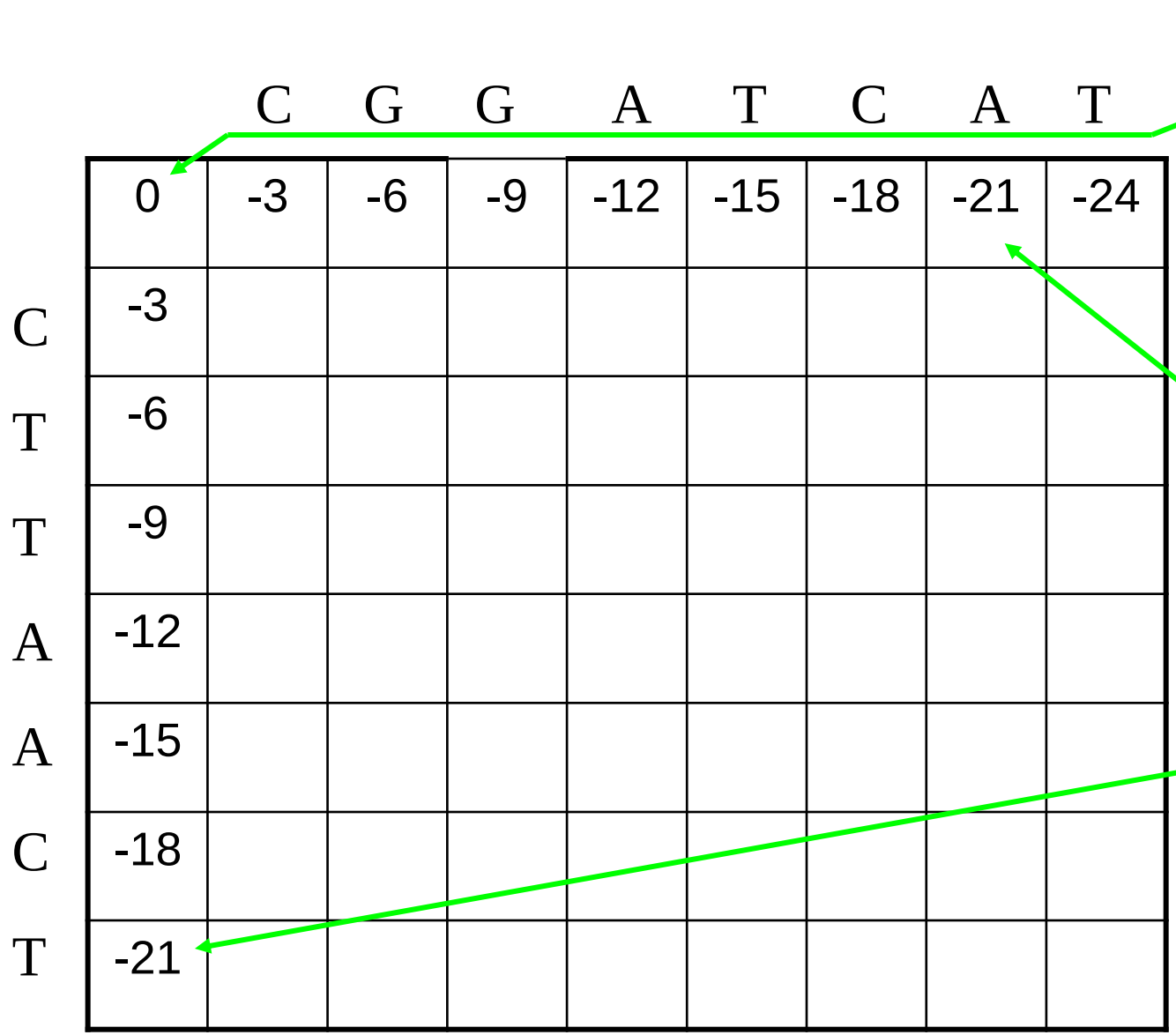
Computing $S_{i,j}$

Take value from left and add gap penalty along the left axis (i)
Take value from above and gap penalty along the top axis (j)
Take value from diagonal element and add match bonus/mismatch penalty (i,j)



Gap symbol: -3

Initializations



$S_{0,0}= 0$

$S_{0,1}=-3$, $S_{0,2}=-6$,

$S_{0,3}=-9$, $S_{0,4}=-12$,

$S_{0,5}=-15$, $S_{0,6}=-18$,

$S_{0,7}=-21$, $S_{0,8}=-24$

$S_{1,0}=-3$, $S_{2,0}=-6$,

$S_{3,0}=-9$, $S_{4,0}=-12$,

$S_{5,0}=-15$, $S_{6,0}=-18$,

$S_{7,0}=-21$

Match: 8

Mismatch: -5

Gap symbol: -3

$S_{1,1} = ?$

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	?							
T	-6								
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

Option 1:

$S_{1,1} = S_{0,0} + w(a_1, b_1)$

$= 0 + 8 = 8$

Option 2:

$S_{1,1} = S_{0,1} + w(a_1, -)$

$= -3 - 3 = -6$

Option 3:

$S_{1,1} = S_{1,0} + w(-, b_1)$

$= -3 - 3 = -6$

Optimal:

$S_{1,1} = 8$

Match: 8

Mismatch: -5

Gap symbol: -3

$$S_{1,2} = ?$$

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	?						
T	-6								
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

Option 1:

$S_{1,2} = S_{0,1} + w(a_1, b_2)$

$= -3 - 5 = -8$

Option 2:

$S_{1,2} = S_{0,2} + w(a_1, -)$

$= -6 - 3 = -9$

Option 3:

$S_{1,2} = S_{1,1} + w(-, b_2)$

$= 8 - 3 = 5$

Optimal:

$S_{1,2} = 5$

Match: 8

Mismatch: -5

Gap symbol: -3

$S_{2,1} = ?$

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5						
T	-6	?							
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

Option 1:

$S_{2,1} = S_{1,0} + w(a_2, b_1)$

$= -3 - 5 = -8$

Option 2:

$S_{2,1} = S_{1,1} + w(a_2, -)$

$= 8 - 3 = 5$

Option 3:

$S_{2,1} = S_{2,0} + w(-, b_1)$

$= -6 - 3 = -9$

Optimal:

$S_{2,1} = 5$

Match: 8

Mismatch: -5

Gap symbol: -3

$$S_{2,2} = ?$$

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5						
T	-6	5	?						
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

Option 1:

$S_{2,2} = S_{1,1} + w(a_2, b_2)$

$= 8 - 5 = 3$

Option 2:

$S_{2,2} = S_{1,2} + w(a_2, -)$

$= 5 - 3 = 2$

Option 3:

$S_{2,2} = S_{2,1} + w(-, b_2)$

$= 5 - 3 = 2$

Optimal:

$S_{2,2} = 3$

$S_{3,5} = ?$

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5	2	-1	-4	-7	-10	-13
T	-6	5	3	0	-3	7	4	1	-2
T	-9	2	0	-2	-5	?			
A	-12								
A	-15								
C	-18								
T	-21								

$S_{3,5} = ?$

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5	2	-1	-4	-7	-10	-13
T	-6	5	3	0	-3	7	4	1	-2
T	-9	2	0	-2	-5	5	-1	-4	9
A	-12	-1	-3	-5	6	3	0	7	6
A	-15	-4	-6	-8	3	1	-2	8	5
C	-18	-7	-9	-11	0	-2	9	6	3
T	-21	-10	-12	-14	-3	8	6	4	14

optimal score

```
C T T A A C – T
C G G A T C A T
```

8 – 5 –5 +8 -5 +8 -3 +8 = 14

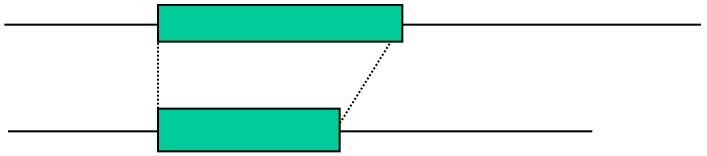
		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5	2	-1	-4	-7	-10	-13
T	-6	5	3	0	-3	7	4	1	-2
T	-9	2	0	-2	-5	5	-1	-4	9
A	-12	-1	-3	-5	6	3	0	7	6
A	-15	-4	-6	-8	3	1	-2	8	5
C	-18	-7	-9	-11	0	-2	9	6	3
T	-21	-10	-12	-14	-3	8	6	4	14

Global Alignment vs. Local Alignment

- global alignment:

All sections are counted

- local alignment:



Only local sections (normally separated by gaps) are counted

Local vs. Global Sequence Alignment:

Example:

DNA sequence a: ATTCTTGC

DNA sequence b: ATCCTATTCTAGC

Local Alignment:

```
        ATTCTTGC
ATCCTATTCTAGC
   /|\
   gap
```

Gaps ignored in local alignments

Global Alignment:

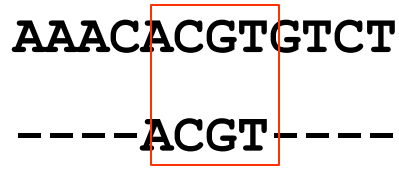
```
AT   TCTT   GC
ATCCTATTCTAGC
/|\       /|\
   gap     gap
```

Gaps counted in global alignments

Semi-global alignment

AAACACGTGTCT

ACGT



Shorter sequence appears entirely within the longer sequence.

Terminal gaps are usually the result of incomplete data acquisition and do not have biological significance.

It is appropriate to treat them separately than internal gaps.

This approach is referred as semi-global alignment.

Local alignment

Semi-global alignment do not afford flexibility needed in a sequence search.

Eg. Find any sub-sequences that are similar to any part of the yeast genome.

All the non-matching residues will be penalized in semi-global alignment.

AACCTATAGCT

GCGATATA

```
AAC-CTATAGCT
-GCGATATA---
```

One sense: fairly a bad alignment;

This alignment reveals the matching region TATA.

The approach to find the best matching sub sequences within the two search sequences is called local alignment.

F. Smith and M. Waterman

Fourth option for alignment with the minimum value "Zero"

An optimal local alignment

$S_{i,j}$: the score of an optimal local alignment ending at a_i and b_j

With proper initializations, $S_{i,j}$ can be computed as follows.

S_{ij} is the maximum among the four values

(i) $S_{i-1,j-1} + w(a_i,b_j)$

(ii) $S_{i-1,j} + w(a_i,-)$

(iii) $S_{i,j-1} + w(-,b_j)$

(iv) 0

Match: 8

Mismatch: -5

Gap symbol: -3

Initializations

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0								
T	0								
T	0								
A	0								
A	0								
C	0								
T	0								

Match: 8

Mismatch: -5

Gap symbol: -3

$$S_{1,1} = ?$$

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	?							
T	0								
T	0								
A	0								
A	0								
C	0								
T	0								

Option 1:

$S_{1,1} = S_{0,0} + w(a_1, b_1)$

$= 0 + 8 = 8$

Option 2:

$S_{1,1} = S_{0,1} + w(a_1, -)$

$= 0 - 3 = -3$

Option 3:

$S_{1,1} = S_{1,0} + w(-, b_1)$

$= 0 - 3 = -3$

Option 4:

$S_{1,1} = 0$

Optimal:

$S_{1,1} = 8$

Match: 8

Mismatch: -5

Gap symbol: -3

local alignment

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	8	5	2	0	0	8	5	2
T	0	5	3	0	0	8	5	3	13
T	0	2	0	0	0	8	5	2	11
A	0	0	0	0	8	5	3	?	
A	0								
C	0								
T	0								

```
A - C - T
A T C A T
8-3+8-3+8 = 18
```

local alignment

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	8	5	2	0	0	8	5	2
T	0	5	3	0	0	8	5	3	13
T	0	2	0	0	0	8	5	2	11
A	0	0	0	0	8	5	3	13	10
A	0	0	0	0	8	5	2	11	8
C	0	8	5	2	5	3	13	10	7
T	0	5	3	0	2	13	10	8	18

Local sequence alignment: Example 2

AACCTATAGCT

GCGATATA

Initializations

Match: 1;

Mismatch: -1;

Gap symbol: -1

		A	A	C	C	T	A	T	A	G	C	T
	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
C	0											
G	0											
A	0											
T	0											
A	0											
T	0											
A	0											

Alignment

Match: 1;

Mismatch: -1;

Gap symbol: -1

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

TATA
TATA

Questions

1. Using the Needleman and Wunsch dynamic programming method, construct the partial alignment score table for the following two sequences, using the scoring parameters: match score: +1; mismatch score: 0 and gap penalty: -1

   ACAGTCGAACG and ACCGTCCG

2. Using the Smith-Waterman method, construct the partial alignment scoring table for a local alignment of the following two sequences:

   ACGTATCGCGTATA and GATGCTCTCGGAAA

   scoring parameters: match score: +1; mismatch score: 0 and gap penalty: -1