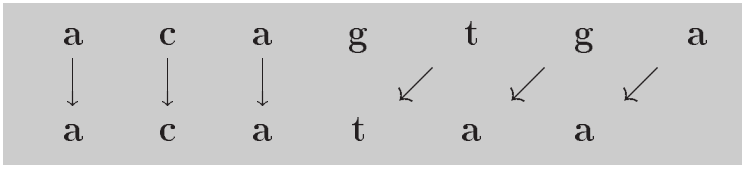
SEQUENCE SIMILARITY AND SEQUENCE ALIGNMENT

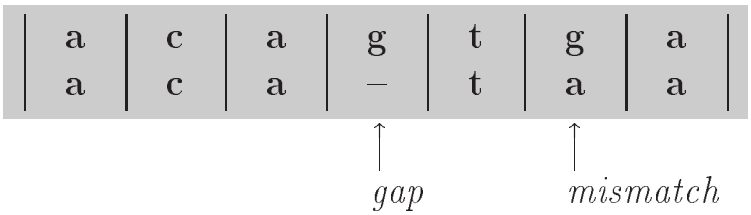
1 SEQUENCE ALIGNMENT

Given two sequences in an alphabet,



$$DNA = \{A, C, G, T\}$$

A sequence alignment is a correspondence between the letters from the two sequences.



2 SEQUENCE SIMILARITY (or DISTANCE)

Given an alignment between two sequences, and a distance matrix, we can compute the distance (similarity) between the two sequences.

	a	c	g	t	-
a	0	1	1	1	1
c	1	0	1	1	1
g	1	1	0	1	1
t	1	1	1	0	1
–	1	1	1	1	0

distance matrix

$\Longleftrightarrow$

$$d(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$$

scoring schema

The distance between two aligned sequences is the sum of the distances between each pair of aligned letters.

a	c	a	g	t	g	a	
a	c	a	–	t	a	a	
0	**0**	**0**	**1**	**0**	**1**	**0**	$\Longrightarrow \underline{\mathbf{2}}$

a	c	a	g	t	g	a	
a	c	a	t	a	–	a	
0	**0**	**1**	**1**	**0**	**1**	**0**	$\Longrightarrow \underline{\mathbf{3}}$

Different alignments between the two same sequences result in different distances

a	c	a	g	t	g	a	
a	c	a	–	t	a	a	
0	**0**	**0**	**2**	**0**	**1**	**0**	$\Longrightarrow \underline{\mathbf{2}}$

The same alignment between the same two sequences may result in different distances, depending on the schoring schema,

match	$d = 0$
mismatch	$d = 1$
gap	$d = 2$

3 OPTIMAL ALIGNMENT (GLOBAL)

Problem: Given two sequences and an scoring schema, find the alignment that minimizes the distance between the sequences (the Optimal Alignment)

• *Why is not possible to explicitlly compute all possible alignments between two sequences?.*
The number of (global) alignments between two sequences of length n is $g(n) \sim 2^{2n}/4\sqrt{n\pi}$. If $n = 100$, then $g(n) > 10^{600}$ (Waterman, 1984)

Dynamic Programming (Needleman and Wunsch, 1970)

	–	**a**	**c**	**a**	**g**	**t**	**g**	**a**	
–	0	1	2	3	4	5	6	7	
a	1	0	1	2	3	4	5	6	
c	2	1	0	1	2	3	4	5	
a	3	2	1	0	1	2	3	4	
t	4	3	2	1	1	1	2	3	
t	5	4	3	2	2	2	2	2	
t	6	5	4	3	3	3	3	2	

In the practive, given two sequences A and B of lengths m and n, we build a matrix D where each element D_{ij} stores the minimal distance between the subsequences A_i and B_j. Such minimal distances are computed, according with the following recurrence

$D_{00} = 0$; $D_{0j} = \sum_{k=1}^{j} d(--, b_k)$; $D_{i0} = \sum_{k=1}^{i} d(a_k, -)$

$$\text{then, } D_{ij} = \min \begin{cases} D_{i-1,j-1} + d(a_i, b_j) \\ D_{i-1,j} + d(a_i, -) \\ D_{i,j-1} + d(-, b_j) \end{cases}$$

• Given an schoring schema, the dynamic programming algorithm garanties that the alignment found minimizes the distance.

• The execution time of these algorithms is proportional to the product $m \times n$.

3.1 SCORING MATRICES

The aligment depends on the scoring matrices. Below two distinct dynamic programming alignments of the same two proteins using two different matrices

```
34.4% identity; Optimized score: 224

               10        20        30        40        50        60
EGR1_H MAAAKAEMQLMSPLQISDPFGSFPHSPTMDNYPKLEEMMLLSNGAPQFLGAAGAPEGSGS
       :..  :: .:  :. .:. .. .:.    . ::  ::.
EGR3_H MTGKLAE-KL--PVTMSSLLNQLPD----NLYP--EEI----------------------
                 10        20

               70        80        90       100       110       120
EGR1_H NSSSSSSGGGGGGGGGSNSSSSSSTFNPQADTGEQPYEHLTAESFPDISLNNEKVLVETS
                        .:. . :.  .:.  . :.....:.  ::.:.:::   : :
EGR3_H --------------------PSALNLFSGSSDSVVH-YNQMATENVMDIGLTNEKPNPELS
                        30        40        50        60

                  130       140       150       160       170
EGR1_H YPSQTTRLP---PITYTGRFSLEPAPNSGNTLW-PEPLFSLVSGLVSMTNPPASSSSAPS
       :..   . :   ..:: :.:......:     : .. ..:.:.. . .. ::::.  : :
EGR3_H YSGSFQPAPGNKTVTYLGKFAFDSPSN-----WCQDNIISLMSAGI-LGVPPASG--ALS
       70        80        90              100       110           120

          180       190       200       210       220       230
EGR1_H PAASSASASQSPPLSCAVPSNDSSPIYSAAPTFPTPNTDIFPEPQSQAFPGSAGTA-LQY
       ...:.::  :.:.       .: ...:.: :.... . :...:: :  :....:.. :.:
EGR3_H TQTSTASMVQPPQ-------GDVEAMYPALPPYSNCG-DLYSEPVS--FHDPQGNPGLAY
             130               140       150       160       170

           240          250       260       270       280       290
EGR1_H PPPAYPAAKGGFQV---PMIPDYLFPQQQGDLGLGTPDQKPFQGLES-RTQQPSLTPLST
       .:..:..:: ...    :::::: . ....:.: ..:..:::::... :...:..:::.:
EGR3_H SPQDYQSAKPALDSNLFPMIPDYNLYHHPNDMG-SIPEHKPFQGMDPIRVNPPPITPLET
             180       190       200       210       220       230

             300       310       320       330       340       350
EGR1_H IKAFATQSGSQDLKALNTSYQSQLIKPSRMRKYPNRPSKTPPHERPYACPVESCDRRFSR
       :::: ..    ... .:  .   .  .:: : ::::::::: :::.:::.:.:::::::
EGR3_H IKAFKDKQIHPGFGSLPQP---PLTLKPIRPRKYPNRPSKTPLHERPHACPAEGCDRRFSR
             240          250       260       270       280

             360       370       380       390       400       410
EGR1_H SDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFACDICGRKFARSDER
       :::::::.:::::.:::::::::::.:::::::::::::::::::::::..::::::::::
EGR3_H SDELTRHLRIHTGHKPFQCRICMRSFSRSDHLTTHIRTHTGEKPFACEFCGRKFARSDER
       290       300       310       320       330       340

             420       430       440       450       460       470
EGR1_H KRHTKIHLRQKDKKADKSVVASSATSSLSSYPSPVATSYPSPVTTSYPSPATTSYPSPVP
       :::.::::.::.:::.:. ..:.....      ::. .   ::.:...
EGR3_H KRHAKIHLKQKEKKAEKGGAPSASSAP------PVSLA---PVVTTCA------------
       350       360       370             380


34.2% identity; Optimized score: 516

               10        20        30        40        50        60
EGR1_H MAAAKAEMQLMSPLQISDPFGSFPHSPTMDNYPKLEEMMLLSNGAPQFLGAAGAPEGSGS
       :..  ::.....                          : .:.:  :  : ... :  :..
EGR3_H MTGKLAEKLPVT-------------------------MSSLLNQLPDNLYPEEIP--SAL
               10                                    20        30

               70        80        90       100       110       120
EGR1_H NSSSSSSGGGGGGGGSNSSSSSSTFNPQADTGEQPYEHLTAESFPDISLNNEKVLVETS
       :.:.::....                      .: ....:.. ::.:.:::   : :
EGR3_H NLFSGSSDSVV------------------------HYNQMATENVMDIGLTNEKPNPELS
              40                                    50        60

              130       140       150       160       170       180
EGR1_H YPSQTTRLPP ITYTGRFSLEPAPNSGNTLWPEPLFSLVSGLVSMTNPPASSSSAPSPAAS
       : : . ...: . :.... . : .: ..... ..::.:. ...  ::::.  :.:. .:
EGR3_H Y-SGSFQPAPGNKTVTYLGKFAFDSPSNWCQDNIISLMSA-GILGVPPASG--ALSTQTS
       70          80        90        100         110         120

              190       200       210       220       230       240
EGR1_H SASASQSPPLSCAVPSNDSSPIYSAAPTFPTPNTDIFPEPQSQAFPGSAGTALQYPPPAY
       .::  : ::.. ... . ..: ::    . .  ...:..::.   ::.: ... :.. :
EGR3_H TASMVQ-PPQGDVEAMYPALPPYSNCGDLYSE-PVSFHDPQG--NPGLAYSPQDYQS-AK
         130       140       150       160       170       180

              250       260       270       280       290
EGR1_H PAAKGGFQVPMIPDYLFPQQQGDLGLGTPDQKPFQGLESRTQ-QPSLTPLSTIKAFATQS
       ::  .. ..:::::: . ... :.:. .:..::::::.....  .:..::: ::::  .
EGR3_H PAL-DSNLFPMIPDYNLYHHPNDMGS-IPEHKPFQGMDPIRVNPPPITPLETIKAFKDKQ
              190       200       210       220       230

     300       310       320       330       340       350
EGR1_H GSQDLKALNTSYQSQLIKPSRMRKYPNRPSKTPPHERPYACPVESCDRRFSRSDELTRHI
         ... .:  . ... .::.: ::::::::::::.::::.:::.:.:::::::::::::::.
EGR3_H IHPGFGSL--PQPPLTLKPIRPRKYPNRPSKTPLHERPHACPAEGCDRRFSRSDELTRHL
       240       250       260       270       280       290

     360       370       380       390       400       410
EGR1_H RIHTGQKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFACDICGRKFARSDERKRHTKIHL
       :::::.:::::::::.:::::::::::::::::::::..:::::::::::::::.::::
EGR3_H RIHTGHKPFQCRICMRSFSRSDHLTTHIRTHTGEKPFACEFCGRKFARSDERKRHAKIHL
        300       310       320       330       340       350

     420       430       440       450       460       470
EGR1_H RQKDKKADKSVVASSATSSLSSYPSPVATSYPSPVTTSYPSPATTSYPSPVPTSFSSPGS
       .::.:::.:.. :.::.:... ...::.:...
EGR3_H KQKEKKAEKGG-APSASSAPPVSLAPVVTTCA----------------------------
        360       370       380
```

pam250

```
;P standard pam250 matrix for proteins
 -12 -4
@ *
A R N D C Q E G H I  L  K  M  F  P  S  T  W  Y  V B Z X
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 2 6 0
  2
 -2  6
  0  0  2
  0 -1  2  4
 -2 -4 -4 -5 12
  0  1  1  2 -5  4
  0 -1  1  3 -5  2  4
  1 -3  0  1 -3 -1  0  5
 -1  2  2  1 -3  3  1 -2  6
 -1 -2 -2 -2 -2 -2 -2 -3 -2  5
 -2 -3 -3 -4 -6 -2 -3 -4 -2  2  6
 -1  3  1  0 -5  1  0 -2  0 -2 -3  5
 -1  0 -2 -3 -5 -1 -2 -3 -2  2  4  0  6
 -4 -4 -4 -6 -4 -5 -5 -5 -2  1  2 -5  0  9
  1  0 -1 -1 -3  0 -1 -1  0 -2 -3 -1 -2 -5  6
  1  0  1  0  0 -1  0  1 -1 -1 -3  0 -2 -3  1  2
  1 -1  0  0 -2 -1  0  0 -1  0 -2  0 -1 -3  0  1  3
 -6  2 -4 -7 -8 -5 -7 -7 -3 -5 -2 -3 -4  0 -6 -2 -5 17
 -3 -4 -2 -4  0 -4 -4 -5  0 -1 -1 -4 -2  7 -5 -3 -3  0 10
  0 -2 -2 -2 -2 -2 -2 -1 -2  4  2 -2  2 -1 -1 -1  0 -6 -2  4
  0 -1  2  3 -4  1  2  0  1 -2 -3  1 -2 -5 -1  0  0 -5 -3 -2  2
  0  0  1  3 -5  3  3 -1  2 -2 -3  0 -2 -5  0  0 -1 -6 -4 -2  2  3
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

genetic code

```
;P genetic code matrix
 -12 -4
@ *
A R N D C Q E G H I L K M F P S T W Y V B Z X
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 3 6 0
  6
 -2  6
 -2 -2  6
  2 -2  2  6
 -2  2 -2 -2  6
 -2  2 -2 -2 -6  6
  2 -2 -2  2 -6  2  6
  2  2 -2  2  2 -2  2  6
 -2  2  2  2 -2  2 -2 -2  6
 -2  2  2 -2 -2 -2 -2 -2 -2  6
 -2  2 -2 -2 -2  2 -2 -2  2  2  6
 -2  2  2 -2 -6  2  2 -2 -2  2 -2  6
 -2  2 -2 -6 -6 -2 -2 -2 -6  2  2  2  6
 -2 -2 -2 -2  2 -6 -6 -2 -2  2  2 -6 -2  6
  2  2 -2 -2 -2  2 -2 -2  2 -2  2 -2 -2 -2  6
  2  2  2 -2  2 -2 -2  2 -2  2  2 -2 -2  2  2  6
  2  2  2 -2 -2 -2 -2 -2 -2  2 -2  2  2 -2  2  2  6
 -2  2 -6 -6  2 -2 -2  2 -6 -6  2 -2 -2 -2 -2  2 -2  6
 -2 -2  2  2  2 -2 -2 -2 -2 -2 -6  2 -2  2 -2 -2  6
  2 -2 -2  2 -2 -2  2  2 -2  2  2 -2  2  2 -2 -2 -2 -2  6
  2 -2  6  6 -2  2  2  2  2  2 -2  2 -2 -2 -2  2  2 -6  2  2  6
  2  2  2  2 -6  6  6  2  2 -2  2  2 -2 -6  2 -2 -2 -2 -2  2  2  6
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

3.2 GAP PENALTIES

a	c	a	g	t	g	a	a	
a	c	a	g	–	–	c	a	
0	**0**	**0**	**0**	**1**	**1**	**1**	**0**	$\Longrightarrow \underline{\mathbf{3}}$

a	c	a	g	t	g	a	a	
a	c	a	g	c	–	a	–	
0	**0**	**0**	**0**	**1**	**1**	**0**	**1**	$\Longrightarrow \underline{\mathbf{3}}$

An scoring schema in which gaps are treated as mismatches is quite unrealistic. For instance, let's assume the following schema

$$d(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$$

According wich such an scoring schema, the two alignments at the right have the same distance. However, from an evolutionary stanpoint the second alignment is less likely than the first one, since it requires two indpendent *insertion/deletion* events, while the first one can be explained by only one.

a	c	a	g	t	g	a	a	
a	c	a	g	–	–	c	a	
0	**0**	**0**	**0**	**2**	**1**	**1**	**0**	$\Longrightarrow \underline{\mathbf{4}}$

a	c	a	g	t	g	a	a	
a	c	a	g	c	–	a	–	
0	**0**	**0**	**0**	**1**	**2**	**0**	**2**	$\Longrightarrow \underline{\mathbf{5}}$

In the *Affine Gap Model* the penalty for gaps has two parts: a penalty for initializing a gap, and another that depends linearly on the length of the gap. That is, gap penalty is $W_g + qW_s$, where W_g and W_s are both constants, and $q \geq 1$ is the length of the gap. If $W_g = 1$ and $W_s = 1$, then the distances of the above two alignments is indeed different.

4 LOCAL ALIGNMENTS

Smith and Waterman (1981)

```
(A) EGR1_HUMAN.fa >EGR1_HUMAN                                        - 543 aa
(B) EGR3_HUMAN.fa >EGR3_HUMAN                                        - 387 aa
 using protein matrix

  47.3% identity in 372 aa overlap; score:  781

          90        100        110        120           130        140
EGR1_H FNPQADTGEQPYEHLTAESFPDISLNNEKVLVETSYPSQTTRLP---PITYTGRFSLEPA
       :.  .:.  . :.....:.  ::.:.:::   : ::..   . :   ..:: :.:.....
EGR3_H FSGSSDSVVH-YNQMATENVMDIGLTNEKPNPELSYSGSFQPAPGNKTVTYLGKFAFDSP
          40         50         60         70         80         90

            150        160        170        180        190        200
EGR1_H PNSGNTLW-PEPLFSLVSGLVSMTNPPASSSSAPSPAASSASASQSPPLSCAVPSNDSSP
       .:     : .. ..::.:. . .. ::::  .: :...:.::  :.:.       .: ..
EGR3_H SN-----WCQDNIISLMSAGI-LGVPPAS--GALSTQTSTASMVQPPQ-------GDVEA
               100        110          120        130

             210        220        230        240          250
EGR1_H IYSAAPTFPTPNTDIFPEPQSQAFPGSAGTA-LQYPPPAYPAAKGGFQV---PMIPDYLF
       .:.: :.... . :...:: :  :....:.. :.:.:..:..:: ...    :::::: .
EGR3_H MYPALPPYSNCG-DLYSEPVS--FHDPQGNPGLAYSPQDYQSAKPALDSNLFPMIPDYNL
     140        150          160        170        180        190

       260        270        280        290        300        310
EGR1_H PQQQGDLGLGTPDQKPFQGLES-RTQQPSLTPLSTIKAFATQSGSQDLKALNTSYQSQLI
        ....:.: ..:..:::::... :...:..:::.:::::  ..   ... .:  .  .  .
EGR3_H YHHPNDMG-SIPEHKPFQGMDPIRVNPPPITPLETIKAFKDKQIHPGFGSLPQP--PLTL
        200         210        220        230        240         250

        320        330        340        350        360        370
EGR1_H KPSRMRKYPNRPSKTPPHERPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRN
       :: : :::::::::::: ::::.:::.:.:::::::::::::::.:::::.::::::::::.
EGR3_H KPIRPRKYPNRPSKTPLHERPHACPAEGCDRRFSRSDELTRHLRIHTGHKPFQCRICMRS
           260        270        280        290        300        310

        380        390        400        410        420        430
EGR1_H FSRSDHLTTHIRTHTGEKPFACDICGRKFARSDERKRHTKIHLRQKDKKADKSVVASSAT
       :::::::::::::::::::::..:::::::::::::::.::::.::.:::.:. ..:...
EGR3_H FSRSDHLTTHIRTHTGEKPFACEFCGRKFARSDERKRHAKIHLKQKEKKAEKGGAPSASS
           320        330        340        350        360        370

        440
EGR1_H SSLSSYPSPVAT
       ..  : .. :.:
EGR3_H APPVSLAPVVTT
           380
```

5 MULTIPLE ALIGNMENTS

```
EGR1_HUMAN      MAAAKAEMQLMSPLQISDPFGSFPHSPTMDNYPKLEEMMLLSNGAPQFLGAAGAPEGSGSN--SSSSSSGGGGGGGGSNSSSSSSTFNPQADTGEQPYEHLTAESFPDISLNNEKVLVE
EGR1_MOUSE      MAAAKAEMQLMSPLQISDPFGSFPHSPTMDNYPKLEEMMLLSNGAPQFLGAAGTPEGSGGN---SSSSTSSGGGGGGGSNSGSS--AFNPQGEPSEQPYEHLTTESFSDIALNNEKAMVE
EGR1_RAT        ----------------------------MDNYPKLEEMMLLSNGAPQFLGAAGTPEGSGGNNSSSSSSSSSGGGGGGGSNSGSS--AFNPQGEPSEQPYEHLTTESFSDIALNNEKALVE
EGR1_BRARE      -MAAAKTEMLLPALQISDPLS-FPHSP-TDNYPKLEEMIMLNSAGTPFLNAT-APEGAVFG-----S--------------------GE-----PGEQ-FDHLAGDTLSEISMEKP----L
EGR2_XENLA      -------MAAKAVDKLPVTFGSFVHQIPEGFYPGEDSTLPASVTIFPNVDLGGP-----------------------------------------LIQMSGVTGDGMISVDMNNDKRSLD
EGR3_HUMAN      -------MTGKLAEKLPVTMSSLLNQLPDNLYP---EEIPSALNLFSGSSDS-----------------------------------------VVHYNQMATENVMDIGLTNEKPNPE

EGR1_HUMAN      TSYPSQTTRLP---PITYTGRFSLEPAPNSGNTLWPEPLFSLVSGLVSMTNPPASS-----------SSAPSPAASSASAS-QSPPLSCAVPSNDSSPIYSAAPTFPTPNTDIFPEPQSQ
EGR1_MOUSE      TSYPSQTTRLP---PITYTGRFSLEPAPNSGNTLWPEPLFSLVSGLVSMTNPPTSS-----------SSAPSPAASSSSSASQSPPLSCAVPSNDSSPIYSAAPTFPTPNTDIFPEPQSQ
EGR1_RAT        TSYPSQTTRLP---PITYTGRFSLEPAPNSGNTLWPEPLFSLVSGLVSMTNPPTSS-----------SSAPSPAASSSSSASQSPPLSCAVPSNDSSPIYSAAPTFPTPNTDIFPEPQSQ
EGR1_BRARE      SDQTYSTQRLP---PISYTGRFTLEPATNCSNSLWAEPLFSLVSGLVGINPPPASIPSSTSQATHPSSSSTSSIPSSSSSSTSSASLSCSVHQSEPNPIYSAAPTYSSASPDIFPE-SGP
EGR2_XENLA      FSYSSNYPLAPRTQPIAYMGKISIDHQ-YSGSGWNTEGIFNLVSAASLLGVPPSS---------------CSSTSSSNASSGSPNLSCSMSHPQSDLEHIYSP-PPYSSCNEIYQDPLRFP
EGR3_HUMAN      LSYSGSFQPAPGNKTVTYLGKFAFD---SP-SNWCQDNIISLMSAG-ILGVPPASG-------------ALSTQTSTASMVQP--------PQGDVEAMYPALPPYSNCGDLYSEPVSFH

EGR1_HUMAN      AFPGSAGTALQYPPPAYPAAKGG---FQVPMIPDYLFPQQQGDLGLGTPDQKPFQGLESR-TQQPSLTPLSTIKAFATQSGSQDLKALNTSYQSQ--LIKPSR-MRKYPNRPSKTPPHER
EGR1_MOUSE      AFPGSAGTALQYPPPAYPATKGG---FQVPMIPDYLFPQQQGDLSLGTPDQKPFQGLENR-TQQPSLTPLSTIKAFATQSGSQDLKALNTTYQSQ--LIKPSR-MRKYPNRPSKTPPHER
EGR1_RAT        AFPGSAGTALQYPPPAYPATKGG---FQVPMIPDYLFPQQQGDLSLGTPDQKPFQGLENR-TQQPSLTPLSTIKAFATQSGSQDLKALNNTYQSQ--LIKPSR-MRKYPNRPSKTPPHER
EGR1_BRARE      NFSTTVGTSLQYSSSTYPSAKTCNPSFSVPMIPDYLFTQQQSEISLVPPDQKPIQTQAG---QQPALTPLHTIKAFATQTGSQ---DLKSVYQSQ--LIKPSR-MRKYPNRPSKTPPHER
EGR2_XENLA      CGSPTA--ASLPPPPSYPSPKGASDGGMFPMIPDYSALFPPQCQRDLHSDRKPFP--------CPRHPSPLSTIRNFTLG-GSSEGPRLASAYSPQNLPLRPILRPRKYPNRPSKTPVHER
EGR3_HUMAN      DPQGNP--GLAYSPQDYQSAKPALDSNLFPMIPDYNLYHHPNDMGSIP-EHKPFQGMDPIRVNPPPITPLETIKAFKD---KQIHPGFGSLPQPP-LTLKPIR-PRKYPNRPSKTPLHER

EGR1_HUMAN      PYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFACDICGRKFARSDERKRHTKIHLRQKDKKADK--SVVASSATSSLSSYPSPVATSYPSPV
EGR1_MOUSE      PYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFACDICGRKFARSDERKRHTKIHLRQKDKKADK--SVVASPAASSLSSYPSPVATSYPS--
EGR1_RAT        PYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFACDICGRKFARSDERKRHTKIHLRQKDKKADK--SVVASSAASSLSSYPSPVATSYPS--
EGR1_BRARE      PYACPVETCDRRFSRSDELTRHIRIHTGQKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFACEICGRKFARSDERKRHTKIHMRQKDKKAEKGATAAVQSSVSNISISASSPVSSYPS--
EGR2_XENLA      PYPCPAEGCDRRFSRSDELTRHIRIHTGHKPFQCRICMRNFSRSDHLTTHIRTHTGEKPFACDYCGRKFARSDERKRHTKIHLRQKERKNS----------ATAAWRQHVARTSLKPQ--
EGR3_HUMAN      PHACPAEGCDRRFSRSDELTRHLRIHTGHKPFQCRICMRSFSRSDHLTTHIRTHTGEKPFACEFCGRKFARSDERKRHAKIHLKQKEKKAEK---------GGAPSASSAPPVSLAP---

EGR1_HUMAN      TTSYPSPATTSYPSPVPTSFSSPGSSTYPSPVHSGFPSPSVATTYSSVPPAFPAQVSSFPSSAVTN---SFSASTGLSDMTATFSPRTIEIC
EGR1_MOUSE      ------PATTSFPSPVPTSYSSPGSSTYPSPAHSGFPSPSVATTFASVPPAFPTQVSSFPSAGVSS---SFSTSTGLSDMTATFSPRTIEIC
EGR1_RAT        ------PATTSFPSPVPTSYSSPGSSTYPSPAHSGFPSPSVATTYASVPPAFPAQVSTFQSAGVSN---SFSTSTGLSDMTATFSPRTIEIC
EGR1_BRARE      -------PITSYPSPVS-SFPSPVNSCYSSPVHTSYPSPSIATTYPSATSTFQTQVATSFPTSVASNIYSSPVTTPLPDMQSALSPRTADIC
EGR2_XENLA      --------SGRDRQPCALLGPAAAHWDSIDPNRTG-----------------------------------------------------------
EGR3_HUMAN      -----------VVTTCA-------------------------------------------------------------------------------
```

Multiple alignment produced by the `clustalw` program (Thompson et al. 1994)

5.1 PROGRESSIVE PAIRWISE ALIGNMENT

On the left, output from the `clustalw` program to produce above alignment.

```
1 EGR1_HUMAN
2 EGR1_MOUSE
3 EGR1_RAT
4 EGR1_BRARE
5 EGR2_XENLA
6 EGR3_HUMAN
```

Given a set of m sequences $A_1, \cdots, A_m$ to be aligned

```
Sequences (1:2) Aligned.  Score:  92
Sequences (1:3) Aligned.  Score:  91
Sequences (1:4) Aligned.  Score:  59
Sequences (1:5) Aligned.  Score:  39
Sequences (1:6) Aligned.  Score:  39
Sequences (2:3) Aligned.  Score:  97
Sequences (2:4) Aligned.  Score:  56
Sequences (2:5) Aligned.  Score:  39
Sequences (2:6) Aligned.  Score:  38
Sequences (3:4) Aligned.  Score:  54
Sequences (3:5) Aligned.  Score:  36
Sequences (3:6) Aligned.  Score:  38
Sequences (4:5) Aligned.  Score:  35
Sequences (4:6) Aligned.  Score:  37
Sequences (5:6) Aligned.  Score:  43
```

step 1. All sequences are aligned against all others. An $m \times m$ symilarity matrix, $S = (s_{ij})$ is constructed, where s_{ij} is the similarity between the sequences A_i and A_j.

```
(
EGR1_HUMAN:0.02653,
(
EGR1_MOUSE:0.00640,
EGR1_RAT:0.01919)
:0.03762,
(
EGR1_BRARE:0.22632,
(
EGR2_XENLA:0.28570,
EGR3_HUMAN:0.28278)
:0.12406)
:0.16395);
```

step 2. Hierarchical clustering is used to construct a guide tree from the similarity matrix S.

hierarchical clustering with single-link—nearest neighbor
of sequences $A_1, \cdots, A_m$ according to matrix S.

iteration 1

(1:2)	92	
(1:3)	91	
(1:4)	59	
(1:5)	39	
(1:6)	39	
(2:3)	97	⇐=
(2:4)	56	
(2:5)	39	
(2:6)	38	
(3:4)	54	
(3:5)	36	
(3:6)	38	
(4:5)	35	
(4:6)	37	
(5:6)	43	

iteration 2

(1:(2:3))	92	⇐=
(4:(2:3))	56	
(5:(2:3))	39	
(6:(2:3))	36	
(1:4)	59	
(1:5)	39	
(1:6)	39	
(4:5)	35	
(4:6)	37	
(5:6)	43	

iteration 3

(4:(1:(2:3)))	59	⇐=
(5:(1:(2:3)))	39	
(6:(1:(2:3)))	39	
(4:5)	35	
(4:6)	37	
(5:6)	43	

iteration 4

(5:(4:(1:(2:3))))	39	
(6:(4:(1:(2:3))))	39	
(5:6)	43	⇐=

iteration 5 (final)

((5:6):(4:(1:(2:3))))	39

step 3. The sequences are aligned in an stepwise maner according to the guide tree. This requires aligning sequences with sequences, sequences with alignments, and alignments with alignments.

If C_i and C_j are two alignments

$$C_i = \begin{Bmatrix} a_1(1) & a_1(2) & \cdots & a_1(l_i) \\ a_2(1) & a_2(2) & \cdots & a_2(l_i) \\ . & . & . & . \\ a_{|i|}(1) & a_{|i|}(2) & \cdots & a_{|i|}(l_i) \end{Bmatrix} \quad C_j = \begin{Bmatrix} b_1(1) & b_1(2) & \cdots & b_1(l_j) \\ b_2(1) & b_2(2) & \cdots & b_2(l_j) \\ . & . & . & . \\ b_{|i|}(1) & b_{|i|}(2) & \cdots & b_{|i|}(l_j) \end{Bmatrix}$$

then the similarity between element h_1 of C_i and element h_2 of C_j is

$$S(c_i(h_1), c_j(h_2)) = \frac{\sum_{r=1}^{|i|} \sum_{s=1}^{|j|} s(w(a_r(h_1), b_s(h_2)))}{|i||j|}$$