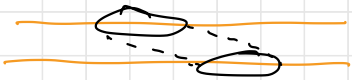
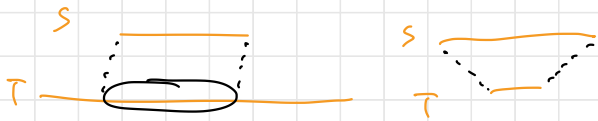
Review:

* Local alignment:

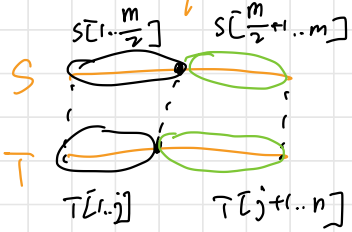


* fit alignment:



* Linear space: Divide & conquer

Find j:



Algorithm: Caculate score($S[1..\frac{m}{2}]$, $T[1..j]$) for all j.

$D[\frac{m}{2}, j]$

Caculate score($S[\frac{m}{2}+1..m]$, $T[j+1..n]$) for all j

Reverse

Score and Significance

Illustration

- Try BLAST the following protein: Note the scores.
- >pdb|6WPT|D Chain D, S309 neutralizing antibody heavy chain
  QVQLVQSGAEVKKPGASVKVSCKASGYPFTSYGISWVRQAPGQGLEWMGWISTYNGNTNYAQKFQGRVTM
  TTDTSTTTGYMELRRLRSDDTAVYYCARDYTRGAWFGESLIGGFDNWGQGTLVTVSS
- See the description of this sequence at
  https://www.ncbi.nlm.nih.gov/protein/6WPT_D

Optimization Problem

- **Instance**: describes the input
- **Feasible Solution**: describes the format of the output
- **Score Function**: measures how good a solution is
- **Objective**: either maximize or minimize the score.

E.g. Sequence Alignment

- **Instance**: two sequences S and T
- **Feasible Solution**: insert gaps into S and T so that they have the same length
- **Score Function**: add up the score of each column
- **Objective**: to maximize the score

Purposes of the Score

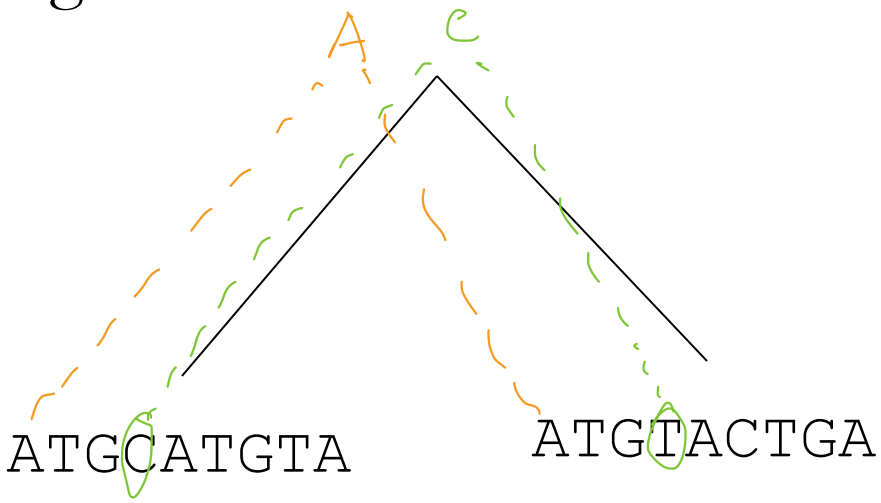
- Purpose 1: It helps us to compare solutions of the same instance.
  - Which alignment is the best for the same input (s,t).
- Purpose 2: It helps us to compare solutions of different instances.
  - Which of (s,t) and (x,y) is more likely to be a homology?
- Purpose 3: It helps us to tell how significant the solution is.
  - Does the alignment between s and t indicate that they are homologous?

Which solution is better? – same instance.

PURPOSE 1

Purpose 1

- We first examine how the scoring function is designed for the first purpose – compare two alignments and tell which one is better.
- Recall that what we really want is to find out homologies.

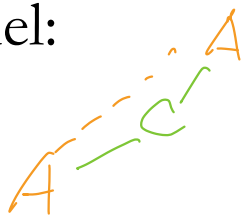



```
ATGCA-TGTA
|||  |  ||  |
ATGTACTG-A
```

match : p
mismatch : q
indel : r

An Simple Evolutionary Model

- We first need an (simplified/oversimplified) evolution model:
  - Only substitution and indel
  - Two mutations do not overlap
    - Guarantee that all evolutionary information but the order is represented by the alignment.
  - Along the path of evolution, p: unchanged, q: substitution, r: indel. (p+q+r=1)



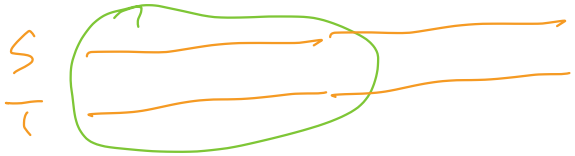

- ```
  ATGCA-TGTA   (S)
  |||  |  ||  |
  ATGTACTG-A   (T)
  ```
  ```

Probability of the Alignment

- Under this model, the probability of the alignment is:
  - $p^7 * q * r^2$
- We want to maximize this probability
  - $\log (p^7 * q * r^2) = 7 \log p + \log q + 2 \log r$
- Let match = log p, mismatch = log q, indel = log r. We get a scoring scheme.
- Maximizing the score is equivalent to maximizing the probability of evolutionary history.

Simple Score

- This is sufficient to compare the alignments of the same two sequences.
- Problems
  - Always negative
  - Local alignment becomes meaningless
  - Repeating the alignment twice make the score lower.
- Useless in comparison of two alignments of different pairs of sequences.
- Question: Can these be solved by adding a minus sign?
- Next let us make this score better.



match = 1

mismatch = -1

indel = -1

Which solution is better? – different instances.

PURPOSE 2

Likelihood Ratio

- Model 1 (homology): the alignment A between S and T reflects evolutionary history.
- Model 2 (random): the alignment A between S and T is merely a random event.
- We want to examine the likelihood ratio
  - Pr(alignment|homology) / Pr(alignment|random)
- If it's much bigger than 1 (such as 100000), it's evidence towards model one being the truth.
- If it's much below 1 (such as 0.00001), it's evidence towards model two being the truth.

Log likelihood Ratio

- Assume for the homology and random models, we have established the probabilities for each column type:
  - Match: p and p'
  - Substitution: q and q'
  - Indel: r and r'
- For the following alignment,
  - ```
    ATGCA-TGTA  (S)
    |||  |  ||  |
    ATGTACTG-A  (T)
    ```
  - Pr(alignment|homology) / Pr(alignment | random) = $(p/p')^7 * (q/q') * (r/r')^2$
- We usually take a logarithm. The sore becomes
  - 7* log (p/p') + log (q/q') + 2 * log (r/r').
- If this is very positive, then homology model explains the alignment better than random model. And vice versa.
    ```

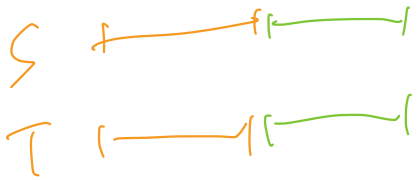
This is a Better Scoring Scheme

- Score scheme prefers column types happens more often in the homology model than in the random model. Usually,
  - p>p', therefore a matching column has a positive score
  - q<q', therefore a mismatching column has a negative score
  - r <r', therefore an indel column has a negative score.

$\log \frac{p}{p'} > 0$

$\log \frac{q}{q'} < 0$

- Avoids the problems we had when only probabilities (not the ratio) were used.
  - Putting two positive alignments together increase the homology chance.
  - Can compare two alignments with different lengths.
- This is indeed the scoring scheme we have seen and have used in practice (in BLAST etc.)



Statistics

- The probability values used in the homology and random models may be obtained by simple counting their frequencies in some "real" alignments and "random" alignments, respectively.
- Often, the statistics is only approximate and does not need to be precise. In particular, the "random" model often uses some (over)-simplied values.
  - For example: Pr(indel)= 0.2, Pr(match) = 1/20 * 0.8, Pr(mismatch) = 19/20 * 0.8.

Substitution Matrix

- The non-indel columns can be further refined to have different scores for different pairs of letters.
- For each pair of letters a and b, assume the probability of seeing (a,b) in a column is p(a,b) for the homology model, and is q(a,b) for the random model. $= Pr(a,b \mid homology)$ $= Pr(a,b \mid random)$.
- Then substitution score is then log (p(a,b)/q(a,b)).
- This is called a substitution matrix.
- Let us assume that q(a,b)=p(a)*p(b). Here p(a) is the frequency of letter a in the sequences. Note that this is an (over)-simplification. But it provides "good enough" values in practice.

Substitution Matrix

- The substitution matrix is particularly important when aligning protein sequences because
  - there are 20 amino acids
  - some of them share significant similarities
  - protein alignments have fewer matching columns.

Alignment of Protein Sequences

```
Conserved domain database 22426:
KOG4652, HORMA domain [Chromatin structure and dynamics]

Conserved domain length = 324 residues, 100% aligned

CT46          15  VFPNKISTEHQSLVLVKRLLAVSVSCITYLRGIFPECAYGTRYLD--DLCVKILREDKNCPG--STQLVKWMLGC
                    PN +  E QSL  + RLL V++S I   RGIFPE  +  RY+D    L + +LR      G   + L K +
KOG4652       1   TLPNGLENEKQSLEFMTRLLYVAISTILRERGIFPEEYFKDRYVDGNLLVMTLLRRQDAPEGRLVSWLEKGV---

CT46          85  YDALQKKYLRMVVLAVYTNPEDPQTISECYQFKFKYTNNGPLMDFISKN------QSNESSMLSTD-TKKASILL
                  +DA+++K L+ + L V T  EDP+ I E Y F F Y   G +   I+        ++ E S LS D T++    L
KOG4652       73  HDAIRQKLLKKLSL-VITESEDPEDI-EVYIFSFVYDEEGSVSARINYGINGQSSKAFELSQLSMDDTRRQFAKL

CT46         154 IRKIYILMQNLGPLPNDVCLTMKLFYYDEVTPPDYQPPGFKDGDCEGVIFEGEPMYLNVGEVSTPFHIFKVKVTT
                  IRK++I  Q L PLP    +     YY E  PPDYQP GFKD           P  +N+G VSTP H   VKV
KOG4652      146 IRKLHICTQLLEPLPQ-GLILSMRLYYTERVPPDYQPEGFKDSTRAFYTLPVNPEQINIGAVSTPHHKGFVKVL-

CT46         229 ERERMENIDSTILSPKQIKTPFQKILRDKDVEDEQEHYTSDDLDIETKMEEQEKNPASSELEEPSLVCEEDEIMR
                                                  SD  D   K E
KOG4652      219 ----------------------------------------SDATDSMEKAER----------------------T

CT46         304 SKESPDLSISHSQVEQLVNKTSELDMSESKTRSGKVFQNKMANGNQPVKSSKENRKRSQHESGR---IVLHHFDS
                  K S D       V+Q +NK+ E D S S+ ++  +  N + N   PV  S+E+  +SQ   G          D
KOG4652      232 DKISDDP-FDLILVQQELNKSEEADKSFSQEKTTSITPNVLGNPLVPVDQSEEDLLKSQDSPGTGRCSCECGLDV

CT46         376 SSQESVPKRRKFSEPKEHI
                 S Q SVPK RK     EH
KOG4652      306 SKQASVPKTRKSCRKTEHG

Homology between CT46 and MGC26710 hypothetical protein

Identities = 136/249 (54%), with conservative changes = 180/249 (72%)

CT46          1   MATAQLQR-----TPMSALVFPNKISTEHQSLVLVKRLLAVSVSCITYLRGIFPECAYGTRYLDDLCVKILREDK
                  MATAQL             VFP++I+ EH+SL +VK+L A S+SCITYLRG+FPE +YG R+LDDL +KILREDK
MGC26710      1   MATAQLSHCITIHKASKETVFPSQITNEHESLKMVKKLFATSISCITYLRGLFPESSYGERHLDDLSLKILREDK

CT46         71   NCPGSTQLVKWMLGCYDALQKKYLRMVVLAVYTNPEDPQTISECYQFKFKYTNNGPLMDF--ISKNQSNESSMLS
                   CPGS  +++W+ GC+DAL+K+YLRM VL +YT+P   + ++E YQFKFKYT  G  MDF   S + S ES   +
MGC26710     76   KCPGSLHIIRWIQGCFDALEKRYLRMAVLTLYTDPMGSEKVTEMYQFKFKYTKEGATMDFDSHSSSTSFESGTNN

CT46         144 TDTKKASILLIRKIYILMQNLGPLPNDVCLTMKLFYYDEVTPPDYQPPGFKDG-DCEGVIFEGEPMYLNVGEVST
                  D KKAS+LLIRK+YILMQ+L PLPN+V LTMKL YY+ VTP DYQP GFK+G +   ++F+ EP+ + VG VST
MGC26710     151 EDIKKASVLLIRKLYILMQDLEPLPNNVVLTMKLHYYNAVTPHDYQPLGFKEGVNSHFLLFDKEPINVQVGFVST

CT46         218 PFHIFKVKVTTERERMENIDSTIL  241
                  FH  KVKV TE  ++ ++++ +
MGC26710     226 GFHSMKVKVMTEATKVIDLENNLF  249
```

ungapped alignment.

Notice the blocks with no indel.

BLOSUM 62

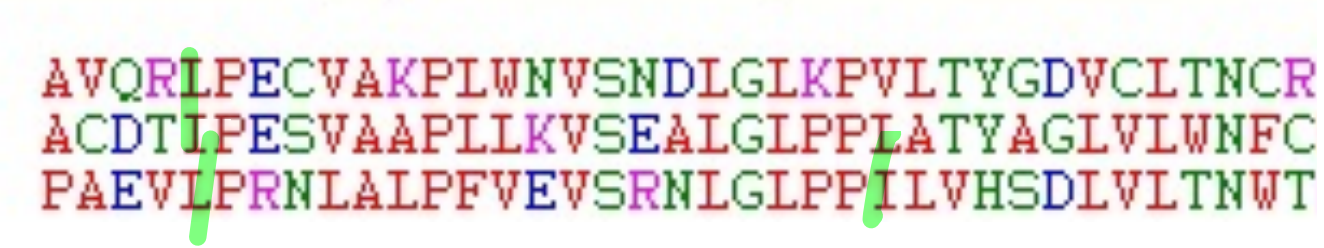
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3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

Non-standard letters

B= D or N
Z= E or Q
X= any
* = translation stop

The most used amino acid substitution matrix. Let's study how this is constructed.

Frequency of AA pairs



- 37 columns, each column 3 pairs. In total 111 pairs.
- For example, the pair I-L occurs 3 times; the pair L-L occurs 13 times

$$P_{IL} = \frac{3}{111}, P_{LL} = \frac{13}{111}$$

$\text{score}(L,L) = 2 \cdot \log_2 \frac{P_{LL}}{P_L \cdot P_L}$

- Total amino acid 111 (a coincident).

$$P_I = \frac{2}{111}, P_L = \frac{21}{111}$$

$\text{score}(I,L) = 2 \log_2 \frac{P_{IL}}{P_I \cdot P_L + P_L \cdot P_I}$

- We can then use log likelihood ratio to calculate scores.
- But we should correctly distinguish the counting when two letters are the same or different.

Blosum

$$score(x,y) = 2\log_2 \frac{P_{xy}}{2P_xP_y}, \text{ if } x \neq y$$

$$score(x,x) = 2\log_2 \frac{P_{xx}}{P_xP_x}$$

In BLOSUM matrices these values are rounded to the nearest integer.