Sequence alignments (1)

Why create them?

How are they stored?

How can they be manipulated?

This presentation

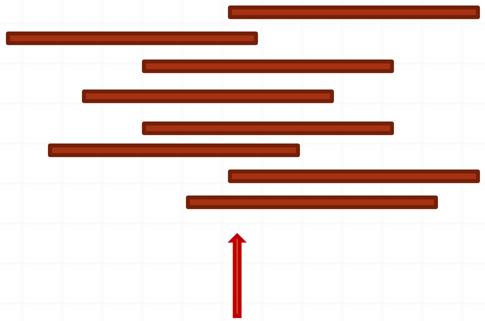
- - is intended to :
  - give a brief overview of types of sequence alignment file formats
  - provide brief background useful for the SAMTOOLs tutorial run in the practical sessions
    - Course notes for tutorial : http://biobits.org/samtools_primer.html
- There will be more detailed presentations / sessions on both the above topics

Sequence alignment data

- What is the purpose of creating and storing a sequence alignment?
- There are multiple purposes
- Different aims
- These aims are associated with different ways of representing the alignment
- And thus different **alignment file formats**

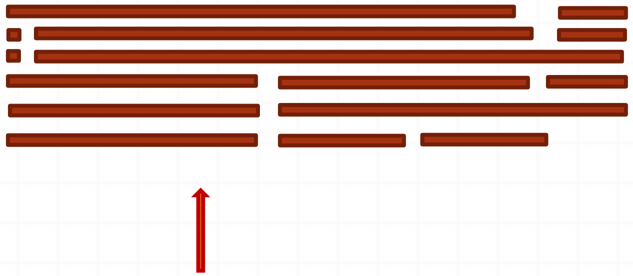
- Related but different aims, meanings and file formats

- Sequence read alignment ("assembly")



- Each nucleotide position (column) represents multiple copies of the same base of an original sequence (e.g. genome sequence)

- Multiple protein or nucleotide sequence alignment



- Each position (column) represents a **homologous** nucleotide (or amino acid).
- Sequences are evolutionarily related (homologous) sequences, typically from different organisms, and/or multiple members of a gene family
- Gaps represent insertions/ deletions

"Traditional" alignment of 2 or more sequences

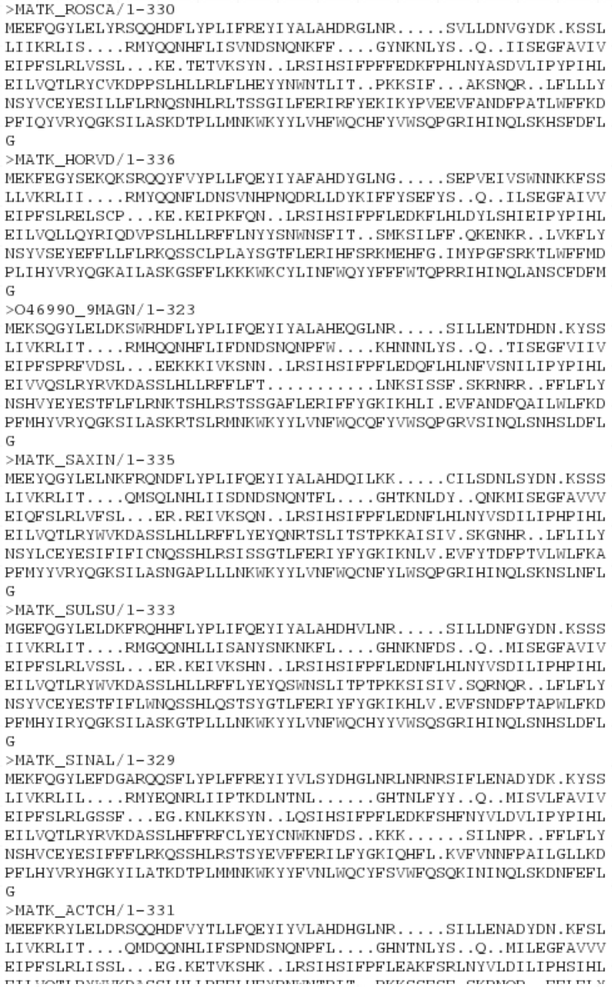
- **Pairwise alignment and multiple alignment**
- From an algorithmic point of view:
- There are some significant differences between methods used to:
  - compare and align only 2 sequences (**pairwise alignment**)
  - compare and align 3 or more sequences (**multiple alignment**)
- Some pairwise alignment tools produce their own particular alignment formats
  - E.g. some programs in the versatile EMBOSS software package

- Example pairwise alignment
  - from EMBOSS `needle`
    - a **global alignment**, i.e. complete lengths of both sequences are aligned

```
#=======================================
#
# Aligned_sequences: 2
# 1: P0A1L1
# 2: Q0PBF7_CAMJE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 265
# Identity:      27/265 (10.2%)
# Similarity:    49/265 (18.5%)
# Gaps:         140/265 (52.8%)
# Score: 32.0
#
#
#=======================================

P0A1L1             0 --------------------------------------------------      0

Q0PBF7_CAMJE       1 MRLLVLFFLILPLYSVELISYNIYDRNDRVDLMLSFDNAYNGKISQKKEK     50

P0A1L1             0 --------------------------------------------------      0

Q0PBF7_CAMJE      51 NLTLLTFSDLTYSKDELKELNSQLVDKISISSKNNNTYIMLQNKQNINLE    100

P0A1L1             1 --------------MMKTEATV-SQPTAPAGS--PLM--QVSGALIG---     28
                                   :.:.:|.: |.||..|.:  .||  ..|.:|.|
Q0PBF7_CAMJE     101 LSSINDKFGVRIRAIEQGKANIESAPTTTANNSQELMPKPKSTSLEGYDY    150

P0A1L1            29 --------IIALILAAAWVIKRMGFAPKGNSVRGLKVSASASLGPRERVV     70
                             |:.::|...|..|:.......|..|...:.....|....::|
Q0PBF7_CAMJE     151 TNYILVMLILVILLIVLWWFKKTMVYKNNNVSRDFTMIFQRFLDKNNQLV    200

P0A1L1            71 IVEVENARLVLGVTASQINLLHTLPPAENDTEAPVAPPADFQNMMKSLLK    120
                     :.:..|.|..:.:..|.:.|.....|.|...:.......:|.:..:...|
Q0PBF7_CAMJE     201 VFDHANKRYTMIIGNSNVLLESIEIPEEQTIKHTEKTEKNFDSFFEENKK    250

P0A1L1           121 RSGRS----------    125
                     |....
Q0PBF7_CAMJE     251 RIQNLIEQRQKGKKS    265
```

“Traditional” alignment of 2 or more sequences

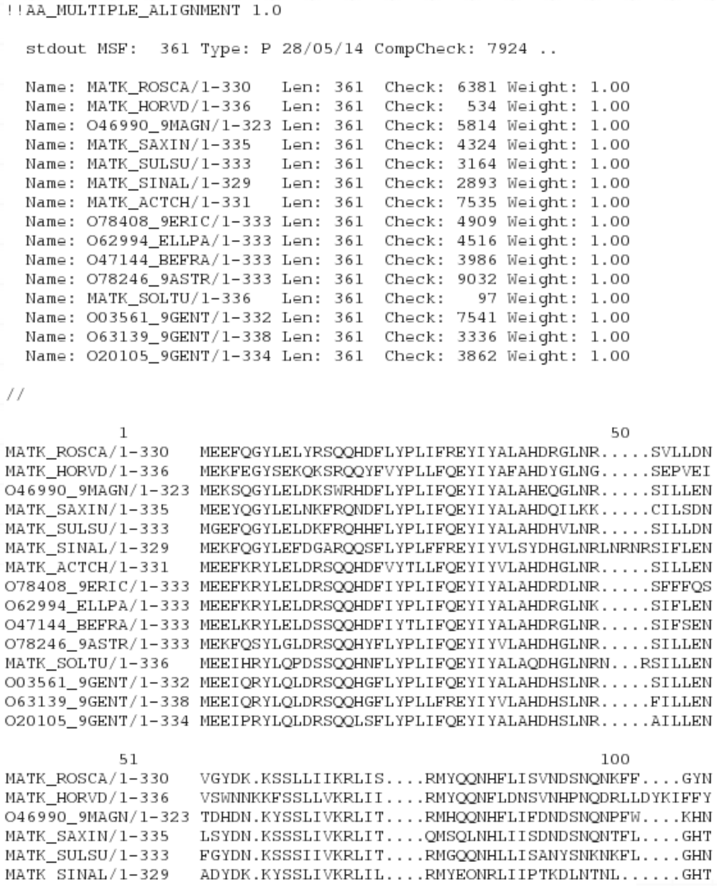
- Typically, DNA sequences representing (complete) genes
- Or protein sequences
- A principle purpose is to identify the parts of the sequence which are the same and which are different
- I.e. identify **conservation** and **divergence**
- This can highlight regions of potential functional importance
- Can be especially informative if you are dealing with several/many sequences (i.e. a **multiple alignment**)



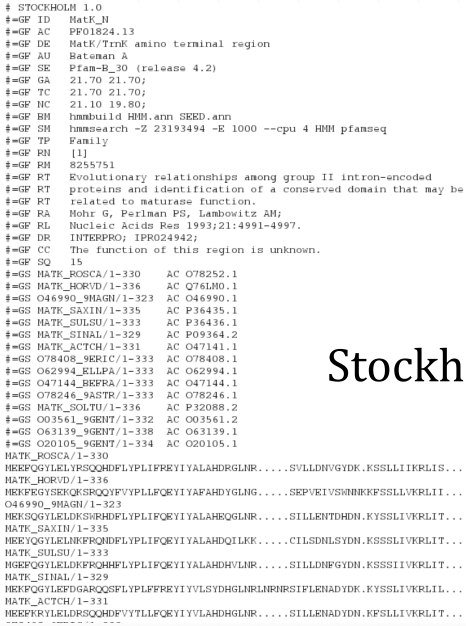
```
>MATK_ROSCA/1-330
MEEFQGYLELYRSQQHDFLYPLIFREYIYALAHDRGLNR.....SVLLDNVGYDK.KSSL
LIIKRLIS....RMYQQNHFLISVNDSNQNKFF....GYNKNLYS..Q..IISEGFAVIV
EIPFSLRLVSSL...KE.TETVKSYN..LRSIHSIFPFFEDKFPHLNYASDVLIPYPIHL
EILVQTLRYCVKDPPSLHLLRLFLHEYYNWNTLIT..PKKSIF...AKSNQR..LFLLLY
NSYVCEYESILLFLRNQSNHLRLTSSGILFERIRFYEKIKYPVEEVFANDFPATLWFFKD
PFIQYVRYQGKSILASKDTPLLMNKWKYYLVHFWQCHFYVWSQPGRIHINQLSKHSFDFL
G
>MATK_HORVD/1-336
MEKFEGYSEKQKSRQQYFVYPLLFQEYIYAFAHDYGLNG.....SEPVEIVSWNNKKFSS
LLVKRLII....RMYQQNFLDNSVNHPNQDRLLDYKIFFYSEFYS..Q..ILSEGFAIVV
EIPFSLRELSCP...KE.KEIPKFQN..LRSIHSIFPFLEDKFLHLDYLSHIEIPYPIHL
EILVQLLQYRIQDVPSLHLLRFFLNYYSNWNSFIT..SMKSILFF.QKENKR..LVKFLY
NSYVSEYEFFLLFLRKQSSCLPLAYSGTFLERIHFSRKMEHFG.IMYPGFSRKTLWFFMD
PLIHYVRYQGKAILASKGSFFLKKKWKCYLINFWQYYFFFWTQPRRIHINQLANSCFDFM
G
>O46990_9MAGN/1-323
MEKSQGYLELDKSWRHDFLYPLIFQEYIYALAHEQGLNR.....SILLENTDHDN.KYSS
LIVKRLIT....RMHQQNHFLIFDNDSNQNPFW....KHNNNLYS..Q..TISEGFVIIV
EIPFSPRFVDSL...EEKKKIVKSNN..LRSIHSIFPFLEDQFLHLNFVSNILIPYPIHL
EIVVQSLRYRVKDASSLHLLRFFLFT..........LNKSISSF.SKRNRR..FFLFLY
NSHVYEYESTFLFLRNKTSHLRSTSSGAFLERIFFYGKIKHLI.EVFANDFQAILWLFKD
PFMHYVRYQGKSILASKRTSLRMNKWKYYLVNFWQCQFYVWSQPGRVSINQLSNHSLDFL
G
>MATK_SAXIN/1-335
MEEYQGYLELNKFRQNDFLYPLIFQEYIYALAHDQILKK.....CILSDNLSYDN.KSSS
LIVKRLIT....QMSQLNHLIISDNDSNQNTFL....GHTKNLDY..QNKMISEGFAVVV
EIQFSLRLVFSL...ER.REIVKSQN..LRSIHSIFPFLEDNFLHLNYVSDILIPHPIHL
EILVQTLRYWVKDASSLHLLRFFLYEYQNRTSLITSTPKKAISIV.SKGNHR..LFLILY
NSYLCEYESIFIFICNQSSHLRSISSGTLFERIYFYGKIKNLV.EVFYTDFPTVLWLFKA
PFMYYVRYQGKSILASNGAPLLLNKWKYYLVNFWQCNFYLWSQPGRIHINQLSKNSLNFL
G
>MATK_SULSU/1-333
MGEFQGYLELDKFRQHHFLYPLIFQEYIYALAHDHVLNR.....SILLDNFGYDN.KSSS
IIVKRLIT....RMGQQNHLLISANYSNKNKFL....GHNKNFDS..Q..MISEGFAVIV
EIPFSLRLVSSL...ER.KEIVKSHN..LRSIHSIFPFLEDNFLHLNYVSDILIPHPIHL
EILVQTLRYWVKDASSLHLLRFFLYEYQSWNSLITPTPKKSISIV.SQRNQR..LFLFLY
NSYVCEYESTFIFLWNQSSHLQSTSYGTLFERIYFYGKIKHLV.EVFSNDFPTAPWLFKD
PFMHYIRYQGKSILASKGTPLLLNKWKYYLVNFWQCHYYVWSQSGRIHINQLSNHSLDFL
G
>MATK_SINAL/1-329
MEKFQGYLEFDGARQQSFLYPLFFREYIYVLSYDHGLNRLNRNRSIFLENADYDK.KYSS
LIVKRLIL....RMYEQNRLIIPTKDLNTNL......GHTNLFYY..Q..MISVLFAVIV
EIPFSLRLGSSF...EG.KNLKKSYN..LQSIHSIFPFLEDKFSHFNYVLDVLIPYPIHL
EILVQTLRYRVKDASSLHFFRFCLYEYCNWKNFDS..KKK......SILNPR..FFLFLY
NSHVCEYESIFFFLRKQSSHLRSTSYEVFFERILFYGKIQHFL.KVFVNNFPAILGLLKD
PFLHYVRYHGKYILATKDTPLMMNKWKYYFVNLWQCYFSVWFQSQKININQLSKDNFEFL
G
>MATK_ACTCH/1-331
MEEFKRYLELDRSQQHDFVYTLLFQEYIYVLAHDHGLNR.....SILLENADYDN.KFSL
LIVKRLIT....QMDQQNHLIFSPNDSNQNPFL....GHNTNLYS..Q..MILEGFAVVV
EIPFSLRLISSL...EG.KETVKSHK..LRSIHSIFPFLEAKFSRLNYVLDILIPHSIHL
```

FASTA

MSF



```
!!AA_MULTIPLE_ALIGNMENT 1.0

 stdout MSF:  361 Type: P 28/05/14 CompCheck: 7924 ..

 Name: MATK_ROSCA/1-330   Len: 361  Check: 6381 Weight: 1.00
 Name: MATK_HORVD/1-336   Len: 361  Check:  534 Weight: 1.00
 Name: O46990_9MAGN/1-323 Len: 361  Check: 5814 Weight: 1.00
 Name: MATK_SAXIN/1-335   Len: 361  Check: 4324 Weight: 1.00
 Name: MATK_SULSU/1-333   Len: 361  Check: 3164 Weight: 1.00
 Name: MATK_SINAL/1-329   Len: 361  Check: 2893 Weight: 1.00
 Name: MATK_ACTCH/1-331   Len: 361  Check: 7535 Weight: 1.00
 Name: O78408_9ERIC/1-333 Len: 361  Check: 4909 Weight: 1.00
 Name: O62994_ELLPA/1-333 Len: 361  Check: 4516 Weight: 1.00
 Name: O47144_BEFRA/1-333 Len: 361  Check: 3986 Weight: 1.00
 Name: O78246_9ASTR/1-333 Len: 361  Check: 9032 Weight: 1.00
 Name: MATK_SOLTU/1-336   Len: 361  Check:   97 Weight: 1.00
 Name: O03561_9GENT/1-332 Len: 361  Check: 7541 Weight: 1.00
 Name: O63139_9GENT/1-338 Len: 361  Check: 3336 Weight: 1.00
 Name: O20105_9GENT/1-334 Len: 361  Check: 3862 Weight: 1.00

//

           1                                                50
MATK_ROSCA/1-330   MEEFQGYLELYRSQQHDFLYPLIFREYIYALAHDRGLNR.....SVLLDN
MATK_HORVD/1-336   MEKFEGYSEKQKSRQQYFVYPLLFQEYIYAFAHDYGLNG.....SEPVEI
O46990_9MAGN/1-323 MEKSQGYLELDKSWRHDFLYPLIFQEYIYALAHEQGLNR.....SILLEN
MATK_SAXIN/1-335   MEEYQGYLELNKFRQNDFLYPLIFQEYIYALAHDQILKK.....CILSDN
MATK_SULSU/1-333   MGEFQGYLELDKFRQHHFLYPLIFQEYIYALAHDHVLNR.....SILLDN
MATK_SINAL/1-329   MEKFQGYLEFDGARQQSFLYPLFFREYIYVLSYDHGLNRLNRNRSIFLEN
MATK_ACTCH/1-331   MEEFKRYLELDRSQQHDFVYTLLFQEYIYVLAHDHGLNR.....SILLEN
O78408_9ERIC/1-333 MEEFKRYLELDRSQQHDFIYPLIFQEYIYALAHDRDLNR.....SFFFQS
O62994_ELLPA/1-333 MEEFKRYLELDRSQQHDFIYPLIFQEYIYALAHDRGLNK.....SIFLEN
O47144_BEFRA/1-333 MEELKRYLELDSSQQHDFIYTLIFQEYIYALAHDRGLNR.....SIFSEN
O78246_9ASTR/1-333 MEKFQSYLGLDRSQQHYFLYPLIFQEYIYVLAHDHGLNR.....SILLEN
MATK_SOLTU/1-336   MEEIHRYLQPDSSQQHNFLYPLIFQEYIYALAQDHGLNRN...RSILLEN
O03561_9GENT/1-332 MEEIQRYLQLDRSQQHGFLYPLIFQEYIYALAHDHSLNR.....SILLEN
O63139_9GENT/1-338 MEEIQRYLQLDRSQQHGFLYPLLFREYIYVLAHDHSLNR.....FILLEN
O20105_9GENT/1-334 MEEIPRYLQLDRSQQLSFLYPLIFQEYIYALAHDHSLNR.....AILLEN

           51                                               100
MATK_ROSCA/1-330   VGYDK.KSSLLIIKRLIS....RMYQQNHFLISVNDSNQNKFF....GYN
MATK_HORVD/1-336   VSWNNKKFSSLLVKRLII....RMYQQNFLDNSVNHPNQDRLLDYKIFFY
O46990_9MAGN/1-323 TDHDN.KYSSLIVKRLIT....RMHQQNHFLIFDNDSNQNPFW....KHN
MATK_SAXIN/1-335   LSYDN.KSSSLIVKRLIT....QMSQLNHLIISDNDSNQNTFL....GHT
MATK_SULSU/1-333   FGYDN.KSSSIIVKRLIT....RMGQQNHLLISANYSNKNKFL....GHN
MATK SINAL/1-329   ADYDK.KYSSLIVKRLIL....RMYEONRLIIPTKDLNTNL......GHT
```



```
# STOCKHOLM 1.0
#=GF ID   MatK_N
#=GF AC   PF01824.13
#=GF DE   MatK/TrnK amino terminal region
#=GF AU   Bateman A
#=GF SE   Pfam-B_30 (release 4.2)
#=GF GA   21.70 21.70;
#=GF TC   21.70 21.70;
#=GF NC   21.10 19.80;
#=GF BM   hmmbuild HMM.ann SEED.ann
#=GF SM   hmmsearch -Z 23193494 -E 1000 --cpu 4 HMM pfamseq
#=GF TP   Family
#=GF RN   [1]
#=GF RM   8255751
#=GF RT   Evolutionary relationships among group II intron-encoded
#=GF RT   proteins and identification of a conserved domain that may be
#=GF RT   related to maturase function.
#=GF RA   Mohr G, Perlman PS, Lambowitz AM;
#=GF RL   Nucleic Acids Res 1993;21:4991-4997.
#=GF DR   INTERPRO; IPR024942;
#=GF CC   The function of this region is unknown.
#=GF SQ   15
#=GS MATK_ROSCA/1-330    AC O78252.1
#=GS MATK_HORVD/1-336    AC Q76LM0.1
#=GS O46990_9MAGN/1-323  AC O46990.1
#=GS MATK_SAXIN/1-335    AC P36435.1
#=GS MATK_SULSU/1-333    AC P36436.1
#=GS MATK_SINAL/1-329    AC P09364.2
#=GS MATK_ACTCH/1-331    AC O47141.1
#=GS O78408_9ERIC/1-333  AC O78408.1
#=GS O62994_ELLPA/1-333  AC O62994.1
#=GS O47144_BEFRA/1-333  AC O47144.1
#=GS O78246_9ASTR/1-333  AC O78246.1
#=GS MATK_SOLTU/1-336    AC P32088.2
#=GS O03561_9GENT/1-332  AC O03561.2
#=GS O63139_9GENT/1-338  AC O63139.1
#=GS O20105_9GENT/1-334  AC O20105.1
MATK_ROSCA/1-330
MEEFQGYLELYRSQQHDFLYPLIFREYIYALAHDRGLNR.....SVLLDNVGYDK.KSSLLIIKRLIS....RMYQQNHFLISVNDSNQNKFF....GYNKNLYS..Q..IISEGFAVIVEIPFSLRLVSSL...KE.TETVKSYN..LRSIHSIFPFFEDKFPHLNYAS
MATK_HORVD/1-336
MEKFEGYSEKQKSRQQYFVYPLLFQEYIYAFAHDYGLNG.....SEPVEIVSWNNKKFSSLLVKRLII....RMYQQNFLDNSVNHPNQDRLLDYKIFFYSEFYS..Q..ILSEGFAIVVEIPFSLRELSCP...KE.KEIPKFQN..LRSIHSIFPFLEDKFLHLDYLS
O46990_9MAGN/1-323
MEKSQGYLELDKSWRHDFLYPLIFQEYIYALAHEQGLNR.....SILLENTDHDN.KYSSLIVKRLIT....RMHQQNHFLIFDNDSNQNPFW....KHNNNLYS..Q..TISEGFVIIVEIPFSPRFVDSL...EEKKKIVKSNN..LRSIHSIFPFLEDQFLHLNFVS
MATK_SAXIN/1-335
MEEYQGYLELNKFRQNDFLYPLIFQEYIYALAHDQILKK.....CILSDNLSYDN.KSSSLIVKRLIT....QMSQLNHLIISDNDSNQNTFL....GHTKNLDY..QNKMISEGFAVVVEIQFSLRLVFSL...ER.REIVKSQN..LRSIHSIFPFLEDNFLHLNYVS
MATK_SULSU/1-333
MGEFQGYLELDKFRQHHFLYPLIFQEYIYALAHDHVLNR.....SILLDNFGYDN.KSSSIIVKRLIT....RMGQQNHLLISANYSNKNKFL....GHNKNFDS..Q..MISEGFAVIVEIPFSLRLVSSL...ER.KEIVKSHN..LRSIHSIFPFLEDNFLHLNYVS
MATK_SINAL/1-329
MEKFQGYLEFDGARQQSFLYPLFFREYIYVLSYDHGLNRLNRNRSIFLENADYDK.KYSSLIVKRLIL....RMYEQNRLIIPTKDLNTNL......GHTNLFYY..Q..MISVLFAVIVEIPFSLRLGSSF...EG.KNLKKSYN..LQSIHSIFPFLEDKFSHFNYVL
MATK_ACTCH/1-331
MEEFKRYLELDRSQQHDFVYTLLFQEYIYVLAHDHGLNR.....SILLENADYDN.KFSLLIVKRLIT....QMDQQNHLIFSPNDSNQNPFL....GHNTNLYS..Q..MILEGFAVVVEIPFSLRLISSL...EG.KETVKSHK..LRSIHSIFPFLEAKFSRLNYVL
```

Stockholm

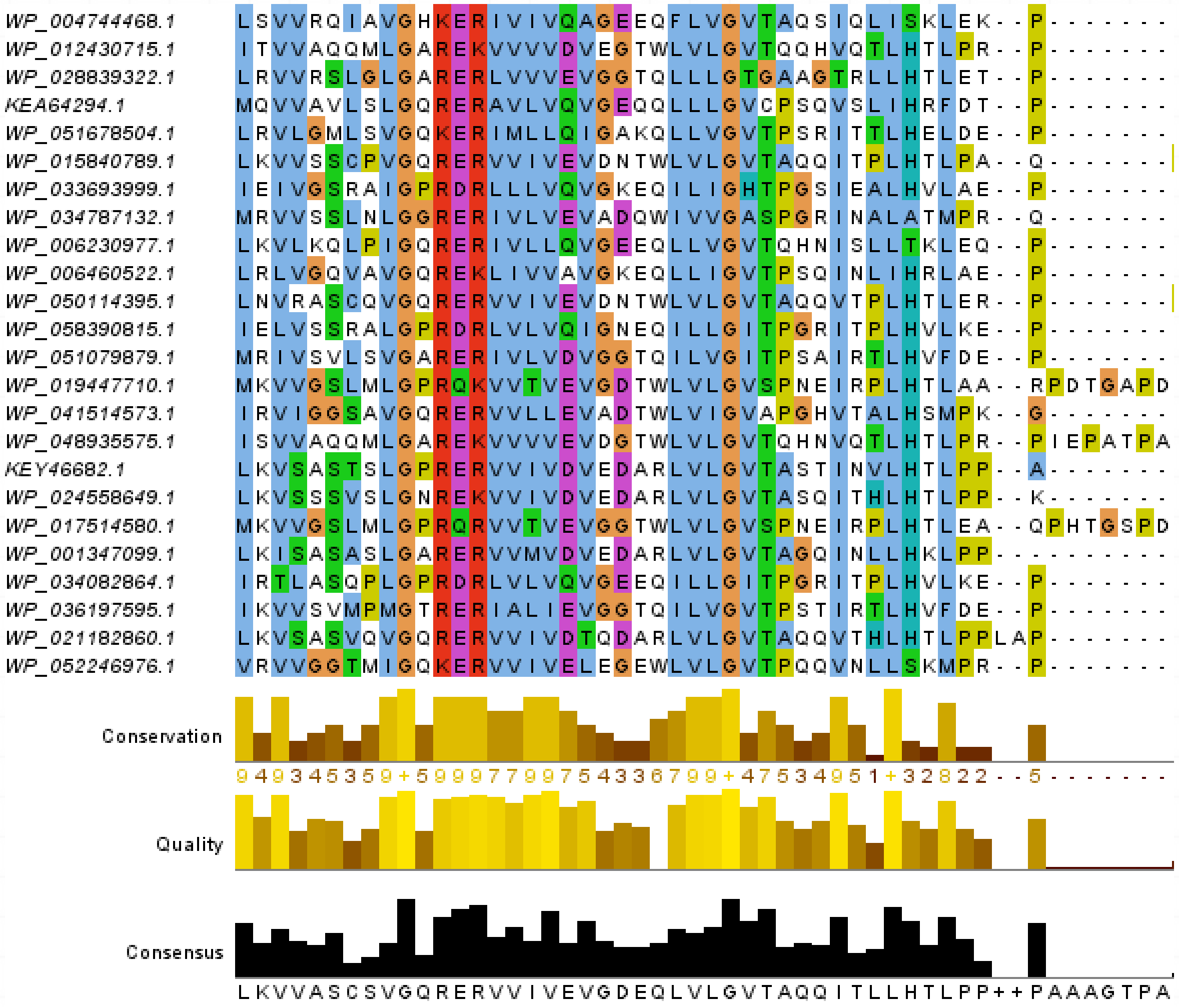
Some multiple sequence alignment formats – these are all **plain text files** (“flatfiles”)

Numerous other formats exist

CLUSTAL format is widely used

Viewing multiple alignments

- Example display from a program (Jalview) which reads the flatfiles and creates a formatted/annotated view



```
WP_004744468.1   LSVVRQIAVGHKERIVIVQAGEEQFLVGVTAQSIQLISKLEK--P-------
WP_012430715.1   ITVVAQQMLGAREKVVVVDVEGTWLVLGVTQQHVQTLHTLPR--P-------
WP_028839322.1   LRVVRSLGLGARERLVVVEVGGTQLLLGTGAAGTRLLHTLET--P-------
KEA64294.1       MQVVAVLSLGQRERAVLVQVGEQQLLLGVCPSQVSLIHRFDT--P-------
WP_051678504.1   LRVLGMLSVGQKERIMLLQIGAKQLLVGVTPSRITTLHELDE--P-------
WP_015840789.1   LKVVSSCPVGQRERVVIVEVDNTWLVLGVTAQQITPLHTLPA--Q-------
WP_033693999.1   IEIVGSRAIGPRDRLLLVQVGKEQILIGHTPGSIEALHVLAE--P-------
WP_034787132.1   MRVVSSLNLGGRERIVLVEVADQWIVVGASPGRINALATMPR--Q-------
WP_006230977.1   LKVLKQLPIGQRERIVLLQVGEEQLLVGVTQHNISLLTKLEQ--P-------
WP_006460522.1   LRLVGQVAVGQREKLIVVAVGKEQLLIGVTPSQINLIHRLAE--P-------
WP_050114395.1   LNVRASCQVGQRERVVIVEVDNTWLVLGVTAQQVTPLHTLER--P-------
WP_058390815.1   IELVSSRALGPRDRLVLVQIGNEQILLGITPGRITPLHVLKE--P-------
WP_051079879.1   MRIVSVLSVGARERIVLVDVGGTQILVGITPSAIRTLHVFDE--P-------
WP_019447710.1   MKVVGSLMLGPRQKVVTVEVGDTWLVLGVSPNEIRPLHTLAA--RPDTGAPD
WP_041514573.1   IRVIGGSAVGQRERVVLLEVADTWLVIGVAPGHVTALHSMPK--G-------
WP_048935575.1   ISVVAQQMLGAREKVVVVEVDGTWLVLGVTQHNVQTLHTLPR--PIEPATPA
KEY46682.1       LKVSASTSLGPRERVVIVDVEDARLVLGVTASTINVLHTLPP--A-------
WP_024558649.1   LKVSSSVSLGNREKVVIVDVEDARLVLGVTASQITHLHTLPP--K-------
WP_017514580.1   MKVVGSLMLGPRQRVVTVEVGGTWLVLGVSPNEIRPLHTLEA--QPHTGSPD
WP_001347099.1   LKISASASLGARERVVMVDVEDARLVLGVTAGQINLLHKLPP----------
WP_034082864.1   IRTLASQPLGPRDRLVLVQVGEEQILLGITPGRITPLHVLKE--P-------
WP_036197595.1   IKVVSVMPMGTRERIALIEVGGTQILVGVTPSTIRTLHVFDE--P-------
WP_021182860.1   LKVSASVQVGQRERVVIVDTQDARLVLGVTAQQVTHLHTLPPLAP-------
WP_052246976.1   VRVVGGTMIGQKERVVIVELEGEWLVLGVTPQQVNLLSKMPR--P-------

    Conservation
                 949345359+5999779975433679 9+4753495 1+32822--5-------

         Quality

       Consensus
                 LKVVASCSVGQRERVVIVEVGDEQLVLGVTAQQITLLHTLPP++PAAAGTPA
```

- Example BLAST output
  - A list of **pairwise** alignments
  - One sequence (the Query) is always the same
  - In each case, the other sequence is the 'hit' (Subject)
  - N.B. these are **local** alignments (i.e. only of matching segments, not necessarily the whole sequence)
- More compact BLAST output formats are available
- These are **flatfiles** (plain text)
- Again, when interpreted and displayed by other software they may appear very different
- E.g. on a website which provides a BLAST service

```
> tr|A0A031GPC8|A0A031GPC8_9BURK Flagellar biosynthesis protein
FliO OS=Janthinobacterium lividum GN=fliO PE=4 SV=1
Length=196

 Score = 45.8 bits (107),  Expect = 0.009, Method: Composition-based stats.
 Identities = 16/114 (14%), Positives = 40/114 (35%), Gaps = 5/114 (4%)

Query  143  TSLEGYDYTNYILVMLILVILLIVLWWFKKTMVYKNNNVSRDFTMIFQRFLDKNNQLVVF  202
             +         I  ++ ++ LLI L WF K    K    +    ++    L    ++V+
Sbjct  79   PASSAGSLLQTIFALMFVLALLIGLAWFMKRYGPKVMGGNNKMRVVSSLNLGGRERIVLV  138

Query  203  DHANKRYTMIIGNSNV-LLESIEIPEEQTIKHTE----KTEKNFDSFFEENKKR  251
            + A++   +      +  L ++   E    +           NF  + ++  ++
Sbjct  139  EVADQWIVVGASPGRINALATMPRQEGDLPQLATAQNGPAAANFSEWLKQTIEK  192


> tr|A0A0B1REZ0|A0A0B1REZ0_9ENTR Flagellar assembly protein FliO
OS=Pantoea rodasii GN=QU24_01715 PE=4 SV=1
Length=131

 Score = 44.6 bits (104),  Expect = 0.009, Method: Composition-based stats.
 Identities = 15/98 (15%), Positives = 40/98 (41%), Gaps = 2/98 (2%)

Query  156  VMLILVILLIVL-WWFKKTMVYKNNNVSRDFTMIFQRFLDKNNQLVVFDHANKRYTMIIG  214
            V+ ++V+L++   W  K+         ++   +     + +  ++V+ D A+ R  + +
Sbjct  29   VLAVIVLLILACGWLAKRLGFAPKTVNTQALKISASVQVGRQERVVIVDTADARLVLGVT  88

Query  215  NSNVL-LESIEIPEEQTIKHTEKTEKNFDSFFEENKKR  251
               +  L S+     + +       ++F   F+   KR
Sbjct  89   AQQITHLHSLPPVPPEELASNSVAPQDFRQLFQNLVKR  126


> tr|A0A090U6M1|A0A090U6M1_9ENTR Flagellar biosynthesis protein
FliO OS=Citrobacter farmeri GTC 1319 GN=fliO PE=4 SV=1
Length=124

 Score = 44.2 bits (103),  Expect = 0.010, Method: Composition-based stats.
 Identities = 14/84 (17%), Positives = 30/84 (36%), Gaps = 0/84 (0%)

Query  168  WWFKKTMVYKNNNVSRDFTMIFQRFLDKNNQLVVFDHANKRYTMIIGNSNVLLESIEIPE  227
            W  K+       + +R   +     L    ++V+ D  + R  + +  SN+ +     P
Sbjct  37   WVIKRLGFSPKGSHTRGLKVSASTSLGPRERVVIVDVEDARLVLGVTASNISVLHTLPPA  96
```

“Traditional” alignment of 2 or more sequences

- Strictly speaking, an alignment is only meaningful if the sequences are homologous (related by descent from a common ancestor)
- This common descent applies to the individual nucleotide (or amino acid) positions –
  - i.e., equivalent bases/ amino acids are lined up
- This also enables inferences of evolutionary events:
  - Substitutions
  - Insertions
  - Deletions

“Traditional” alignment of 2 or more sequences

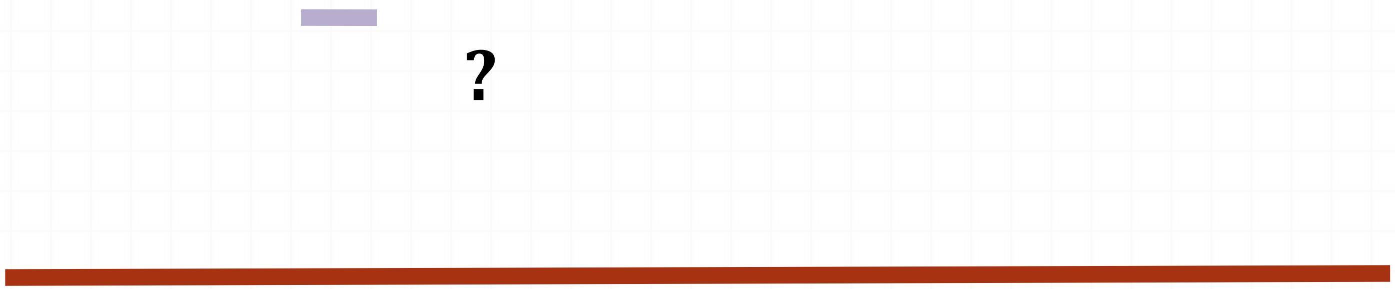
- But the process of alignment can help to **determine whether the sequences are homologous or not**
- This principle applies to **sequence similarity search** methods
  - E.g. **BLAST**
- - in which a single query sequence is compared, to many sequences in a database, one at a time
- i.e. **many <u>pairwise alignments</u>** (all involving the query sequence)
- Each alignment has a score
- Only those with a sufficiently high score are treated as **<u>hits</u>**
- The rest are ignored

- One way of very briefly summarising an alignment is to quote a single metric such as:
  - ‘percentage identity’
  - ‘percentage similarity’
- However, these are not absolutes, and depend on **how** the sequences are aligned
  - I.e., which method and scoring parameters are used
- These, and other aspects of alignment will be described further in a future session

Principles of sequence alignment

Other reasons for pairwise alignment

- Comparing 2 DNA sequences (or 1 RNA sequence and 1 DNA sequence) –
- where one is expected to be a (often very small) fragment of the other
- E.g.:
- Genome **<u>re</u>**sequencing
    - - and *de novo genome sequencing*
- RNASeq mapping



Genome resequencing

- You already have a **reference genome sequence** of organism X
- The new project is to sequence the genome of organism Y
  - Might be a closely related species
  - Or a different strain of the same species
- Assumption is that the 2 genome sequences are the same or very similar along most of their length
- ▬ Each sequence read from genome sequence Y can be “mapped” to the equivalent position in sequence X

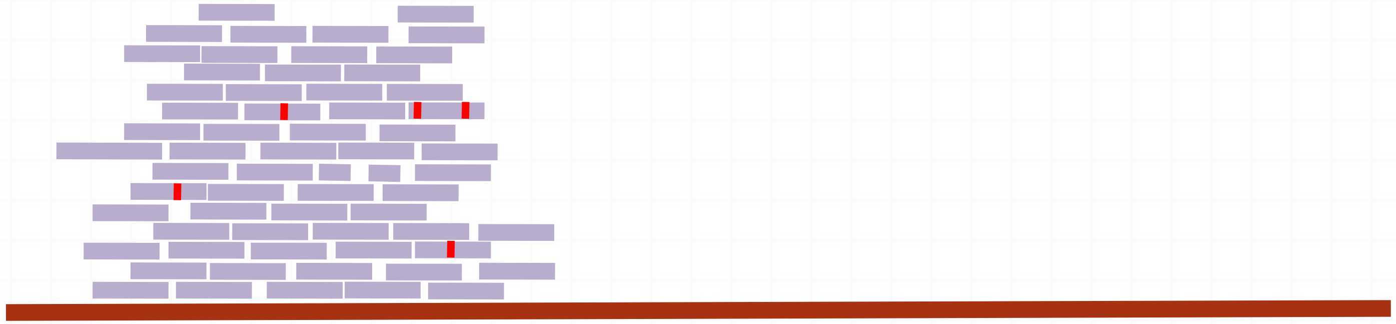
Genome resequencing

- N.B. depending on the circumstances, it may be just as good (or preferable) to do *de novo* assembly of genome Y
  - And then compare the whole of Y to X
  - This “comparative genomics” can be rather more complex than just a pairwise alignment of X and Y
- Resequencing of many genomes – or particular parts of genomes – helps to identify variations between strains in a species
- And variation among individuals in a population
- E.g. identification of SNPs
  - → ***variant calling***

RNASeq

- mRNA transcripts are sampled and sequenced
- The object is to map each “read” to a reference genome sequence
- Some reads won’t match perfectly, due to sequencing errors
  - - and even some biological errors
- This is more complicated in eukaryotes than prokaryotes
  - Due to the presence of introns (present in reference genome, absent from reads) and splice variants
- One approach to “community RNASeq” (metatranscriptomics) involves mapping reads to several/many reference genomes
  - Some reads will originate from genomes which are absent from the reference set

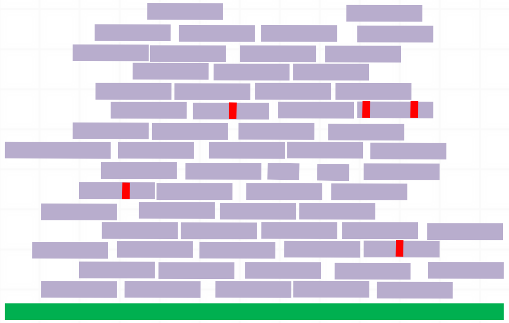
Storing information on mapped reads - economically



- *Is there any point in explicitly recording all the bases which are the **same** as the reference?*

- Each read can be recorded by specifying:
- Its start position relative to the reference genome sequence
- Its length
- All the differences between the read and the reference

De novo genome sequence assembly



- *Again, it should be necessary to record only positions and differences of each read, relative to the consensus*

- Similar principles
- But no reference genome sequence to compare the reads with
- Various algorithms, which in essence compare the reads with each other
- Produces “contigs”
- A consensus sequence can be produced from each contig

How can mapping locations and differences be encoded?

- Various approaches
- Some have ‘evolved’ beyond their original purpose
- **CIGAR** format (“Compact Idiosyncratic Gapped Alignment Report”)
  - (and it *is* idiosyncratic)
- CIGAR is used in **SAM** format (“Sequence Alignment/Map”) files
- SAM, has a “binary” (compressed) equivalent, BAM
  - **SAM and BAM are still very widely used**
- A more recent development is **CRAM** format
  - Stores only differences compared to reference – **minimal storage space**
  - CRAM files are **not** plain text; you need utilities (in the CRAMTOOLS software) to view them – displayed in e.g. FASTQ or SAM format

What do you need to know?

- **So how much do you need to know about these formats?**
  - A little awareness of the concepts will be useful.
  - Do not be concerned with understanding all the details.
    - (Probably not many people do understand them all…
    - the full SAM specification has many details, some more documented than others)
  - There have been many approaches to **compressing** this kind of sequence data, due to the ever-increasing sizes of data sets
    - Compression of SAM/BAM in particular
    - Compression of FASTQ data in general
    - Beyond the scope of this topic

CIGAR and SAM in brief

A quick overview for now.

There are some more details in another presentation

SAM, pileup and related formats

CIGAR format

- One CIGAR ‘string’ per alignment (read ↔ reference):
  - Example: ‘**6=1X7=2X6=**’
- Original CIGAR specification was designed to specify for each alignment:
  - **How long the alignment is**
  - **Where the insertions and deletions (indels) are**
  - And some other attributes like ‘clipping’, ‘padding’ (ignore this now)
- Later additions to CIGAR also permitted specification of:
  - ***Where*** **mismatching bases occur**
  - But ***NOT*** what those mismatches actually are
  - -it was simply not the original purpose of the tools which used CIGAR

SAM format

- Each read is represented on one line (1 line = 1 ‘record’)
- Each line has a field specifying the position on the reference sequence
- And a field in CIGAR format
- Then it gets complicated…
- ***Also additional <u>optional</u> fields*** – these can be used to state:
  - explicit base differences (the optional **`MD` field**)
  - (and many other things)
- SAM format **can** explicitly state the sequence of each read
  - And often does
  - But does **not** have to

That all seems a bit messy….?

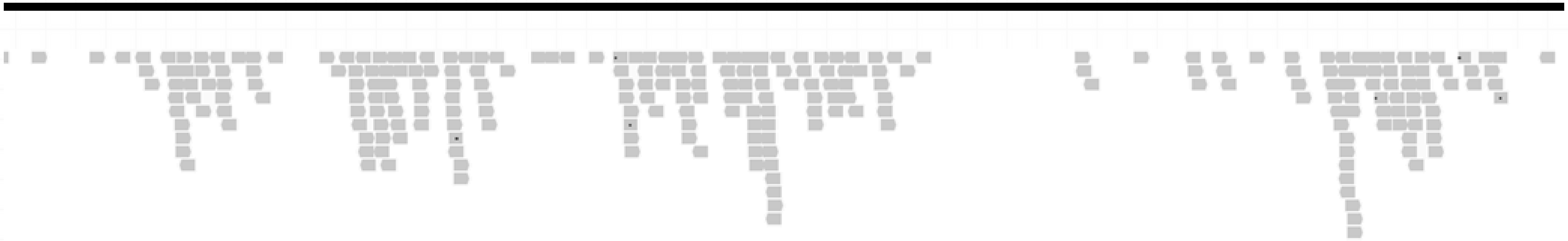
So why is it like this?

Evolution of bioinformatics formats

- This is an example of bioinformatics format specifications which have become revised over the years
- The original design may have been to achieve something very specific
- Additional functionality has been enabled by additional information included in the data files
- New fields “bolted on” to a simpler, earlier spec
- Complete revisions are generally uncommon because of the need to retain backward compatibility

Some context

- Example: You want to know **how many reads** are aligned to each part of the genome



Some context

- Example: You want to know **how many reads** are aligned to each part of the genome
- E.g. transcriptomics (RNAseq) – thus, gene expression levels
  - Which genes are most represented?
  - You don’t care if there are a small number of mismatches due to sequence errors
  - You don’t even care what the sequences of the alignments in each region actually ***are***
  - **All you want to know is where the alignment/mapping program (whatever that may be) mapped the reads**
  - **Only minimal CIGAR/SAM information is required for this**
  - (N.B. SAM format can optionally store scores produced by the alignment program)

Summary (1)

- There are many different file formats to store alignment data
- These can be read and created by many different software tools
- Nearly all of these formats are plain text (“flatfile”)
  - BAM is not, but is a compressed version of SAM
  - Various other compressed, “binary” formats exist, e.g. CRAM
- These file formats are used for a variety of purposes
  - Which data types are stored in any one file can depend on the purpose
- It is ***not*** necessary to understand the detailed specifications of these data file formats
- But it is important to know:
  - ***what kind of data*** **can/cannot be contained in these files**

Summary (2)

- SAM and BAM data files are principally manipulated with the SAMTOOLS software package
- Various other software can read SAM/BAM files
  - E.g. visual browsers, like the Integrated Genome Viewer
- External link to SAMTOOLS tutorial at BIOBITS:
  - http://biobits.org/samtools_primer.html
- If you are interested in learning more details of SAM, CIGAR etc., these are provided in a further presentation

SAM, pileup and related formats