Basics of Population Genetics and Phylogenetic Reconstruction

-> *Alignments* <-

C. Kiefer[1] M. Kiefer[1]

[1]COS Heidelberg
Dept. Biodiversity and Plant Systematics

Summer 2019

Outline

Centre for Organismal Studies Heidelberg

Outline

Alignment. . . ?

What?

What does the internet say?

Alignment. . . ?

What?

What does the internet say?

Archaeology a linear arrangement of megalithic standing stones.

Alignment. . . ?

What?

What does the internet say?

Archaeology a linear arrangement of megalithic standing stones.

Astronomy a straight line configuration of three celestial bodies.

Alignment. . . ?

What?

What does the internet say?

Archaeology a linear arrangement of megalithic standing stones.

Astronomy a straight line configuration of three celestial bodies.

RPG the moral and ethical perspective of characters, monsters, and societies.

Alignment. . . ?

What?

What does the internet say?

Archaeology a linear arrangement of megalithic standing stones.

Astronomy a straight line configuration of three celestial bodies.

RPG the moral and ethical perspective of characters, monsters, and societies.

Linguistics system used to distinguish between the arguments of transitive and intransitive verbs.

Alignment. . . ?

What?

What does the internet say?

Archaeology a linear arrangement of megalithic standing stones.

Astronomy a straight line configuration of three celestial bodies.

RPG the moral and ethical perspective of characters, monsters, and societies.

Linguistics system used to distinguish between the arguments of transitive and intransitive verbs.

Bioinformatics arranging the sequences of DNA, RNA, or protein to identify similarities

Alignment. . . ?

What?

What does the internet say?

Archaeology a linear arrangement of megalithic standing stones.

Astronomy a straight line configuration of three celestial bodies.

RPG the moral and ethical perspective of characters, monsters, and societies.

Linguistics system used to distinguish between the arguments of transitive and intransitive verbs.

Bioinformatics arranging the sequences of DNA, RNA, or protein to identify similarities

Basically it's just a datamatrix.

What does it consist of?

Anatomy of an alignment

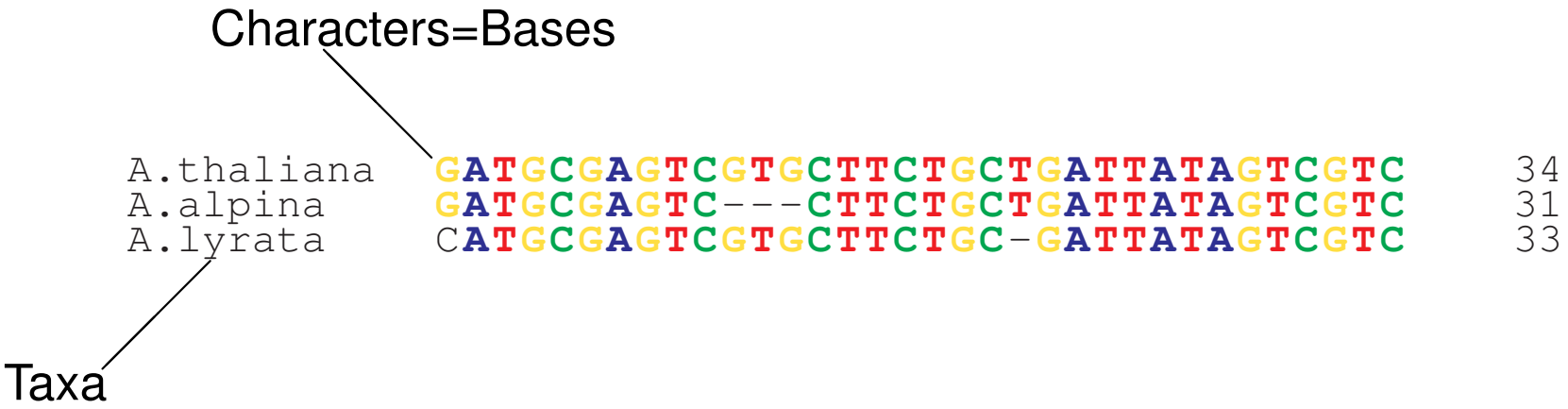
```
A.thaliana   GATGCGAGTCGTGCTTCTGCTGATTATAGTCGTC   34
A.alpina     GATGCGAGTC---CTTCTGCTGATTATAGTCGTC   31
A.lyrata     CATGCGAGTCGTGCTTCTGC-GATTATAGTCGTC   33
```

What does it consist of?

Anatomy of an alignment

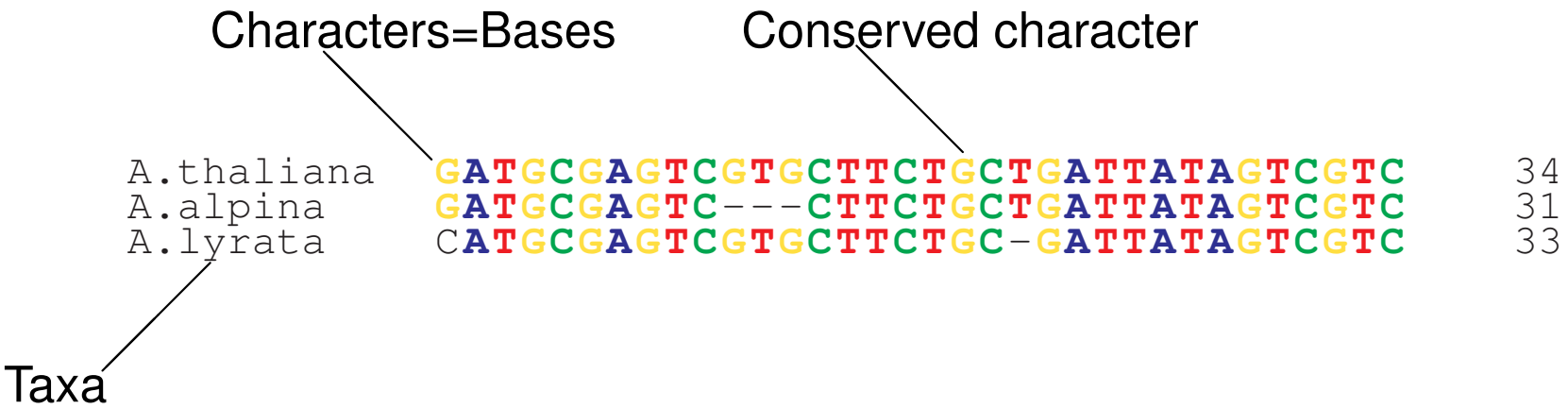
What does it consist of?

Anatomy of an alignment



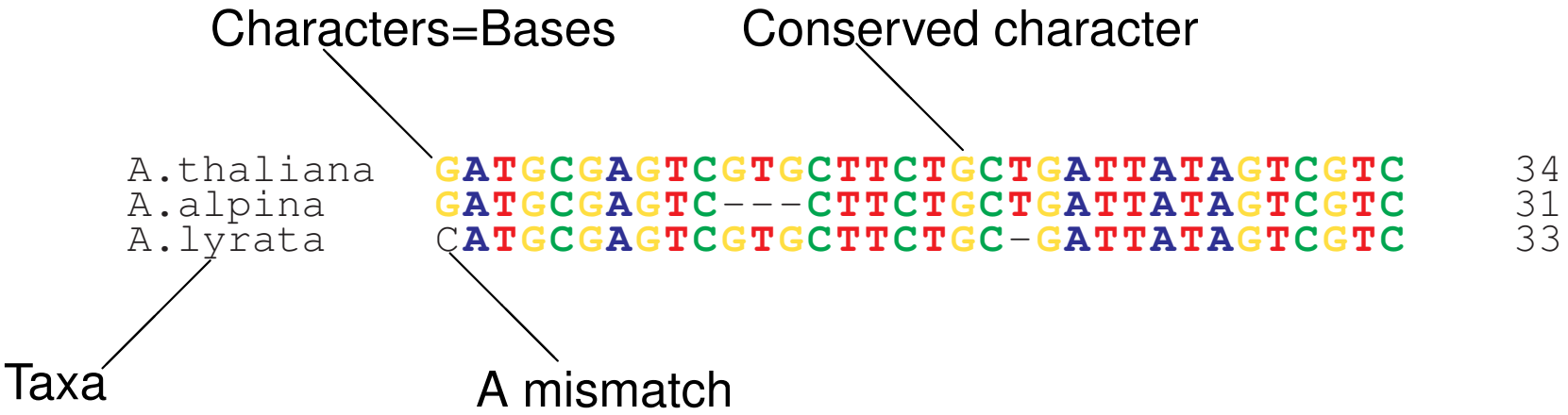
What does it consist of?

Anatomy of an alignment



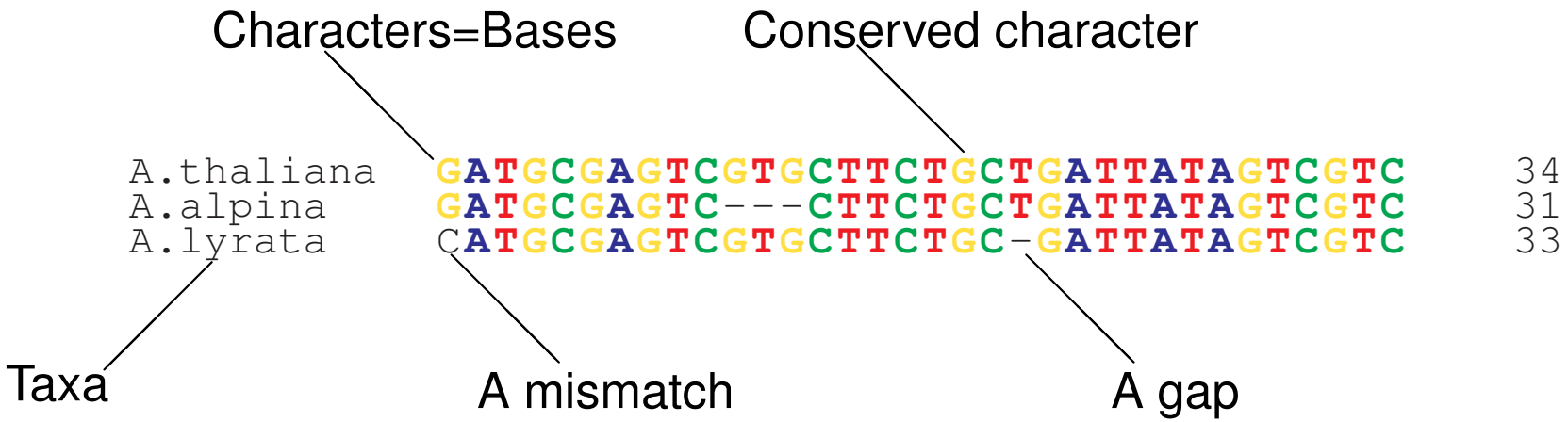
What does it consist of?

Anatomy of an alignment



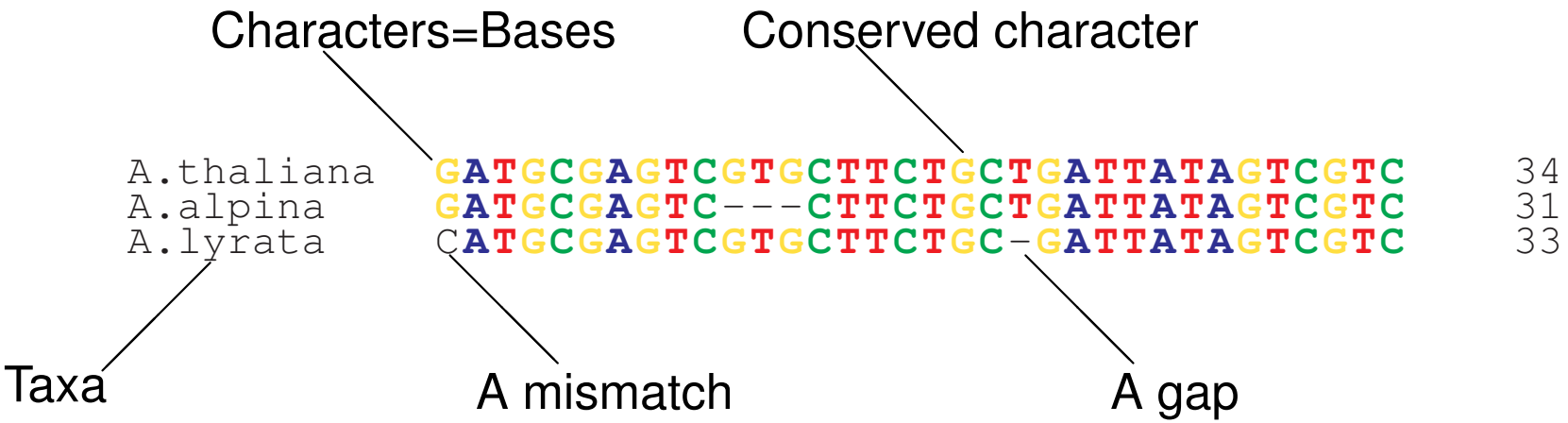
What does it consist of?

Anatomy of an alignment



What does it consist of?

Anatomy of an alignment



- Characters in columns
- Taxa in rows
- Base substitutions (SNPs) as mismatches
- Indels represented as gaps (-)

Alignments

What to do?

The task:

Make as many characters fit to their homologous counterparts in other taxa as possible.

Alignments

What to do?

The task:

Make as many characters fit to their homologous counterparts in other taxa as possible.

The problem:

Sequences differ when there are mutations.

SNP A base substitution, 1 bp, represented by a mismatch.

InDel An insertion or deletion, n bp, represented by a gap.

Alignments

What to do?

The task:

Make as many characters fit to their homologous counterparts in other taxa as possible.

The problem:

Sequences differ when there are mutations.

SNP A base substitution, 1 bp, represented by a mismatch.

InDel An insertion or deletion, n bp, represented by a gap.

The method:

Place as many clever gaps in your sequences as reasonable.

Sequence variability

What's the difference?

Variability

It wouldn't make much sense to compare identical sequences, now would it?
So what almost any sequence analysis is after, are those variations.

Sequence variability

What's the difference?

Variability

It wouldn't make much sense to compare identical sequences, now would it?
So what almost any sequence analysis is after, are those variations.

Substitutions

Sometimes bases just get exchanged against others. (Terms: SNP, transition, transversion. . .)
This does not affect the length of a given DNA stretch. A real substitution is usually just "accepted" in an alignment, without "improvement".

Sequence variability

What's the difference?

Variability

It wouldn't make much sense to compare identical sequences, now would it?
So what almost any sequence analysis is after, are those variations.

Substitutions

```
Seq1   ATCGTACGTATC   12
Seq2   ATCGAACGTATC   12
```

Sequence variability

What's the difference?

Variability

It wouldn't make much sense to compare identical sequences, now would it?
So what almost any sequence analysis is after, are those variations.

InDels

A stretch of DNA can get deleted from the genome (deletion) or something new can be inserted into it (insertion).
This changes the length of a given DNA stretch and has to be compensated with a gap (in this or the other sequence(s)). Only *one* mutation.

Sequence variability

What's the difference?

Variability

It wouldn't make much sense to compare identical sequences, now would it?
So what almost any sequence analysis is after, are those variations.

InDels

```
Seq1   ATCGTACGTATC   12
Seq2   ATCG---GTATC    9
```

Sequence variability

What's the difference?

Variability

It wouldn't make much sense to compare identical sequences, now would it?
So what almost any sequence analysis is after, are those variations.

Inversions

A stretch of DNA is deleted and re-inserted as its reverse (complement).
Especially pesky because it is notoriously difficult to spot. It looks like a funny bunch of odd SNPs but is really only *one* mutation, again.

Sequence variability

What's the difference?

Variability

It wouldn't make much sense to compare identical sequences, now would it?
So what almost any sequence analysis is after, are those variations.

Inversions

```
Seq1   ATCGTACGTATC    12
Seq2   ATCGCGTATATC    12
```

Variability choices

It's all about location.

Mutation impact

Can have different impact in different DNA.
SNPs in 3^{rd} base in codon position are more deleterious.
InDels that shift the reading frame are more deleterious.

Variability choices

It's all about location.

Mutation impact

Can have different impact in different DNA.
SNPs in 3^{rd} base in codon position are more deleterious.
InDels that shift the reading frame are more deleterious.
. . . as long as we're talking coding regions.

Variability choices

It's all about location.

Mutation impact

Can have different impact in different DNA.
SNPs in 3^{rd} base in codon position are more deleterious.
InDels that shift the reading frame are more deleterious.
. . . as long as we're talking coding regions.

Location

Variability usually rises from *coding* to *regulatory* to "*junk*" regions in the genome.

Variability choices

It's all about location.

Mutation impact

Can have different impact in different DNA.
SNPs in 3^{rd} base in codon position are more deleterious.
InDels that shift the reading frame are more deleterious.
. . . as long as we're talking coding regions.

Location

Variability usually rises from *coding* to *regulatory* to "*junk*" regions in the genome.

Sampling

Not every genomic region is suitable for every analysis.

What is an alignment, really?

Quite a lot, actually...

When aligning sequences you make assumptions about the course of evolution.

What is an alignment, really?

Quite a lot, actually...

When aligning sequences you make assumptions about the course of evolution.
Every gap you insert into your set of sequences assumes that a mutation has happened in one or more of the sequences, having more or less dramatic consequences on an organism and/or his lineage of offspring.

What is an alignment, really?

Quite a lot, actually. . .

When aligning sequences you make assumptions about the course of evolution.
Every gap you insert into your set of sequences assumes that a mutation has happened in one or more of the sequences, having more or less dramatic consequences on an organism and/or his lineage of offspring.
So gaps shouldn't be placed lightly. Think parsimoniously.

Alignments

Try it!

http://cumulus.cos.uni-heidelberg.de/alignapp/

Outline

Alignment

What is it good for?

Phylogenetic reconstruction

Though you can infer phylogenies from morphological data, most are reconstructed from molecular (DNA or protein) data. Thus, alignments are always the first step, here.

Alignment

What is it good for?

Phylogenetic reconstruction

Though you can infer phylogenies from morphological data, most are reconstructed from molecular (DNA or protein) data. Thus, alignments are always the first step, here.

Sequence identification

One Word:

Alignment

What is it good for?

Phylogenetic reconstruction

Though you can infer phylogenies from morphological data, most are reconstructed from molecular (DNA or protein) data. Thus, alignments are always the first step, here.

Sequence identification

One Word:
“Basic local alignment search tool”

Alignment

What is it good for?

Phylogenetic reconstruction

Though you can infer phylogenies from morphological data, most are reconstructed from molecular (DNA or protein) data. Thus, alignments are always the first step, here.

Sequence identification

One Word:
“Basic local alignment search tool”

Cloning experiments

To find overlapping regions when constructing Plasmids and stuff. . .

Alignment

What is it good for?

Assembly

When constructing contigs from raw sequencing reads. . .

```
ctagtcagctgatcta------------gatgctatcagctact
            |||||              |||||
            atctagcatgctgagcaggatgc
```

Alignment

What is it good for?

Assembly

When constructing contigs from raw sequencing reads. . .

```
ctagtcagctgatcta------------gatgctatcagctact
            |||||              |||||
            atctagcatgctgagcaggatgc
```

Finding patterns

Finding common patterns in differnt sequences.

Alignment

What is it good for?

Assembly

When constructing contigs from raw sequencing reads. . .

```
ctagtcagctgatcta------------gatgctatcagctact
            |||||              |||||
            atctagcatgctgagcaggatgc
```

Finding patterns

Finding common patterns in differnt sequences.

Designing PCR primers

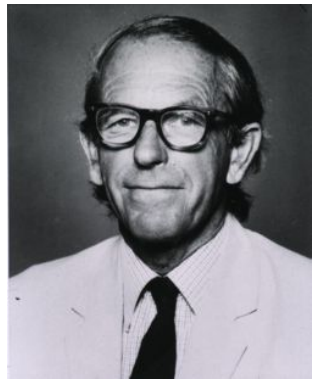
Finding regions of high conservation helps in picking robust primers.

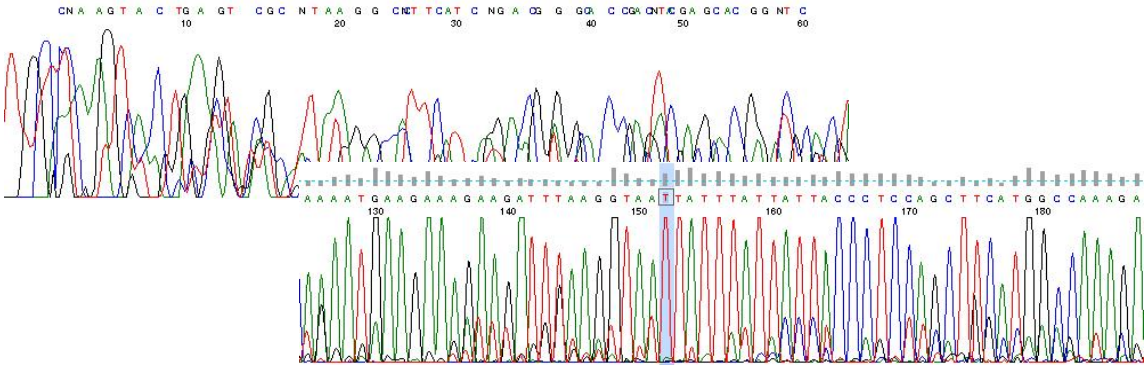
Outline

Classical Sequencing

Good ol' Sanger

- Didesoxy-sequencing...

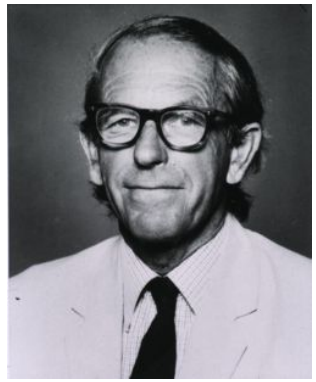


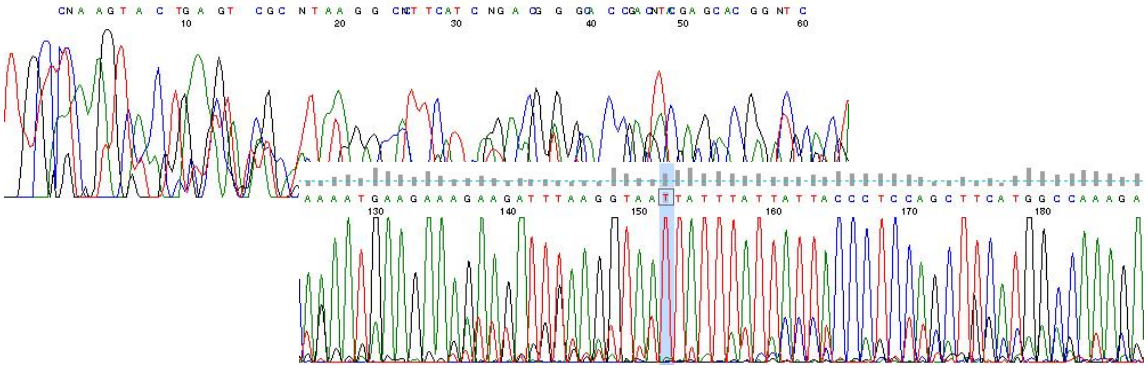


Classical Sequencing

Good ol' Sanger

- Didesoxy-sequencing...
- Approx. 1 kbp read length

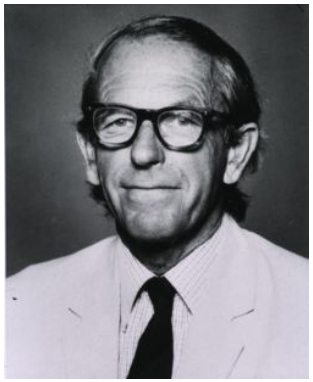


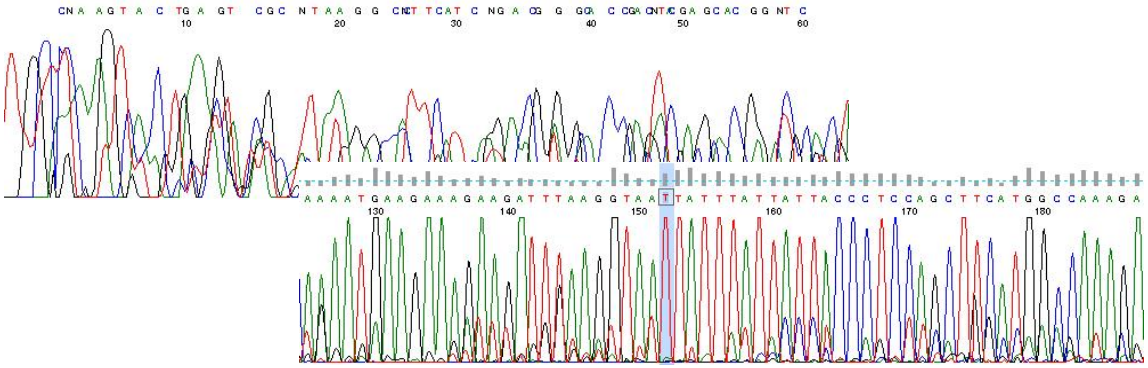


Classical Sequencing

Good ol' Sanger

- Didesoxy-sequencing...
- Approx. 1 kbp read length
- Cheap for single sequences
- Outrageously expensive for genomes

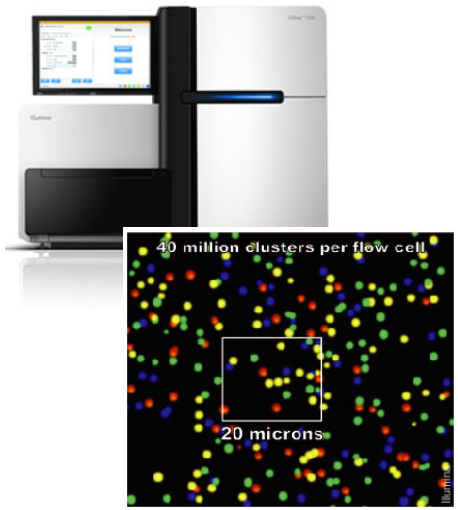




Next Generation Sequencing

Truckloads of data

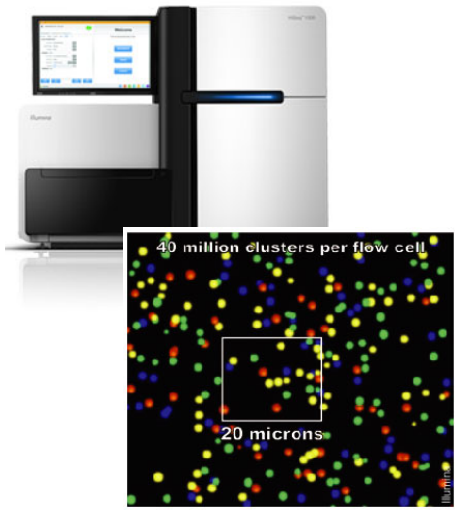
- Sequencing by synthesis, reversible terminators



Next Generation Sequencing

Truckloads of data

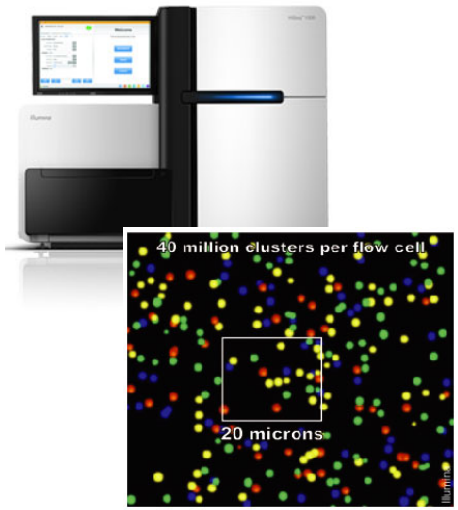
- Sequencing by synthesis, reversible terminators
- Read length approx. 200 bp, billions of those



Next Generation Sequencing

Truckloads of data

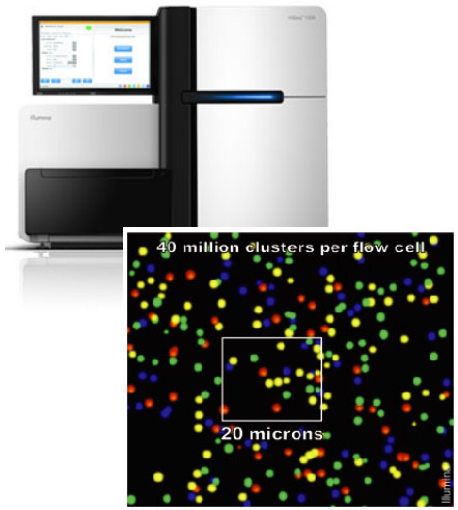
- Sequencing by synthesis, reversible terminators
- Read length approx. 200 bp, billions of those
- Affordable for genomes



Next Generation Sequencing

Truckloads of data

- Sequencing by synthesis, reversible terminators
- Read length approx. 200 bp, billions of those
- Affordable for genomes
- Bioinformatically slightly more taxing



Databases

Avoid wheel-reinventions

- Billions of basepairs have already been sequenced.

Databases

Avoid wheel-reinventions

- Billions of basepairs have already been sequenced.
- Most are readily available, why not use them?

Databases

Avoid wheel-reinventions

- Billions of basepairs have already been sequenced.
- Most are readily available, why not use them?
- Virtually “priceless”. . .

Databases

Avoid wheel-reinventions

- Billions of basepairs have already been sequenced.
- Most are readily available, why not use them?
- Virtually “priceless”...
- Choose genome regions for your analysis that are well used and available in databases.
  Complement by only a couple of own sequences.

Sequence Formats

ABI/SCF

- proprietary formats
- trace files generated by DNA sequencers
- special software needed to view or edit
- sequence data can be extracted
- usefull in the first phase of an analysis only, for quality control

Sequence Formats

EMBL

```
ID   SomeIDNumber
AC   AY32154
DE   Some gene with some product,complete CDS
XX
FT   some feature
SQ   315bp;80G,85C,46T,57A
     GGCCGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCGATATC CCCCGATATC      60
     GGCTGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCTGAATC CCCCGATATC     120
     GGCCGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCCTGATC CCCCGATATC     180
     GGCCGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCCTTATC CCCCGATATC     240
     GGCCGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCCAAACC CCCCGATATC     300
     ACGTCTGCCT GGGTG                                                     315
//
ID   SomeotherID
...
```

Sequence Formats

Genbank

```
LOCUS       SomeIDNumber       265 bp
DEFINITION  Some gene with some product, complete CDS
ORIGIN
       1  GGCCGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCGATATC
      51  GGCTGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCTGAATC
     101  GGCCGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCCTGATC
     151  GGCCGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCCTTATC
     201  GGCCGAGGGC ACGTCTGCCT GGGTGTCACA AATCGTCGTC CCCCCAAACC
     251  ACGTCTGCCT GGGTG
//
LOCUS       NextIDnumber       1024 bp
DEFINITION  Some other gene
ORIGIN
       1  GATCGATCTG ATCGTATCAA TAGCTACGTA TACGACTAGG TAGCTAGCTA
      51  TACGATCGAA CTAGCTACGA TCGATCGATC GATCGATCGA CTAGCTACGA
...
```

Sequence Formats

FASTA

```
>some gene|AY64738
GCTGCGcCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAGCTACGATCATCGACTAC
GCTGCGATCTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
GCTGCGATCTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
>some other gene
GCCAGTATCTAAAGTCTAGAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG ACCAGTATCTAAAGTCTAACACGACTCTCGGCA
>yet another gene
GACTAGCTACGTACGATCGACTAGCTAGCTACGTACGTACGATCGACAGCTAGCTACGTACGTACCGATCAGCTAC
TACGATCGACTAGCTAGCAGCTACGATCGATCGACTAGCACGTACGATCGATCGACTGACTGACTAGCTAGCTACG
GATCAGCGAGGCGGCGGGGAGCATCATTATTCGGCTACGATCGATTACGGCATCGATCGGGCGCGCATTATGCGAGC
...
```

Sequence Formats

FASTA

```
>some gene|AY64738
GCTGCGcCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAGCTACGATCATCGACTAC
GCTGCGATCTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
GCTGCGATCTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
>some other gene
GCCAGTATCTAAAGTCTAGAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG ACCAGTATCTAAAGTCTAACACGACTCTCGGC
>yet another gene
GACTAGCTACGTACGATCGACTAGCTAGCTACGTACGTACGATCGACAGCTAGCTACGTACGTACCGATCAGCTAC
TACGATCGACTAGCTAGCAGCTACGATCGATCGACTAGCACGTACGATCGATCGACTGACTGACTAGCTAGCTACG
GATCAGCGAGGCGGCGGGGAGCATCATTATTCGGCTACGATCGATTACGGCATCGATCGGGCGCGCATTATGCGAGC
...
```

Old, simple and very widely used format.
(Deep seq data ist delivered in a similar format: „FASTQ“, consisting of billions of short reads with quality annotation per base.)

Centre for Organismal Studies Heidelberg

Alignment Formats

(Gapped) FASTA

```
>some gene|AY64738
GCTGCGcCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAGCTACGATCATCGACTAC
GCTGCGATCTAAAGTCTAA--------------ACGGATATCTCGGCTCTCGCATCGATG
GCTGCGATCTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
>some other gene
GCCAGTATCTAAAGTCTAGAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
ACCAGTATCTAAAGTCTAACACGACTCTCGGCAACGGATATCT--------GCATCGATG
GCTGCGATCTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
>yet another gene
GACTAGCTACGTACGATCGACTAGCTAGCTACGTACGTACGATCGACAGCTAGCTACGTA
TACGATCGACTAGCTAGCAGCTACGATCGATCGACTAGCACGTACGATCGATCGACTGAC
GATC--CGAGGCGGCGGGGAGCATCATTATTCGGCTACGATCGATTACGGCATCGATCGG
...
```

Alignment Formats

(Gapped) FASTA

```
>some gene|AY64738
GCTGCGcCTCGGCAACGGATATCTCGGCTCTCGCATCGATGAGCTACGATCATCGACTAC
GCTGCGATCTAAAGTCTAA--------------ACGGATATCTCGGCTCTCGCATCGATG
GCTGCGATCTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
>some other gene
GCCAGTATCTAAAGTCTAGAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
ACCAGTATCTAAAGTCTAACACGACTCTCGGCAACGGATATCT--------GCATCGATG
GCTGCGATCTAAAGTCTAAAACGACTCTCGGCAACGGATATCTCGGCTCTCGCATCGATG
>yet another gene
GACTAGCTACGTACGATCGACTAGCTAGCTACGTACGTACGATCGACAGCTAGCTACGTA
TACGATCGACTAGCTAGCAGCTACGATCGATCGACTAGCACGTACGATCGATCGACTGAC
GATC--CGAGGCGGCGGGGAGCATCATTATTCGGCTACGATCGATTACGGCATCGATCGG
...
```

Still old, simple and very widely used.

Alignment Formats

Phylip

```
  5    42
Tax1       AAGCTNGGGC ATTTCAGGGT
Tax2       TAAGCCTTGG CAGTGCAGGG
Tax3       ACCGGTTGGC CGTTCAGGGT
Tax4       AAACCCTTGC CGTTACGCTT
Tax5       AAACCCTTGC CGGTACGCTT

GAGCCCGGGC AATACAGGGT AT
GAGCCGTGGC CGGGCACGGT AT
ACAGGTTGGC CGTTCAGGGT AA
AAACCGAGGC CGGGACACTC AT
AAACCATTGC CGGTACGCTT AA
```

Alignment Formats

Phylip

```
  5    42
Tax1       AAGCTNGGGC ATTTCAGGGT
Tax2       TAAGCCTTGG CAGTGCAGGG
Tax3       ACCGGTTGGC CGTTCAGGGT
Tax4       AAACCCTTGC CGTTACGCTT
Tax5       AAACCCTTGC CGGTACGCTT

GAGCCCGGGC AATACAGGGT AT
GAGCCGTGGC CGGGCACGGT AT
ACAGGTTGGC CGTTCAGGGT AA
AAACCGAGGC CGGGACACTC AT
AAACCATTGC CGGTACGCTT AA
```

The native format of Joe Felsenstein's Phylip package. Worth mentioning (only) because RAxML uses it.

Alignment Formats

Clustal ALN

```
CLUSTAL W (1.82) multiple sequence alignment


FOSB_MOUSE      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
                ************************************************************

FOSB_MOUSE      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
FOSB_HUMAN      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 120
                ********************************.***************:*.**:******

FOSB_MOUSE      GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT 180
FOSB_HUMAN      GGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT 180
                ****** ***** .*********************************************

FOSB_MOUSE      DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD 240
FOSB_HUMAN      DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD 240
                ************************************************************
```

Alignment Formats

Clustal ALN

```
CLUSTAL W (1.82) multiple sequence alignment


FOSB_MOUSE      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
                ************************************************************

FOSB_MOUSE      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
FOSB_HUMAN      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 120
                ********************************.***************:*.**:****** 

FOSB_MOUSE      GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT 180
FOSB_HUMAN      GGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT 180
                ****** ***** .********************************************** 

FOSB_MOUSE      DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD 240
FOSB_HUMAN      DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD 240
                ************************************************************
```

Of limited usefulness. But it's the default output of Clustal...

Alignment Formats

Nexus

```
#NEXUS
  begin data;
    dimensions ntax=5 nchar=36;
    format datatype=dna missing=? gap=-;
    matrix
    A ATCGATGCTGGTAGGCTAGCGTATGCTGCGACTGA
    B ATCGATGCTGGTAGGC--GCGTATGCTGCGACTGA
    C ATCGATGCTGGTAGGCTAGCGTATGCTGCGACTGA
    D ATCGATGCTGGTAGGCTAGCGTATGATGCGACTGA
    E ---GATGCTGGTAGGCTAGCGTATGCTGCGACTGA
    ;
  end;
  [Let's make an MP analysis. ]
  begin paup;
    hsearch;
    bootstrap nrep=100;
    contree;
  end;
```

Alignment Formats

Nexus

```
#NEXUS
  begin data;
    dimensions ntax=5 nchar=36;
    format datatype=dna missing=? gap=-;
    matrix
    A ATCGATGCTGGTAGGCTAGCGTATGCTGCGACTGA
    B ATCGATGCTGGTAGGC--GCGTATGCTGCGACTGA
    C ATCGATGCTGGTAGGCTAGCGTATGCTGCGACTGA
    D ATCGATGCTGGTAGGCTAGCGTATGATGCGACTGA
    E ---GATGCTGGTAGGCTAGCGTATGCTGCGACTGA
    ;
  end;
  [Let's make an MP analysis. ]
  begin paup;
    hsearch;
    bootstrap nrep=100;
    contree;
  end;
```

We have seen this before. . .

Useful because it's relatively widely used, and has the ability for batch analyses.

Outline

Scoring

How good is my alignment?

To do something computationally we usually have to find a way to quantify it.

Scoring

How good is my alignment?

To do something computationally we usually have to find a way to quantify it.
A numerical scoring scheme for alignments includes:

Scoring

How good is my alignment?

To do something computationally we usually have to find a way to quantify it.
A numerical scoring scheme for alignments includes:

Match score A positive score for matching residues.

Scoring

How good is my alignment?

To do something computationally we usually have to find a way to quantify it.
A numerical scoring scheme for alignments includes:

Match score A positive score for matching residues.

Mismatch score A negative penalty for mismatching residues.

Scoring

How good is my alignment?

To do something computationally we usually have to find a way to quantify it.
A numerical scoring scheme for alignments includes:

Match score A positive score for matching residues.

Mismatch score A negative penalty for mismatching residues.

Gap opening penalty A penalty for opening a gap. . .

Scoring

How good is my alignment?

To do something computationally we usually have to find a way to quantify it.
A numerical scoring scheme for alignments includes:

Match score A positive score for matching residues.

Mismatch score A negative penalty for mismatching residues.

Gap opening penalty A penalty for opening a gap. . .

Gap extension pen. A variable (lower) penalty for enlarging gaps.

Scoring

How good is my alignment?

To do something computationally we usually have to find a way to quantify it.
A numerical scoring scheme for alignments includes:

Match score A positive score for matching residues.

Mismatch score A negative penalty for mismatching residues.

Gap opening penalty A penalty for opening a gap. . .

Gap extension pen. A variable (lower) penalty for enlarging gaps.

Why is the GEP lower than the GOP?

Scoring

How good is my alignment?

To do something computationally we usually have to find a way to quantify it.
A numerical scoring scheme for alignments includes:

Match score A positive score for matching residues.

Mismatch score A negative penalty for mismatching residues.

Gap opening penalty A penalty for opening a gap. . .

Gap extension pen. A variable (lower) penalty for enlarging gaps.

Why is the GEP lower than the GOP?

Why do we have to score gaps at all?

Scoring Matrices

Match/mismatch reloaded

DNA is easy

$$\begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{pmatrix} \end{array}$$

This is nothing more than the wellknown match/mismatch scheme.

Scoring Matrices

Match/mismatch reloaded

DNA is easy

$$\begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{pmatrix} \end{array}$$

This is nothing more than the wellknown match/mismatch scheme.

What if you took transition/transversion into account, too?

Scoring Matrices

Match/mismatch reloaded

DNA is easy

$$\begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{pmatrix} \end{array}$$

This is nothing more than the wellknown match/mismatch scheme.

What if you took transition/transversion into account, too?

Proteins are. . . not.

$$\begin{array}{c} \\ C \\ S \\ T \\ P \\ A \end{array} \begin{array}{c} \begin{array}{ccccc} C & S & T & P & A \end{array} \\ \begin{pmatrix} 9 & -1 & -1 & -3 & 0 \\ -1 & 4 & -1 & -1 & 1 \\ -1 & 1 & 4 & 1 & -1 \\ -3 & -1 & 1 & 7 & -1 \\ 0 & 1 & -1 & -1 & 4 \end{pmatrix} \end{array}$$

This is part of BLOSUM62, a common protein scoring matrix, including “functional similarity” of amino acids.

Centre for Organismal Studies Heidelberg

Brute force

. . . and ignorance.

Though it seems easy to just evaluate every possible alignment like that, the overwhelming number of possible gapped alignments makes that approach impossible.

Even small tasks with 2 sequences of very moderate length would make you wait a couple of decades. (e.g. 10^{88} possibilities for 2×300 characters)

Brute force

. . . and ignorance.

Though it seems easy to just evaluate every possible alignment like that, the overwhelming number of possible gapped alignments makes that approach impossible.

Even small tasks with 2 sequences of very moderate length would make you wait a couple of decades. (e.g. 10^{88} possibilities for 2×300 characters)

And it'd be *really* inelegant, too.

Dynamic Programming (DP)

Needleman and Wunsch

Saul B. Needleman, Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, **1970**

- "Programming" in this context means s.t. like "optimization".

Dynamic Programming (DP)

Needleman and Wunsch

Saul B. Needleman, Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, **1970**

- "Programming" in this context means s.t. like "optimization".
- Solving big problems by solving smaller, overlapping subproblems one by one

Dynamic Programming (DP)

Needleman and Wunsch

Saul B. Needleman, Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, **1970**

- "Programming" in this context means s.t. like "optimization".
- Solving big problems by solving smaller, overlapping subproblems one by one
- Here: create subalignments and assemble the optimal complete alignment according to their scores

Dynamic Programming (DP)

The algorithm

Let's begin with two short sequences:

5′-ATTGG-3′

5′-ATGC-3′

Dynamic Programming (DP)

The algorithm

Let's begin with two short sequences:

5′-ATTGG-3′

5′-ATGC-3′

And a scoring scheme:

$$S_{x,y} = max \begin{cases} S_{x,y-1} - 2 \\ S_{x-1,y-1} + M_{x,y} \\ S_{x-1,y} - 2 \end{cases}$$

With M = +1 or -1 for match or mismatch.

Dynamic Programming (DP)

The algorithm

With this we will create a matrix that holds the scores for every (sub)alignment of two fragments of our sequences.
This will allow us to finally construct the complete alignment.

Dynamic Programming (DP)

The algorithm

So, create a matrix:

	-	A	T	T	G	G
-	0	-2				
A	-2					
T						
G						
C						

Dynamic Programming (DP)

The algorithm

This means:

In each new cell, check which of the three predecessors ($\downarrow$, $\rightarrow$, $\searrow$) maximizes the score for that cell.

$\downarrow$ and $\rightarrow$ stand for a gap in one of the two sequences (no new base is used), so a gap penalty of -2 is applied.

$\searrow$ stands for a match or mismatch, so +1 or -1 is applied, respectively.

		-		A
	-	0		-2
			$\searrow$	$\downarrow$
	A	-2	$\rightarrow$	**1**

Dynamic Programming (DP)

The algorithm

This means:

In each new cell, check which of the three predecessors ($\downarrow$, $\rightarrow$, $\searrow$) maximizes the score for that cell.

$\downarrow$ and $\rightarrow$ stand for a gap in one of the two sequences (no new base is used), so a gap penalty of -2 is applied.

$\searrow$ stands for a match or mismatch, so +1 or -1 is applied, respectively.

		-		A
	-	0		-2
			$\searrow$	$\downarrow$
	A	-2	$\rightarrow$	**1**

$\downarrow$ means: $-2 - 2 = -4$
$\rightarrow$ means: $-2 - 2 = -4$
$\searrow$ is a match here and thus means: $0 + 1 = 1$
So $\searrow$ is best.

Dynamic Programming (DP)

The algorithm

	-		A		T		T		G		G
-	0	→	-2	→	-4	→	-6	→	-8	→	-10
	↓	↘									
A	-2		1								
	↓										
T	-4										
	↓										
G	-6										
	↓										
C	-8										

Dynamic Programming (DP)

The algorithm

	-		A		T		T		G		G
-	0	→	-2	→	-4	→	-6	→	-8	→	-10
	↓	↘									
A	-2		1								
	↓										
T	-4										
	↓										
G	-6										
	↓										
C	-8										

Try filling out a few.

Dynamic Programming (DP)

The algorithm

	-		A		T		T		G		G
-	**0**	→	-2	→	-4	→	-6	→	-8	→	-10
	↓	↘									
A	-2		**1**	→	-1	→	-3	→	-5	→	-7
	↓		↓	↘		↘					
T	-4		-1		**2**	→	0	→	-2	→	-4
	↓		↓		↓	↘					
G	-6		-3		0		**1**	→	-1		-1
	↓		↓		↓	↘	↓	↘			
C	-8		-5		-2		-1		**2**	→	**0**

Dynamic Programming (DP)

The algorithm

	-		A		T		T		G		G
-	**0**	→	-2	→	-4	→	-6	→	-8	→	-10
	↓	↘									
A	-2		**1**	→	-1	→	-3	→	-5	→	-7
	↓		↓	↘		↘					
T	-4		-1		**2**	→	0	→	-2	→	-4
	↓		↓		↓	↘					
G	-6		-3		0		**1**	→	-1		-1
	↓		↓		↓	↘	↓	↘			
C	-8		-5		-2		-1		**2**	→	**0**

Follow the arrows backward and pick bases and gaps as necessary to build the alignment.

Try it out on: http://cumulus.cos.uni-heidelberg.de/dp/

Alignments

Global or local

Global

- Use all characters.
- Find an optimum that includes the complete sequences.
- Needleman-Wunsch algorithm

Alignments

Global or local

Global

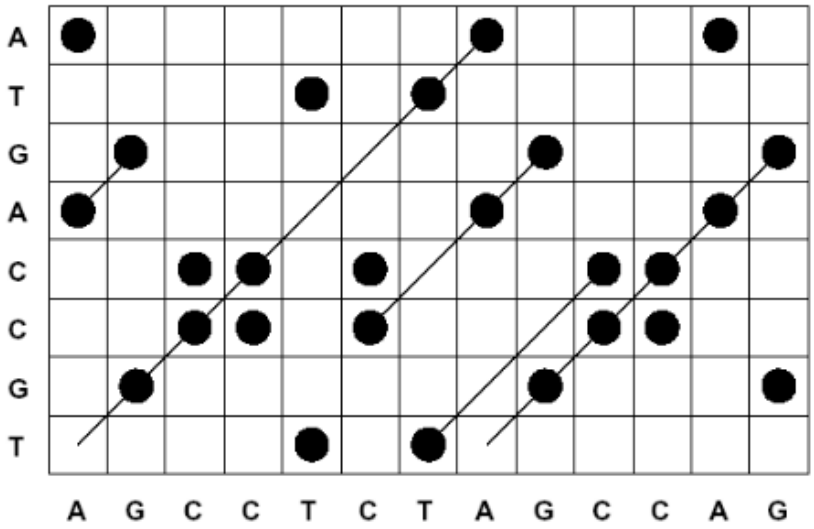
- Use all characters.
- Find an optimum that includes the complete sequences.
- Needleman-Wunsch algorithm

Local

- Use only the most homologous parts of the sequences.
- Allowed to skip the rest if necessary.
- Smith-Waterman algorithm (also DP).
- Not necessarily just part of a global one.

Dotplots

A quite visual algorithm



Create a matrix from the sequences, place “dots” in match-cells.

Diagonals are local (sub)alignments.
Try to get longer diagonals by opening gaps.

Dotplots

Exercise

AACTGTACTCGTGAGCGATGGGCAAT

AACTGTTCGTGAGCGATGCAAAG

Create a global alignment with a dotplot.

Outline

Centre for Organismal Studies Heidelberg

Multisequence alignments (MSA)

Why not n-dimensional DP?

n-dimensional DP

Is possible with n-dimensional matrices, but (again) just too computationally expensive.

Progressive Alignment

- Approximation, not necessarily optimal.
- Depends on correct estimation of sequence similarity.
- Reasonably fast.
- Widely used and easily available.

Progressive Alignment

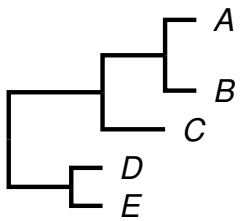
First: guide tree

- do optimal pairwise NW-alignments with every possible combination of sequences
- build distance matrix from results
- build “guide tree” from distance matrix

Progressive Alignment

First: guide tree

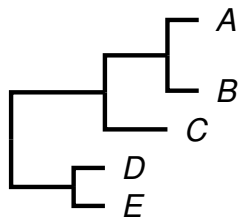
- do optimal pairwise NW-alignments with every possible combination of sequences
- build distance matrix from results
- build “guide tree” from distance matrix
  - an NJ algorithm ist used
  - the guide tree is saved as a dendrogramm file (*.dnd)
  - premade trees can be fed to the program instead



Progressive Alignment

Second: Guided pairwise Alignments

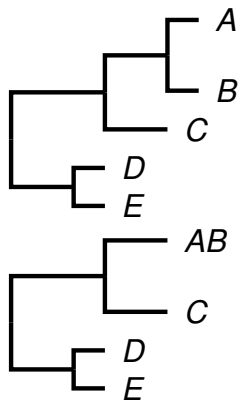
- Align most similar pair of sequences, using DP algorithm, again.



Progressive Alignment

Second: Guided pairwise Alignments

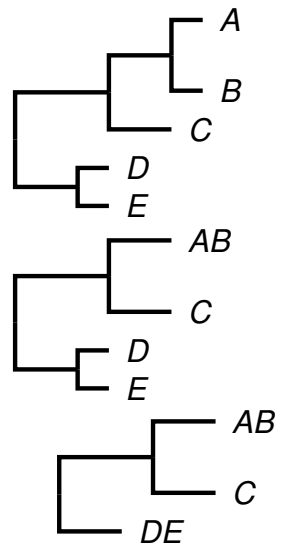
- Align most similar pair of sequences, using DP algorithm, again.
- Replace pair in tree by consensus sequence.



Progressive Alignment

Second: Guided pairwise Alignments

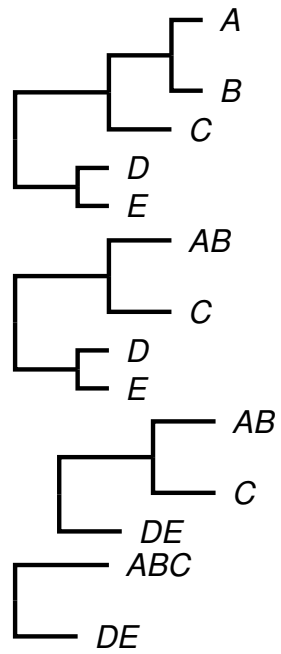
- Align most similar pair of sequences, using DP algorithm, again.
- Replace pair in tree by consensus sequence.
- Go on with next pair and so on, treating consensus sequences just as normal ones.



Progressive Alignment

Second: Guided pairwise Alignments

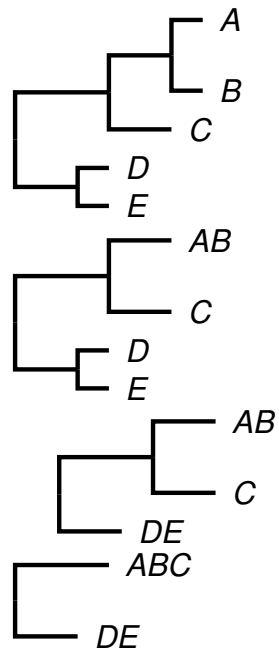
- Align most similar pair of sequences, using DP algorithm, again.
- Replace pair in tree by consensus sequence.
- Go on with next pair and so on, treating consensus sequences just as normal ones.



Progressive Alignment

Second: Guided pairwise Alignments

- Align most similar pair of sequences, using DP algorithm, again.
- Replace pair in tree by consensus sequence.
- Go on with next pair and so on, treating consensus sequences just as normal ones.
- Keep track of inserted gaps.
- Build MSA from gapped sequences.



Multisequence alignments (MSA)

Consensus sequences

How to create consensus sequences?

Democratic Most abundant character per position.

Multisequence alignments (MSA)

Consensus sequences

How to create consensus sequences?

Democratic Most abundant character per position.

Pattern List of possibilities per position
(Ambiguity codes, RegExs)

Multisequence alignments (MSA)

Consensus sequences

How to create consensus sequences?

Democratic Most abundant character per position.

Pattern List of possibilities per position (Ambiguity codes, RegExs)

Profile List of probabilities per position

```
ATGGCT
ACGGCT
AYGGCT A[TC]GGCT A(T:0.5,C:0.5)GGCT
```

Multisequence alignments (MSA)

MSA Scores

MSA-Score

Usually just the sum of the pairwise scores.

Multisequence alignments (MSA)

MSA Scores

MSA-Score

Usually just the sum of the pairwise scores.

Comparability

Not easily possible to compare “goodness” of different alignments by their scores.

When might that actually make sense?

Outline

“Database Alignments”

The needle and the haystack

The problem

Find sequences in a large database by homology, using –of course– local alignments.

“Database Alignments”

The needle and the haystack

The problem

Find sequences in a large database by homology, using –of course– local alignments.

Depending on the database’s size pairwise DP-alignments are too costly.

“Database Alignments”

The needle and the haystack

The problem

Find sequences in a large database by homology, using –of course– local alignments.

Depending on the database’s size pairwise DP-alignments are too costly.

The method

Do not align everything, only use promising candidates filtered by lookups in a pre-built index.

“Database Alignments”

The needle and the haystack

The problem

Find sequences in a large database by homology, using –of course– local alignments.

Depending on the database’s size pairwise DP-alignments are too costly.

The method

Do not align everything, only use promising candidates filtered by lookups in a pre-built index.

A “promising candidate” for alignment has to contain several small fragments from the query sequence to qualify for a closer look.

“Database Alignments”

. . . are useful, why?

Because it is easy, that way, to find more sequences to include into an analysis within the scope of this course.

Find DQ060111.1 on https://www.ncbi.nlm.nih.gov/nuccore.

What is it? Run a BLAST search on it. Evaluate the result list for usability in a phylogenetic reconstruction. What would you include, what rather not?

Outline

Programs for Pairwise Alignments

Needle and Water

In Principal, every Alignment program –be it for MSA or pairwise– will create a decent pairwise alignment in a very short time.

Specialized pairwise alignment software *will* be quicker, which has to be taken into account when you plan on doing large amounts of pairs.

There's many implementations for this task to be found on the web, mostly based on EMBOSS's tools.

Programs for Pairwise Alignments

Needle and Water

In Principal, every Alignment program –be it for MSA or pairwise– will create a decent pairwise alignment in a very short time.

Specialized pairwise alignment software *will* be quicker, which has to be taken into account when you plan on doing large amounts of pairs.

There's many implementations for this task to be found on the web, mostly based on EMBOSS's tools.

Create two (random) sequences using `nano` on the server (bioinf2) and align them with `needle` and with `water`.

Outline

A Software List

In no particular order

Clustal	Trusty PA
Pileup	Old PA
T-Coffee	Probability-enhanced PA
Dialign	Dotplots with anchoring
Mafft	Versatile and fast PA & more
Prank	Works well, I've been told...

And many others.

Clustal

W or X or Ω

ClustalW

Long-time standard version of Clustal. Commandline-based.
A little slow, maybe.

Clustal

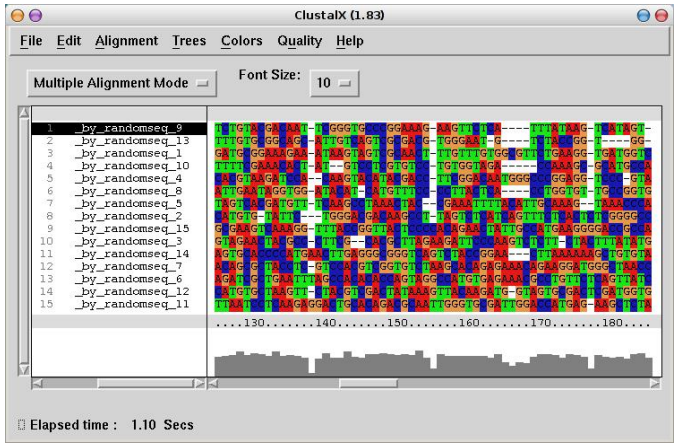
W or X or Ω

ClustalW

Long-time standard version of Clustal. Commandline-based.
A little slow, maybe.

ClustalX

Just a graphical user interface (GUI) for ClustalW.
Can do tricks like partial realignment or combining alignments (“profile mode”).



Clustal

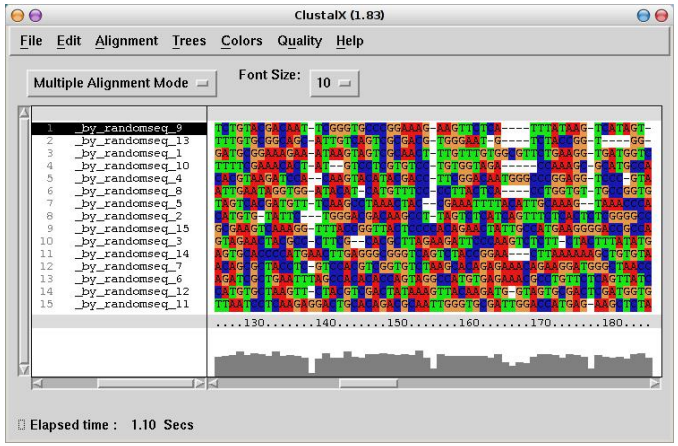
W or X or Ω

ClustalW

Long-time standard version of Clustal. Commandline-based.
A little slow, maybe.

ClustalX

Just a graphical user interface (GUI) for ClustalW.
Can do tricks like partial realignment or combining alignments (“profile mode”).



ClustalΩ

The “final” version of Clustal. (Supposedly for proteins.)

Mafft

The complete package

- Choose between speed and accuracy.
- Choose between several working modes.
- Can make use of multi-cored computers.
- Web-based versions available.

Multialignment Software

Exercise

Retrieve a dozen trnLF-IGS sequences from Genbank as FASTA.
Align them with ClustalX on Windows.
Align them on the server using:
`time clustalw trnlf.fasta` and
`time mafft trnlf.fasta > out.fasta`
Use `alan file.fasta` to check for differences.

Outline

Alignment Editors

Back to the roots

Sometimes you don't want to rely on automatically generated alignments.

You definitively want to have a closer look on the handywork of your alignment program.

Especially FASTA-Alignments are difficult to make sense of.

Alignment Editors

Back to the roots

Sometimes you don't want to rely on automatically generated alignments.

You definitively want to have a closer look on the handywork of your alignment program.

Especially FASTA-Alignments are difficult to make sense of.

So: *Use an alignment editor.*

Alignment Editors

A list

Alignment Editors

A list

- PhyDE

Alignment Editors

A list

- PhyDE

Period.

Alignment Editors

A list

- PhyDE

Period.
(There used to be a few others like GeneDoc or BioEdit.)

PhyDE

An alignment editor

- A Java program, running (equally well or crappy) on all major platforms.
- Developed at Bonn University.
- Get it anytime for free from: http://www.phyde.de/
- Works with DNA and Peptides, even with a mixture of both.
- Quite convenient tool for the task.

PhyDE

An alignment editor

- A Java program, running (equally well or crappy) on all major platforms.
- Developed at Bonn University.
- Get it anytime for free from: http://www.phyde.de/
- Works with DNA and Peptides, even with a mixture of both.
- Quite convenient tool for the task.

Load your unaligned FASTA sequences into PhyDE and familiarize yourself with it. Insert a few gaps to see if you can get it aligned (partially) by hand.