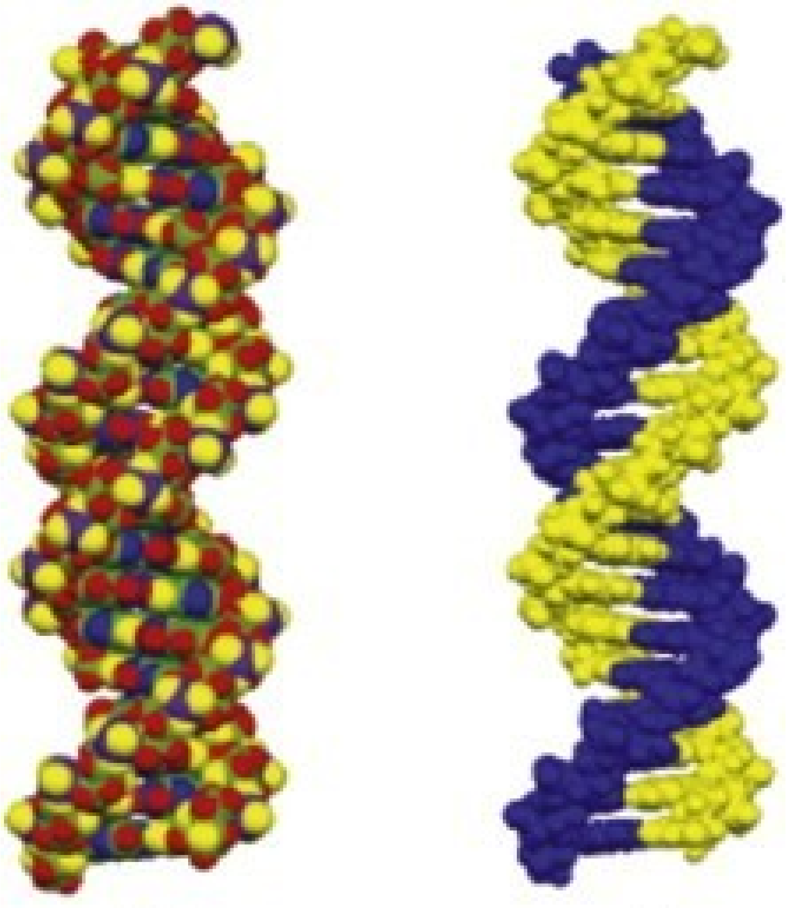
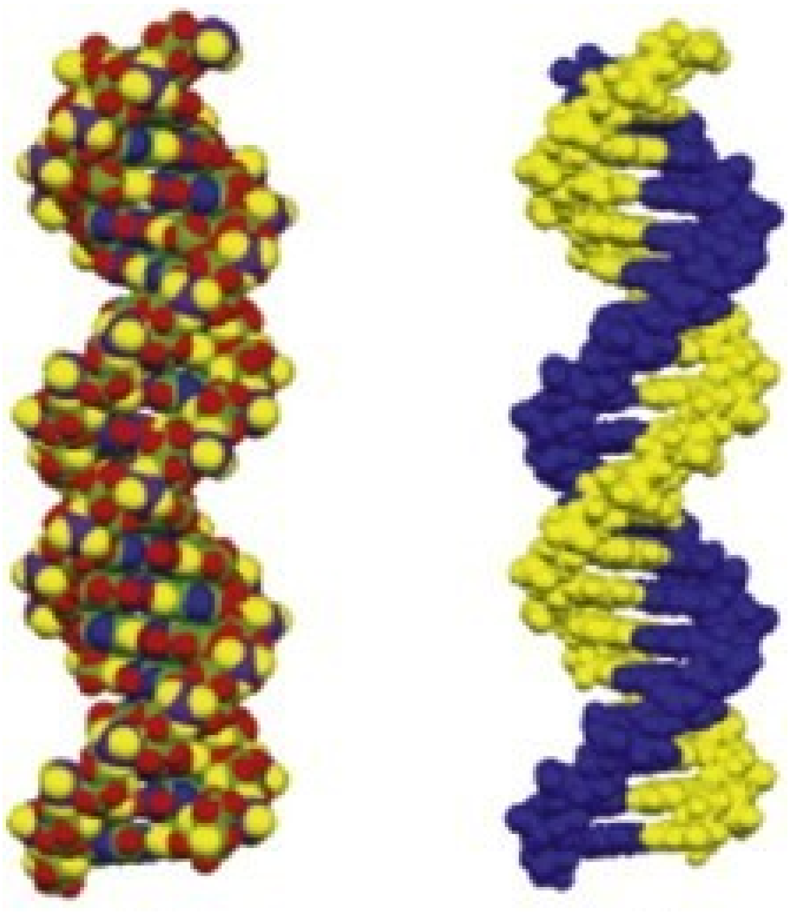
Lab closing
Bioinformatics 1 students only



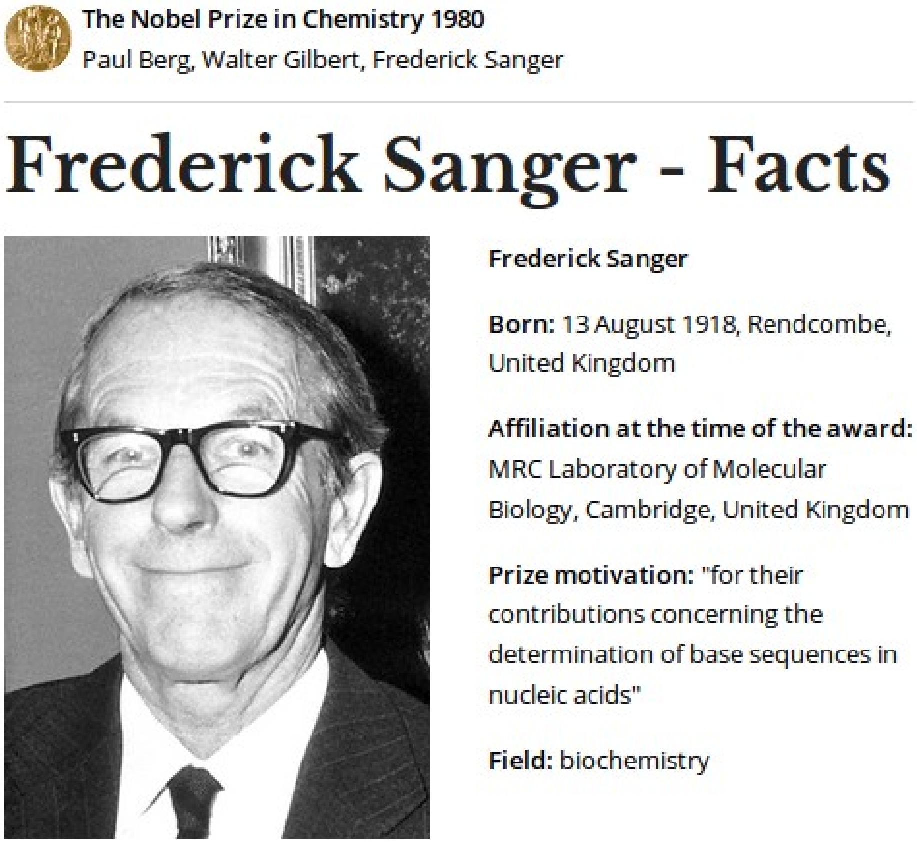
Bioinformatics 1 Lecture 3 Sequencing and Sequence Alignment Basics



Today's topics

- DNA sequencing
- Sequence alignment, fundamentals
- BLAST

DNA Sequencing



The Nobel Prize in Chemistry 1980

Paul Berg, Walter Gilbert, Frederick Sanger

Frederick Sanger - Facts

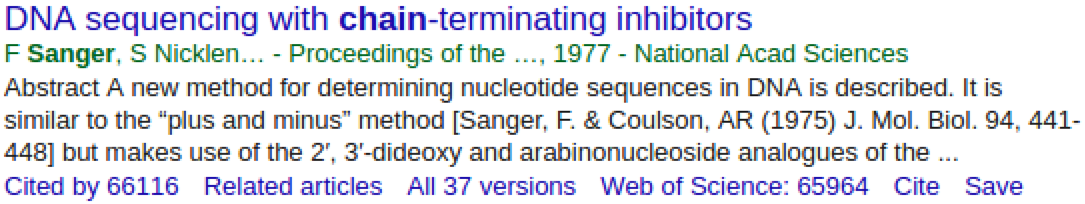
Frederick Sanger

Born: 13 August 1918, Rendcombe, United Kingdom

Affiliation at the time of the award: MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

Prize motivation: "for their contributions concerning the determination of base sequences in nucleic acids"

Field: biochemistry



DNA sequencing with **chain**-terminating inhibitors

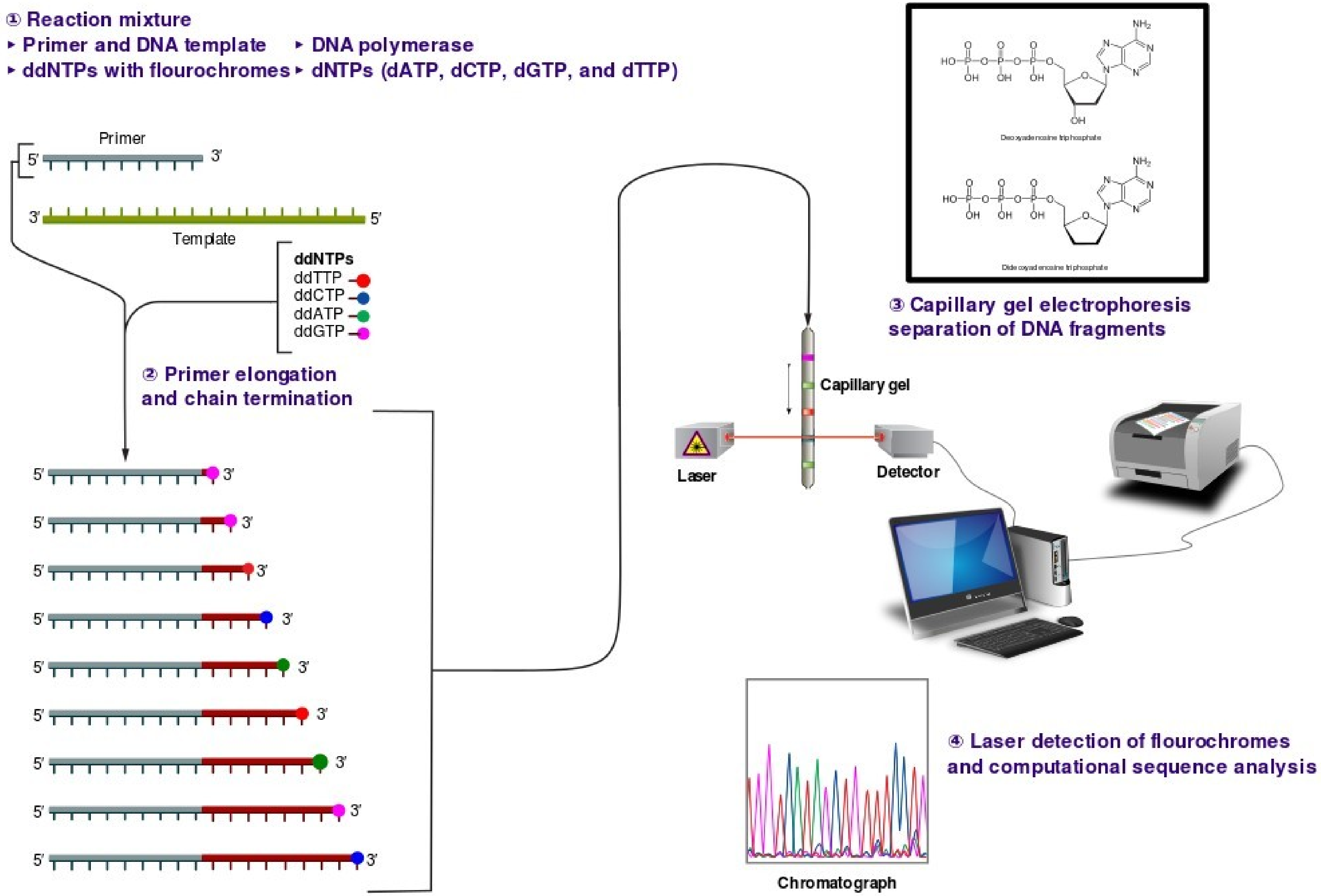
F **Sanger**, S Nicklen… - Proceedings of the …, 1977 - National Acad Sciences

Abstract A new method for determining nucleotide sequences in DNA is described. It is similar to the “plus and minus” method [Sanger, F. & Coulson, AR (1975) J. Mol. Biol. 94, 441-448] but makes use of the 2′, 3′-dideoxy and arabinonucleoside analogues of the ...

Cited by 66116 Related articles All 37 versions Web of Science: 65964 Cite Save

http://www.dnalc.org/view/15923-Cycle-sequencing.html

Sanger chain termination method



① Reaction mixture
- Primer and DNA template
- ddNTPs with flourochromes
- DNA polymerase
- dNTPs (dATP, dCTP, dGTP, and dTTP)

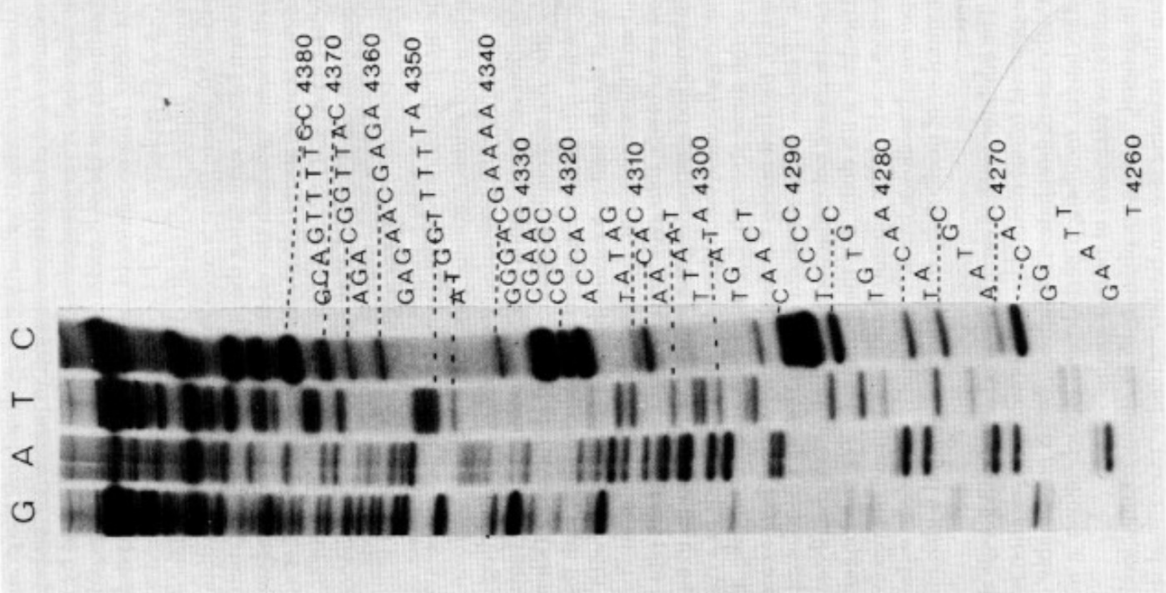
② Primer elongation and chain termination

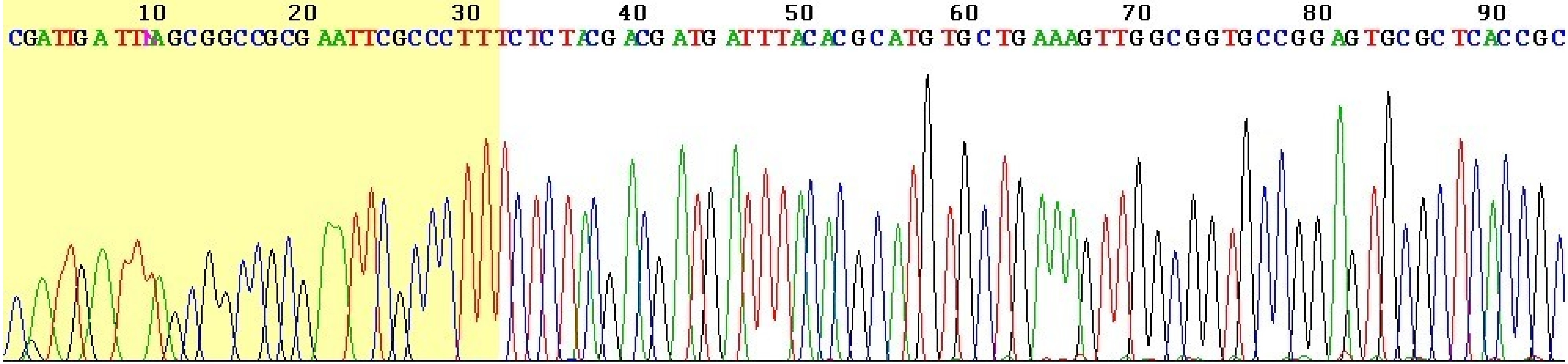
③ Capillary gel electrophoresis separation of DNA fragments

④ Laser detection of flourochromes and computational sequence analysis

https://en.wikipedia.org/wiki/Sanger_sequencing

Sanger sequencing





Problem: only short stretches (1000bp) of DNA are sequenced in any one run

The problem of **sequence assembly** can be compared to taking many copies of a book, passing each of them through a **shredder** with a different cutter, and piecing the text of the book back together just by looking at the shredded pieces.
Besides the obvious difficulty of this task, there are some **extra practical issues**: the original may have many repeated paragraphs, and some shreds may be modified during shredding to have typos. Excerpts from another book may also be added in, and some shreds may be completely unrecognisable.

https://en.wikipedia.org/wiki/Sequence_assembly

Main strategies

- **HGP**
  - create short (100-300kb) Bacterial Artificial Chromosomes (BAC) with overlapping DNA and amplify
  - Use repeated shotgun Sequencing: cut DNA up and sequence

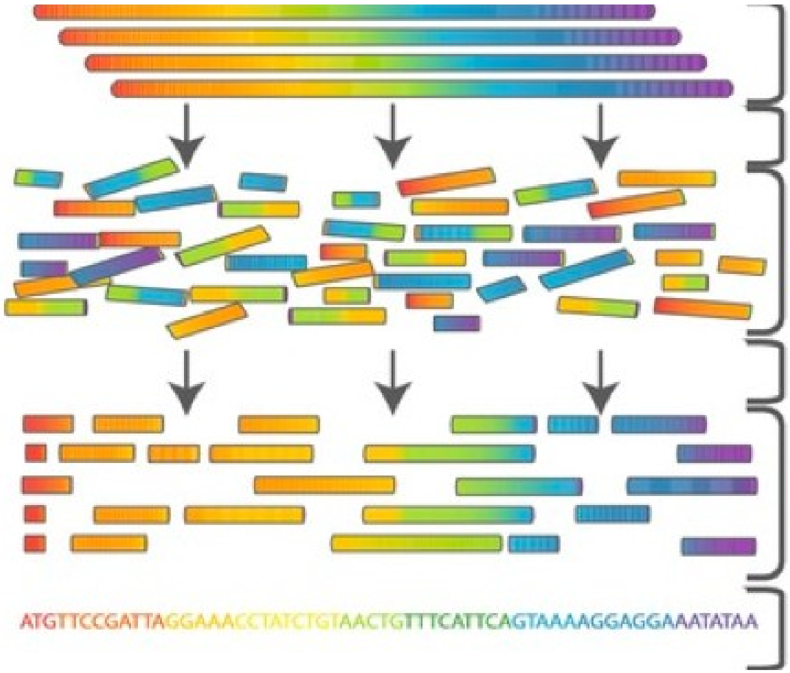
    ++ reliable - - slow - - expensive ($3bn)

- **Celerea/Venter**
  - blast whole genome into 2-10kb fragments and re-assemble in computer

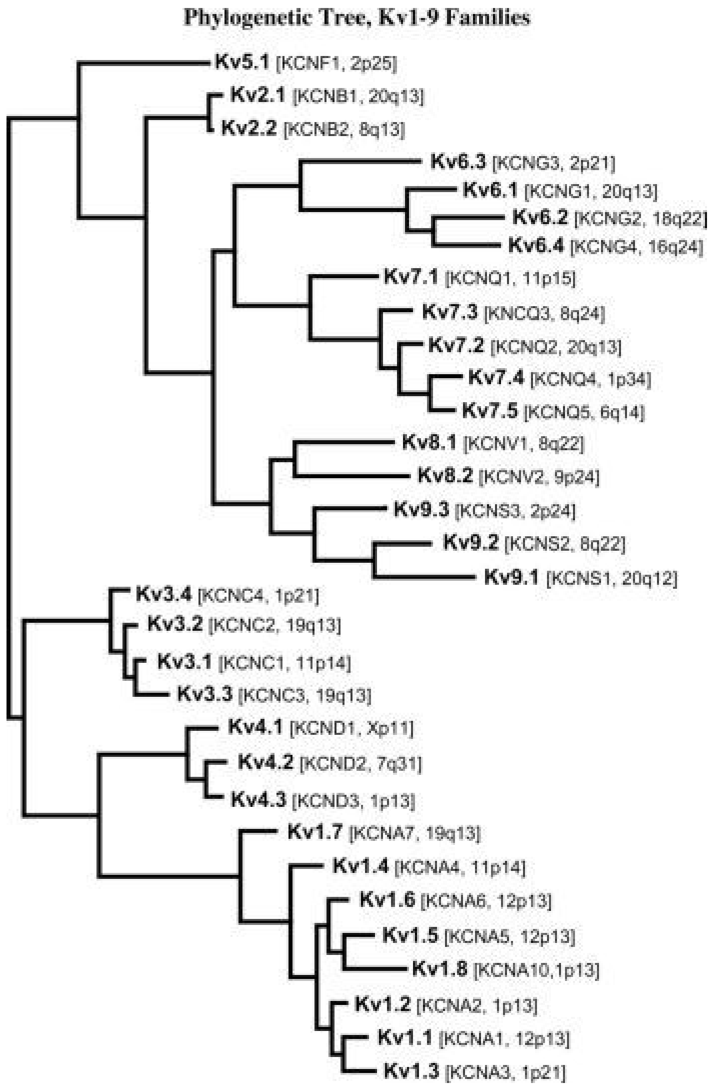
    - - error prone/requires extra data ++ cheap ($300Mio)

http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml

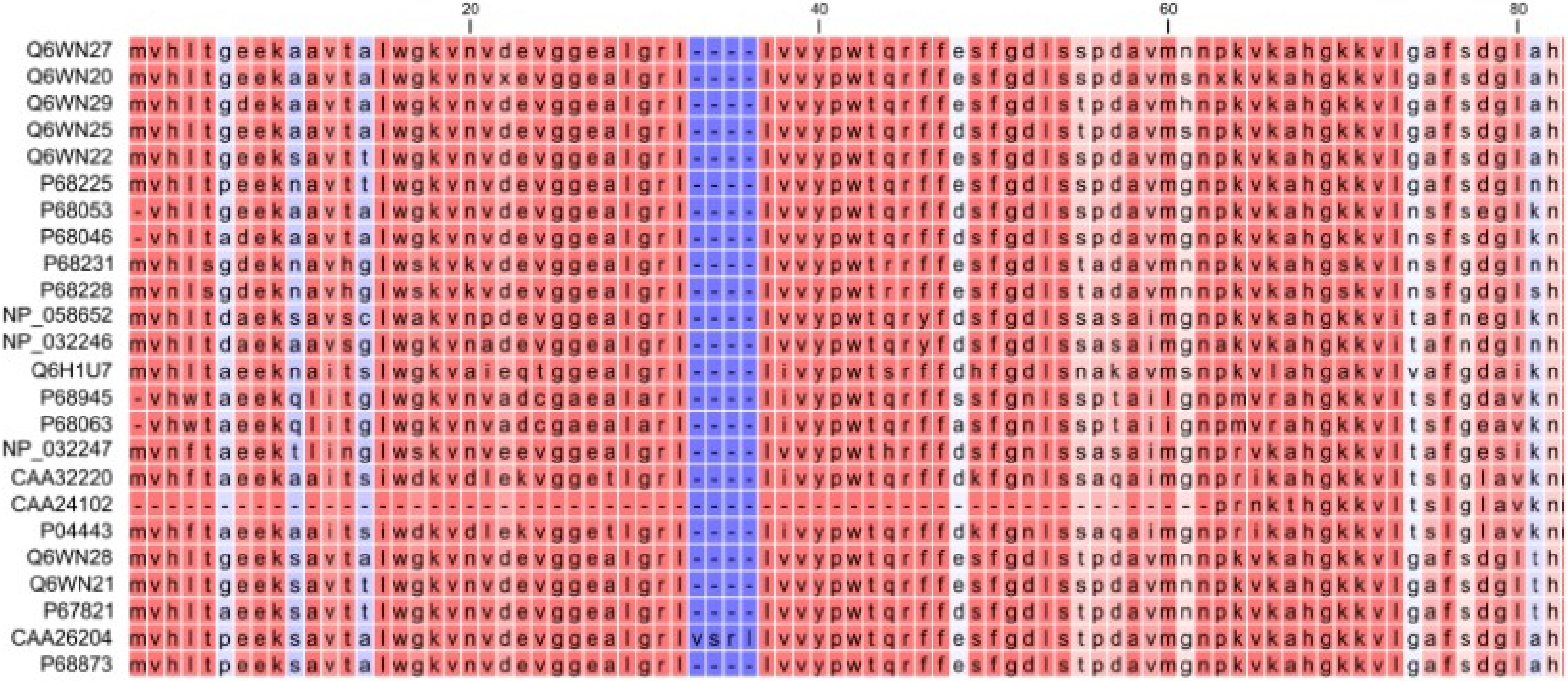
Sequence alignment



Sequence alignment


Phylogenetic Tree, Kv1-9 Families

Multiple sequence alignments



```
                            20                  40                  60                  80
                             |                   |                   |                   |
Q6WN27    mvhltgeekaavtalwgkvnvdevggealgrl----lvvypwtqrffesfgdlsspdavmnnpkvkahgkkvlgafsdglah
Q6WN20    mvhltgeekaavtalwgkvnvxevggealgrl----lvvypwtqrffesfgdlsspdavmsnxkvkahgkkvlgafsdglah
Q6WN29    mvhltgdekaavtalwgkvnvdevggealgrl----lvvypwtqrffesfgdlstpdavmhnpkvkahgkkvlgafsdglah
Q6WN25    mvhltgeekaavtalwgkvnvdevggealgrl----lvvypwtqrffdsfgdlstpdavmsnpkvkahgkkvlgafsdglah
Q6WN22    mvhltgeeksavttlwgkvnvdevggealgrl----lvvypwtqrffesfgdlsspdavmgnpkvkahgkkvlgafsdglah
P68225    mvhltpeeknavttlwgkvnvdevggealgrl----lvvypwtqrffesfgdlsspdavmgnpkvkahgkkvlgafsdglnh
P68053    -vhltgeekaavtalwgkvnvdevggealgrl----lvvypwtqrffdsfgdlsspdavmgnpkvkahgkkvlnsfseglkn
P68046    -vhltadekaavtalwgkvnvdevggealgrl----lvvypwtqrffdsfgdlsspdavmgnpkvkahgkkvlnsfsdglkn
P68231    mvhlsgdeknavhglwskvkvdevggealgrl----lvvypwtrrffesfgdlstadavmnnpkvkahgskvlnsfgdglnh
P68228    mvnlsgdeknavhglwskvkvdevggealgrl----lvvypwtrrffesfgdlstadavmnnpkvkahgskvlnsfgdglsh
NP_058652 mvhltdaeksavsclwakvnpdevggealgrl----lvvypwtqryfdsfgdlssasaimgnpkvkahgkkvitafneglkn
NP_032246 mvhltdaekaavsglwgkvnadevggealgrl----lvvypwtqryfdsfgdlssasaimgnakvkahgkkvitafndglnh
Q6H1U7    mvhltaeeknaitslwgkvaieqtggealgrl----livypwtsrffdhfgdlsnakavmsnpkvlahgakvlvafgdaikn
P68945    -vhwtaeekqlitglwgkvnvadcgaealarl----livypwtqrffssfgnlssptailgnpmvrahgkkvltsfgdavkn
P68063    -vhwtaeekqlitglwgkvnvadcgaealarl----livypwtqrffasfgnlssptaiignpmvrahgkkvltsfgeavkn
NP_032247 mvnftaeektlinglwskvnveevggealgrl----lvvypwthrffdsfgnlssasaimgnprvkahgkkvltafgesikn
CAA32220  mvhftaeekaaitsiwdkvdlekvggetlgrl----livypwtqrffdkfgnlssaqaimgnprikahgkkvltslglavkn
CAA24102  --------------------------------------------------------------prnkthgkkvltslglavkn
P04443    mvhftaeekaaitsiwdkvdlekvggetlgrl----livypwtqrffdkfgnlssaqaimgnprikahgkkvltslglavkn
Q6WN28    mvhltgeeksavtalwgkvnvdevggealgrl----lvvypwtqrffesfgdlstpdavmnnpkvkahgkkvlgafsdglth
Q6WN21    mvhltgeeksavttlwgkvnvdevggealgrl----lvvypwtqrffesfgdlsspdavmnnpkvkahgkkvlgafsdglth
P67821    mvhltaeeksavttlwgkvnvdevggealgrl----lvvypwtqrffdsfgdlstpdavmnnpkvkahgkkvlgafsdglth
CAA26204  mvhltpeeksavtalwgkvnvdevggealgrlvsrllvvypwtqrffesfgdlstpdavmgnpkvkahgkkvlgafsdglah
P68873    mvhltpeeksavtalwgkvnvdevggealgrl----lvvypwtqrffesfgdlstpdavmgnpkvkahgkkvlgafsdglah
```

24 Hemoglobin protein sequences

Interpretation of alignments

- **Sequence identity**: exactly the same sequence
- **Similarity**: sufficiently high score, but scoring should have some biological relevance
- **Homology**: two sequences have a common ancestor

“Homology is the central concept for all of biology. Whenever we say that a mammalian hormone is the ‘same’ hormone as a fish hormone, that a human gene sequence is the ‘same’ as a sequence in a chimp or a mouse, that a HOX gene is the ‘same’ in a mouse, a fruit fly, a frog, and a human - even when we argue that discoveries about a worm, a fruit fly, a frog, a mouse, or a chimp have relevance to the human condition - we have made a bold and direct statement about homology. The aggressive confidence of modern biomedical science implies that we know what we are talking about. But a deeper reflection shows that this confidence is based more on hope than on certainty.”

Wake, Comparative terminology. Science 1994, 265:268-269

Graphical sequence comparison (preparation)

Programmatic access to sequences:

```
from Bio import Entrez

gene_id = "some_id"

record = Entrez.efetch(db="nucleotide", id=gene_id1, rettype="gb",
retmode="text")
```

The `record` object has a number of attributes:

```
    – id          – Identifier such as a locus tag (string)

    – seq         – The sequence itself (Seq object or similar)

    – name        – Sequence name, e.g. gene name (string)

    – description – Additional text (string)

    – dbxrefs     – List of database cross references

    – features    – Any (sub)features defined

    – annotations – Further information about the whole sequence.

      Most entries are strings, or lists of strings.
```

Graphical sequence comparison

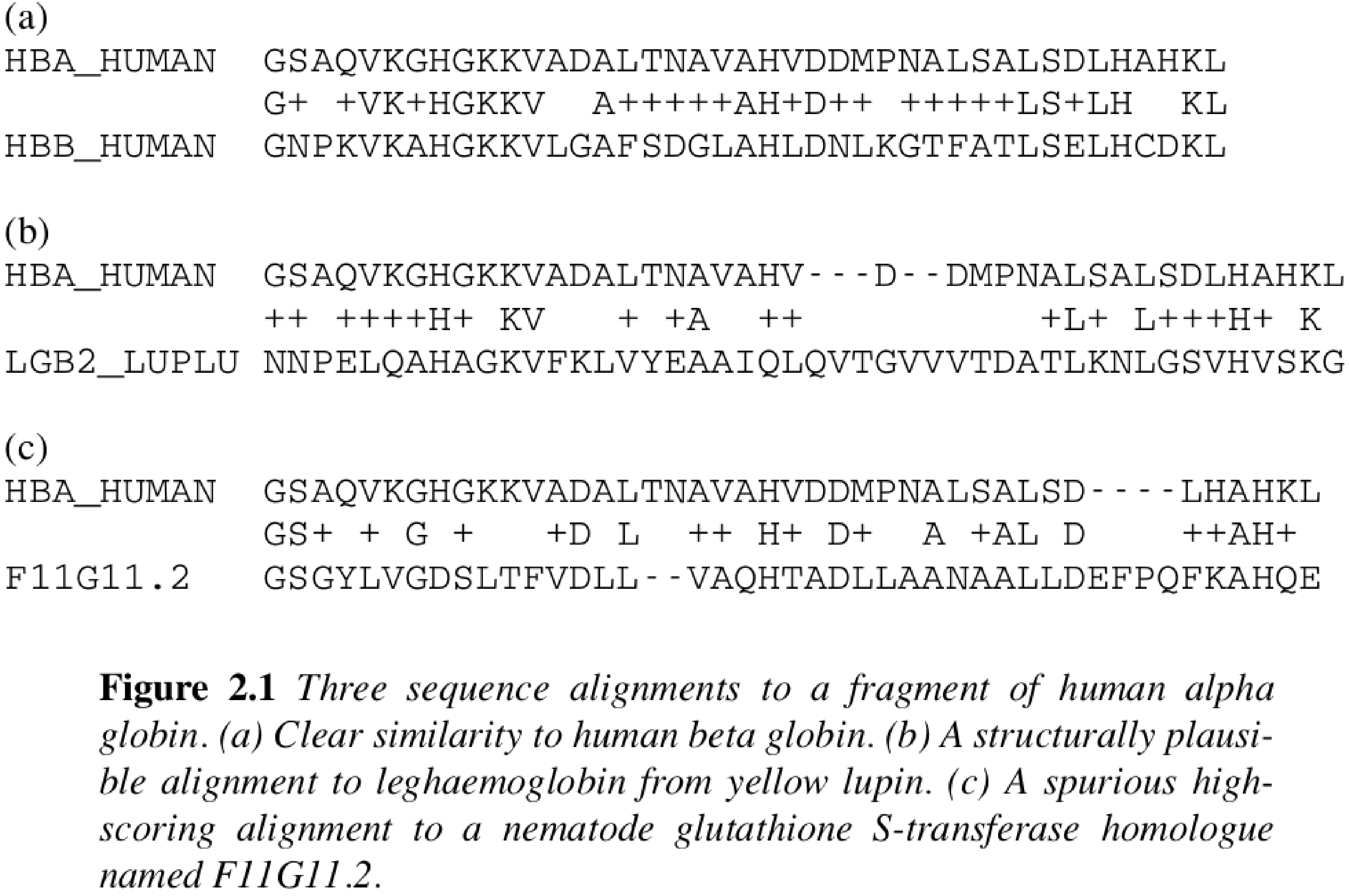
- Find a pair of genes to compare. You can choose two species, or two different genes. Get their **NCBI IDs**.
- Insert these into cell #2 in this Jupyter notebook:

https://www.inf.ed.ac.uk/teaching/courses/bio1/lectures17/Bio1Lecture3GraphAlign.ipynb

HTML version:

https://www.inf.ed.ac.uk/teaching/courses/bio1/lectures17/Bio1Lecture3GraphAlign.html

Alignment scoring



```
(a)
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
           G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN  GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

(b)
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
           ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
LGB2_LUPLU NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

(c)
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
           GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
F11G11.2   GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

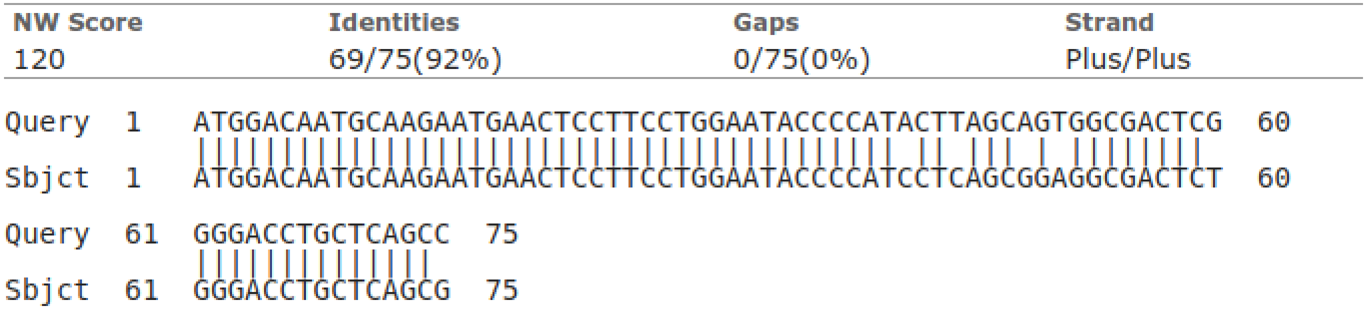
Figure 2.1 *Three sequence alignments to a fragment of human alpha globin. (a) Clear similarity to human beta globin. (b) A structurally plausible alignment to leghaemoglobin from yellow lupin. (c) A spurious high-scoring alignment to a nematode glutathione S-transferase homologue named F11G11.2.*

Alignment scoring

- Match: identical residues
- Substitution: different residues
- Insert/deletion (indel, gap): a gap in one of the two sequences
- Sequence alignment attempts to identify sequences originating from a common ancestor
- The functional implications of each sequence difference is important
- Appropriate scoring schemes take this into account (more on this later in the course)

Alignment scoring

- Example scoring:
  - Match +2
  - Substitution -3
  - Gap -4



NW Score	Identities	Gaps	Strand
120	69/75(92%)	0/75(0%)	Plus/Plus

```
Query  1    ATGGACAATGCAAGAATGAACTCCTTCCTGGAATACCCCATACTTAGCAGTGGCGACTCG  60
            ||||||||||||||||||||||||||||||||||||||||| || ||| | |||||||| 
Sbjct  1    ATGGACAATGCAAGAATGAACTCCTTCCTGGAATACCCCATCCTCAGCGGAGGCGACTCT  60

Query  61   GGGACCTGCTCAGCC  75
            ||||||||||||||
Sbjct  61   GGGACCTGCTCAGCG  75
```

Homo sapiens homeobox A1 (HOXA1), transcript variant 1, mRNA.

vs.

Elephantulus edwardii homeobox A1 (HOXA1), mRNA (Cape elephant shrew)

Optimal Sequence Alignment

- Small sequences can be compared exactly using dynamic programming.

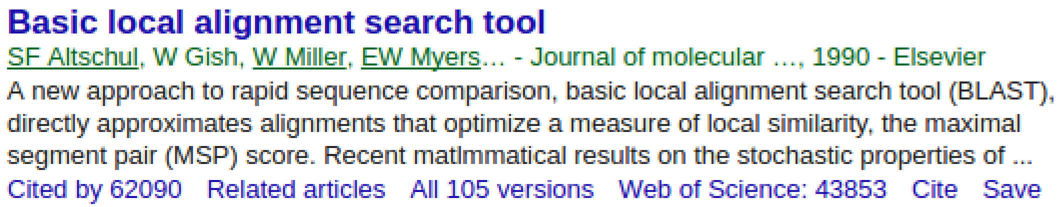
  → more on this next week

- When searching multiple genomes, exact methods are computationally too expensive.

Solution: heuristic methods

Basic **L**ocal **A**lignment **S**earch **T**ool

- Compares a query sequence with a full sequence database
- Avoids full alignments through heuristics
- **Main assumption**: 'true' matches have at least some short high-scoring stretches
- Developed by Altschul, Gish, Miller and Myers at NIH in 1980s



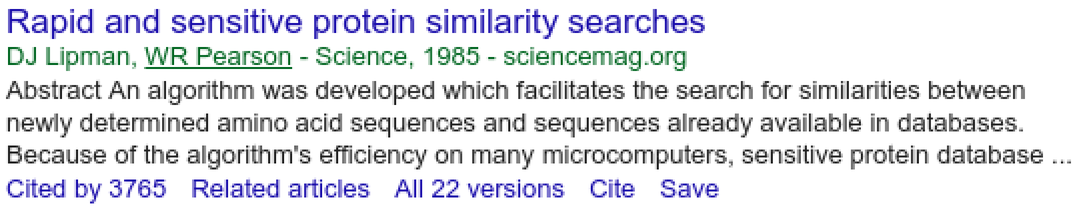
Basic local alignment search tool

SF Altschul, W Gish, W Miller, EW Myers… - Journal of molecular …, 1990 - Elsevier

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent matlmmatical results on the stochastic properties of ...

Cited by 62090 Related articles All 105 versions Web of Science: 43853 Cite Save

- Based on FASTA, which combined DP alignment with heuristics



Rapid and sensitive protein similarity searches

DJ Lipman, WR Pearson - Science, 1985 - sciencemag.org

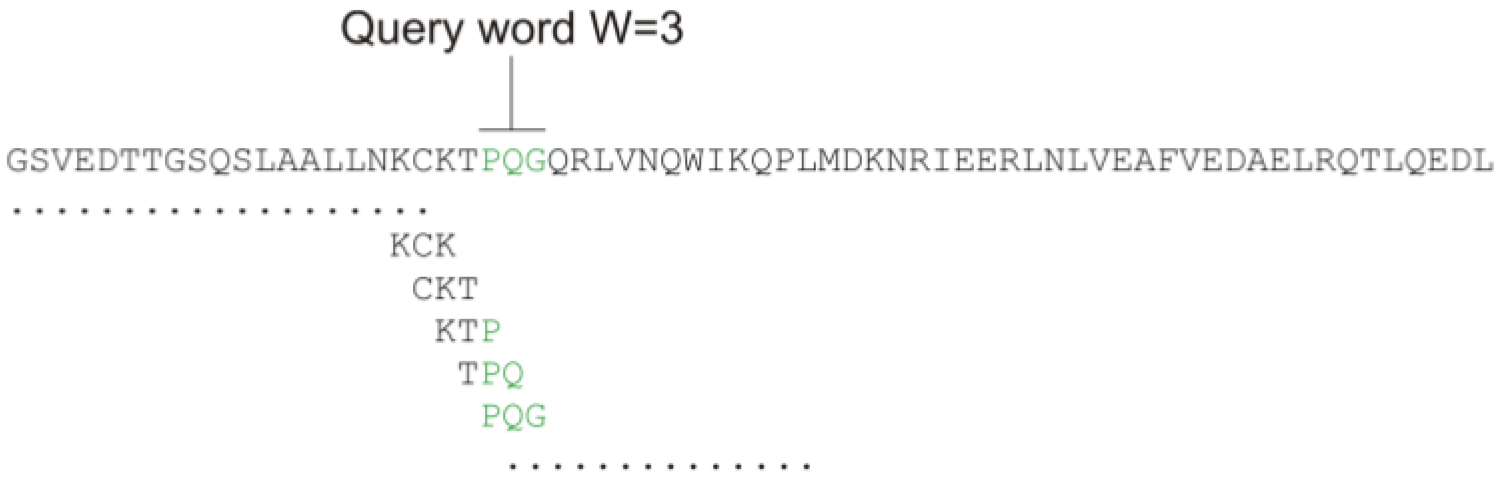
Abstract An algorithm was developed which facilitates the search for similarities between newly determined amino acid sequences and sequences already available in databases. Because of the algorithm's efficiency on many microcomputers, sensitive protein database ...

Cited by 3765 Related articles All 22 versions Cite Save

Strategy

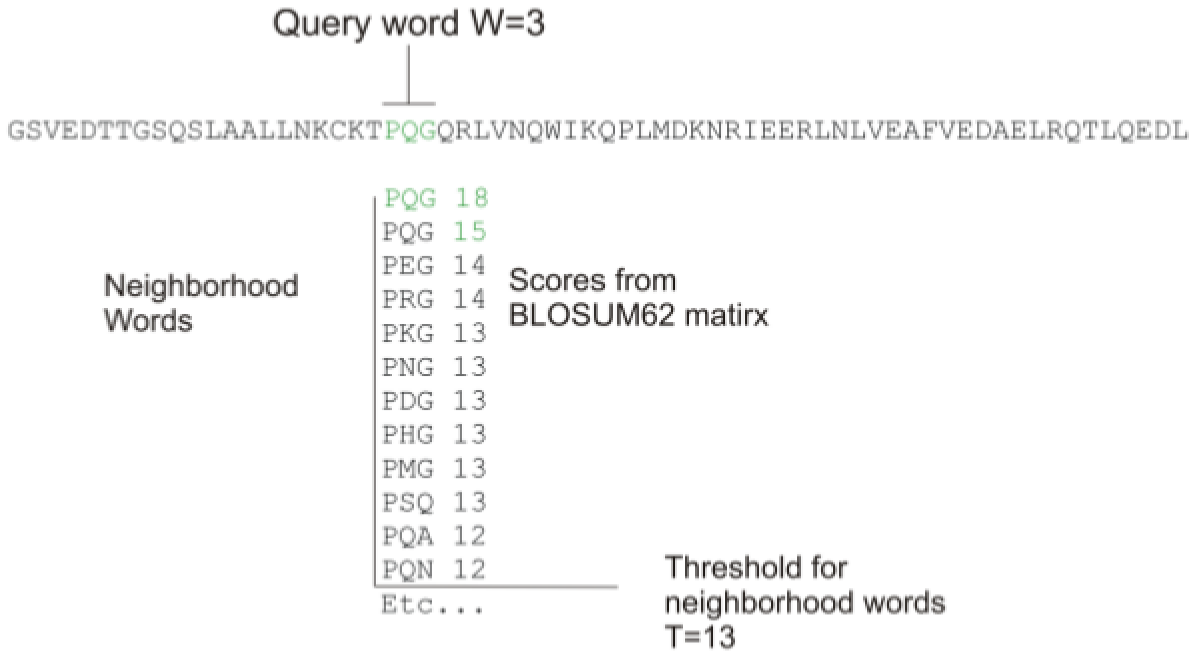
- Remove redundant/low complexity regions from query (SEG or DUST)
- Create set of short words
- Scan database for matching words (seeds)
- Matching words are extended (gap-free), similar to Smith-Waterman
- Extensions below specified threshold score are removed
- Remaining extensions are completed also scoring indels

Step 1: creating short words


Query word W=3
GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
....
KCK
CKT
KTP
TPQ
PQG
....

(n=3 for amino acids, n=11 for nucleotides)

Step 1.1: create neighbourhood words



Query word W=3

GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

Neighborhood Words

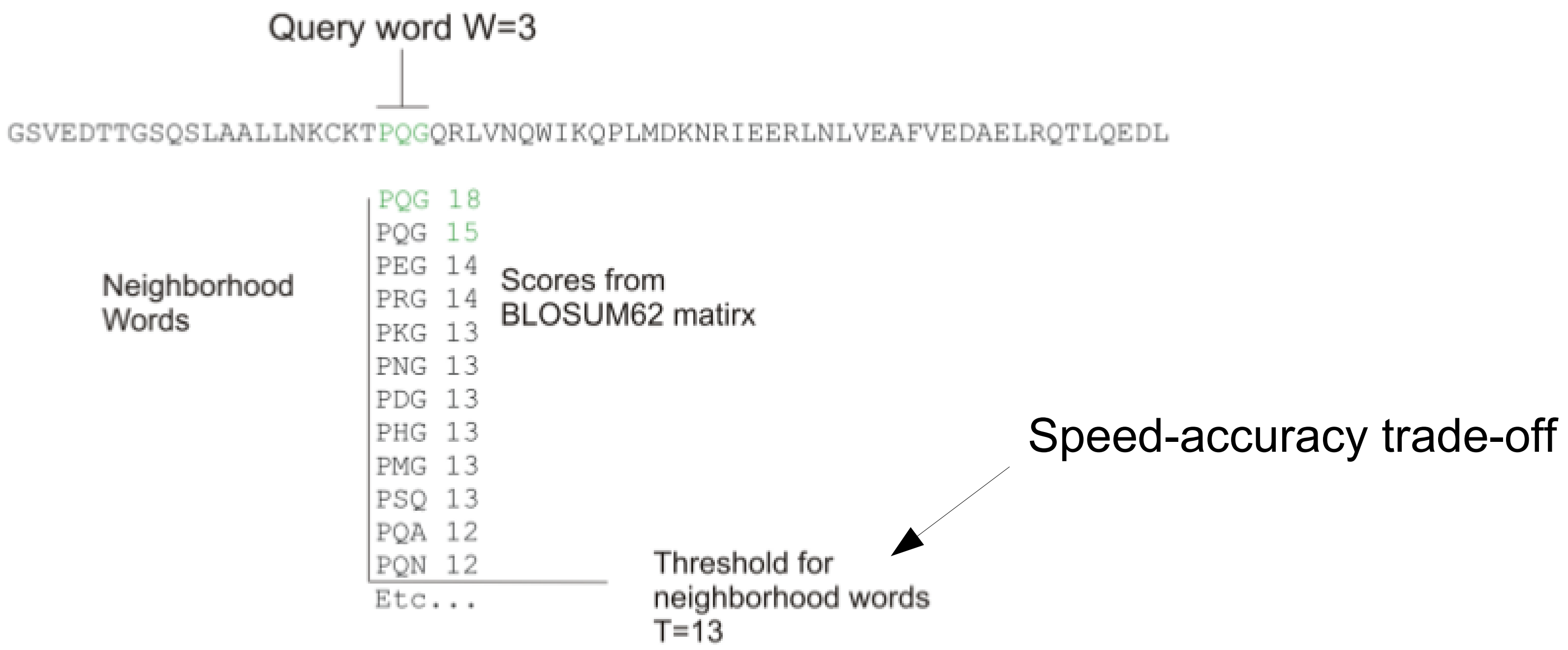
PQG	18
PQG	15
PEG	14
PRG	14
PKG	13
PNG	13
PDG	13
PHG	13
PMG	13
PSQ	13
PQA	12
PQN	12

Etc...

Scores from BLOSUM62 matirx

Threshold for neighborhood words T=13

Step 2: search words in database



Query word W=3

GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

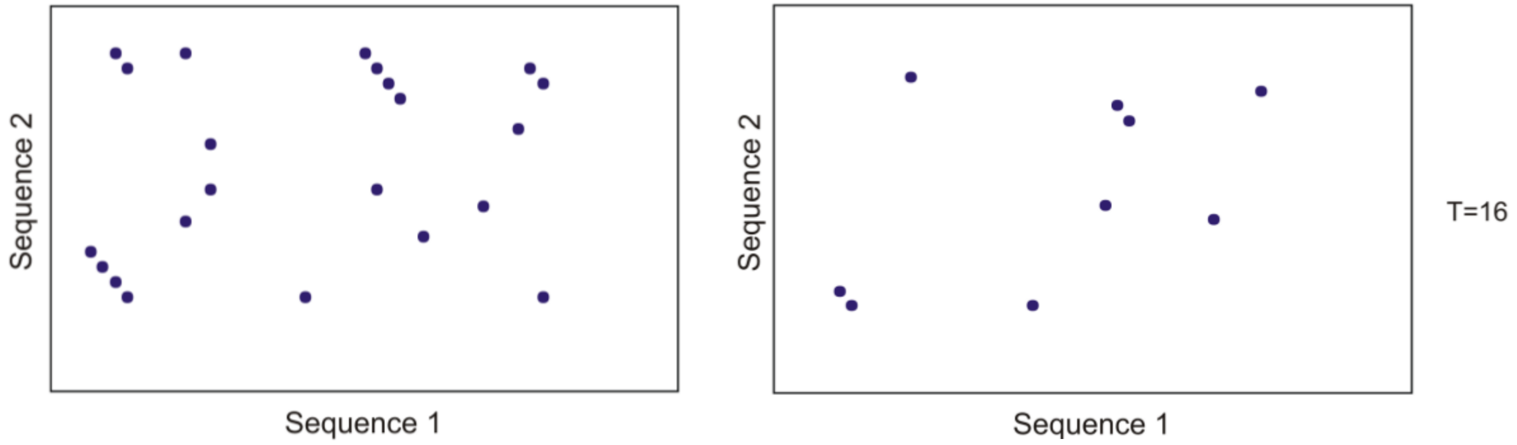
Neighborhood Words

PQG	18
PQG	15
PEG	14
PRG	14
PKG	13
PNG	13
PDG	13
PHG	13
PMG	13
PSQ	13
PQA	12
PQN	12
Etc...	

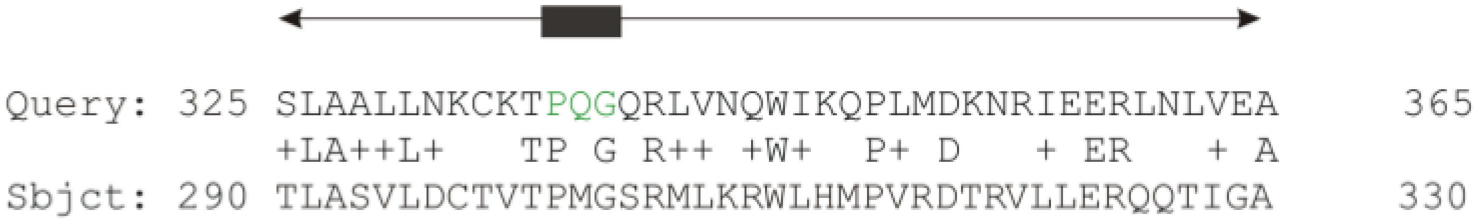
Scores from BLOSUM62 matirx

Speed-accuracy trade-off

Threshold for neighborhood words T=13



Step 3: seed extension



```
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA     365
           +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA     330
```

Scoring without indels, extend until threshold reached.

Step 4: thresholding and scoring

- Compare scores with random surrogates of same length, select a significance threshold
- Keep Maximal Scoring Pairs
- Compute significance through comparison with random sequence (using Gumbel distribution that models number of extreme samples in a distribution)
- Output:
  - Raw scores
  - Bit score (sometimes)
  - E value

Scoring

- Bit score
  a scaled version of alignment score S

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

  λ and K are parameters of the Grumbel distribution, which describes score probabilities for S-W Algorithm.

- E-value

$$E = mn2^{-S'}$$

  n query size, m database size

  Yields expected number chance alignments with score S’ or better = estimate of false positives.

Uses

- Identify species or homologous species in new data
- Search for protein domains
- Phylogeny
- Map chromosome locations
- Map annotations between species

BLAST flavours

- MegaBLAST: similar N-N searches
- BLASTN: more dissimilar N-N
- BLASTP: P-P comparison
- BLASTX: N-P comparison (translation done on the fly)
- BLASTN: P-N comparison (back-translation done on the fly)

NCBI BLAST

- Use this code to run a BLAST search:

https://www.inf.ed.ac.uk/teaching/courses/bio1/lectures17/Bio1Lecture3BLAST.ipynb

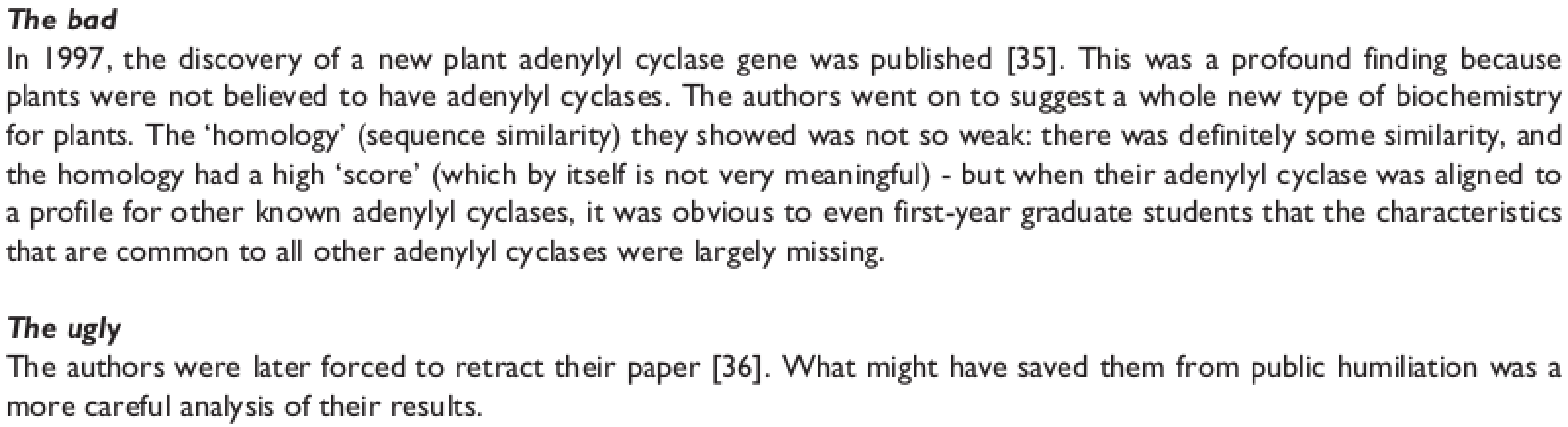
https://www.inf.ed.ac.uk/teaching/courses/bio1/lectures17/Bio1Lecture3BLAST.html

- Try manipulating the query sequence, e.g. make indels, reverse the sequence, or insert a piece of different sequence.
- The notebook runs a nucleotide BLAST. Try to modify it to run a protein BLAST.
- Inspect to what extent the E-value reflects biologically plausibility.

Summary

Exact alignment algorithms are too slow to deal with big genomic data search problems.

Algorithms combined with heuristics are highly efficient, but less precise.



The bad

In 1997, the discovery of a new plant adenylyl cyclase gene was published [35]. This was a profound finding because plants were not believed to have adenylyl cyclases. The authors went on to suggest a whole new type of biochemistry for plants. The 'homology' (sequence similarity) they showed was not so weak: there was definitely some similarity, and the homology had a high 'score' (which by itself is not very meaningful) - but when their adenylyl cyclase was aligned to a profile for other known adenylyl cyclases, it was obvious to even first-year graduate students that the characteristics that are common to all other adenylyl cyclases were largely missing.

The ugly

The authors were later forced to retract their paper [36]. What might have saved them from public humiliation was a more careful analysis of their results.

Pertsemlidis, Alexander, John W. Fondon, and W. John. "Having a BLAST with bioinformatics (and avoiding BLASTphemy)." Genome Biol 2.10 (2001).

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC138974/