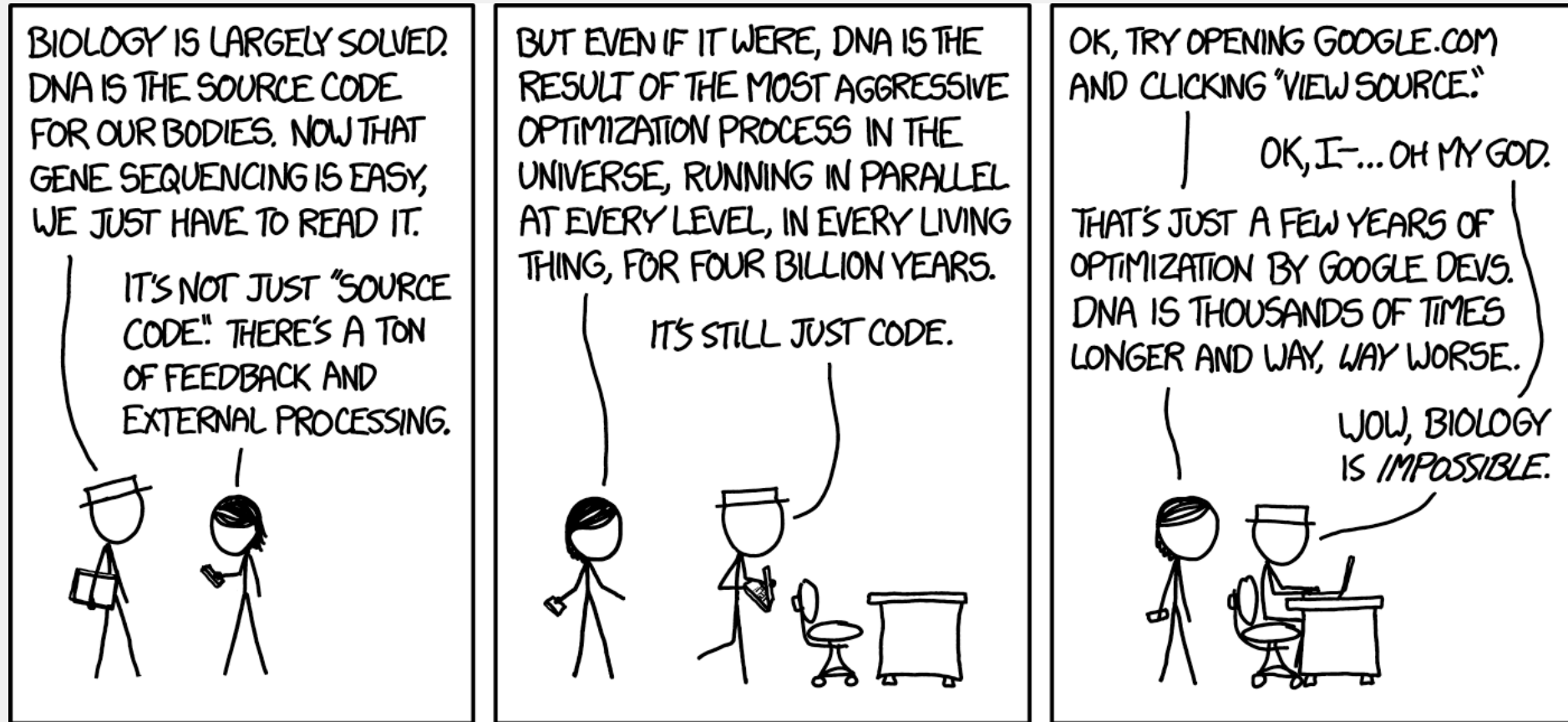


Sequence analysis

Bioinformatics · Universitat Politècnica de Catalunya

Alexandre Perera-Lluna



xkcd 1605 "DNA" — Randall Munroe, xkcd.com/1605, CC BY-NC 2.5.

- Sequence properties and profiles
- Homology and pairwise alignment
- Scoring: substitution matrices and gaps
- Optimal alignment by dynamic programming: global and local
- Statistical significance, BLAST and FASTA
- Multiple sequence alignment

Prediction of biological and structural properties from sequences is a key goal in computational biology.

- Recognise transmembrane proteins from sequence alone (about 25% of all proteins).
- Classify proteins into classes of similar function from sequence similarity.
- These proteins mediate communication and transport between cells, and are central to recent pharmaceutical developments.

Per-Sequence Analysis

- Limited information.
- Analyzed solely from sequence information (residues).
- Characteristics of residues
 - Physical, chemical: hydrophobicity (amino acids)
 - Statistical: GC islands (DNA); codon usage

Between Sequences

- More informative.
- Mainly based on homology.
- Common ancestry and sequence similarity.
- Homology indicates common ancestry. Shared function requires additional evidence.
- The base of homology recognition is the similarity concept between two sequences.
- Genes come in families -> multiple sequence analysis.

Amino acid Features

- Size.
- Flexibility in the backbone.
- Charge.
- ...

Argos

Argos (1987). A sensitive Procedure to Compare Amino acid Sequences. J. Mol. 193:385-396

- Hydrophobicity.
- Turn Preference.
- Residue Bulk or Shape.
- Refractivity Index.
- Antiparallel strand preference.

Most computed in a sliding window approach.

Features on individual amino acids play some role in protein early secondary structure formation

- Glutamate is frequently found in alpha helices
- Valine has a preference for beta strands.
- Proline is strongly avoided in helices



Hydrophobicity

Physico-chemical property

Kyte-Doolittle

Widely applied scale for delineating hydrophobic character of a protein. with values above 0 are hydrophobic in character.

Hopp-Woods

Designed for predicting potentially antigenic regions.

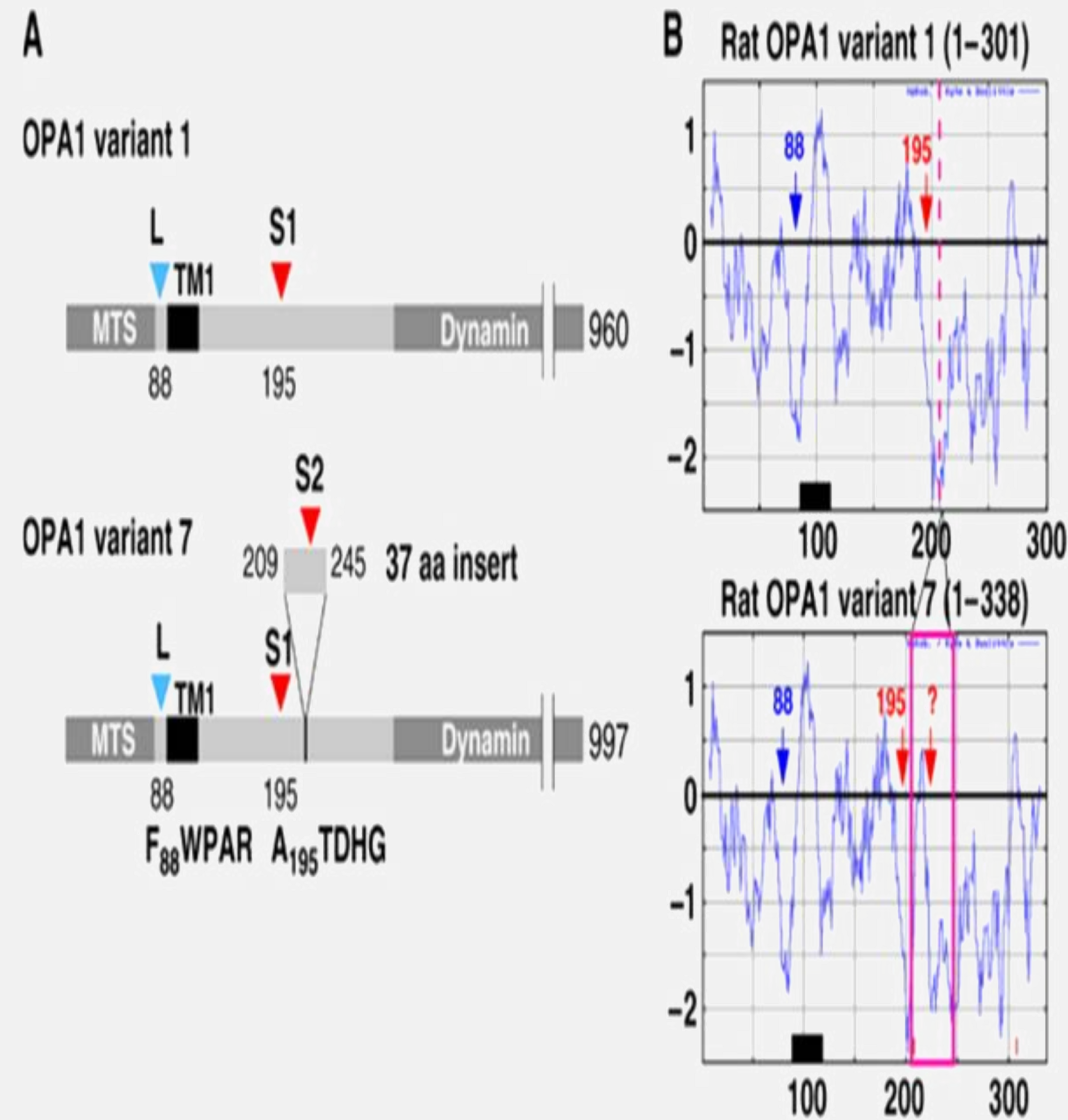
Values greater than 0 are hydrophilic and thus likely to be exposed on the surface of a folded protein.

Both are computed on a sliding window, like the Argos profiles above.

Hydrophobicity: the two scales side by side

	Kyte-Doolittle	Hopp-Woods		Kyte-Doolittle	Hopp-Woods
Alanine	1.8	-0.5	Leucine	3.8	-1.8
Arginine	-4.5	3.0	Lysine	-3.9	3.0
Asparagine	-3.5	0.2	Methionine	1.9	-1.3
Aspartic acid	-3.5	3.0	Phenylalanine	2.8	-2.5
Cysteine	2.5	-1.0	Proline	-1.6	0.0
Glutamine	-3.5	0.2	Serine	-0.8	0.3
Glutamic acid	-3.5	3.0	Threonine	-0.7	-0.4
Glycine	-0.4	0.0	Tryptophan	-0.9	-3.4
Histidine	-3.2	-0.5	Tyrosine	-1.3	-2.3
Isoleucine	4.5	-1.8	Valine	4.2	-1.5

Values from Black, S.D. and Mould, D.R. (1991) Anal. Biochem. 193, 72–82. Sign conventions differ: Kyte-Doolittle > 0 marks hydrophobic residues; Hopp-Woods > 0 marks hydrophilic, exposed ones.



Rat OPA1 windowed hydropathy: TM1 ($\approx$ residue 100) is the only clear hydrophobic peak; variant 7 adds a 37-aa hydrophilic insert.

Evolution suggests

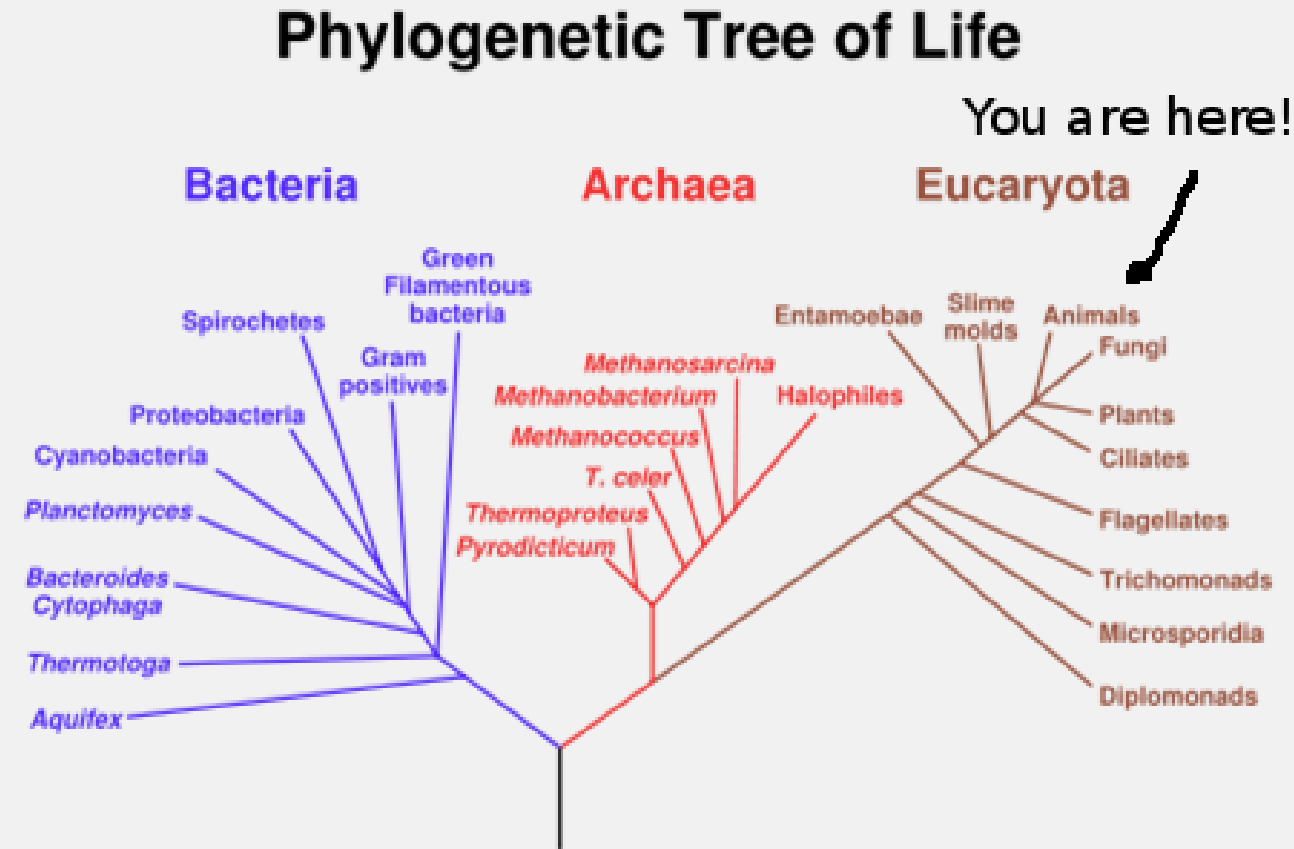
- All gene sequences may have derived from common ancestral sequences.
- Trace the evolutionary history of mutations will help us to understand functionality and structure

Comparing two sequences

- Which is this mutation path, using:
 - Insertions.
 - Deletions.
 - Mutations.
- ... to turn a sequence into another?

Distance between sequences.

- May provide information whether sequence contains protein-coding gene.
- Similarity on sequence may hint first clues on the type of protein the gene codes.
- Protein Structure Prediction.
- Compare DNA information between organisms -> Phylogenetic tree.



Based on 16S rRNA genetic sequences. Woese C, Fox G (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms". Proc Natl Acad Sci USA 74 (11): pp. 5088–90

Homology

Similarity between characters that is due to their shared ancestry — a qualitative relationship, not a percentage.

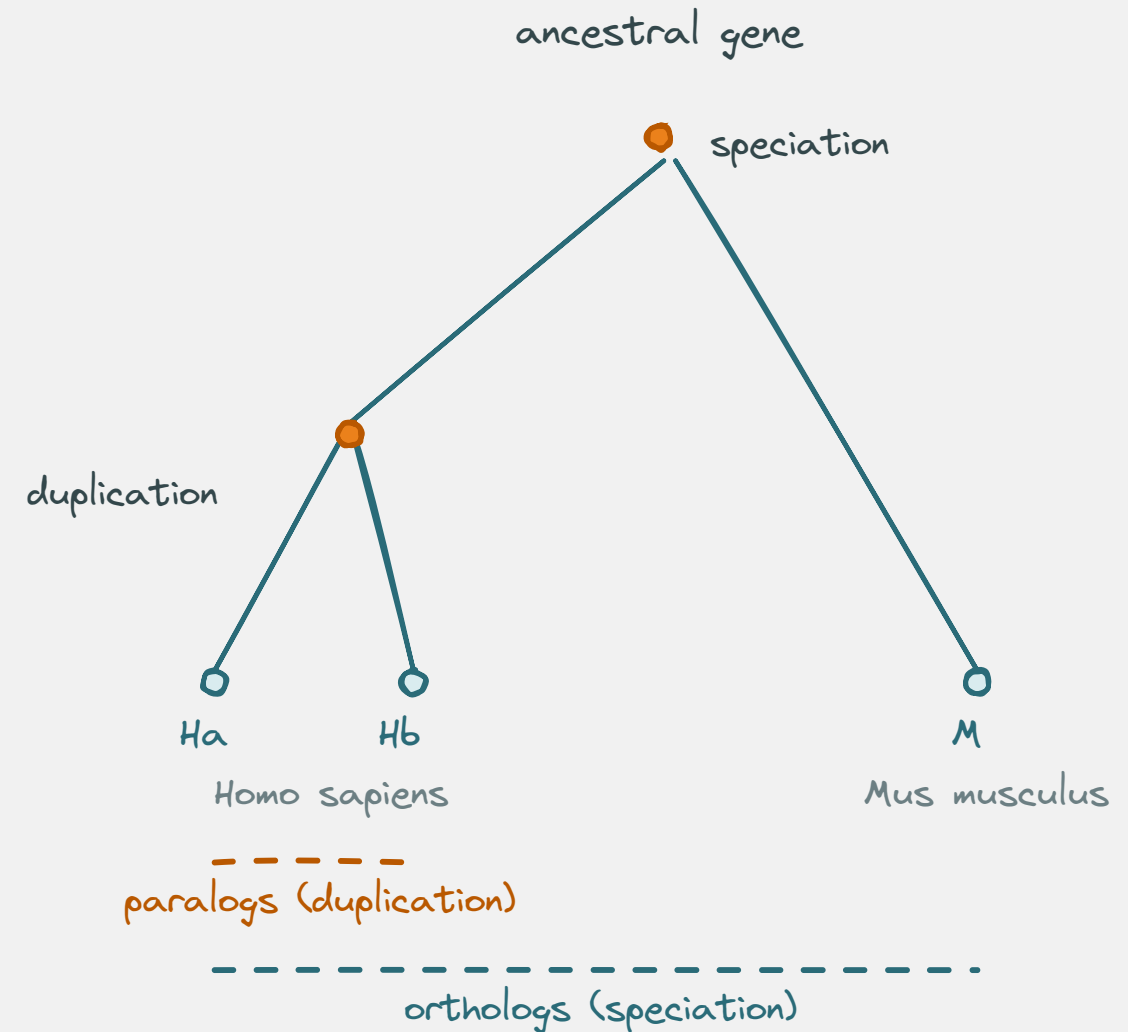
Orthology

Genes in different species whose last shared event is **speciation** (orthologs).

Paralogy

Genes whose last shared event is a **gene duplication** (paralogs): a duplicated gene occupying a second position in the same genome.

Paralogs often keep similar functions; freed from the original selective pressure, one copy may mutate and acquire a new function.



- Alignment places residues into corresponding positions under a scoring model.
- Identical residues can arise through common ancestry or by chance.
- Statistically significant similarity can provide evidence for homology.



Convergent Evolution

Sequence alignment is:

A task of locating equivalent regions of two or more sequences to maximize their similarity

T	H	I	S	S	E	Q	U	E	N	C	E
T	H	A	T	S	E	Q	U	E	N	C	E

Better on AA or NN sequences?

Is easier to detect homology when comparing protein sequences than when comparing nucleic acid sequences.

1) Just 4 symbols in DNA, 20 in AA chain

Given the same sequence length,

- AAs sequence provides with more information
- Similarity by chance is much more probable in Nucleotide sequences (for the same sequence length).

The Genetic Code is Redundant

There are two or more different codons for most amino acids.

AAA: K (Lys) AAG: K (Lys) AAT: N (Asn) AAC: N (Asn)	GAA: E (Glu) GAG: E (Glu) GAT: D (Asp) GAC: D (Asp)	TAA: Stop TAG: Stop TAT: Y (Tyr) TAC: Y (Tyr)	CAA: Q (Gln) CAG: Q (Gln) CAT: H (His) CAC: H (His)
AGA: R (Arg) AGG: R (Arg) AGT: S (Ser) AGC: S (Ser)	GGA: G (Gly) GGG: G (Gly) GGT: G (Gly) GGC: G (Gly)	TGA: Stop TGG: W (Trp) TGT: C (Cys) TGC: C (Cys)	CGA: R (Arg) CGG: R (Arg) CGT: R (Arg) CGC: R (Arg)
ATA: I (Ile) ATG: M (Met) ATG = Start ATT: I (Ile) ATC: I (Ile)	GTA: V (Val) GTG: V (Val) GTT: V (Val) GTC: V (Val)	TTA: L (Leu) TTG: L (Leu) TTT: F (Phe) TTC: F (Phe)	CTA: L (Leu) CTG: L (Leu) CTT: L (Leu) CTC: L (Leu)
ACA: T (Thr) ACG: T (Thr) ACT: T (Thr) ACC: T (Thr)	GCA: A (Ala) GCG: A (Ala) GCT: A (Ala) GCC: A (Ala)	TCA: S (Ser) TCG: S (Ser) TCT: S (ser) TCC: S (Ser)	CCA: P (Pro) CCG: P (Pro) CCT: P (Pro) CCC: P (Pro)

Local and global similarity

Local similarity: similar regions within sequences that may differ elsewhere.

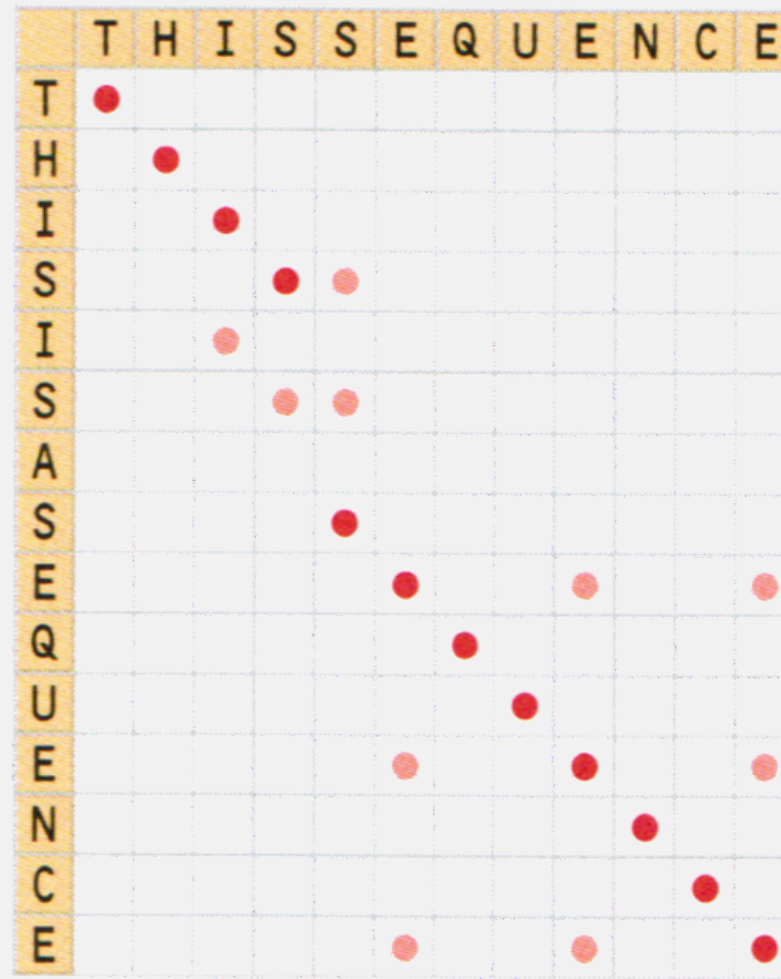
Global similarity: similarity across the full lengths of the sequences.

Why use a dot plot?

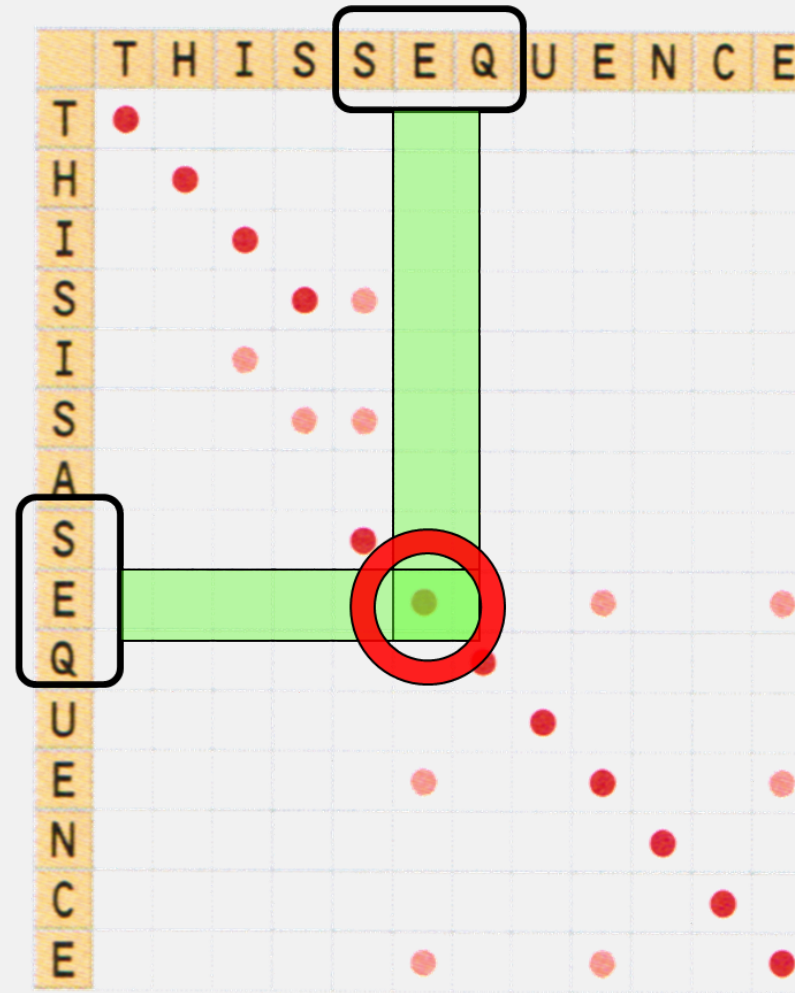
- Visualise similarities without first choosing a local or global alignment.
- Identify matching regions between two sequences.
- Detect repeats within a sequence.

Global	T	F	T	T	A	L	I	L	L	A	V	A	V	
	F	-	-	T	A	L	-	L	L	A	-	A	V	
Local	T	F	T	T	A	L	I	L	L	-	A	V	A	V
	-	-	F	T	A	L	-	L	L	A	A	V	-	-

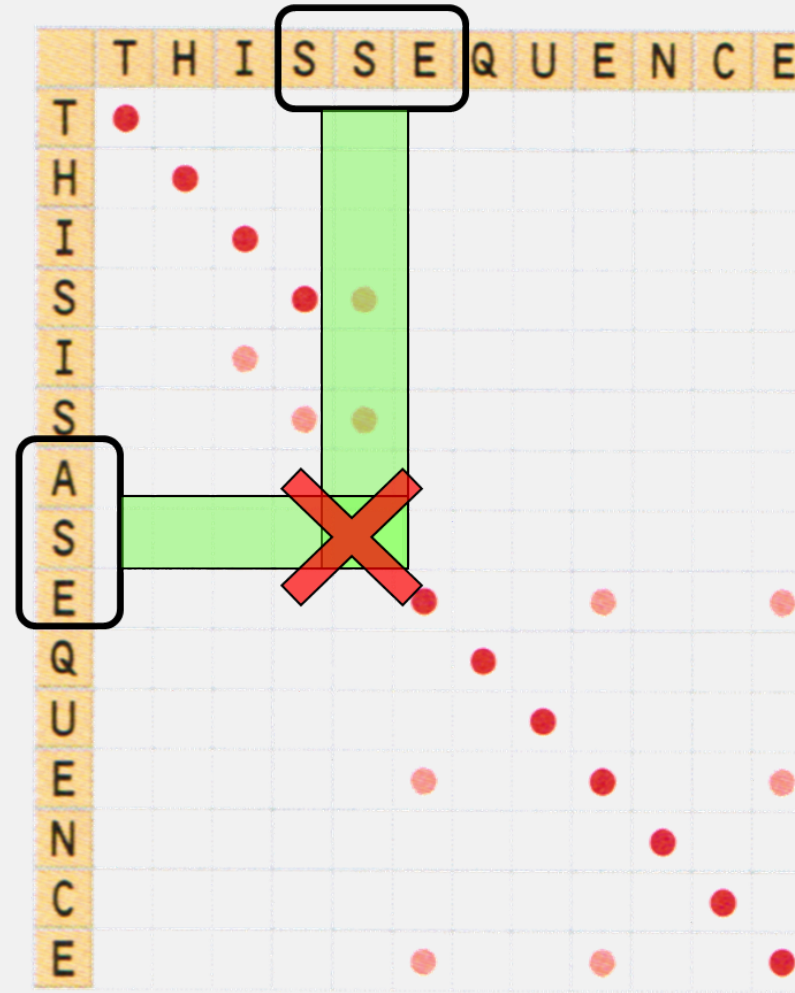
The same pair, aligned globally (upper rows: gaps added where the ends are forced together) and locally (lower rows: only the best-scoring stretch is kept, gaps at the trimmed ends).



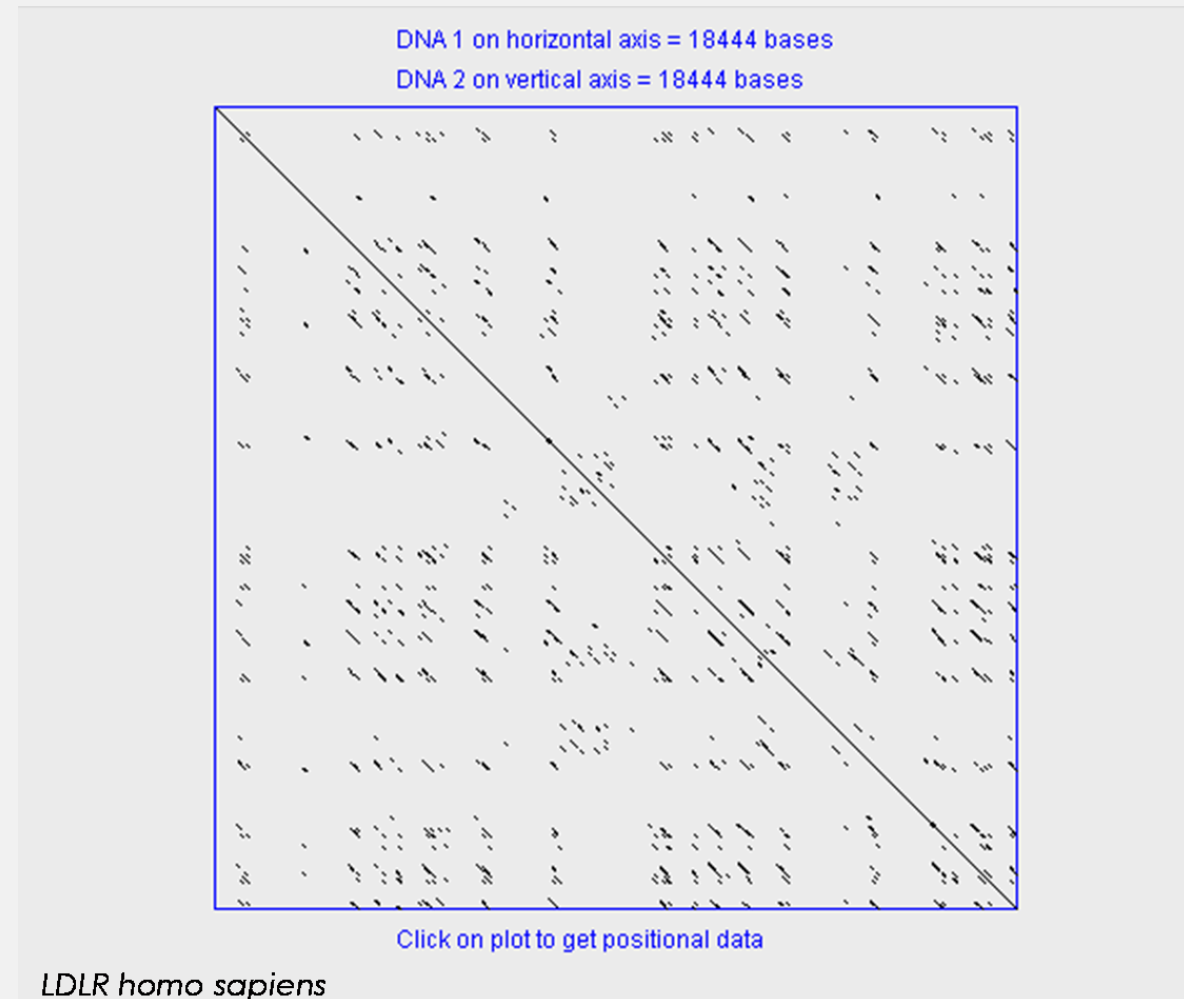
Raw single-residue matches: the diagonal shows through, but isolated chance dots bury the pattern. Sequences here: THISISASEQUENCE vs THISSEQUENCE.



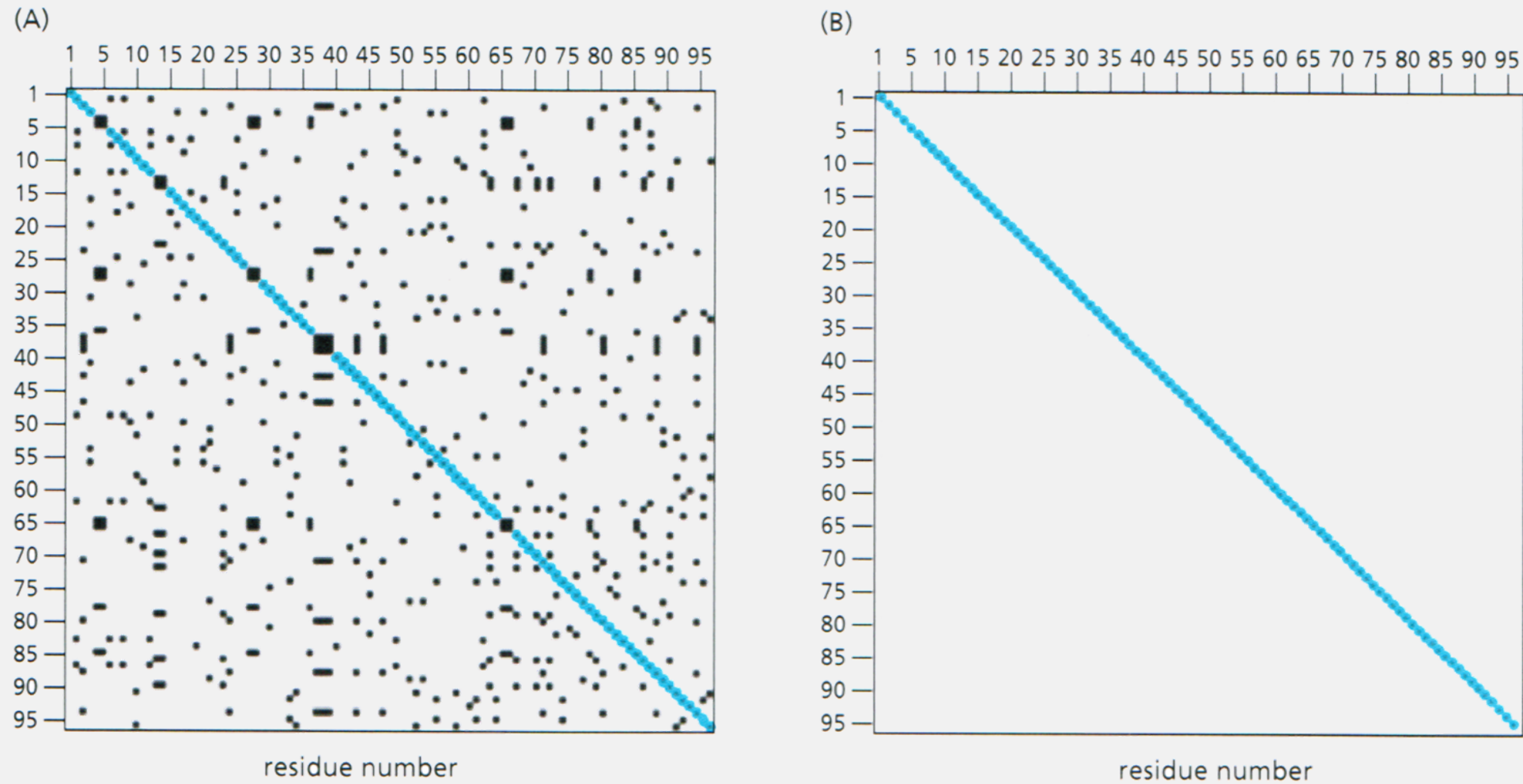
With a 3-residue window, matches are judged over whole windows: the SEQ / SEQ match is confirmed (green) and isolated dots no longer count.



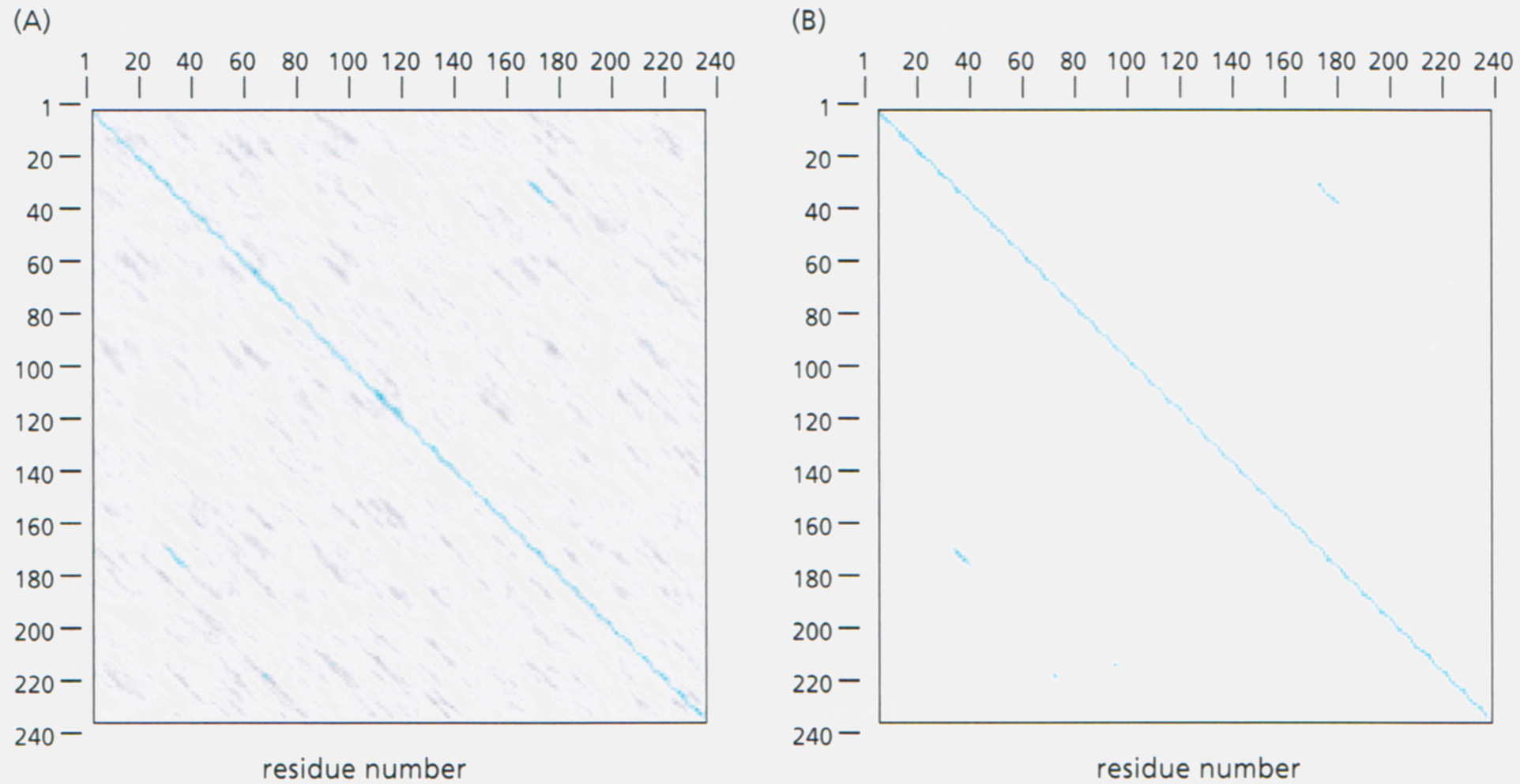
The same rule rejects the wrong residues: SSE / ASE shares too little inside the window — no match.



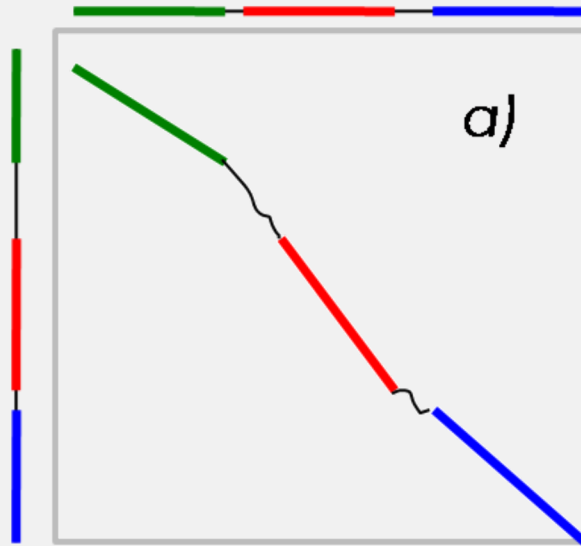
The human LDL-receptor gene (18,444 bases) against itself: beyond the main diagonal, entire families of repeats populate the introns — the reality of whole-gene dot plots.



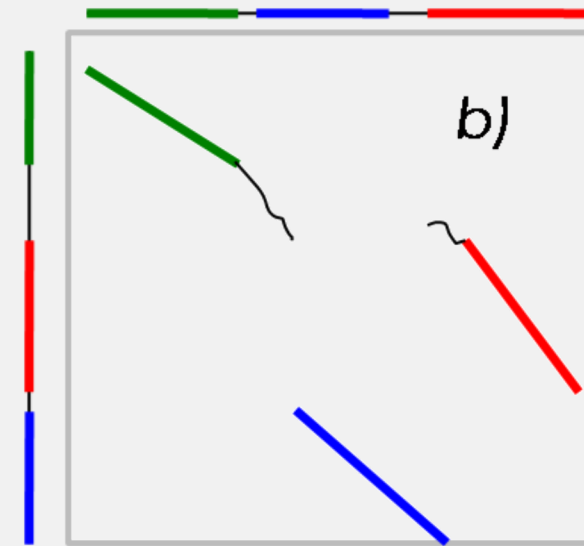
Self-comparison of a 95-residue sequence: single-residue matches (A) are mostly noise; a window filter with higher stringency (B) keeps the main diagonal alone.



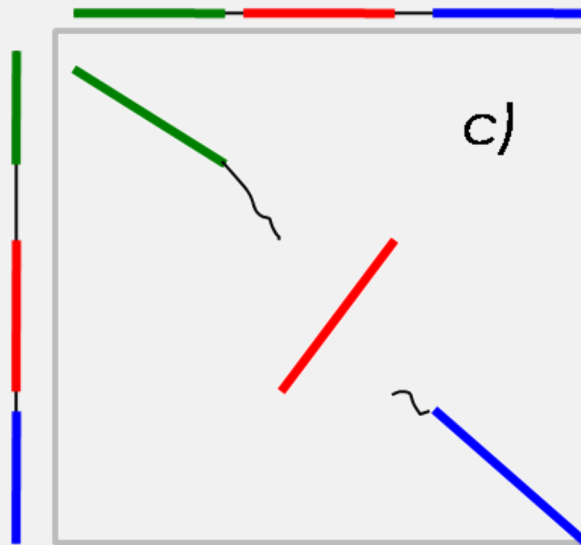
A 240-residue SH2-containing protein: filtering reveals the main diagonal — plus two short off-diagonal matches (internal similarity).



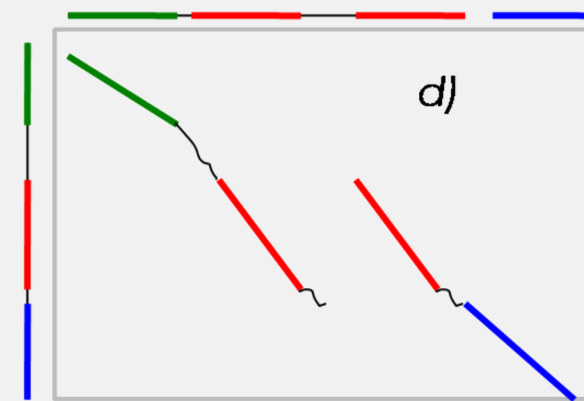
No rearrangement — one continuous diagonal



Translocation — segments meet at unrelated offsets



Inversion — an anti-parallel (reversed) segment



Duplication — one segment matches twice

AA Sequence Alignment

Responds to the question whether any two sequences are related.

Four Main Steps

- What sort of alignment should be considered?
- What scoring system will rank alignments?
- Which algorithm will seek good scoring alignments?
- How do we assess the significance of an alignment score?



What the procedure can yield: a BLAST report with a 338-bit alignment ($E = 4 \times 10^{-94}$) of the aminotrans_1 query to a Pfam class-I aminotransferase family.

Homologous sequences can have different length.

Insertions, deletions, indels..... Marked as “-”

T	H	I	S	I	S	A	S	E	Q	U	E	N	C	E
T	H	A	T	-	-	-	S	E	Q	U	E	N	C	E

(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.*

A simple alignment score

For an aligned pair of sequences, assign a score to each column:

Column	Score
Match	0
Mismatch	-1
Gap character	-G

The obvious strategy fails

1. Assume an alignment.
2. Score every column (table above).
3. Sum over all positions to get a score.
4. Repeat for **all possible alignments** — astronomically many, even for two moderately sized sequences.

There is a better way!

Dynamic programming reuses intermediate prefix scores instead of enumerating every alignment.

The total score is the sum of column terms for aligned pairs, plus gaps.

Identities and conservative substitutions

- More likely in real alignments; positive scores.

Non-conservative changes

- Less frequent in real alignments; negative scores.

An Additive score system

- Assumes mutations at different sites occurs independently.
- This assumption is reasonable for DNA and proteins.
- Could be not suitable for some RNA sequences.



position i	1	2	3	4	5	6
x	S	E	N	D	M	K
y	A	E	N	D	L	Q

Two aligned, gap-free sequences of equal length — the situation the models assume. The index i counts alignment columns.

Each column is judged against two models

Highlighted column, $i = 4$: $x_4 = D$ and $y_4 = D$.

- **Chance (background) model:** $q_{x_4} = q_D$ and $q_{y_4} = q_D$ are just background residue frequencies — $q_D \approx 5.4\%$ across a protein database. The column's chance probability is $q_{x_4} q_{y_4} = q_D^2 \approx 0.3\%$.
- **Related (pair) model:** $p_{x_4 y_4} = p_{DD}$ = how often D sits opposite D inside conserved blocks — usually much more than 0.3%.

Same arithmetic for the mismatch column $i = 5$: $q_{x_5} q_{y_5} = q_M q_L \approx 0.25\%$. The substitution score (next) compares related vs chance, column by column.

For aligned, ungapped sequences of equal length, let q_a denote a background residue probability and p_{ab} an aligned-pair probability.

Independent background model

$$P(x, y \mid R) = \prod_i q_{x_i} q_{y_i}$$

Aligned-pair model

$$P(x, y \mid M) = \prod_i p_{x_i y_i}$$

The models assume independence between alignment columns.

The odds ratio compares the two models:

$$\frac{P(x, y \mid M)}{P(x, y \mid R)} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}.$$

Taking logarithms gives an additive score:

$$s(a, b) = \log \frac{p_{ab}}{q_a q_b}, \quad S = \sum_i s(x_i, y_i).$$

The logarithm base and scaling determine the score units.

$s(a, b)$ defines a matrix, known as score matrix or substitution matrix (SM).

- A symmetric matrix for 20 amino acids has **210 distinct residue-pair entries**, including the diagonal; gap penalties are separate.

Two main Substitution Matrices are usually employed:

- PAM matrices model accepted point mutations.
- BLOSUM, or BLOck SUBstitution Matrices, are built from gap-less conserved blocks.

- Computed by Margaret Dayhoff in 1978 (also known as Dayhoff matrices): 1572 accepted point mutations in 71 families of closely related proteins.
- PAM1 models an average of 1 accepted mutation per 100 residues; **PAM n** is PAM1 raised to the power n — PAM250 models strongly diverged sequences.
- After many cycles a residue may mutate several times — even back to its original state — so **PAM distance is not proportional to evolutionary time**: families evolve at different rates.

- Henikoff and Henikoff, 1992: built from **local** alignments — the BLOCKS database of gap-less conserved blocks of protein families.
- High numbers (BLOSUM80): closely related sequences; low numbers (BLOSUM45): more divergent ones.
- **BLOSUM62** is the *de facto* standard in protein database searches and sequence alignments.

	PAM (Dayhoff, 1978)	BLOSUM (Henikoff, 1992)
Built from	closely related families	gap-less conserved blocks
Alignment used	global	local
Numbering	higher = more diverged (PAM250)	higher = closer (BLOSUM80)
Typical choice	distant relatives	searches: 80, 62, 45

Match the matrix to the expected divergence: there is no universally correct matrix.

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

Residue 1 Residue 2

$$s(S, A) = 1$$

Choose two residues, or click a score.

Positive: favoured under the related-sequence model.

Negative: favoured under the background model.

Gap penalties are separate.

The score depends on **which residues** are aligned—not only whether they match.

Residue pair	BLOSUM62 score	Interpretation
W / W	+11	A tryptophan match
N / N	+6	An asparagine match
S / A	+1	A favourable non-identical pair
E / A	−1	An unfavourable pair
W / D	−4	A more strongly penalised pair

A substitution score is a log-odds score, **not a mutation probability**. Add the scores over aligned pairs; account for gaps separately.

For a gap of length k , subtract the cost d once per gap character:

$$\gamma(k) = -kd, \quad d > 0.$$

Gap length	At $d = 8$	Contribution to the score
1	—	−8
2	— —	−16
3	— — —	−24

All worked examples in this section use BLOSUM62 and a linear gap cost of 8, unless you change it in the laboratory.

Opening one long gap can cost less than opening several short ones. Here o is the opening cost and e is the extension cost:

$$\gamma(k) = -o - (k - 1)e, \quad k \geq 1.$$

With $o = 8$ and $e = 1$:

Pattern	Calculation	Score contribution
One gap of length 3	$-8 - 2 \times 1$	-10
Three separate gaps of length 1	$3 \times (-8)$	-24

favoured — one long gap: $-8 - 6 \times 1 = -14$

ATGTAGTGTATAGTACATGCA

ATGTAG---TACATGCA

penalised — three openings: $-3 \times 8 - (1 + 1 + 2) \times 1 = -28$

ATGTAGTGTATAGTACATGCA

ATGTA-G-TA-CATGCA

The simple recurrence we will use assumes **linear** costs. Affine costs require additional states, as in Gotoh’s algorithm.

Column	1	2	3	4
x	S	E	N	D
y	A	—	N	D
Score	+1	−8	+6	+6

Total = $1 - 8 + 6 + 6 = 5$.

This scores one proposed alignment. How do we find the best alignment without trying every possibility?

A cell stores the best score for two **prefixes** of the sequences.

For example, $F(2, 3)$ aligns **SEN** with **AN**: the first three residues of x and the first two of y .

The last alignment column must contain a residue pair, a gap in x , or a gap in y . Each choice extends a smaller prefix alignment whose best score is already known.

Compute each cell once: mn interior cells for lengths m and n . This gives $O(mn)$ time for the linear-gap model.

Global alignment consumes every residue of both sequences, including their ends.

Stage	What to do
Initialise	Charge for aligning a prefix against an empty sequence.
Fill	Choose the best of three ways to enter each cell.
Trace back	Start at the bottom-right cell; stop at the origin.

The final score belongs to the **bottom-right cell**, even if another cell contains a larger value.

Rows consume y_i ; columns consume x_j . Let $F(i, j)$ be the best prefix-alignment score.

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_j, y_i) & \text{align the two residues} \\ F(i-1, j) - d & \text{gap in } x \\ F(i, j-1) - d & \text{gap in } y \end{cases}$$

$$F(0, 0) = 0, \quad F(i, 0) = -id, \quad F(0, j) = -jd.$$

Store **every tied best predecessor**. Several optimal alignments can share the same score.

Traceback move	Consume from x?	Consume from y?	Append this column
↖ Diagonal	Yes	Yes	residue / residue
↑ Up	No	Yes	– / residue
← Left	Yes	No	residue / –

Tracing backwards discovers the alignment **from right to left**. Reverse the collected columns to display it in sequence order.

The score table and the traceback path describe the same alignment. Its column scores must sum to the final cell score.

x → S E N D

y ↓ A N D

These are short **protein sequences**, written with one-letter amino-acid codes.

- Scoring matrix: **BLOSUM62**.
- Linear gap cost: **8 per gap character**.
- A 4×5 score grid includes the empty-prefix row and column.

We will use these same inputs for global and local alignment.

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap} -8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	-8 ←	-16 ←	-24 ←	-32 ←
A	-8 ↑	.	.	.	.
N	-16 ↑	.	.	.	.
D	-24 ↑	.	.	.	.

↖ residue pair · ↑ gap in x · ← gap in y

Cell F(1, 1): S with A

BLOSUM62[S, A] = 1

Predict the maximum of the diagonal, upper and left candidates.

The next slide shows the first calculation.

0 / 12 interior cells computed. Fill left to right, one row at a time.

First cell: compare S with A

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap} -8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	-8 ←	-16 ←	-24 ←	-32 ←
A	-8 ↑	1 ↖	.	.	.
N	-16 ↑	.	.	.	.
D	-24 ↑	.	.	.	.

↖ residue pair · ↑ gap in x · ← gap in y

Cell F(1, 1): S with A

$\text{BLOSUM62}[S, A] = 1$

↖ **Diagonal: align residues**
 $0 + 1 = 1 \checkmark$

↑ **Up: gap in x**
 $-8 - 8 = -16$

← **Left: gap in y**
 $-8 - 8 = -16$

Maximum = 1

1 / 12 interior cells computed. Fill left to right, one row at a time.

Second cell: a gap can be the best choice

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap} -8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	-8 ←	-16 ←	-24 ←	-32 ←
A	-8 ↑	1 ↖	-7 ←	.	.
N	-16 ↑	.	.	.	.
D	-24 ↑	.	.	.	.

↖ residue pair · ↑ gap in x · ← gap in y

Cell F(1, 2): E with A

$\text{BLOSUM62}[E, A] = -1$

↖ **Diagonal: align residues**
 $-8 - 1 = -9$

↑ **Up: gap in x**
 $-16 - 8 = -24$

← **Left: gap in y**
 $1 - 8 = -7 \checkmark$

Maximum = -7

2 / 12 interior cells computed. Fill left to right, one row at a time.

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap} -8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	-8 ←	-16 ←	-24 ←	-32 ←
A	-8 ↑	1 ↖	-7 ←	-15 ←	-23 ←
N	-16 ↑	-7 ↖↑	1 ↖	.	.
D	-24 ↑	.	.	.	.

↖ residue pair · ↑ gap in x · ← gap in y

6 / 12 interior cells computed. Fill left to right, one row at a time.

Cell F(2, 2): E with N

$\text{BLOSUM62}[E, N] = 0$

↖ **Diagonal: align residues**
 $1 + 0 = 1 \checkmark$

↑ **Up: gap in x**
 $-7 - 8 = -15$

← **Left: gap in y**
 $-7 - 8 = -15$

Maximum = 1

Complete the global score grid

x → SEND · y ↓ AND · BLOSUM62 · gap -8

y / x	∅	S	E	N	D
∅	0	-8 ←	-16 ←	-24 ←	-32 ←
A	-8 ↑	1 ↖	-7 ←	-15 ←	-23 ←
N	-16 ↑	-7 ↖↑	1 ↖	-1 ↖	-9 ←
D	-24 ↑	-15 ↑	-5 ↖	2 ↖	5 ↖

↖ residue pair · ↑ gap in x · ← gap in y

Cell F(3, 4): D with D

BLOSUM62[D, D] = 6

↖ **Diagonal: align residues**
-1 + 6 = 5 ✓

↑ **Up: gap in x**
-9 - 8 = -17

← **Left: gap in y**
2 - 8 = -6

Maximum = 5

12 / 12 interior cells computed. Click any cell to inspect its calculation; then trace back.

Trace back from the bottom-right cell

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap} -8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	-8 ←	-16 ←	-24 ←	-32 ←
A	-8 ↑	1 ↖	-7 ←	-15 ←	-23 ←
N	-16 ↑	-7 ↖↑	1 ↖	-1 ↖	-9 ←
D	-24 ↑	-15 ↑	-5 ↖	2 ↖	5 ↖

↖ residue pair · ↑ gap in x · ← gap in y

Global score: 5 · Traceback complete · start bottom-right; stop at (0, 0)

x SEND

y A-ND

Column scores: $1 - 8 + 6 + 6 = 5$

Cell F(3, 4): D with D

$\text{BLOSUM62}[\text{D}, \text{D}] = 6$

↖ **Diagonal: align residues**
 $-1 + 6 = 5 \checkmark$

↑ **Up: gap in x**
 $-9 - 8 = -17$

← **Left: gap in y**
 $2 - 8 = -6$

Maximum = 5

Recover the alignment and check its score

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap } -8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	-8 ←	-16 ←	-24 ←	-32 ←
A	-8 ↑	1 ↖	-7 ←	-15 ←	-23 ←
N	-16 ↑	-7 ↖↑	1 ↖	-1 ↖	-9 ←
D	-24 ↑	-15 ↑	-5 ↖	2 ↖	5 ↖

↖ residue pair · ↑ gap in x · ← gap in y

Global score: 5 · Traceback complete · start bottom-right; stop at (0, 0)

x SEND

y A-ND

Column scores: $1 - 8 + 6 + 6 = 5$

Cell F(3, 4): D with D

BLOSUM62[D, D] = 6

↖ Diagonal: align residues
 $-1 + 6 = 5 \checkmark$

↑ Up: gap in x
 $-9 - 8 = -17$

← Left: gap in y
 $2 - 8 = -6$

Maximum = 5

A score cell can have **two or three equally good predecessors**.

For SEND / AND with BLOSUM62 and $d = 8$, $F(2, 1)$ has a tie:

$$\max\{-8 + 1, 1 - 8, -16 - 8\} = -7.$$

Both diagonal and up are optimal ways to reach this cell.

A tie in a cell does not necessarily create multiple optimal **complete** alignments: the final traceback may never visit that cell.

The interactive grid shows every tied arrow. Its demonstration follows diagonal, then up, then left when tied.

x → columns · y ↓ rows · BLOSUM62 · gap -8

y / x	∅	S	E	N	D
∅	0	-8 ←	-16 ←	-24 ←	-32 ←
A	-8 ↑	1 ↖	-7 ←	-15 ←	-23 ←
N	-16 ↑	-7 ↖↑	1 ↖	-1 ↖	-9 ←
D	-24 ↑	-15 ↑	-5 ↖	2 ↖	5 ↖

↖ residue pair · ↑ gap in x · ← gap in y

Cell F(3, 4): D with D

BLOSUM62[D, D] = 6

↖ **Diagonal: align residues**
-1 + 6 = 5 ✓

↑ **Up: gap in x**
-9 - 8 = -17

← **Left: gap in y**
2 - 8 = -6

Maximum = 5

Global score: 5 · Traceback complete · start bottom-right; stop at (0, 0)

x SEND

y A-ND

Column scores: 1 - 8 + 6 + 6 = 5

Local alignment: find the strongest matching region

Keep **SEND / AND**, BLOSUM62 and the gap cost of 8. Change only the alignment objective.

Global: Needleman–Wunsch

Local: Smith–Waterman

Align all residues

Align the best-scoring subsequences

SEND / A-ND

ND / ND

$$1 - 8 + 6 + 6 = 5$$

$$6 + 6 = 12$$

Local alignment can omit unfavourable flanking regions. It does not charge gaps for the omitted residues outside the selected subsequences.

The extra zero allows an alignment to start at the current position.

$$F(i, j) = \max \begin{cases} 0 & \text{restart} \\ F(i-1, j-1) + s(x_j, y_i) & \text{align the residues} \\ F(i-1, j) - d & \text{gap in } x \\ F(i, j-1) - d & \text{gap in } y \end{cases}$$

Initialise the first row and column to **zero**. Trace back from a **highest-scoring cell anywhere** and stop at zero.

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap} -8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	0	0	0	0
A	0	1 ↖	0	.	.
N	0	.	.	.	.
D	0	.	.	.	.

↖ residue pair · ↑ gap in x · ← gap in y

Cell F(1, 2): E with A

$\text{BLOSUM62}[E, A] = -1$

↖ **Diagonal: align residues**
 $0 - 1 = -1$

↑ **Up: gap in x**
 $0 - 8 = -8$

← **Left: gap in y**
 $1 - 8 = -7$

Restart here = 0 ✓

Maximum = 0

2 / 12 interior cells computed. Fill left to right, one row at a time.

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap} - 8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	0	0	0	0
A	0	1 ↖	0	0	0
N	0	1 ↖	1 ↖	6 ↖	1 ↖
D	0	0	3 ↖	2 ↖	12 ↖

↖ residue pair · ↑ gap in x · ← gap in y

Cell F(3, 4): D with D

$\text{BLOSUM62}[D, D] = 6$

↖ **Diagonal: align residues**
 $6 + 6 = 12 \checkmark$

↑ **Up: gap in x**
 $1 - 8 = -7$

← **Left: gap in y**
 $2 - 8 = -6$

Restart here = 0

Maximum = 12

12 / 12 interior cells computed. Click any cell to inspect its calculation; then trace back.

$x \rightarrow \text{SEND} \cdot y \downarrow \text{AND} \cdot \text{BLOSUM62} \cdot \text{gap} -8$

y / x	$\emptyset$	S	E	N	D
$\emptyset$	0	0	0	0	0
A	0	1 ↖	0	0	0
N	0	1 ↖	1 ↖	6 ↖	1 ↖
D	0	0	3 ↖	2 ↖	12 ↖

↖ residue pair · ↑ gap in x · ← gap in y

Cell F(3, 4): D with D

$\text{BLOSUM62}[D, D] = 6$

↖ **Diagonal: align residues**
 $6 + 6 = 12 \checkmark$

↑ **Up: gap in x**
 $1 - 8 = -7$

← **Left: gap in y**
 $2 - 8 = -6$

Restart here = 0

Maximum = 12

Best local score: 12 · Traceback complete · start at a highest cell; stop at zero

x ND

y ND

Column scores: $6 + 6 = 12 \cdot x[3-4], y[2-3]$

A longer protein example: explore both objectives

x → columns · y ↓ rows · BLOSUM62 · gap -8

y / x	∅	H	E	A	G	A	W	G	H	E	E
∅	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	4↖	0	4↖	0	0	0	0	0
W	0	0	0	0	2↖	0	15↖	7←	0	0	0
H	0	8↖	0	0	0	0	7↑	13↖	15↖	7←	0
E	0	0	13↖	5←	0	0	0	5↖↑	13↖	20↖	12↖←
A	0	0	5↑	17↖	9←	4↖	0	0	5↑	12↖↑	19↖
E	0	0	5↖	9↑	15↖	8↖	1↖	0	0	10↖	17↖

↖ residue pair · ↑ gap in x · ← gap in y

Best local score: 20 · Traceback complete · start at a highest cell; stop at zero

x AWGHE

y AW-HE

Column scores: 4 + 11 - 8 + 8 + 5 = 20 · x[5-9], y[2-5]

Cell F(5, 9): E with E

BLOSUM62[E, E] = 5

↖ Diagonal: align residues
15 + 5 = 20 ✓

↑ Up: gap in x
7 - 8 = -1

← Left: gap in y
13 - 8 = 5

Restart here = 0

Maximum = 20

Decision	Needleman–Wunsch	Smith–Waterman
First row and column	Accumulated gap costs	Zero
Candidate moves	Diagonal, up, left	Diagonal, up, left, zero
Traceback starts	Bottom-right cell	A highest-scoring cell
Traceback stops	Origin	A zero-score cell

Predict: for **W** versus **D**, BLOSUM62 gives -4 and the gap cost is 8. What does each algorithm return?

Global: score -4 , alignment W / D. **Local:** score 0, no positive-scoring alignment.

A high score under a scoring model is not yet evidence of homology — quantify it first (next section).

For sufficiently long random sequences and an appropriate scoring model, optimal local-alignment scores follow an extreme-value law:

$$E = K m n e^{-\lambda S}$$

$$P = 1 - e^{-E} \approx E \text{ (small } E)$$

- m, n : sequence lengths; S : alignment score.
- K, λ : calibrated by the scoring system and the residue background model — not from Smith–Waterman itself.

Illustrative E-values

With approximate BLOSUM62 constants
 $K = 0.13, \lambda = 0.318, m = 1000, n = 100$:

Score S	E -value
40	3.9×10^{-2}
50	1.6×10^{-3}
60	6.7×10^{-5}

Every +10 in score divides the expected count of chance matches by ≈ 24 .

The E -value is an **expected count of chance matches**, not the probability that an alignment is homologous.

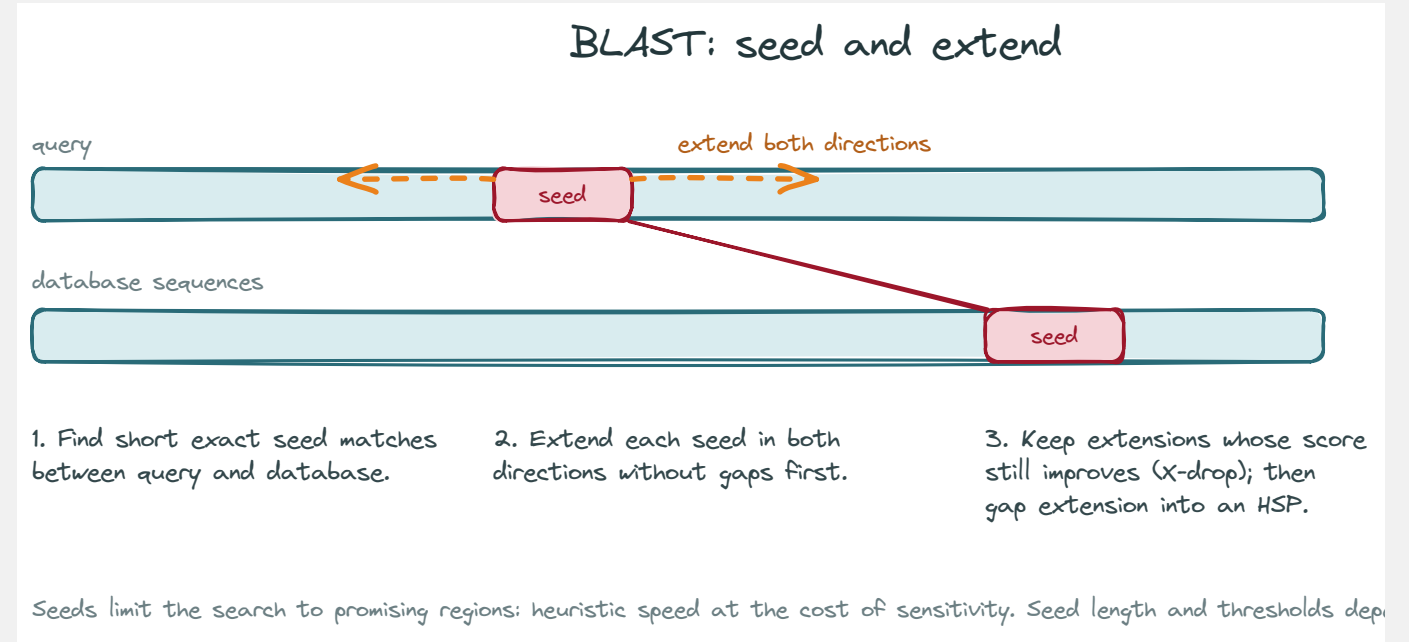
- **Percent identity:** identical residue pairs divided by the stated alignment-length denominator, multiplied by 100.
- **Percent similarity:** residue pairs satisfying a specified similarity rule, often a positive substitution score, divided by the stated denominator.
- State whether gaps are included and which scoring matrix defines similarity.

Worked example — **SEND** / **A-ND**: 3 identical residue pairs. Over **all 4 alignment columns**: 75% identity. Excluding the gap column: 100% of residue–residue columns. The denominator changes the number.

These percentages describe an alignment. Homology describes common ancestry.

Heuristic search: seed and extend

- BLAST — Altschul et al., 1990.
- Finds short **exact seed matches**, extends them in both directions while the score improves, and keeps high-scoring segment pairs (HSPs) with bit scores and *E*-values.



Why not full dynamic programming?

- Smith–Waterman needs $O(nm)$ cells: a 1 000-residue query against a 10^8 -residue database $\approx 10^{11}$ updates — infeasible.
- Seeding limits the work to each seed's neighbourhood, trading some sensitivity.

FASTA

- Pearson and Lipman, 1988: heuristic local alignment.
- Uses a lookup table of short words (**ktup**: 2 for proteins, 6 for DNA by default) to find dense diagonals of similarity.
- Rebuilds the top diagonals with gap-aware, optimized local alignment.
- Higher ktup: faster, less sensitive; lower ktup: slower, more sensitive.
- <https://fasta.bioch.virginia.edu/> (redirects to HTTPS; verified 2026-10-01)

		DATABASE SEQUENCES	
		Nucleotide	Protein
QUERY	Nucleotide	blastn Compare nucleotide sequences tblastx Translate both query and database	blastx Translate the nucleotide query
	Protein	tblastn Translate the nucleotide database	blastp Compare protein sequences

Translated searches consider all six reading frames of the nucleotide sequence.

Multiple Sequence Alignment

To align three or more biological sequences

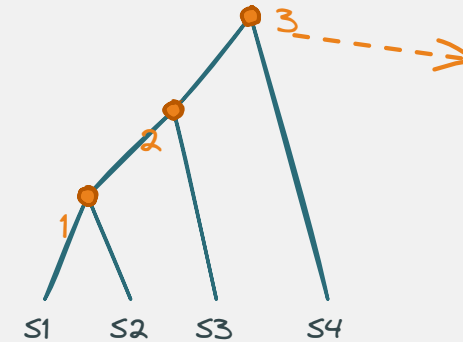
We could think in an extension of dynamic programming methods for N individual sequences:

- Involves the construction of an N -way or N -dimensional equivalent of the F matrix in a similar way than pairwise sequence alignment.
- Not as computational efficient with large N . Combinatorial explosion as the number of sequences increases!!!
- Methods have been published to reduce the search space of the algorithm.
- Dynamic programming is used with limitations on MSA.

- Compute pairwise similarities or distances.
- Build a guide tree or another ordering of the sequences.
- Align sequences and profiles progressively in that order.
- Early alignment errors can influence later steps.

A **guide tree** organises alignment; it is not automatically a validated phylogenetic inference.

1. Score every sequence pair
2. Build a guide tree (most similar pairs join first)



a guide tree is a merge order, not a validated phylogenetic inference

3. Merge sequences and profiles in tree order

final MSA (all rows, same length)

S1	M	A	E	K	L	V	-	N
S2	M	A	Q	K	L	V	I	N
S3	-	A	Q	K	L	V	I	N
S4	-	A	Q	R	L	V	I	N

gap columns (-) stay in the alignment after they are introduced

Alignment programs

- MUSCLE, Clustal Omega, MAFFT

- Kalign and T-Coffee

Viewers and editors

- Jalview

- SeaView

The screenshot illustrates the original course software comparison. Its entries and interfaces are historical.

