Lecture – Sequence analysis

Global Alignment and Local Alignment

Dot Matrix Method

Dynamic Programming Method

Sequence analysis

Global Alignment and Local Alignment

Alignment algorithms

Sequence alignment Dynamic Programming Method

- Dynamic programming is a method that determines optimal alignment by matching two sequences for all possible pairs of characters between the two sequences.
- It is fundamentally similar to the dot matrix method in that it also creates a two-dimensional alignment grid.
- However, it finds alignment in a more quantitative way by converting a dot matrix into a scoring matrix to account for matches and mismatches between sequences.
- By searching for the set of highest scores in this matrix, the best alignment can be accurately obtained.

Method:

- Dynamic programming works by first constructing a two-dimensional matrix whose axes are the two sequences to be compared.
- The residue matching is according to a particular scoring matrix.
- The scores are calculated one row at a time.
- This starts with the first row of one sequence, which is used to scan through the entire length of the other sequence, followed by scanning of the second row.

	A	T	T	G	C
A	1	0	0	0	0
G					
G					
C					

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Dynamic Programming Method for Global Alignment

Sequence alignment

Needleman–Wunsch algorithm

- The classical global pairwise alignment algorithm using dynamic programming is the Needleman–Wunsch algorithm.
- In this algorithm, an optimal alignment is obtained over the entire lengths of the two sequences.
- It must extend from the beginning to the end of both sequences to achieve the highest total score.
- In other words, the alignment path has to go from the bottom right corner of the matrix to the top left corner.
- The drawback of focusing on getting a maximum score for the full-length sequence alignment is the risk of missing the best local similarity.
- This strategy is only suitable for aligning two closely related sequences that are of the same length.
- For divergent sequences or sequences with different domain structures, the approach does not produce optimal alignment.
- One of the few web servers dedicated to global pairwise alignment is EMBOSS.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Dynamic Programming Method for Local Alignment

Sequence alignment

Smith-Waterman algorithm

- In regular sequence alignment, the divergence level between the two sequences to be aligned is not easily known.
- The sequence lengths of the two sequences may also be unequal.
- In such cases, identification of regional sequence similarity may be of greater significance than finding a match that includes all residues.
- The first application of dynamic programming in local alignment is the Smith–Waterman algorithm.
- In this algorithm, positive scores are assigned for matching residues and zeros for mismatches.
- No negative scores are used.
- A similar tracing-back procedure is used in dynamic programming.
- However, the alignment path may begin and end internally along the main diagonal.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Dynamic Programming Method for Local Alignment

Sequence alignment

Smith-Waterman algorithm

- However, the alignment path may begin and end internally along the main diagonal.
- It starts with the highest scoring position and proceeds diagonally up to the left until reaching a cell with a zero.
- Gaps are inserted if necessary.
- Occasionally, several optimally aligned segments with best scores are obtained.
- As in the global alignment, the final result is influenced by the choice of scoring systems used.
- The goal of local alignment is to get the highest alignment score locally, which may be at the expense of the highest possible overall score for a full-length alignment.
- This approach maybe suitable for aligning divergent sequences or sequences with multiple domains that may be of different origins.
- BLAST is the most commonly used pairwise local alignment web servers.

Source: Essential Bioinformatics by Jin Xiong

Lecture – Sequence analysis

Global Alignment and Local Alignment

Dot Matrix Method

Dynamic Programming Method

Scoring Matrices

Sequence alignment

- In the dynamic programming algorithm presented, the alignment procedure has to make use of a scoring system, which is a set of values for quantifying the likelihood of one residue being substituted by another in an alignment.
- The scoring systems is called a substitution matrix and is derived from statistical analysis of residue substitution data from sets of reliable alignments of highly related sequences.

For nucleotide sequences

- Scoring matrices for nucleotide sequences are relatively simple.
- A positive value or high score is given for a match and a negative value or low score for a mismatch.
- This assignment is based on the assumption that the frequencies of mutation are equal for all bases.
- However, this assumption may not be realistic;
  - observations show that transitions (substitutions between purines and purines or between pyrimidines and pyrimidines) occur more frequently than transversions (substitutions between purines and pyrimidines).
- Therefore, a more sophisticated statistical model with different probability values to reflect the two types of mutations is needed.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Sequence alignment

For amino acid sequences

Amino acid substitutions

- Scoring matrices for amino acids are more complicated because scoring has to reflect the physicochemical properties of amino acid residues, as well as the likelihood of certain residues being substituted among true homologous sequences.
- Certain amino acids with similar physicochemical properties can be more easily substituted than those without similar characteristics.
- Substitutions among similar residues are likely to preserve the essential functional and structural features.
- However, substitutions between residues of different physicochemical properties are more likely to cause disruptions to the structure and function.
- This type of disruptive substitution is less likely to be selected in evolution because it renders nonfunctional proteins.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Sequence alignment

For amino acid sequences Example of: Amino acid substitutions

1. For example, phenylalanine, tyrosine, and tryptophan all share aromatic ring structures.
   - ✓ Because of their chemical similarities, they are easily substituted for each other without perturbing the regular function and structure of the protein.
2. Similarly, arginine, lysine, and histidine are all large basic residues and there is a high probability of them being substituted for each other.
3. Aspartic acid, glutamic acid, asparagine, and glutamine belong to the acid and acid amide groups and can be associated with relatively high frequencies of substitution.
4. The hydrophobic residue group includes methionine, isoleucine, leucine, and valine.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Sequence alignment

For amino acid sequences Example of: Amino acid substitutions

5. Small and polar residues include serine, threonine, and cysteine.
   - ✓ Residues within these groups have high likelihoods of being substituted for each other.
   - ✓ However, cysteine contains a sulfhydryl group that plays a role in metal binding, active site, and disulfide bond formation.
   - ✓ Substitution of cysteine with other residues therefore often abolishes the enzymatic activity or destabilizes the protein structure.
   - ✓ It is thus a very infrequently substituted residue.
6. The small and nonpolar residues such as glycine and proline are also unique in that their presence often disrupts regular protein secondary structures.
   - ✓ Thus, substitutions with these residues do not frequently occur.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Sequence alignment

For amino acid sequences

Example of: **Amino acid substitutions**

Amino Acid Group	Amino Acid Name	Three- and One-Letter Code	Main Functional Features
Small and nonpolar	Glycine Alanine Proline	Gly, G Ala, A Pro, P	Nonreactive in chemical reactions; Pro and Gly disrupt regular secondary structures
Small and polar	Cysteine Serine Threonine	Cys, C Ser, S Thr, T	Serving as posttranslational modification sites and participating in active sites of enzymes or binding metal
Large and polar	Glutamine Asparagine	Gln, Q Asn, N	Participating in hydrogen bonding or in enzyme active sites
Large and polar (basic)	Arginine Lysine Histidine	Arg, R Lys, K His, H	Found in the surface of globular proteins providing salt bridges; His participates in enzyme catalysis or metal binding
Large and polar (acidic)	Glutamate Aspartate	Glu, E Asp, D	Found in the surface of globular proteins providing salt bridges
Large and nonpolar (aliphatic)	Isoleucine Leucine Methionine Valine	Ile, I Leu, L Met, M Val, V	Nonreactive in chemical reactions; participating in hydrophobic interactions
Large and nonpolar (aromatic)	Phenylalanine Tyrosine Tryptophan	Phe, F Tyr, Y Trp, W	Providing sites for aromatic packing interactions; Tyr and Trp are weakly polar and can serve as sites for phosphorylation and hydrogen bonding

Note: Each amino acid is listed with its full name, three- and one-letter abbreviations, and main functional roles when serving as amino acid residues in a protein. Properties of some amino acid groups overlap.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Sequence alignment

Amino acid substitution matrices

- Amino acid substitution matrices, which are 20 × 20 matrices, have been devised to reflect the likelihood of residue substitutions.
- There are essentially two types of amino acid substitution matrices.
  1. One type is based on interchangeability of the genetic code or amino acid properties,
  2. and the other is derived from empirical studies of amino acid substitutions.
- Although the two different approaches coincide to a certain extent, the first approach, which is based on the genetic code or the physicochemical features of amino acids, has been shown to be less accurate than the second approach, which is based on surveys of actual amino acid substitutions among related proteins.
- Thus, the empirical approach has gained the most popularity in sequence alignment applications and is the focus of our next discussion.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices PAM and BLOSUM

- The empirical matrices, which include PAM and BLOSUM matrices, are derived from actual alignments of highly similar sequences.
- By analyzing the probabilities of amino acid substitutions in these alignments, a scoring system can be developed by giving a high score for a more likely substitution and a low score for a rare substitution.

- For a given substitution matrix, a positive score means that the frequency of amino acid substitutions found in a data set of homologous sequences is greater than would have occurred by random chance.
- They represent substitutions of very similar residues or identical residues.
- A zero score means that the frequency of amino acid substitutions found in the homologous sequence data set is equal to that expected by chance.
- In this case, the relationship between the amino acids is weakly similar at best in terms of physicochemical properties.
- A negative score means that the frequency of amino acid substitutions found in the homologous sequence data set is less than would have occurred by random chance.
- This normally occurs with substitutions between dissimilar residues.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices PAM and BLOSUM

- The substitution matrices apply logarithmic conversions to describe the probability of amino acid substitutions.
- The converted values are the so-called log-odds scores (or log-odds ratios), which are logarithmic ratios of the observed mutation frequency divided by the probability of substitution expected by random chance.
- The conversion can be either to the base of 10 or to the base of 2.

Scoring example

For example, in an alignment that involves ten sequences, each having only one aligned position, nine of the sequences are F (phenylalanine) and the remaining one I (isoleucine).

- ✓ The observed frequency of I being substituted by F is one in ten (0.1),
- ✓ whereas the probability of I being substituted by F by random chance is one in twenty (0.05).

Thus, the ratio of the two probabilities is 2 (0.1/0.05).

After taking this ratio to the logarithm to the base of 2, this makes the log odds equal to 1.

This value can then be interpreted as the likelihood of substitution between the two residues being 2^1, which is two times more frequently than by random chance.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **PAM matrices**

- The PAM matrices (also called Dayhoff PAM matrices) were first constructed by Margaret Dayhoff, who compiled alignments of seventy-one groups of very closely related protein sequences.
- PAM stands for "point accepted mutation" (although "accepted point mutation" or APM may be a more appropriate term, PAM is easier to pronounce).
- Because of the use of very closely related homologs, the observed mutations were not expected to significantly change the common function of the proteins.
- Thus, the observed amino acid mutations are considered to be accepted by natural selection.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **PAM matrices**

- These protein sequences were clustered based on phylogenetic reconstruction (using maximum parsimony).
- The PAM matrices were subsequently derived based on the evolutionary divergence between sequences of the same cluster.
- One PAM unit is defined as 1% of the amino acid positions that have been changed.
- To construct a PAM1 substitution table, a group of closely related sequences with mutation frequencies corresponding to one PAM unit is chosen.
- Based on the collected mutational data from this group of sequences, a substitution matrix can be derived.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **PAM matrices**

PAM1 corresponds to 1% observed mutational rates

PAM1 used for identical sequences

PAM250 corresponds to high observed mutational rates (80%)

PAM250 used for divergent sequences

Correspondence of PAM numbers with observed amino acid mutational rates

PAM Number	Observed Mutation Rate (%)	Sequence Identity (%)
0	0	100
1	1	99
30	25	75
80	50	50
110	40	60
200	75	25
250	80	20

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **PAM matrices** Construction of the PAM1 matrix

- Construction of the PAM1 matrix involves alignment of full-length sequences and subsequent construction of phylogenetic trees (using the parsimony principle).
- This allows computation of ancestral sequences for each internal node of the trees.
- Ancestral sequence information is used to count the number of substitutions along each branch of a tree.
- The PAM score for a particular residue pair is derived from a multistep procedure involving
  - calculations of relative mutability (which is the number of mutational changes from a common ancestor for a particular amino acid residue divided by the total number of such residues occurring in an alignment),
  - normalization of the expected residue substitution frequencies by random chance,
  - and logarithmic transformation to the base of 10 of the normalized mutability value divided by the frequency of a particular residue.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **PAM matrices** **Construction of the PAM1 matrix**

- **The resulting value is rounded to the nearest integer and entered into the substitution matrix, which reflects the likelihood of amino acid substitutions.**
- **This completes the log-odds score computation.**
- **After compiling all substitution probabilities of possible amino acid mutations, a 20 × 20 PAM matrix is established.**
- **Positive scores in the matrix denote substitutions occurring more frequently than expected among evolutionarily conserved replacements.**
- **Negative scores correspond to substitutions that occur less frequently than expected.**

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **PAM matrices** Other PAM matrices

- Other PAM matrices with increasing numbers for more divergent sequences are extrapolated from PAM1 through matrix multiplication.
- For example, PAM80 is produced by values of the PAM1 matrix multiplied by itself eighty times.
- The mathematical transformation accounts for multiple substitutions having occurred in an amino acid position during evolution.
- For example, when a mutation is observed as F replaced by I, the evolutionary changes may have actually undergone a number of intermediate steps before becoming I, such as in a scenario of F → M → L → I.
- For that reason, a PAM80 matrix only corresponds to 50% of observed mutational rates.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **PAM matrices** **Other PAM matrices**

Correspondence of PAM numbers with observed amino acid mutational rates

PAM1 corresponds to 1% observed mutational rates

PAM1 used for identical sequences

PAM Number	Observed Mutation Rate (%)	Sequence Identity (%)
0	0	100
1	1	99
30	25	75
80	50	50
110	40	60
200	75	25
250	80	20

PAM250 corresponds to high observed mutational rates (80%)

PAM250 used for divergent sequences

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **PAM matrices** **Other PAM matrices**

- A PAM unit is defined as 1% amino acid change or one mutation per 100 residues.
- The increasing PAM numbers correlate with increasing PAM units and thus evolutionary distances of protein sequences (Table 3.1).
- For example, PAM250, which corresponds to 20% amino acid identity, represents 250 mutations per 100 residues.
- In theory, the number of evolutionary changes approximately corresponds to an expected evolutionary span of 2,500 million years.
- Thus, the PAM250 matrix is normally used for divergent sequences.
- Accordingly, PAM matrices with lower serial numbers are more suitable for aligning more closely related sequences.

PAM Number	Observed Mutation Rate (%)	Sequence Identity (%)
0	0	100
1	1	99
30	25	75
80	50	50
110	40	60
200	75	25
250	80	20

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **BLOSUM Matrices**

- In the PAM matrix construction, the only direct observation of residue substitutions is in PAM1, based on a relatively small set of extremely closely related sequences.
- Sequence alignment statistics for more divergent sequences are not available.

- To fill in the gap, a new set of substitution matrices have been developed.
- This is the series of blocks amino acid substitution matrices (BLOSUM), all of which are derived based on direct observation for every possible amino acid substitution in multiple sequence alignments.
- These were constructed based on more than 2,000 conserved amino acid patterns representing 500 groups of protein sequences.
- **BLOCKS**: The sequence patterns, also called blocks, are ungapped alignments of less than sixty amino acid residues in length.
- The frequencies of amino acid substitutions of the residues in these blocks are calculated to produce a numerical table, or block substitution matrix.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **BLOSUM Matrices**

- **Instead of using the extrapolation function, the BLOSUM matrices are actual percentage identity values of sequences selected for construction of the matrices.**
- **For example, BLOSUM62 indicates that the sequences selected for constructing the matrix share an average identity value of 62%.**
- **Other BLOSUM matrices based on sequence groups of various identity levels have also been constructed.**
- **In the reversing order as the PAM numbering system, the lower the BLOSUM number, the more divergent sequences they represent.**

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices **BLOSUM Matrices**

- The BLOSUM score for a particular residue pair is derived from the log ratio of observed residue substitution frequency versus the expected probability of a particular residue (similar to defined earlier).
- The log odds is taken to the base of 2 instead of 10 as in the PAM matrices.
- The resulting value is rounded to the nearest integer and entered into the substitution matrix.
- As in the PAM matrices, positive and negative values correspond to substitutions that occur more or less frequently than expected among evolutionarily conserved replacements.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Scoring Matrices

Amino acid scoring matrices

Amino acid substitution matrices

Sequence alignment

Empirical matrices

PAM versus BLOSUM Matrices

- ✓ There are a number of differences between PAM and BLOSUM.
- ✓ The principal difference is that the PAM matrices, except PAM1, are derived from an evolutionary model whereas the BLOSUM matrices consist of entirely direct observations.
  - ➢ Thus, the BLOSUM matrices may have less evolutionary meaning than the PAM matrices.
  - ➢ This is why the PAM matrices are used most often for reconstructing phylogenetic trees.
- ✓ However, because of the mathematical extrapolation procedure used, the PAM values may be less realistic for divergent sequences.
- ✓ The BLOSUM matrices are entirely derived from local sequence alignments of conserved sequence blocks, whereas the PAM1 matrix is based on the global alignment of full-length sequences composed of both conserved and variable regions.
  - ➢ This is why the BLOSUM matrices may be more advantageous in searching databases (local alignment) and finding conserved domains in proteins.

Source: Essential Bioinformatics by Jin Xiong

PAM versus BLOSUM Matrices

- ✓ **Several empirical tests have shown that the BLOSUM matrices outperform the PAM matrices in terms of accuracy of local alignment.**
  - ➢ **This could be largely because the BLOSUM matrices are derived from a much larger and more representative dataset than the one used to derive the PAM matrices.**
  - ➢ **This renders the values for the BLOSUM matrices more reliable.**

Source: Essential Bioinformatics by Jin Xiong

Lecture – Sequence analysis

Database searching and pairwise alignment

BLAST

Sequence analysis

Database searching and pairwise alignment

- A main application of pairwise alignment is retrieving biological sequences in databases based on similarity.
- This process involves submission of a query sequence and performing a pairwise comparison of the query sequence with all individual sequences in a database.
- Thus, database similarity searching is pairwise alignment on a large scale.
- This type of searching is one of the most effective ways to assign putative functions to newly determined sequences.
- However, the dynamic programming method described earlier is slow and impractical to use in most cases.
- Special search methods are used to speed up the computational process of sequence comparison.

Source: Essential Bioinformatics by Jin Xiong

Factors influencing database search

- ✓ sensitivity,
- ✓ selectivity,
- ✓ and speed in database searches

- There are unique requirements for implementing algorithms for sequence database searching.
- The first criterion is sensitivity, which refers to the ability to find as many correct hits as possible.
  - It is measured by the extent of inclusion of correctly identified sequence members of the same family.
  - These correct hits are considered "true positives" in the database searching exercise.
- The second criterion is selectivity, also called specificity, which refers to the ability to exclude incorrect hits.
  - These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered "false positives."
- The third criterion is speed, which is the time it takes to get results from database searches. Depending on the size of the database, speed sometimes can be a primary concern.

Source: Essential Bioinformatics by Jin Xiong

Factors influencing database search

- Ideally, one wants to have the greatest sensitivity, selectivity, and speed in database searches.
- However, satisfying all three requirements is difficult in reality.
  - What generally happens is that an increase in sensitivity is associated with decrease in selectivity.
  - A very inclusive search tends to include many false positives.
  - Similarly, an improvement in speed often comes at the cost of lowered sensitivity and selectivity.
- A compromise between the three criteria often has to be made.

Source: Essential Bioinformatics by Jin Xiong

Factors influencing database search

Solution → 1. Exhaustive search
2. Heuristic search

- In database searching, as well as in many other areas in bioinformatics, are two fundamental types of algorithms.

1. One is the exhaustive type, which uses a rigorous algorithm to find the best or exact solution for a particular problem by examining all mathematical combinations.
   - Dynamic programming is an example of the exhaustive method and is computationally very intensive.
2. Another is the heuristic type, which is a computational strategy to find an empirical or near optimal solution by using rules of thumb.
   - Essentially, this type of algorithms take shortcuts by reducing the search space according to some criteria.
   - However, the shortcut strategy is not guaranteed to find the best or most accurate solution.
   - It is often used because of the need for obtaining results within a realistic time frame without significantly sacrificing the accuracy of the computational output.

Source: Essential Bioinformatics by Jin Xiong

Heuristic database search

- Searching a large database using the dynamic programming methods, such as the Smith–Waterman algorithm, although accurate and reliable, is too slow and impractical when computational resources are limited.
- Thus, speed of searching became an important issue.
- To speed up the comparison, heuristic methods have to be used.
- The heuristic algorithms perform faster searches because they examine only a fraction of the possible alignments examined in regular dynamic programming.

Two major heuristic algorithms for performing database searches include:

- ✓ BLAST
- ✓ FASTA

- These methods are not guaranteed to find the optimal alignment or true homologs but are 50–100 times faster than dynamic programming.
- The increased computational speed comes at a moderate expense of sensitivity and specificity of the search, which is easily tolerated by working molecular biologists.
- Both programs can provide a reasonably good indication of sequence similarity by identifying similar sequence segments.

Source: Essential Bioinformatics by Jin Xiong

Heuristic database search

Two major heuristic algorithms for performing database searches include:

- ✓ BLAST
- ✓ FASTA

- Both BLAST and FASTA use a heuristic word method for fast pairwise sequence alignment.
- This is the third method of pairwise sequence alignment.
- It works by finding short stretches of identical or nearly identical letters in two sequences.
- These short strings of characters are called words, which are similar to the windows used in the dot matrix method (discussed earlier).
- The basic assumption is that two related sequences must have at least one word in common.
- By first identifying word matches, a longer alignment can be obtained by extending similarity regions from the words.
- Once regions of high sequence similarity are found, adjacent high-scoring regions can be joined into a full alignment.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Database searching and pairwise alignment

BLAST → **Basic Local Alignment Search Tool**

Heuristic algorithms

- **The BLAST program was developed by Stephen Altschul of NCBI in 1990 and has since become one of the most popular programs for sequence analysis.**
- **BLAST uses heuristics to align a query sequence with all sequences in a database.**
- **The objective is to find high-scoring ungapped segments among related sequences.**
- **The existence of such segments above a given threshold indicates pairwise similarity beyond random chance, which helps to discriminate related sequences from unrelated sequences in a database.**

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

BLAST

Steps overview

- ✓ BLAST procedure using a hypothetical query sequence matching with a hypothetical database sequence.
- ✓ The alignment scoring is based on the BLOSUM62 matrix.
- ✓ The example of the word match is highlighted in the box.

1. Query: MRDPYNKLIS
2. Scan every three residues to be used in searching BLAST word database.
3. Assuming one of the words finds matches in the database.

Query	PYN	PYN	PYN	PYN	...
Database	PYN	PFN	PFQ	PFE	...

4. Calculate sums of match scores based on BLOSUM62 matrix.

Query	PYN	PYN	PYN	PYN	...
Database	PYN	PFN	PFQ	PFE	...
Sum of score	20	16	10	10	...

5. Find the database sequence corresponding to the best word match and extend alignment in both directions.

Query	M	R	D	PYN	K	L	I	S
Database	M	H	E	PYN	D	V	P	W

extension to left ← → extension to right

6. Determine high scored segment above threshold (22).

Query	M	R	D	PYN	K	L	I	S
Database	M	H	E	PYN	D	V	P	W
	5	0	2	20	-1	1	-3	-3

HSP, total score 24

Source: Essential Bioinformatics by Jin Xiong

BLAST Steps

BLAST performs sequence alignment through the following steps.

1. Create a list of words from the query sequence:

- The first step is to create a list of words from the query sequence.
- Each word is typically three residues for protein sequences and eleven residues for DNA sequences.
- The list includes every possible word extracted from the query sequence.
- This step is also called ***seeding***.

2. Search a sequence database for the occurrence of the query words:

- The second step is to search a sequence database for the occurrence of these words.
- This step is to identify database sequences containing the matching words.
- The matching of the words is scored by a given substitution matrix (such as BLOSUM62).
- A word is considered a match if it is above a threshold.

Source: Essential Bioinformatics by Jin Xiong

BLAST Steps

BLAST performs sequence alignment through the following steps.

3. Pairwise alignment by extending from the words:

- The next step involves pairwise alignment by extending from the words in both directions while counting the alignment score using the same substitution matrix.
- The extension continues until the score of the alignment drops below a threshold due to mismatches (the drop threshold is twenty-two for proteins and twenty for DNA).
- The resulting contiguous aligned segment pair without gaps is called high-scoring segment pair (HSP).
- They are also called maximum scoring pairs.

BLAST Steps

Gapped alignment:

- ✓ BLAST has the ability to provide gapped alignment.
- ✓ In gapped BLAST, the highest scored segment is chosen to be extended in both directions using dynamic programming where gaps may be introduced.
- ✓ The extension continues if the alignment score is above a certain threshold; otherwise, it is terminated.
- ✓ However, the overall score is allowed to drop below the threshold only if it is temporary and rises again to attain above threshold values.
- ✓ Final trimming of terminal regions is needed before producing a report of the final alignment.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Database searching and pairwise alignment

BLAST Types ⇩ https://blast.ncbi.nlm.nih.gov/Blast.cgi

1.

Known as BLASTN

3.

4.

2.

Known as BLASTP

5. TBLASTX

Source: https://blast.ncbi.nlm.nih.gov/Blast.cgi

Sequence analysis

Database searching and pairwise alignment

BLAST Types ⇩ https://blast.ncbi.nlm.nih.gov/Blast.cgi

1.

Known as BLASTN

⇨ **BLASTN queries nucleotide sequences with a nucleotide sequence database.**

Sequence analysis

Database searching and pairwise alignment

BLAST Types ⇩ https://blast.ncbi.nlm.nih.gov/Blast.cgi

BLASTP uses protein sequences as queries to search against a protein sequence database.

2.

Known as BLASTP

Source: https://blast.ncbi.nlm.nih.gov/Blast.cgi Essential Bioinformatics by Jin Xiong

Sequence analysis

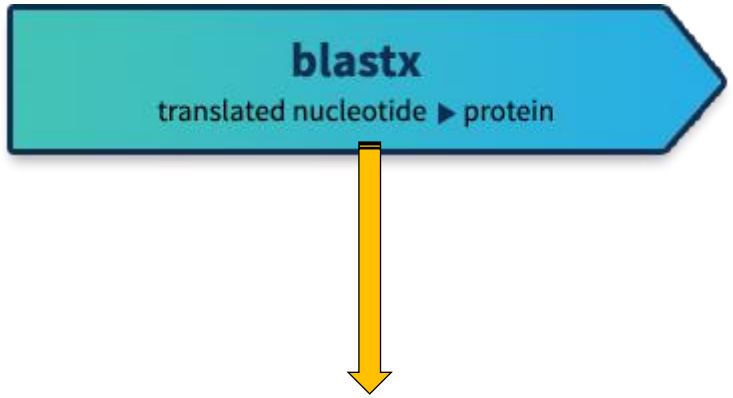
Database searching and pairwise alignment

BLAST Types ⬇ https://blast.ncbi.nlm.nih.gov/Blast.cgi

1.

Known as BLASTN

3.



2.

Known as BLASTP

BLASTX uses nucleotide sequences as queries and translates them in all six reading frames to produce translated protein sequences, which are used to query a protein sequence database.

Nucleotide query ➡ **translates in all six reading frames (translated protein)** ➡ **Search against protein database**

Source: https://blast.ncbi.nlm.nih.gov/Blast.cgi Essential Bioinformatics by Jin Xiong

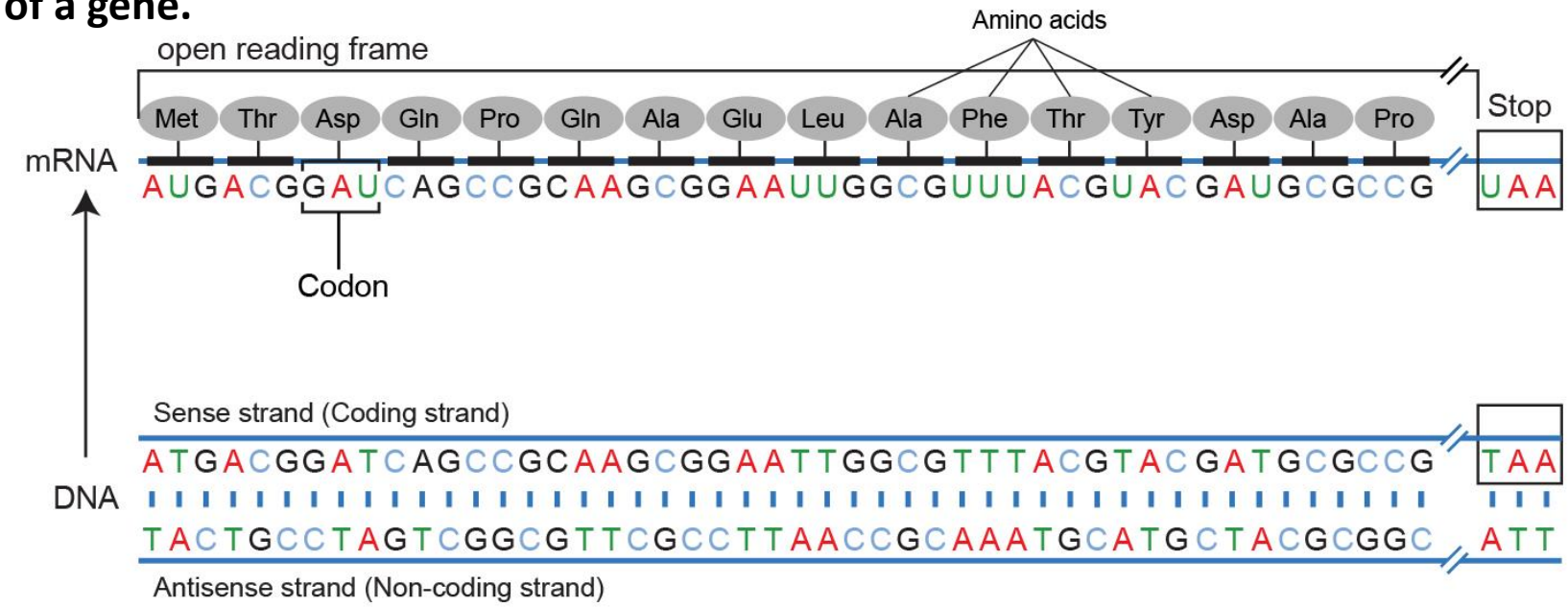
Read about reading frames from: https://www.genome.gov/genetics-glossary/Open-Reading-Frame

Sequence analysis

Database searching and pairwise alignment

BLAST **ORF**

- ✓ **An open reading frame (ORF) is a portion of a DNA molecule that, when translated into amino acids, contains no stop codons.**
- ✓ **The genetic code reads DNA sequences in groups of three base pairs, which means that a double-stranded DNA molecule can read in any of six possible reading frames:**
  - **three in the forward direction (on positive or sense strand)**
  - **and three in the reverse direction (on negative or antisense strand) .**
- ✓ **A long open reading frame is likely part of a gene.**



Source: https://www.genome.gov/genetics-glossary/Open-Reading-Frame

Read about reading frames from:
https://www.genome.gov/genetics-glossary/Open-Reading-Frame

Sequence analysis

BLAST

ORF

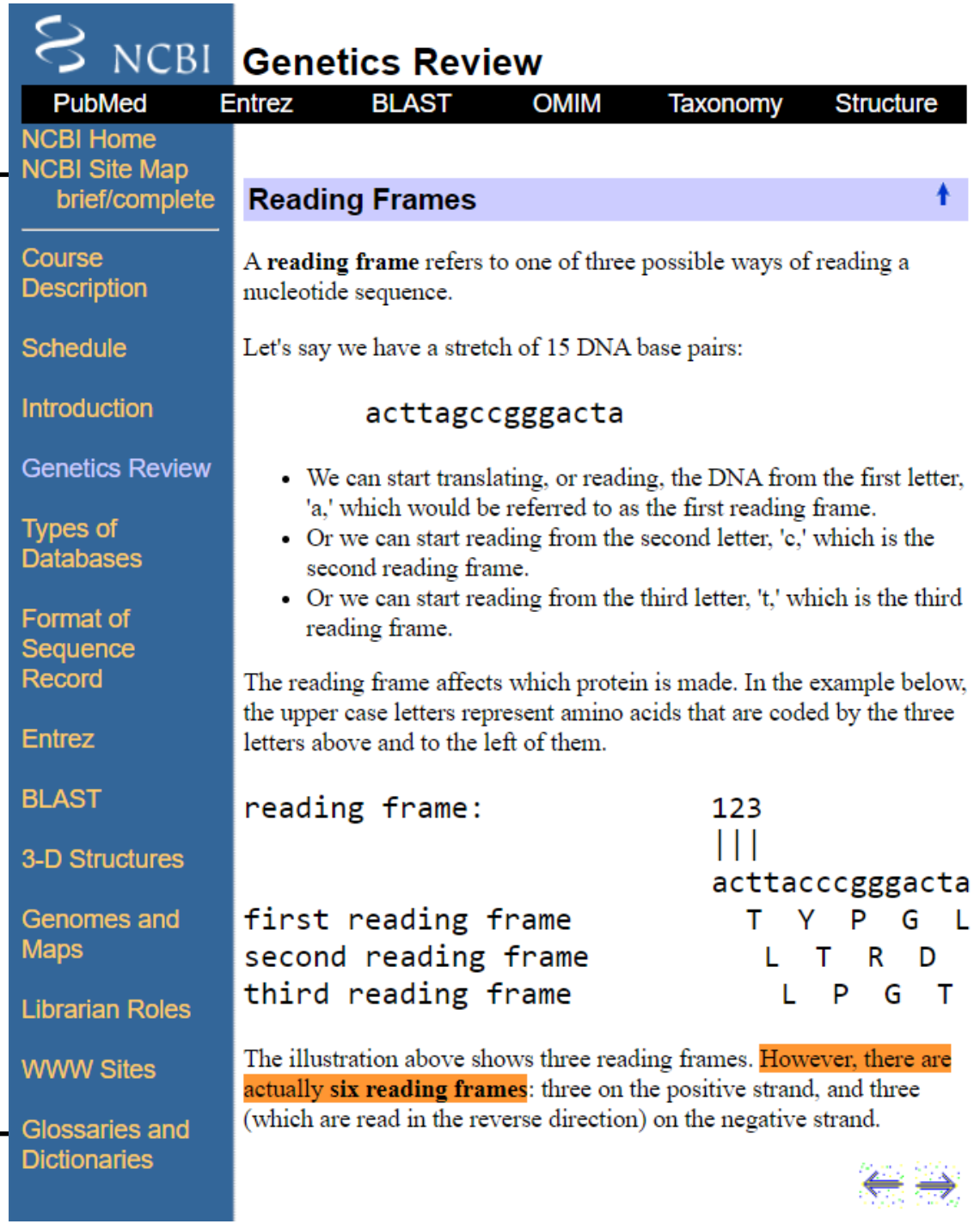
Open Reading Frame Finder (server):

https://www.ncbi.nlm.nih.gov/orffinder/

Read about reading frames from:

Source:

https://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/readingframe.html



NCBI Genetics Review

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI Home
NCBI Site Map brief/complete

Course Description
Schedule
Introduction
Genetics Review
Types of Databases
Format of Sequence Record
Entrez
BLAST
3-D Structures
Genomes and Maps
Librarian Roles
WWW Sites
Glossaries and Dictionaries

Reading Frames

A **reading frame** refers to one of three possible ways of reading a nucleotide sequence.

Let's say we have a stretch of 15 DNA base pairs:

acttagccgggacta

- We can start translating, or reading, the DNA from the first letter, 'a,' which would be referred to as the first reading frame.
- Or we can start reading from the second letter, 'c,' which is the second reading frame.
- Or we can start reading from the third letter, 't,' which is the third reading frame.

The reading frame affects which protein is made. In the example below, the upper case letters represent amino acids that are coded by the three letters above and to the left of them.

```
reading frame:          123
                        |||
                        acttacccgggacta
first reading frame       T   Y   P   G   L
second reading frame        L   T   R   D
third reading frame           L   P   G   T
```

The illustration above shows three reading frames. However, there are actually **six reading frames**: three on the positive strand, and three (which are read in the reverse direction) on the negative strand.

Sequence analysis

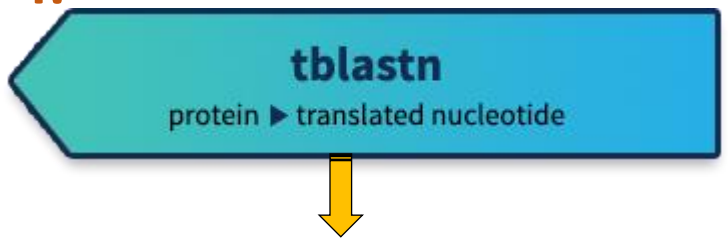
Database searching and pairwise alignment

BLAST Types ⇩ https://blast.ncbi.nlm.nih.gov/Blast.cgi

1.

Known as BLASTN

4.



2.

Known as BLASTP

TBLASTN queries protein sequences to a nucleotide sequence database with the sequences translated in all six reading frames.

Protein query ⇨ **translates in all six reading frames (translated nucleotide)** ⇨ **Search against a nucleotide database**

Source: https://blast.ncbi.nlm.nih.gov/Blast.cgi Essential Bioinformatics by Jin Xiong

Sequence analysis

Database searching and pairwise alignment

BLAST Types ⇩ https://blast.ncbi.nlm.nih.gov/Blast.cgi

TBLASTX uses nucleotide sequences, which are translated in all six frames, to search against a nucleotide sequence database that has all the sequences translated in six frames.

⇧

5. TBLASTX

Nucleotide query ⇨ **translates in all six reading frames (translated protein)** ⇨ **Search against nucleotide database with sequences translated in six frames**

Source: https://blast.ncbi.nlm.nih.gov/Blast.cgi Essential Bioinformatics by Jin Xiong

BLAST

Type of sequences:

- ✓ The choice of the type of sequences also influences the sensitivity of the search.
- ✓ Generally speaking, there is a clear advantage of using protein sequences in detecting homologs.
  - This is because DNA sequences only comprise four nucleotides, whereas protein sequences contain twenty amino acids.
  - This means that there is at least a five-fold increase in statistical complexity for protein sequences.
  - More importantly, amino acid substitution matrices incorporate subtle differences in physicochemical properties between amino acids, meaning that protein sequences are far more informative and sensitive in detection of homologs.
  - This is why searches using protein sequences can yield more significant matches than using DNA sequences.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Database searching and pairwise alignment

BLAST

E-value

(expectation value)

- ✓ The E-value provides information about the likelihood that a given sequence match is purely by chance.
- ✓ The lower the E-value, the less likely the database match is a result of random chance and therefore the more significant the match is.

- The BLAST output provides a list of pairwise sequence matches ranked by statistical significance.
- The significance scores help to distinguish evolutionarily related sequences from unrelated ones.
- Generally, only hits above a certain threshold are displayed.
- Deriving the statistical measure is slightly different from that for single pairwise sequence alignment; the larger the database, the more unrelated sequence alignments there are.
- This necessitates a new parameter that takes into account the total number of sequence alignments conducted, which is proportional to the size of the database.
- In BLAST searches, this statistical indicator is known as the E-value (expectation value), and it indicates the probability that the resulting alignments from a database search are caused by random chance.
- The E-value is related to the P-value used to assess significance of single pairwise alignment.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Database searching and pairwise alignment

BLAST

Low Complexity Regions

These elements in query sequences can cause spurious database matches and lead to artificially high alignment scores with unrelated sequences.

- ✓ For both protein and DNA sequences, there may be regions that contain highly repetitive residues, such as short segments of repeats, or segments that are overrepresented by a small number of residues.
- ✓ These sequence regions are referred to as low complexity regions (LCRs).

- To avoid the problem of high similarity scores owing to matching of LCRs that obscure the real similarities, it is important to filter out the problematic regions in both the query and database sequences to improve the signal-to-noise ratio, a process known as masking.
- There are two types of masking: hard and soft.
  1. Hard masking involves replacing LCR sequences with an ambiguity character such as *N* for nucleotide residues or *X* for amino acid residues.
     - ➢ The ambiguity characters are then ignored by the BLAST program, preventing the use of such regions in alignments and thus avoiding false positives.
     - ➢ However, the drawback is that matching scores with true homologs may be lowered because of shortened alignments.
  2. Soft masking involves converting the problematic sequences to lower case letters, which are ignored in constructing the word dictionary, but are used in word extension and optimization of alignments.

Source:
Essential Bioinformatics by Jin Xiong

Sequence analysis

Database searching and pairwise alignment

BLAST

Tutorial

https://blast.ncbi.nlm.nih.gov/Blast.cgi

1.

Known as BLASTN

3.

4.

2.

Known as BLASTP

5. TBLASTX

Source: https://blast.ncbi.nlm.nih.gov/Blast.cgi

Lecture – Sequence analysis

Database searching and pairwise alignment

BLAST

FASTA

Sequence analysis

Database searching and pairwise alignment

FASTA ➡ **FAST ALL** https://www.ebi.ac.uk/Tools/sss/fasta/

- **FASTA (FAST ALL) was in fact the first database similarity search tool developed, preceding the development of BLAST.**
- **FASTA uses a "hashing" strategy to find matches for a short stretch of identical residues with a length of k.**
- **The string of residues is known as ktuples or ktups, which are equivalent to words in BLAST, but are normally shorter than the words.**
- **Typically, a ktup is composed of two residues for protein sequences and six residues for DNA sequences.**

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

FASTA https://www.ebi.ac.uk/Tools/sss/fasta/

Method

- ✓ **The procedure of ktup identification using the hashing strategy by FASTA.**
- ✓ **Identical offset values between residues of the two sequences allow the formation of ktups.**

1. Given two amino acid sequences for comparision:

	1 2 3 4 5 6 7
sequence 1	A M P S D G L
sequence 2	G P S D N A T

2. Construct a hashing table:

amino acid	sequence position seq 1	sequence position seq 2	offset
A	1	6	-5 (1-6=)
D	5	4	1 (5-4=)
G	6	1	5 (So on...)
L	7	–	–
M	2	–	–
N	–	5	–
P	3	2	1
S	4	3	1
T	–	7	–

3. Identify residues with the same offset values (highlighted in grey).

4. Find the matching word of three residues in the order of 3, 4 and 5 in one sequence and 2, 3,and 4 in the other.

Align:

Positions 3, 4, 5 of sequence 1

Positions 2, 3, 4 of sequence 2

5. This allows establishment of alignment between the two sequences.

	1 2 3 4 5 6 7
sequence 1	A M P S D G L –
	| | |
sequence 2	– G P S D N A T
	1 2 3 4 5 6 7

Source: Essential Bioinformatics by Jin Xiong

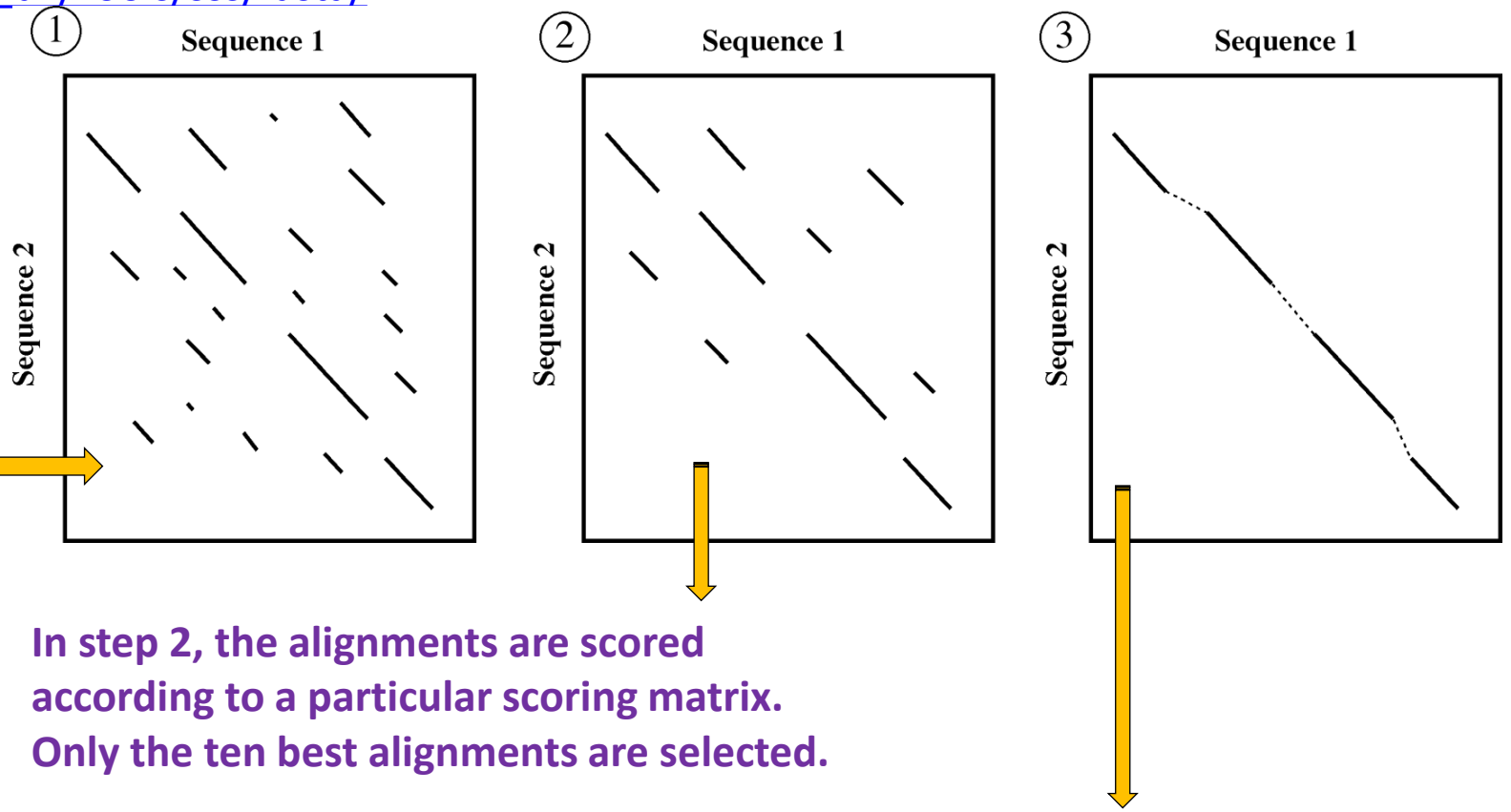
Sequence analysis

Database searching and pairwise alignment

FASTA ➡ **FAST ALL** https://www.ebi.ac.uk/Tools/sss/fasta/

Method – Steps:

Steps of the FASTA alignment procedure.



In step 1, all possible ungapped alignments are found between two sequences with the hashing method.

In step 2, the alignments are scored according to a particular scoring matrix. Only the ten best alignments are selected.

In step 3, the alignments in the same diagonal are selected and joined to form a single gapped alignment, which is optimized using the dynamic programming approach.

Source: Essential Bioinformatics by Jin Xiong

Sequence analysis

Database searching and pairwise alignment

FASTA ➡ **FAST ALL** https://www.ebi.ac.uk/Tools/sss/fasta/

Method – Steps:

1. Identify ktups between two sequences by using the hashing strategy

- **The first step in FASTA alignment is to identify ktups between two sequences by using the hashing strategy.**
- **This strategy works by constructing a lookup table that shows the position of each ktup for the two sequences under consideration.**
- **The positional difference for each word between the two sequences is obtained by subtracting the position of the first sequence from that of the second sequence and is expressed as the offset.**
- **The ktups that have the same offset values are then linked to reveal a contiguous identical sequence region that corresponds to a stretch of diagonal in a two-dimensional matrix**

Source: Essential Bioinformatics by Jin Xiong

FASTA ➡ **FAST ALL** https://www.ebi.ac.uk/Tools/sss/fasta/

Method – Steps:

2. Narrow down the high similarity regions between the two sequences

- The second step is to narrow down the high similarity regions between the two sequences.
- Normally, many diagonals between the two sequences can be identified in the hashing step.
- The top ten regions with the highest density of diagonals are identified as high similarity regions.
- The diagonals in these regions are scored using a substitution matrix.
- Neighboring high-scoring segments along the same diagonal are selected and joined to forma single alignment.
- This step allows introducing gaps between the diagonals while applying gap penalties.
- The score of the gapped alignment is calculated again.

Source: Essential Bioinformatics by Jin Xiong

FASTA **FAST ALL** https://www.ebi.ac.uk/Tools/sss/fasta/

Method – Steps:

3. Refinement of the gapped alignment

- **In step 3, the gapped alignment is refined further using the Smith–Waterman algorithm to produce a final alignment.**

Sequence analysis

Database searching and pairwise alignment

FASTA ➡ **FAST ALL** https://www.ebi.ac.uk/Tools/sss/fasta/

Method – Steps:

4. Statistical Significance

- The last step is to perform a statistical evaluation of the final alignment as in BLAST, which produces the E-value.
- FASTA also uses E-values and bit scores.
- Estimation of the two parameters in FASTA is essentially the same as in BLAST.
- However, the FASTA output provides one more statistical parameter, the Z-score.
- This describes the number of standard deviations from the mean score for the database search.
- Because most of the alignments with the query sequence are with unrelated sequences, the higher the Z-score for a reported match, the further away from the mean of the score distribution, hence, the more significant the match.
- For a Z-score > 15, the match can be considered extremely significant, with certainty of a homologous relationship.
- If Z is in the range of 5 to 15, the sequence pair can be described as highly probable homologs.
- If Z < 5, their relationships is described as less certain.

Source: Essential Bioinformatics by Jin Xiong