**Bioinformatics – Sequence Alignment**

**Transcription of DNA:**

* DNA transcribed into RNA
* RNA exits as a single-strand unit and as a double-helix as well
* RNA consist of A, C, G and U (uracil)

**Types of RNA:**

* Messenger RNA – mRNA
* Transfer RNA – tRNA
* Ribosomal RNA – rRNA

**Translation of mRNA:**

* mRNA is translated into protein

**Proteins:**

* linear polymers built from amino acids

**The transfer of information from DNA to specific protein via RNA takes place according to the genetic code.**

* The RNA sequence is divided into block of three letters – codon
* Each codon corresponds to the specific amino acid
* There are 20 amino acids and 64 = 4\*4\*4 different codons, some codons are redundant and some have special function – to terminate the translation process

**Example:**

DNA: ATAGGA**ATG**CTCGTTACTTCTTCAAAT

RNA: AUAGGA**AUG**CUCGUUACUUCUUCAAAU

Codons used in translation process: AUG CUC GUU ACU UCU UCA AAU

Protein: MLVTFLN

 AUG - methionine – M - met

 CUG - leucine - L - leu

 GUU - valine - V - val

 ACU - threonine - T - thr

 UCU - phenylalanine - F - phe

 UCA - leucine - L - leu

 AAU - asparagine - N - asn

**Sequence Alignment:**

* An **alignment** between two sequences is simply a pairwise match between the characters of each sequence.
* An **alignment** of nucleotide or amino acid sequences is one that reflects the evolutionary relationship between two or more **homologs** (sequences that share a common ancestor)
* An **alignment** between two or more genetic sequences represents a hypothesis about the evolutionary path by which they diverged from a common ancestor.
* Using matrices, we can define **alignment** as a two-row matrix, where the first row is the first string and the second row is the second string
* Columns that contain the same letters in both rows are called **matches**
* Columns that different letters in both rows are called **mismatches** (mutations)
* The columns of the alignment that contains one space ate called **indels**: if the space is in the top row – **insertion**, and if the space is in the bottom row – **deletion**
* ***A***TG (match) A***T***G (mutation) A – G (insertion) A***T***G (deletion)

 ***A***ACA***A***C A***A***C A **-** C

* **Why align sequences?**
	+ **Find the gene purpose**
	+ **Find similar gene in another species**
		- **Align sequences with known genes**
		- **Find the gene with the “best” match**
* Example: consider two sequences: s1: AATCTATA and s2: AAGATA if no gaps are allowed we can align them in the following ways:

 AATCTATA

AAGATA

 AATCTATA

AAGATA

 AATCTATA

AAGATA

* Must to decide how to evaluate or score each alignment
* In the simple, gap-free alignment the score is defined as (where is ***n*** is a length of the longest sequence):

 

 Assume: match score = 1, mismatch score = 0

* Alignment with gaps:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A | A | T | C | T | A | T | A |
| A | A | G | - | A | T | - | A |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | A | T | C | T | A | T | A | - |
| A | A | - | G | A | - | - | T | A |

* A **simple** score for the alignment that allows gaps is defined as: 

Assume: match = 1, mismatch = 0, gap = -1

* **Origination and Length Penalties**
	+ **We want to find alignments that are *evolutionarily likely.***
	+ **We can achieve this by penalizing more for a new gap, than for extending an existing gap**
	+ **Match/mismatch score: +1/+0**
	+ **Origination/length penalty: –2/–1**

Example:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **A** | **A** | **T** | **C** | **T** | **A** | **T** | **A** |
| **A** | **A** | **G** | **-** | **A** | **T** | **-** | **A** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **A** | **A** | **T** | **C** | **T** | **A** | **T** | **A** |
| **A** | **A** | **-** | **G** | **-** | **A** | **T** | **A** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **A** | **A** | **T** | **C** | **T** | **A** | **T** | **A** |
| **A** | **A** | **-** | **-** | **G** | **A** | **T** | **A** |

**See links to learn more about mutations**

**Not all mutations have the same frequencies**

**Substitution Mutations:**

**Type 1:**

[**Transitions**](http://en.wikipedia.org/wiki/Transition_%28genetics%29)are interchanges of two-ring **purines** (**A** **G**) or of one-ring **pyrimidines** (**C** **T**): they therefore involve bases of similar shape. More likely.

**Type 2:** [**Transversions**](http://en.wikipedia.org/wiki/Transversion)are interchanges of **purine for pyrimidine bases**, which therefore involve exchange of **one-ring** and **two-ring** structures: AC or G T. Less likely.

 Transitions are less likely to result in amino acid substitutions ("silent substitutions" or “silent mutations”)