Similarity: de mate waarin 2 sequenties verwant zijn.

Positives: de mate waarin 2 sequenties fysiek op elkaar lijken. **Similarity + identity**

Bit-score: een ruwe alignement score. Ze zijn genormaliseerd waardoor ze makkelijk met andere alignements met andere parameters vergeleken kunnen worden. **Zijn relatief.**

Quantitive score: kwaliteitsscore van een alignement.

E-value: het aantal sequenteis dat je bij dezelfde querylengte en database grootte verwacht te vinden die een zelfde of hogere score hebben dan jou alignemnt. **De E-value is absoluut.**

Hoge E-value 🡪 veel sequenties die randomly een zelfde of hogere score hebben.

Hoge E-value 🡪 hogere kans op insignificant alignements.

Expect threshold: de cut-off E-value. Vanaf welke E-value vertrouw je de resultaten nog?

Gaps: zijn indel events.

**Gap insertion** 🡪 de hogere score van de gap insertino zorgt er voor dat een lange gap sneller gebruikt wordt dan meerdere kleine gaps.

**Gap elongation** 🡪 het verlengen van een gap telt minder zwaar dan het openen van een gap.

Er worden 2 soorten gap penalties gebruikt om alignments en de verdeling van gaps daarin homogener te maken. **Op deze manier is de alignment representatiever evolutionair gezien.**

Query coverage: welk gedeelte an je query matchte met een andere alignement.

Gapsize:

Streng 🡪 de 2 sequenties moeten erg op elkaar lijken.

**Weinig gaps.**

Soepel 🡪 de 2 sequenties hoeven niet erg op elkaar te lijken.

**Veel gaps.**

Homologe sequenties 🡪 2 sequenties met een gemeenschappelijke voorouder. 30% identity is een veilige grens.

**Lager kan ook nog steeds homoloog zijn 🡪 lage conservatie van letters, vorm blijft behouden.**

Orthologen sequenties 🡪 2 homologe sequenties in verschillende soorten, ontstaan door soortvorming.

**De evolutionaire druk op 1 gen is groter, mag niet muteren.**

Paralogen sequenties 🡪 verschillende ‘versies’ van 1 gen in 1 organisme.

**De evolutionaire druk op deze genen is niet groot, er is ruimte voor mutaties.**

PAM: punt accepted matrix, verwante sequenties

PAM1 🡪 1 mutatie op 100 amnizuren

PAM10 🡪 10 mutaties op 100 aminzuren

PAM250 🡪 250 mutaties op 100 aminozren, 1 plek kan meerdere mutaties plaats vinden.

BLOSUM: block substitution matrix, minder tot geen verwante sequenties. Berekend door sequenties met een identity tot 62%.

**Het doel van de blast algoritmes is om een zo hoog mogelijke score te verkrijgen.**

Pairwise alignement : 2 sequenties zo manipuleren dat er een maximale identity bereikt wordt **en conservatie voor eiwitten**. Zo wordt er een simularity bepaald en mogelijk homology. **Gaps worden toegevoegd voor een maximale score te bereiken.**

Proteïne sequenties kunnen informatiever zijn doordat eiwitten beter geconserveerd zijn door silent mutaties. Ook kunnen 2 verschillende eiwitten erg overeenkomen in functie, vorm, lading of hydrofoobiciteit waardoor ze toch verwantschap delen.

Wanneer er meerdere duplicaties van een gen aanwezig zijn wordt de **druk om te muteren kleiner**. Bacteriën hebben geen duplicaties van genen en hebben daarom alleen orthologen sequenties.

Ortholge genen hebben meestal een gelijk moleculaire functie en vaak ook een gelijksoortige biologische rol. Paraloge eiwitten verschillen mogelijk licht in moleculaire functie maar in ieder geval verschillen ze in biologische rol.

Wanneer een sequentie meerdere keren met 1 chromosoom kan alignen (repeterende sequentie) dan zal de score een veelvoud zijn van de max-score.

Matrix scores:

Positief 🡪 grotere kans op een betekenis volle match

Negatief 🡪 grote kans op een random match

Score = 0 🡪 eve grote kans op een betekenis volle match als op een random match.

Je E-value hangt van een paar factoren af:

* Grootte van je database
* Grootte van je query
* Hoogte van je score

Local alignment 🡪 smith waterman, niet onder de 0

Global alinemnt 🡪 neelde man wunsh, wel onder de 0

BLASTX 🡪 nuc tegen eiwit

BLSATn 🡪 nuc tegen nuc

tBLASTx 🡪 zoeken in de nucleotide sequentie database m.b.v. een eiwitsequentie, zoeken in een niet geannotterd/gecalled genoom. **Voor het alsnog identificeren van een gen wat codeert voor een bekend al eiwit.**

BLAST algoritme:

* De sequentie wordt in woorden opgesplitst.
* Deze aparte woorden worden gemuteerd.
* Met de originele woorden en de gemuteerde woorden wordt er in de database naar matches gezocht.
* Wanneer 2 woorden in een sequentie machten, worden de woorden uitgebreid.
* Wanneer de match minder wordt stopt blast met blasten en returnt de hit.

Tools die een transmembraan regio voorspellen zijn gebaseerd op hydrofobiciteit en secundaire structuur (alpha helix) en de lengte van de helix.

Een ontologie is typisch een datastructuur die alle relevante entiteiten en hun onderlinge relaties en regels binnen dat domein bevat. Het is dus een hierarchisch geordend woordenboek van relevante woorden binnen een bepaald domein. Je zou een ontologie willen gebruiken om:

* De functionele hierarchie in de relatie tussen verschillende eiwitten te ontdekken,
* om makkelijk eiwitten te zoeken in een database of
* om gegevens uit verschillende databases aan elkaar te koppelen.