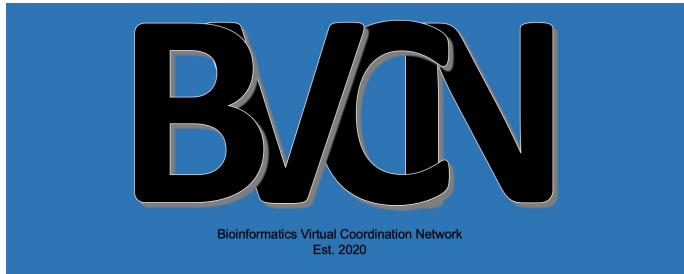
Functional Annotation Lesson 2

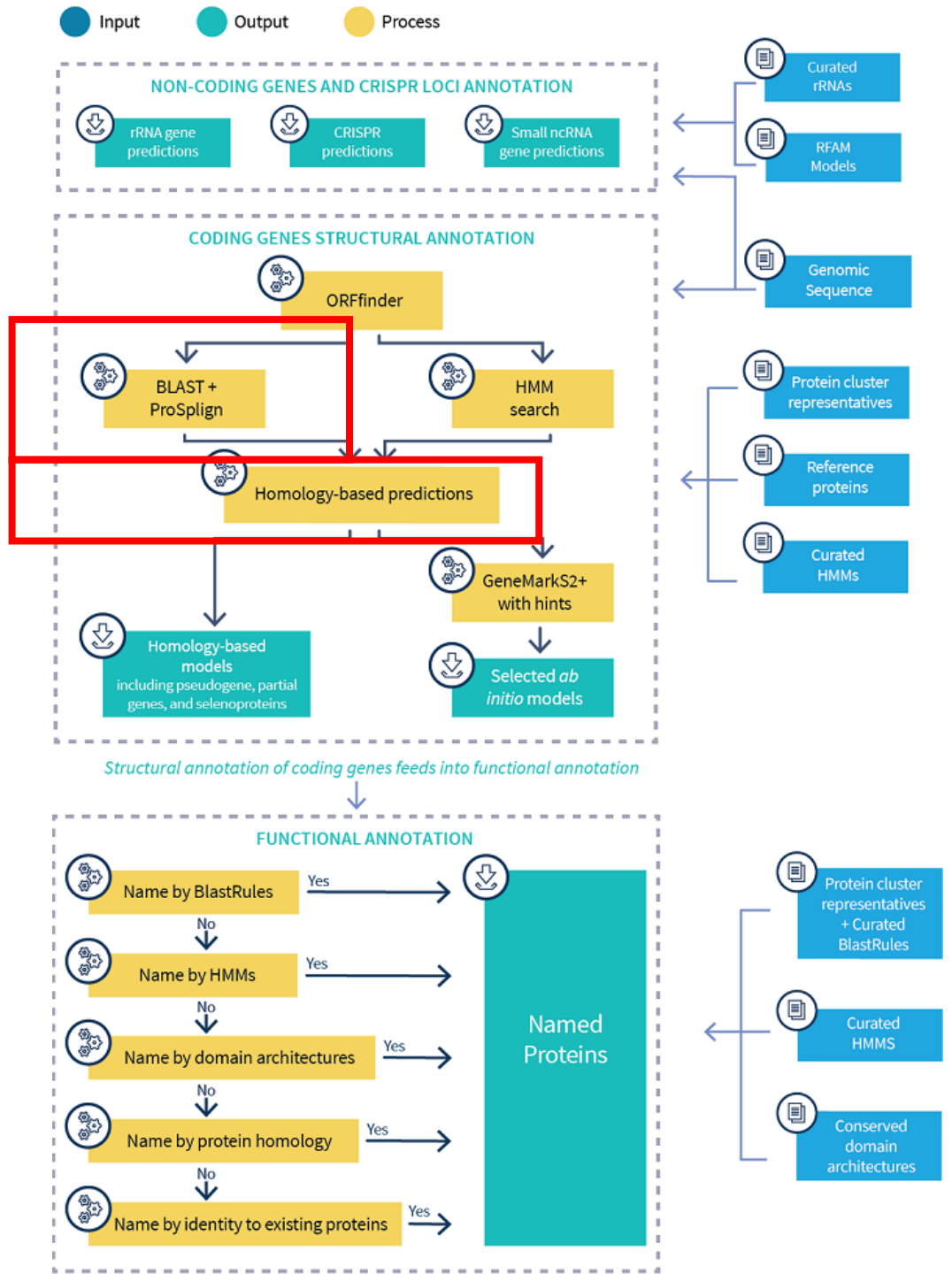
Inferring function from homologous matches

- Limitations of ortholog match
- BLAST+
- DIAMOND

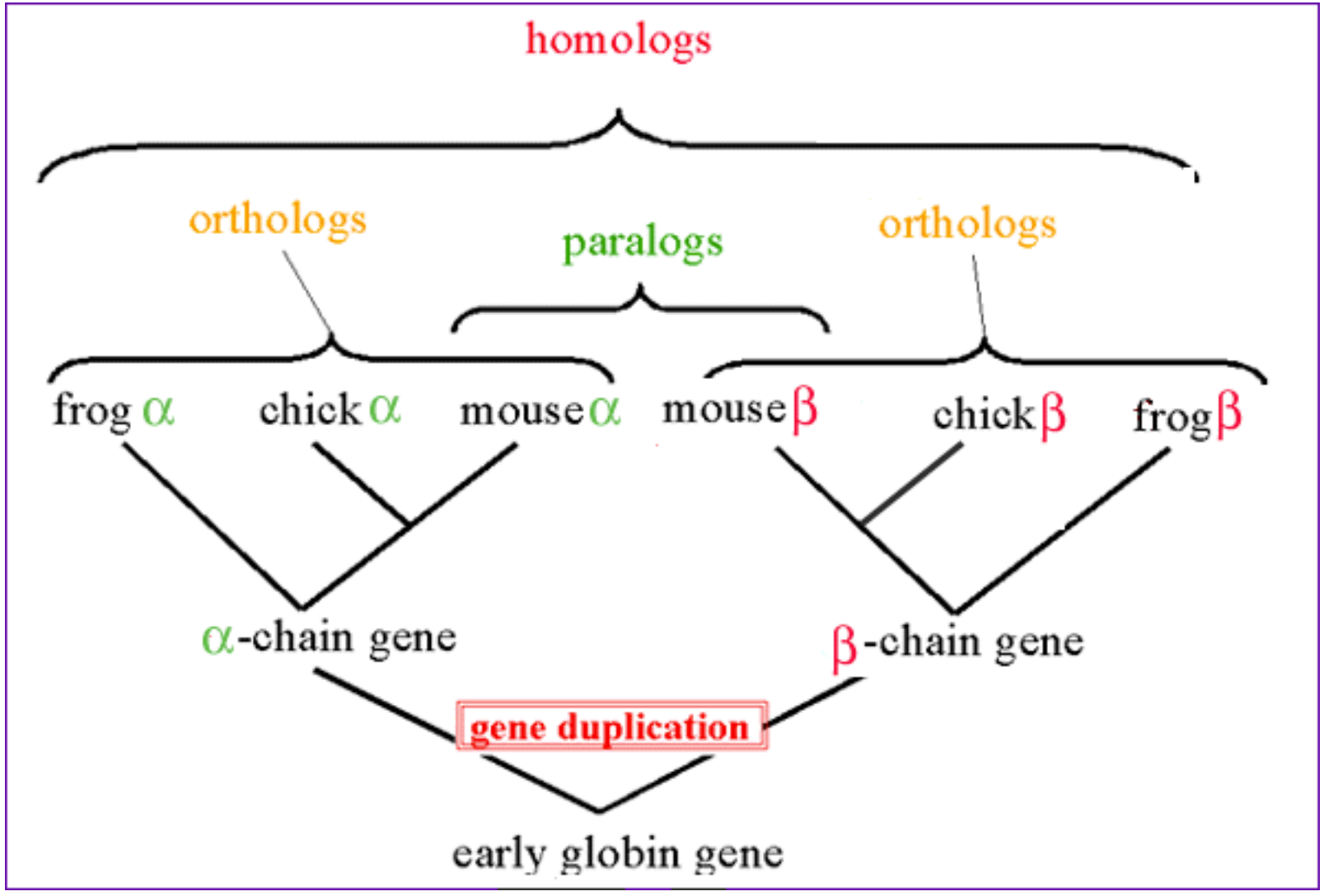


NCBI PGAP

Prokaryotic Genome Annotation Pipeline

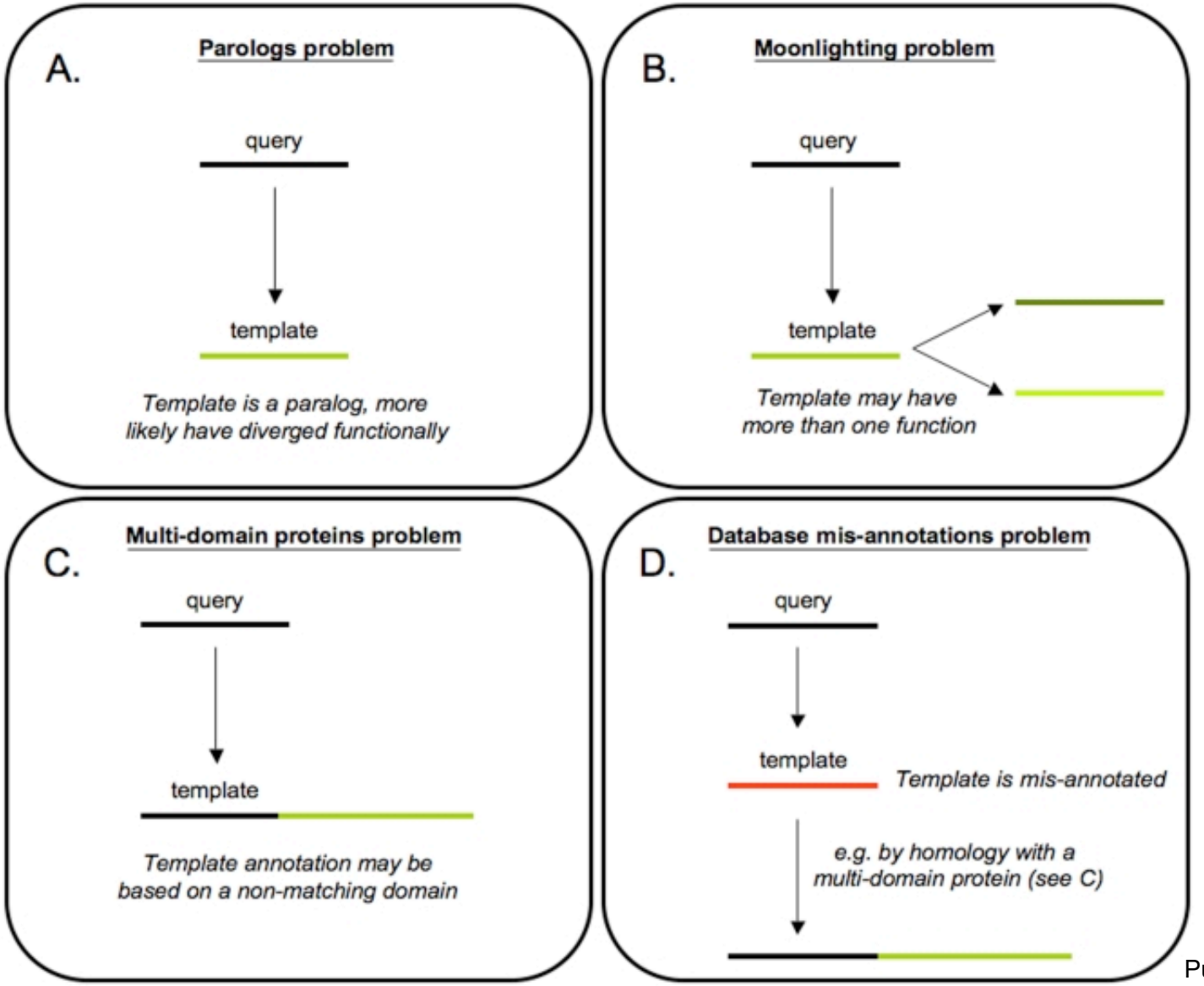


Homology ≠ similarity in function



- No scoring schema provides "biological truth"
- With enough modification - gap inclusions, similarity scores, etc. - ANY pair of sequences can be aligned
- Proteins can have similar functions at 30% amino acid identity
- Proteins can have different functions at 95% amino acid identity
- Context of the match is important

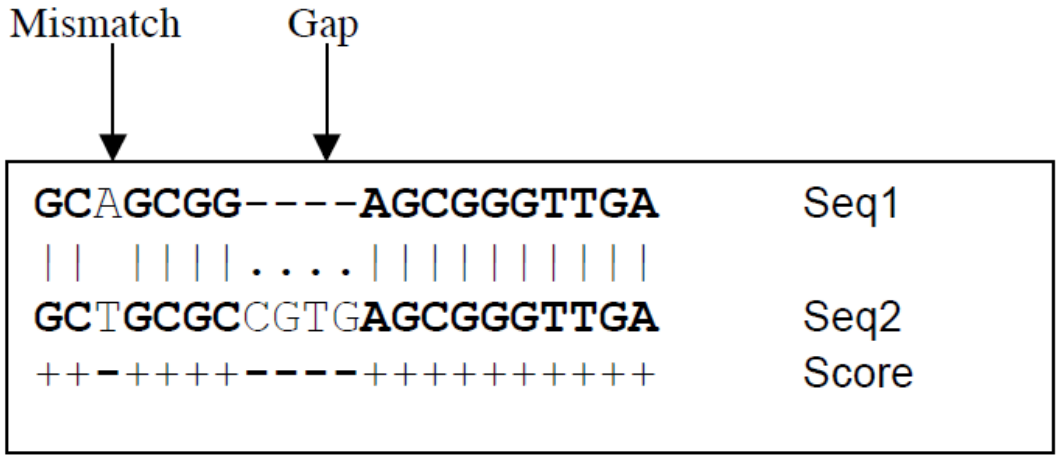
Why homology ≠ similarity in function



Punta & Ofran. PLOS Comp Biol. 2008

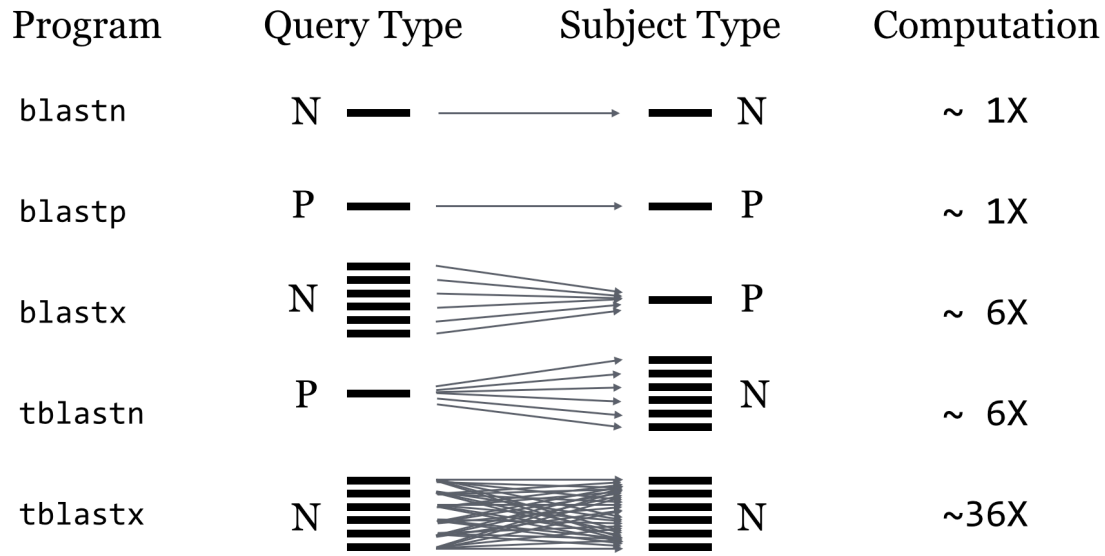
BLAST+

BLAST (Basic Local Alignment Search Tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences (Wikipedia)



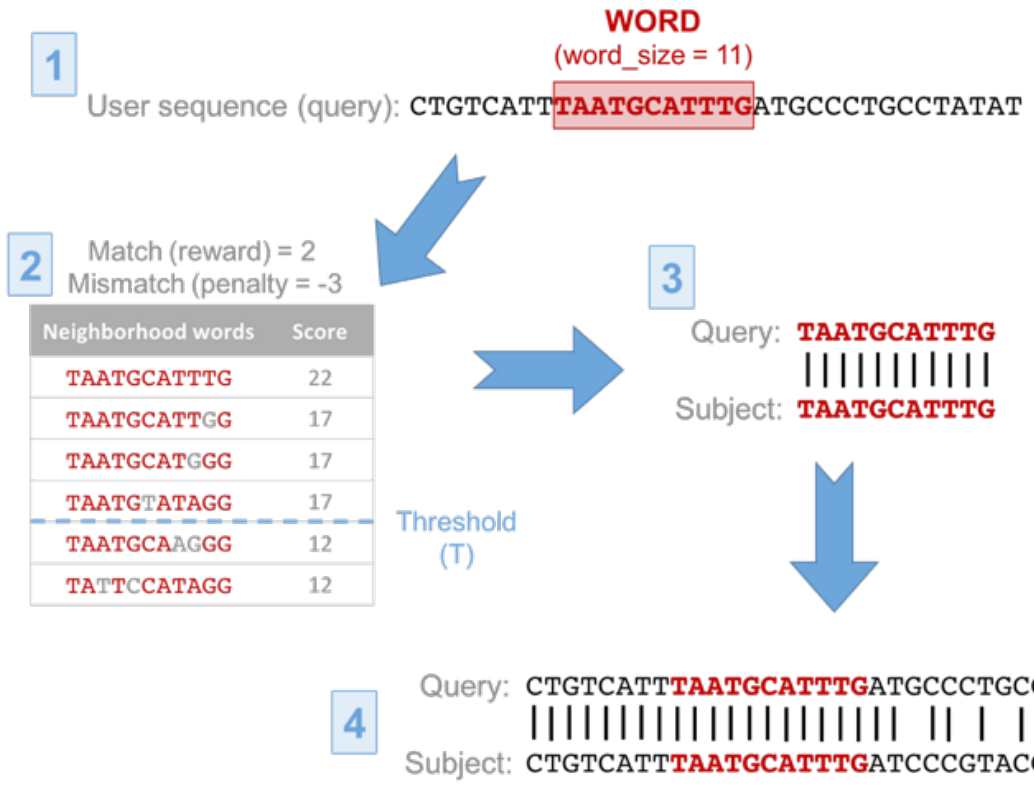
http://www.hypothesisjournal.com/wp-content/uploads/2011/08/boutros-3-1-fig2.png

Program	Input format	Database
blastn	nucleotide	nucleotide
blastp	protein	protein
blastx	translated nucleotide	protein
tblastn	protein	translated nucleotide
tblastx	translated nucleotide	translated nucleotide



https://open.oregonstate.education/computationalbiology/chapter/command-line-blast/

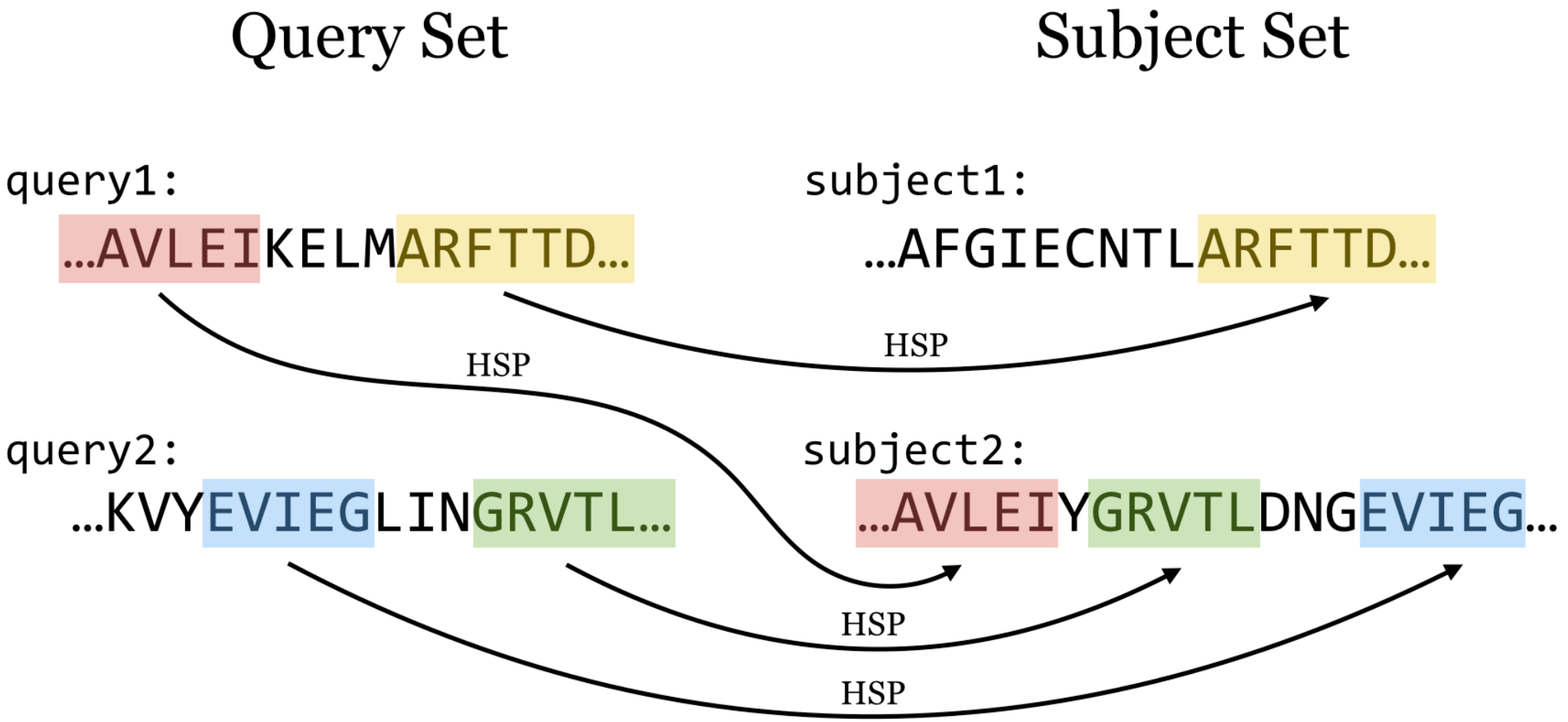
BLAST+



1. BLAST splits the user sequence (**query**) into smaller segments called **words**.
2. These words are then used to search the database in a process called **seeding**. For each word, a series of identical or similar matches (neighborhood words) are retrieved (no gaps allowed) and given a **score**. For nucleotides, this score is based on **match** (**reward**) and **mismatch** (**penalty**) values and substitution matrices for proteins. All matches with scores above the neighborhood **threshold** (**T**) are then kept for extension.
3. An alignment between the user sequence (**query**) and database match (**subject**) is generated.
4. Matches are extended in both directions and scored using the match/mismatch/substitution values with gap open and gap extension penalties.

Task	Description
blastn	Traditional blastn search
blastn-short	Optimized for queries <50bp
megablast	Optimized for seqs with high similarity
dc-megablast	Optimized for distant seqs
blastp	Traditional blastp search
blastp-short	Optimized for queries <30aa

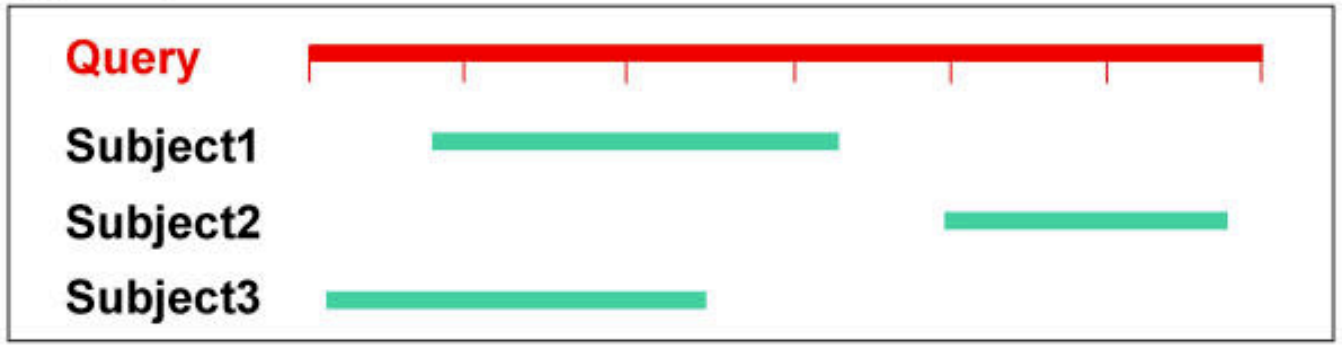
Interpreting local alignments



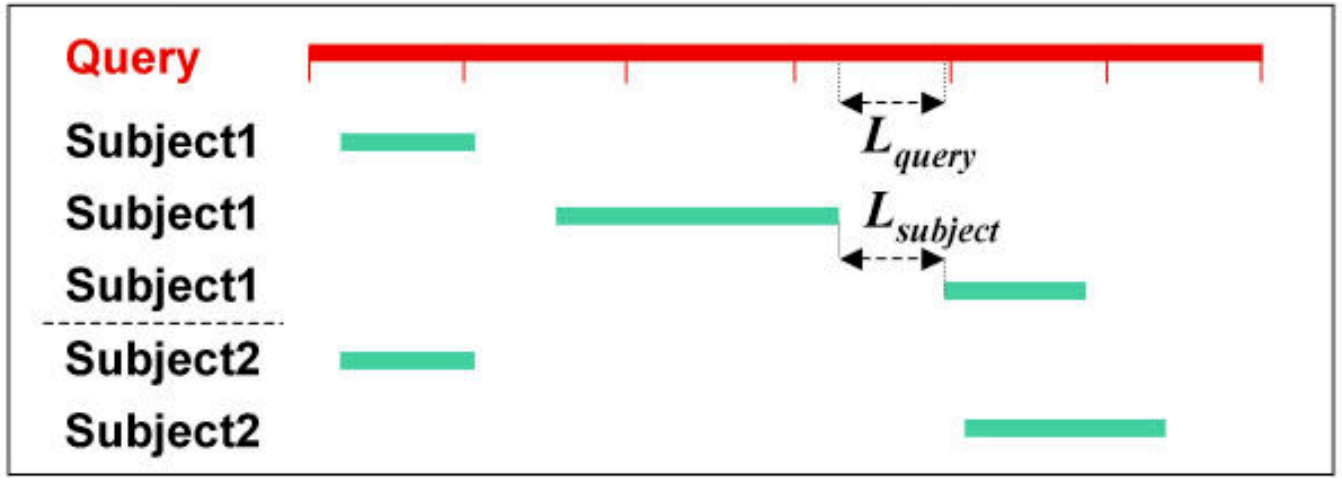
https://open.oregonstate.education/computationalbiology/chapter/command-line-blast/

Interpreting local alignments

Type1 alignment: continuous match



Type2 alignment: discontinuous matches in the same subject



Talla (2003)

Alternative local alignment tools

BLAST (1990)

- Only supports local alignments
- Default e-value 10 – high error rate
- For each query sequence, the reference sequences with the best matches are returned
- Requires BLAST formatted database

USEARCH (2010)

- Supports local and global alignments
  - Global alignments make sense for searches where the similarity of the whole sequence is important
  - e.g. 16S rRNA
- USEARCH for top hit(s) at higher identities
- UBLAST, which is slower but sensitive to lower identities
- Reduces search time by returning only a few high-quality matches rather than considering all possible matches

DIAMOND (2015)

- Aligns short sequence reads at approximately 20,000 times the speed of BLAST and has a **similar** level of sensitivity.
- Constructs a double index to traverse query and reference seeds more quickly
- An 'all mapper' that attempts to determine exhaustively all significant alignments for a given query
- Requires DIAMOND formatted database

Lesson 3

Predicting function using position-sensitive models