Escuela Profesional de
Ciencia de la Computación

ICC Fase 1

Bioinformatics

BLAST

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín de Arequipa

July 4, 2020

Table of Contents

Table of Contents

1 Introduction
- Objectives
- Requeriments of database searching

2 BLAST
- Problem
- BLAST algorithm

vmachacaa@unsa.edu.pe

Introduction

Objectives

- Understand the importance of sequence alignment in Bioinformatics.

Introduction

Objectives

- Understand the importance of sequence alignment in Bioinformatics.
- Understand and implement BLAST algorithm.

Table of Contents

1. Introduction
   - Objectives
   - Requeriments of database searching
2. BLAST
   - Problem
   - BLAST algorithm

vmachacaa@unsa.edu.pe

Requeriments of database searching

- **Sensitivity**.- Ability to find as many correct hits possible (true positives).
- **Selectivity/Specificity**.- Ability to exclude incorrect hits (false positives).
- **Speed**.- Time to take the results.

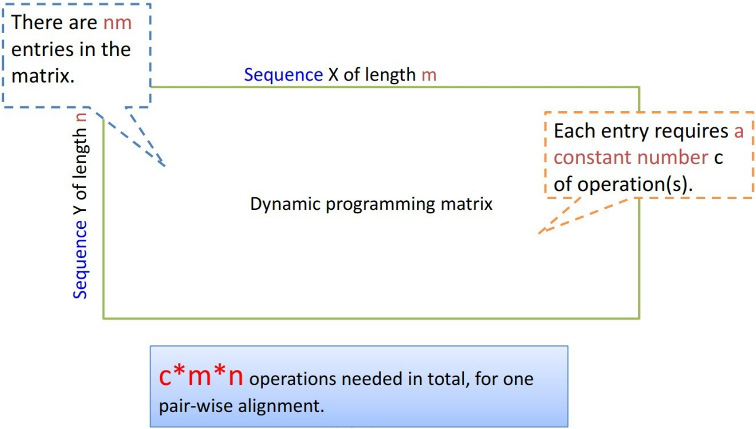
Table of Contents

1 Introduction
- Objectives
- Requeriments of database searching

2 BLAST
- Problem
- BLAST algorithm

vmachacaa@unsa.edu.pe

Problem



There are nm entries in the matrix.

Sequence X of length m

Sequence Y of length n

Dynamic programming matrix

Each entry requires a constant number c of operation(s).

c*m*n operations needed in total, for one pair-wise alignment.

Table of Contents

1. Introduction
   - Objectives
   - Requeriments of database searching
2. BLAST
   - Problem
   - BLAST algorithm

BLAST

Definition

Basic Local alignment Search Tool

Proposed by Altschul in 1990 [1], it use heuristics to reduce time processing in dynamic programming.

BLAST

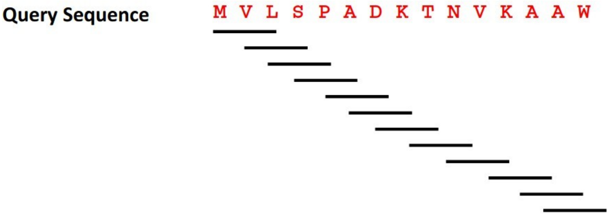
Algorithm

- Given query sequence Q, compile the list of possible words.
- For each word, compute a list of neighbors based on a similarity matrix.
- Scan database for exact matching (hits) with the list of neighbors.
- Extending hits.
- Evaluating significance of extended hits.

BLAST

Seeding

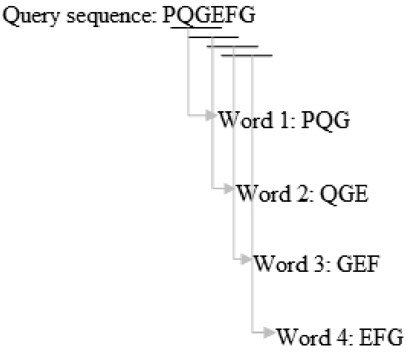
For a given word length *w* (usually 3 for proteins and 11 for nucleotides), slicing the query sequence into multiple continuous **seed words**.



BLAST

Seeding

Examples of words:



Query sequence: PQGEFG

Word 1: PQG

Word 2: QGE

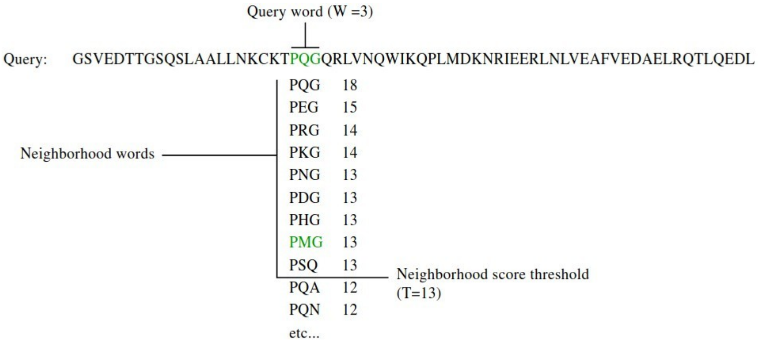
Word 3: GEF

Word 4: EFG

BLAST

Seeding

For each word, compute neighbors (20^3 possibilities). Then score the neighbors (BLOSUM62) and choose the ones that its scores are bigger than T ($T = 13$).



Query: GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

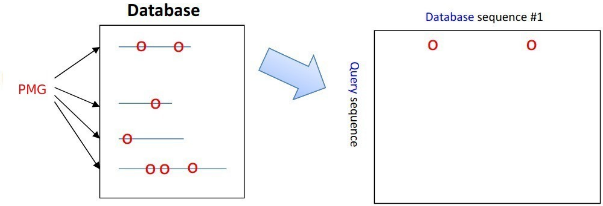
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSQ	13
PQA	12
PQN	12
etc...	

BLAST

Search a sequence database

In sequences database, locate the neighbors. This matches are named: hits.

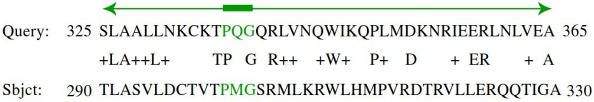
- HashTable: direct addressing method.
- Deterministic finite automaton/finite state machine: much faster.



BLAST

Extending

Extend the hit until the score of the alignment drops below a threshold (22 for proteins and 20 for DNA). The resulting alignment is called high-scoring segment pair (HSP)



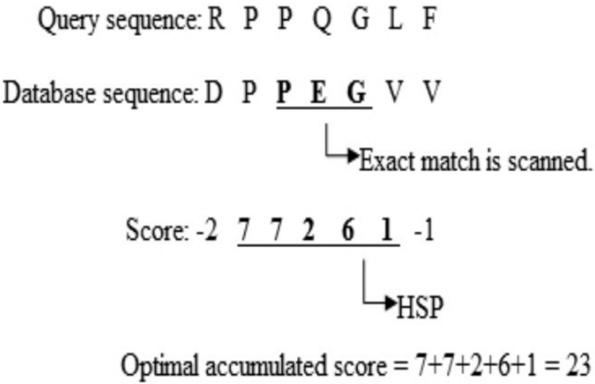
```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

BLAST

Extending

Another example of Extending.



Query sequence: R P P Q G L F

Database sequence: D P $\underline{\text{P E G}}$ V V

Exact match is scanned.

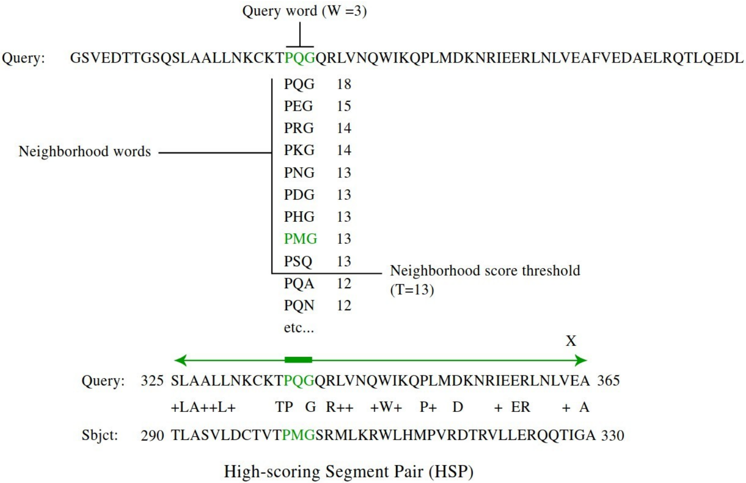
Score: -2 $\underline{7\ 7\ 2\ 6\ 1}$ -1

HSP

Optimal accumulated score = 7+7+2+6+1 = 23

BLAST

Algorithm



Query word (W =3)

Query: GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

Neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSQ	13
PQA	12
PQN	12
etc...	

Neighborhood score threshold (T=13)

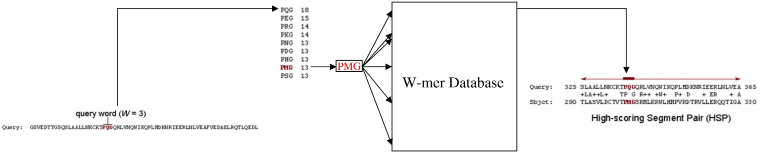
X

```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+     TP  G R++  +W+  P+  D    + ER    + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

BLAST

Algorithm



```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

```
PQG  18
PEG  15
PRG  14
PKG  14
PNG  13
PDG  13
PHG  13
PMG  13
PSG  13
```

```
Query:   325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
             +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:   290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

BLAST

E-value: Significance evaluation

The **E-value** provides information about the likelihood that a given sequence match is purely by chance.

$$E = m * n * p$$

where:

- m: Is the total number of residues in a database.
- n: Is the number of residues in the query sequence.
- p: Is the probability that an HSP alignment is a result of random chance.

BLAST

E-value: Significance evaluation

E	Description
$10 < E$	The sequences under consideration are either unrelated or related by extremely distant relationships.
$0.01 < E < 10$	The match is considered not significant, but may hint at a tentative remote homology relationship.
$1x10^{-50} < E < 0.01$	The match can be considered a result of homology.
$E < 1x10^{-50}$	There should be an extremely high confidence that the database match is a result of homologous relationships.

BLAST

Bit score: Significance evaluation

The **bit score (S')** measures sequence similarity independent of query sequence length and database size and is normalized based on the raw pairwise alignment score.

$$S' = (\lambda * S - lnK)/ln2$$

where:

- λ: Gumble distribution constant.
- S: The raw alignment score.
- K: Constant associated with the scoring matrix used.

The higher the bit score, the more highly significant the match is.

Questions?

References I

- S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.
- S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.
- T. F. Smith, M. S. Waterman *et al.*, “Identification of common molecular subsequences,” *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.
- J. Xiong, *Essential bioinformatics*. Cambridge University Press, 2006.