Bioinformatics course

Short presentation

Adolfo Hoyos[1]

[1] Department of Electronics and Computer Science
Pontificia Universidad Javeriana, Cali

Class work, 2017

Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion
2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion
2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

What to do when we have alignment results and scores?

- Alignment score:
    - By chance?
- Compare with a random sequence.
    - No better? → By chance

What to do when we have alignment results and scores?

- Alignment score:
    - By chance?
- Compare with a random sequence.
    - No better? → By chance

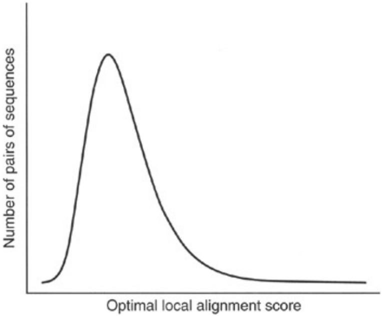
What to do when we have alignment results and scores?

- Alignment score:
  - By chance?
- Compare with a random sequence.
  - No better? → By chance

What to do when we have alignment results and scores?

- Alignment score:
  - By chance?
- Compare with a random sequence.
  - No better? → By chance

Typical scores for random recuences



Comparisson by mean and std. dev.

$$Z - score = \frac{score - mean}{standard\ \ deviation}$$

p-value

E : number of matches with equal similarity.

Similarity ratio

- $> 45\%$: similar structures, common or similar functions
- $> 25\%$: similar folding pattern
- $> 18\%, < 25\%$: "twilight zone"
- $< 18\%$: alignments don't provide much information

Similarity ratio

- $> 45\%$: similar structures, common or similar functions
- $> 25\%$: similar folding pattern
- $> 18\%, < 25\%$: “twilight zone”
- $< 18\%$: alignments don’t provide much information

Similarity ratio

- $> 45\%$: similar structures, common or similar functions
- $> 25\%$: similar folding pattern
- $> 18\%, < 25\%$: “twilight zone”
- $< 18\%$: alignments don’t provide much information

Similarity ratio

- $> 45\%$: similar structures, common or similar functions
- $> 25\%$: similar folding pattern
- $> 18\%, < 25\%$: “twilight zone”
- $< 18\%$: alignments don't provide much information

twilight zone

- “Icebergs”
- shared ligands
- known structures

twilight zone

- “Icebergs”
- shared ligands
- known structures

twilight zone

- “Icebergs”
- shared ligands
- known structures

Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion

2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

Control populations

E : number of matches with equal similarity.

Randomized permutations

E : number of matches with equal similarity.

z-score

How unusual is our original match is?

$$Z - score = \frac{score - mean}{standard\ \ deviation}$$

- $= 0$: No better
- ≥ 5 : Significant

z-score

How unusual is our original match is?

$$Z - score = \frac{score - mean}{standard\ \ deviation}$$

- $= 0$: No better
- ≥ 5 : Significant

z-score

How unusual is our original match is?

$$Z - score = \frac{score - mean}{standard\ \ deviation}$$

- $= 0$: No better
- ≥ 5 : Significant

p-value

is our match worse than random?

- $P \leq 10^{-100}$: exact match
- $10^{-100} < P \leq 10^{-50}$: nearly identical
- $10^{-50} < P \leq 10^{-10}$: closely related
- $10^{-5} < P \leq 10^{-1}$: distant relatives
- $P > 10^{-1}$: probably insignificant

p-value

is our match worse than random?

- $P \leq 10^{-100}$: exact match
- $10^{-100} < P \leq 10^{-50}$: nearly identical
- $10^{-50} < P \leq 10^{-10}$: closely related
- $10^{-5} < P \leq 10^{-1}$: distant relatives
- $P > 10^{-1}$: probably insignificant

p-value

is our match worse than random?

- $P \leq 10^{-100}$: exact match
- $10^{-100} < P \leq 10^{-50}$: nearly identical
- $10^{-50} < P \leq 10^{-10}$: closely related
- $10^{-5} < P \leq 10^{-1}$: distant relatives
- $P > 10^{-1}$: probably insignificant

p-value

is our match worse than random?

- $P \leq 10^{-100}$: exact match
- $10^{-100} < P \leq 10^{-50}$: nearly identical
- $10^{-50} < P \leq 10^{-10}$: closely related
- $10^{-5} < P \leq 10^{-1}$: distant relatives
- $P > 10^{-1}$: probably insignificant

p-value

is our match worse than random?

- $P \leq 10^{-100}$: exact match
- $10^{-100} < P \leq 10^{-50}$: nearly identical
- $10^{-50} < P \leq 10^{-10}$: closely related
- $10^{-5} < P \leq 10^{-1}$: distant relatives
- $P > 10^{-1}$: probably insignificant

e-value

How many seq.s have the same z-score?

- $E \leq 0.02$: probably homologous
- $0.02 < E \leq 1$: cannot be ruled out
- $E > 1$: this could happen by chance

e-value

How many seq.s have the same z-score?

- $E \leq 0.02$: probably homologous
- $0.02 < E \leq 1$: cannot be ruled out
- $E > 1$: this could happen by chance

e-value

How many seq.s have the same z-score?

- $E \leq 0.02$: probably homologous
- $0.02 < E \leq 1$: cannot be ruled out
- $E > 1$: this could happen by chance

Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion

2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

Conclussion

- Statistics are a useful guide
- not a substitute for
    - thinking carefully about the results
    - analysis of ones that look promising

Conclussion

- Statistics are a useful guide
- not a substitute for
  - thinking carefully about the results
  - analysis of ones that look promising

Conclussion

- Statistics are a useful guide
- not a substitute for
  - thinking carefully about the results
  - analysis of ones that look promising

Conclussion

- Statistics are a useful guide
- not a substitute for
  - thinking carefully about the results
  - analysis of ones that look promising

Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion
2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

Not the best

- Big databases
- Not always feasible to find the best
- Heuristics

Not the best

- Big databases
- Not always feasible to find the best
- Heuristics

Not the best

- Big databases
- Not always feasible to find the best
- Heuristics

Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion
2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

BLAST

- Aminoacid sequences (proteins)
- Maximal ungapped local alignments
- Finds subsequences similiar to others in the query

BLAST

- Aminoacid sequences (proteins)
- Maximal ungapped local alignments
- Finds subsequences similiar to others in the query

BLAST

- Aminoacid sequences (proteins)
- Maximal ungapped local alignments
- Finds subsequences similiar to others in the query

BLASTP algorithm

- Partition of query into all posible sequential words
  - default length $= 4$
- Discard words made of mostly common aminoacids
- For every word
  - look for matches
  - when there is a match
    - extend matching until the matching score falls below a given threshold.

BLASTP algorithm

- Partition of query into all posible sequential words
    - default length $= 4$
- Discard words made of mostly common aminoacids
- For every word
    - look for matches
    - when there is a match
        - extend matching until the matching score falls below a given threshold.

BLASTP algorithm

- Partition of query into all posible sequential words
  - default length $= 4$
- Discard words made of mostly common aminoacids
- For every word
  - look for matches
  - when there is a match
    - extend matching until the matching score falls below a given threshold.

BLASTP algorithm

- Partition of query into all posible sequential words
  - default length $= 4$
- Discard words made of mostly common aminoacids
- For every word
  - look for matches
  - when there is a match
    - extend matching until the matching score falls below a given threshold.

BLASTP algorithm

- Partition of query into all posible sequential words
    - default length $= 4$
- Discard words made of mostly common aminoacids
- For every word
    - look for matches
    - when there is a match
        - extend matching until the matching score falls below a given threshold.

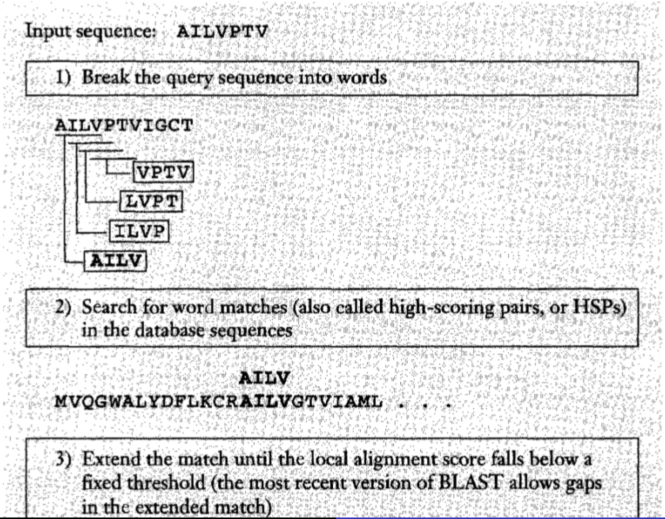
BLASTP algorithm

- Partition of query into all posible sequential words
    - default length $= 4$
- Discard words made of mostly common aminoacids
- For every word
    - look for matches
    - when there is a match
        - extend matching until the matching score falls below a given threshold.

BLASTP algorithm

- Partition of query into all posible sequential words
    - default length $= 4$
- Discard words made of mostly common aminoacids
- For every word
    - look for matches
    - when there is a match
        - extend matching until the matching score falls below a given threshold.

BLASTP algorithm



Input sequence: AILVPTV

1) Break the query sequence into words

2) Search for word matches (also called high-scoring pairs, or HSPs) in the database sequences

3) Extend the match until the local alignment score falls below a fixed threshold (the most recent version of BLAST allows gaps in the extended match)

Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion
2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

FASTA

- Find gapped matches
- Slower than BLAST
    - Considered more sentivie

FASTA

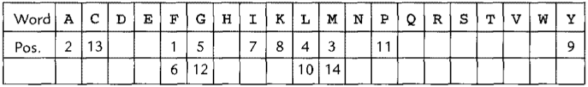
- Find gapped matches
- Slower than BLAST
    - Considered more sentivie

FASTA

- Find gapped matches
- Slower than BLAST
  - Considered more sentivie

FASTA's steps

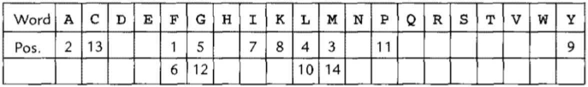
- Partition of sequences in all possible sequential words
  - Generally words of length 4 to 6
- Table the positions of aminoacids from the query sequence



Word	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Pos.	2	13			1	5		7	8	4	3		11							9
					6	12				10	14									

FASTA's steps

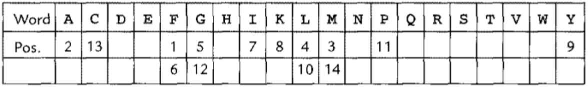
- Partition of sequences in all possible sequential words
  - Generally words of length 4 to 6
- Table the positions of aminoacids from the query sequence



Word	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Pos.	2	13			1	5		7	8	4	3		11							9
					6	12				10	14									

FASTA's steps

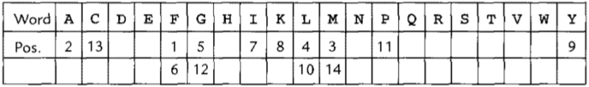
- Partition of sequences in all possible sequential words
  - Generally words of length 4 to 6
- Table the positions of aminoacids from the query sequence



Word	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Pos.	2	13			1	5		7	8	4	3		11							9
					6	12				10	14									

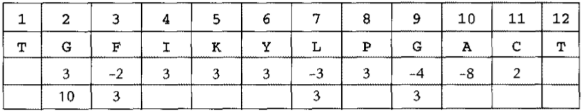
FASTA

- Subtract the positions of aminoacids of the target sequence from the values of those aminoacids in the query sequence tabble
    - Query sequence table:



Word	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Pos.	2	13			1	5		7	8	4	3		11							9
					6	12				10	14									

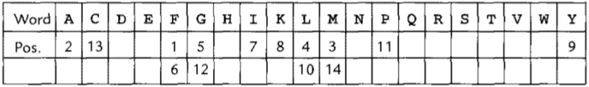
    - Target sequence table:



1	2	3	4	5	6	7	8	9	10	11	12
T	G	F	I	K	Y	L	P	G	A	C	T
	3	-2	3	3	3	-3	3	-4	-8	2	
	10	3				3		3			

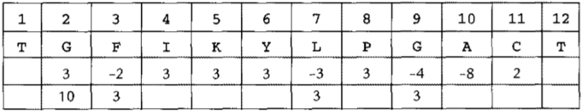
FASTA

- Subtract the positions of aminoacids of the target sequence from the values of those aminoacids in the query sequence tabble
    - Query sequence table:



Word	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Pos.	2	13			1	5		7	8	4	3		11							9
					6	12				10	14									

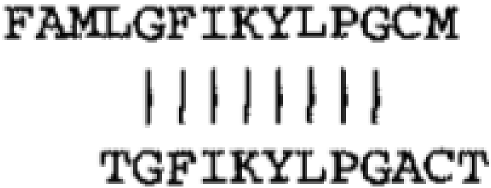
-
    - Target sequence table:



1	2	3	4	5	6	7	8	9	10	11	12
T	G	F	I	K	Y	L	P	G	A	C	T
	3	-2	3	3	3	-3	3	-4	-8	2	
	10	3				3		3			

FASTA

- Count the ocurrence of specific numbers
- The most recurrent represent shifts that bring good alignments between the sequences



Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion
2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

Alignment scores

- Cannot be asumed to be related
- Alignment score are not sufficient
    - Results by chance
        - p-values, e-values, etc.

Alignment scores

- Cannot be asumed to be related
- Alignment score are not sufficient
  - Results by chance
    - p-values, e-values, etc.

Alignment scores

- Cannot be asumed to be related
- Alignment score are not sufficient
  - Results by chance
    - p-values, e-values, etc.

Alignment scores

- Cannot be asumed to be related
- Alignment score are not sufficient
  - Results by chance
    - p-values, e-values, etc.

Outline

1. Significance of Alignments
   - The basics
   - A guide to not getting burnt
   - Conclussion
2. Databases Searches
   - The basics
   - BLAST algorithms
   - FASTA algorithms
   - Significance of results
   - Conclussion

Conclussion

- Brute force algorithms may not be feasible
- Different algorithms can work better than others depending on application
- Aligment scores do not give definitive answers

Conclussion

- Brute force algorithms may not be feasible
- Different algorithms can work better than others depending on application
- Aligment scores do not give definitive answers

Conclussion

- Brute force algorithms may not be feasible
- Different algorithms can work better than others depending on application
- Aligment scores do not give definitive answers

References I

- Krane, Dan E.
  Fundamental concepts of bioinformatics.
  Pearson Education India, 2003.

- Lesk, Arthur.
  Introduction to bioinformatics.
  Oxford University Press, 2013.