

Prof. João Carlos Setubal

QBQ 204 – Aula 7 (biomol)

Bioinformática



Universidade de São Paulo
Instituto de Química

Montagem de DNA

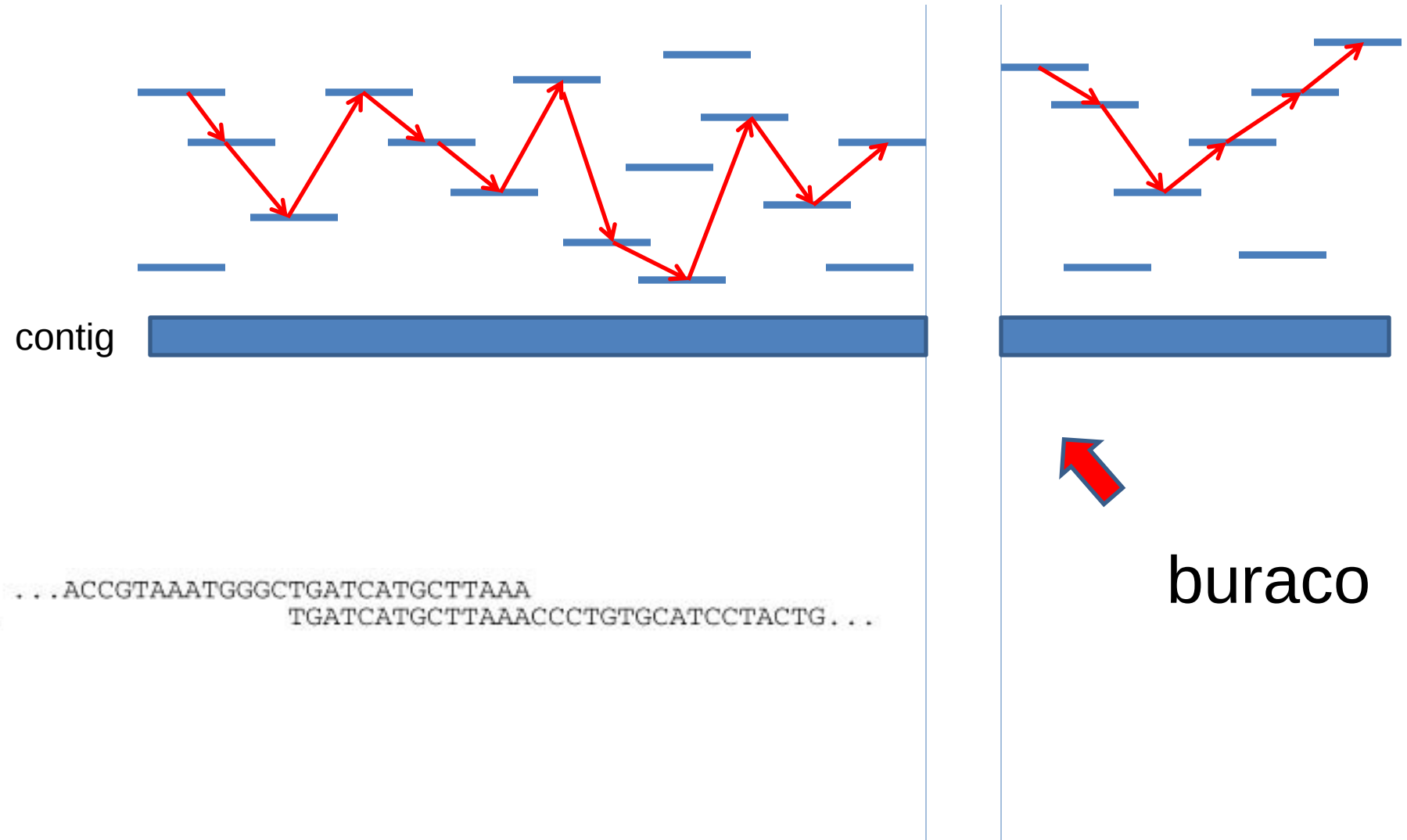
Passo básico: determinar **sobreposição**
entre fragmentos por **comparação de**
sequencias

Quebra-cabeças **linear**

```
...ACCGTAAATGGGCTGATCATGCTTAAA  
TGATCATGCTTAAACCCTGTGCATCCTACTG...
```

Supõe grande redundância no sequenciamento e cortes aleatórios

Montagem de DNA



Comparação de sequências

- Montagem exige um **algoritmo** que seja capaz de comparar sequências entre si
- Algoritmos são implementados em **programas**
- **Bioinformática** é o nome que se usa para o desenvolvimento e aplicação de algoritmos e programas para problemas em biologia, especialmente biologia molecular

Anotação

- Achar genes que codificam proteína
- Achar genes de RNA
- Inferir função das proteínas

Método (rudimentar) para achar genes em procariotos

Ache todas as ORFs com pelo menos 900 bp

DNA de bactéria

...

AGCTCGCGCTCCGCATCCATCCAGTAGGGTTCGGTGTGACGAGCGTGCC
GTCCATATCCCAGAAGACGGCGGCCGGCATCGCGTGCGGAGTCAGTTCGG
TCACGGCTGACAAGTCTATCCCGGCGGCCCGGGCCTATTCTTGAGGGAC
GGCGTCCTGACCGGTGCGCGGATGAAAGGACCAGAACGCCCCGTGACTGA
CGCGAACAGCATCCTCGGAGGGCGCATCCTCGTGGTGGCCTTCGAAGGGT
GGAACGACGCTGGCGAGGCCGCCAGCGGGGCCGTCAAGACGCTCAAGGAC
CAGCTGGATGTGCTCCCGGTGCGCGAGGTGATCCCGAGCTGTACTTCGA
CTTCCAGTTCAACCGGCCGGTCTGTCGCGGACGACGACGGCCGCCGGCGCC
TCATCTGGCCGTCCGCGGAGATCCTGGGCCAGCTCGCCCCGGCGACACC
GGCGATGCGCGCCTGGACGCCACCGGCCCAACGCGGGCAATATCTTCCT
TCTCCTCGGCACCGAGCCGTGCGCGAGCTGGCGCAGCTTCACCGCGGAGA
TCATGGATGCGGCCCTGGCCTCCGACATCGGCGCCATCGTCTTCCTCGGT
GCGATGCTGGCGGACGTACCGCACACCCGCCCATCTCCATCTTCGCTTC
GAGCGAGAACGCGGCCGTCCGTGCGGAGCTCGGCATCGAACGCTCTTCGT
ACGAGGGGCCGGTTCGGTATCCTGAGCGCGCTCGCCGAAGGGGCGGAGGAC
GTGGGCATTCCGACCATCTCCATCTGGGCGTTCGGTTCCGCACTATGTCCA
CAATGCGCCCAGCCCGAAGGCGGTGCTCGCACTGATCGACAAGCTCGAAG
AGCTGGTGAATGTCACCATCCCGCGTGGCTCGCTGGTGGAGGAGGCCACG
GCCTGGGAAGCCGGGATCGACGCGTGGCTCTGGACGACGACGAGATGGC
TACGTACATCCAGCAGCTGGAGCAGGCACGCGACACCGTGGACTCCCCTG
AGGCCAGCGGCGAGGCGATCGCCAGGAGTTCGAGCGCTACCTCCGCCGC
CGCGACGGCCGCGCCGGCGATGACCCCCGCGTGGCTGACGTCACCCCCT
CTCTGCGTCCGCCGTCTCTGTTCCCCCGCTCGGCCTCCCCTGAGGCCG
AGGAGTCGCGCCACATGCCGAAACTCCTCCTTTCCTGACTTTCTGGAG ...

Um gene (CDS)

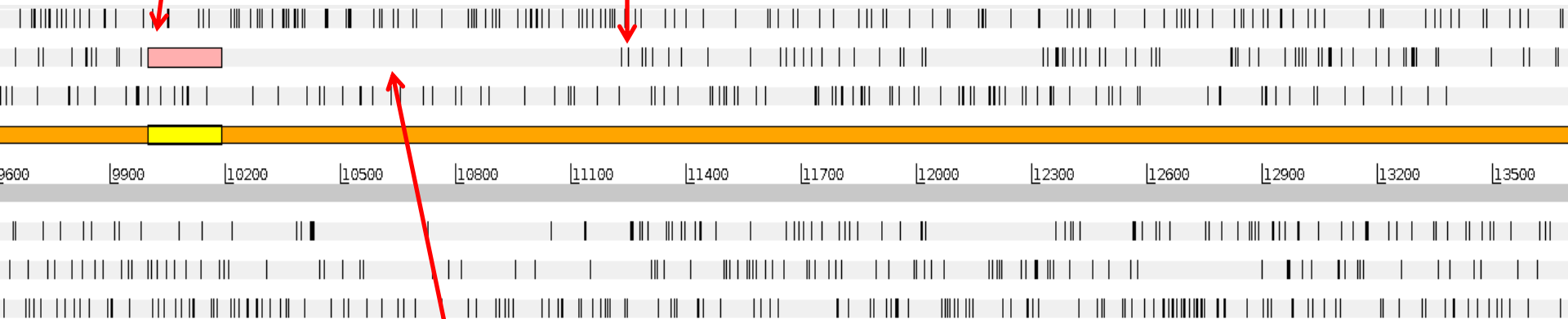
...

AGCTCGCGCTCCGCATCCATCCAGTAGGGTTCGGTGTGACGAGCGTGCC
GTCCATATCCCAGAAGACGGCGGCCGGCATCGCGTGCGGAGTCAGTTCGG
TCACGGCTGACAAGTCTATCCCGGCGGCCCGGGCCTATTCTTGAGGGAC
GGCGTCCTGACCGGTGCGCGGATGAAAGGACCAGAACGCCCGTGACTGA
CGCGAACAGCATCCTC**GGAGG**GCGCATCCTC**GTGGTGGCCTTCGAAGGGT**
GGAACGACGCTGGCGAGGCCGCCAGCGGGGCCGTCAAGACGCTCAAGGAC
CAGCTGGATGTCGTCCCGGTGCGCGAGGTGCATCCCGAGCTGTACTTCGA
CTTCCAGTTCAACCGGCCGGTTCGTGCGGACGACGACGGCCGCCGGCGCC
TCATCTGGCCGTCCGCGGAGATCCTGGGCCAGCTCGCCCCGGCGACACC
GGCGATGCGCGCCTGGACGCCACCGGCCGCAACGCGGGCAATATCTTCCT
TCTCCTCGGCACCGAGCCGTGCGCGAGCTGGCGCAGCTTCACCGCGGAGA
TCATGGATGCGGCCCTGGCCTCCGACATCGGCGCCATCGTCTTCCTCGGT
GCGATGCTGGCGGACGTACCGCACACCCGCCCATCTCCATCTTCGCTTC
GAGCGAGAACGCGGCCGTCCGTGCGGAGCTCGGCATCGAACGCTCTTCGT
ACGAGGGGCCGGTTCGGTATCCTGAGCGCGCTCGCCGAAGGGGCGGAGGAC
GTGGGCATTCCGACCATCTCCATCTGGGCGTCGGTTCCGCACTATGTCCA
CAATGCGCCCAGCCCGAAGGCGGTGCTCGCACTGATCGACAAGCTCGAAG
AGCTGGTGAATGTCACCATCCCGCGTGGCTCGCTGGTGGAGGAGGCCACG
GCCTGGGAAGCCGGGATCGACGCGTGGCTCTGGACGACGACGAGATGGC
TACGTACATCCAGCAGCTGGAGCAGGCACGCGACACCGTGGACTCCCCTG
AGGCCAGCGGCGAGGCGATCGCCAGGAGTTCGAGCGCTACCTCCGCCGC
CGCGACGGCCGCGCCGGCGATGACCCCCGCCGTGGCTGACGTCACCCCT
CTCTGCGTCCGCCGTCTCTGTTCCCCCGCTCGGCCTCCCCTGAGGCCG
AGGAGTCGCGCCACATGCCGAACTCCTCCTTTCTGACTTTCTGGAG ...

início

fim

Tracinhos verticais são stops



L F N A F I + D I L I # D K R N Y F I * # S K F A N H F L L F Y N S I I D N H I F Y Y I F G * F I S # K N S N * Y Y D Q F C H I I
F L M R L F R I F L S K I K G I I L S D K A S L L I I F C Y S I I P L L I I I F F I I S L V D L Y H K K I A I D T T I N S V I L
T F # C V Y L G Y S Y L R # K E L F Y I K Q V C # S F F V I L # F H Y * # S Y F L L Y L W L I Y I I K K + Q L I L R S I L S Y Y
CTTTTAAATGCGTTTATTAGGATATCTTATCTAAGATAAAAGGAATTATTTATCTGATAAAGCAAGTTTGTCTAATCATTTTTTGTATTCTATAATCCATTATTGATAATCATATTTTATTATATCTTTGGTTGATTATATCATAAAAAATAGCAATTGATACTACGATCAATTCTGTCATATTAT
10000 10020 10040 10060 10080 10100 10120 10140 10160 10180
GAAAAATTACGCAATAAATCCTATAAGAATAGATTCTATTTCTTAATAAAATAGACTATTTCTGTTCAAAAGATTAGTAAAAACAATAAGATATTAAGGTAATAACTATTAGTATAAAAAATAATATAGAAACCACTAAATATAGTATTTTTTATCGTTAACTATGATGCTAGTTAAGACAGTATAATA
K K I R K N L I N K D L I F P I I K D S L A L K S I M K Q # E I I G N N I I M N K I I D K T S K Y * L F I A I S V V I L E T M N D
K # H T # K P Y E # R L Y F S N N # R I F C T Q + D N K T I R Y N W # Q Y D Y K K N Y R Q N I # I M F F Y C N I S R D I R D Y # I
S K L A N I # S I R I + S L L F # K I Q Y L L N A L * K K N N + L E M I S L * I K # # I K P Q N I D Y F F L L Q Y + S * N Q * I I

Exemplo de uma ORF bacteriana

Método (um pouco melhor) para achar genes em procariotos

1. Ache todas ORFs
2. Traduza cada uma usando o código genético
3. Compare cada uma com seqüências de genes conhecidos
 - Se achar algum *hit* estatisticamente significativo, guarde; senão jogue fora
4. Resolva sobreposições

Comparação de sequencias

- Similaridade “suficiente”
- O que é similaridade?
- O que é “suficiente”?
- Google das sequencias: BLAST
- Basic Local Alignment Search Tool
- Altschul et al., 1990, 1997

[blastn](#)**[blastp](#)**[blastx](#)[tblastn](#)[tblastx](#)

Enter Query Sequence

BLASTP programs search protein subjects using a protein query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) ?

```
>s
MQLNLAMGAVADGDRAPKACDAACSEAAGDKSAMMHDALFERFSARLKAQVGPEVYASWFA
RLKLHTVSKSVVRFTVPTITFLKSWINNRYMDLITSLVQSEDPPDLKVEILVRSASRPVRPA
QTEERAQPFVQEVGAAPRNKSFIPQSATAPAAQPMQAQATLRQGGSGPLFGSPLDTRFTED
TFVEGSSNRVALAAAKTIAAGAGAVRFNPLFIHAGVGLKTHLLQAIANAATDSRRNPFV
VYLTAEFMWRFFATAIRDNDALTLDKTLRNIDLLVIDDMQFLOGKMIQHEFCHLLNMLDS
```

[Clear](#)

Query subrange ?

From

To

Or, upload file

Browse...

Job Title

s

Enter a descriptive title for your BLAST search ?

☒ Align two or more sequences ?

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence ?

[Clear](#)

Subject subrange ?

```
>t
MBSRGISACIQENNYETPETNADABCLETTCEELFKNVSSKLEDOVGSDDVYASWFORLKEB
SVSHNIVVLSVPTNFKANIKNRYIDTITKLQESISSIQGEVRIIVRSAALMPSETSSSSA
IAHTTAKPRIINTGKISTIQGKOSINPVFGSPLDSEKVFVSNFIEGPNRVALAAAHITAEF
NSSSCTVRFNPLFIHASVGLGKTHLLQAIANAATKKONNLRVVYLTAEYFMWRFFATAIRDN
YALNFKDCLRNIDLLLLIDDMQFLOGKLIQHEFCHLLNSLLDSAKQIVAAADRPPSELESID
```

From

To

Or, upload file

Browse...

Program Selection

Algorithm

☒ blastp (protein-protein BLAST)

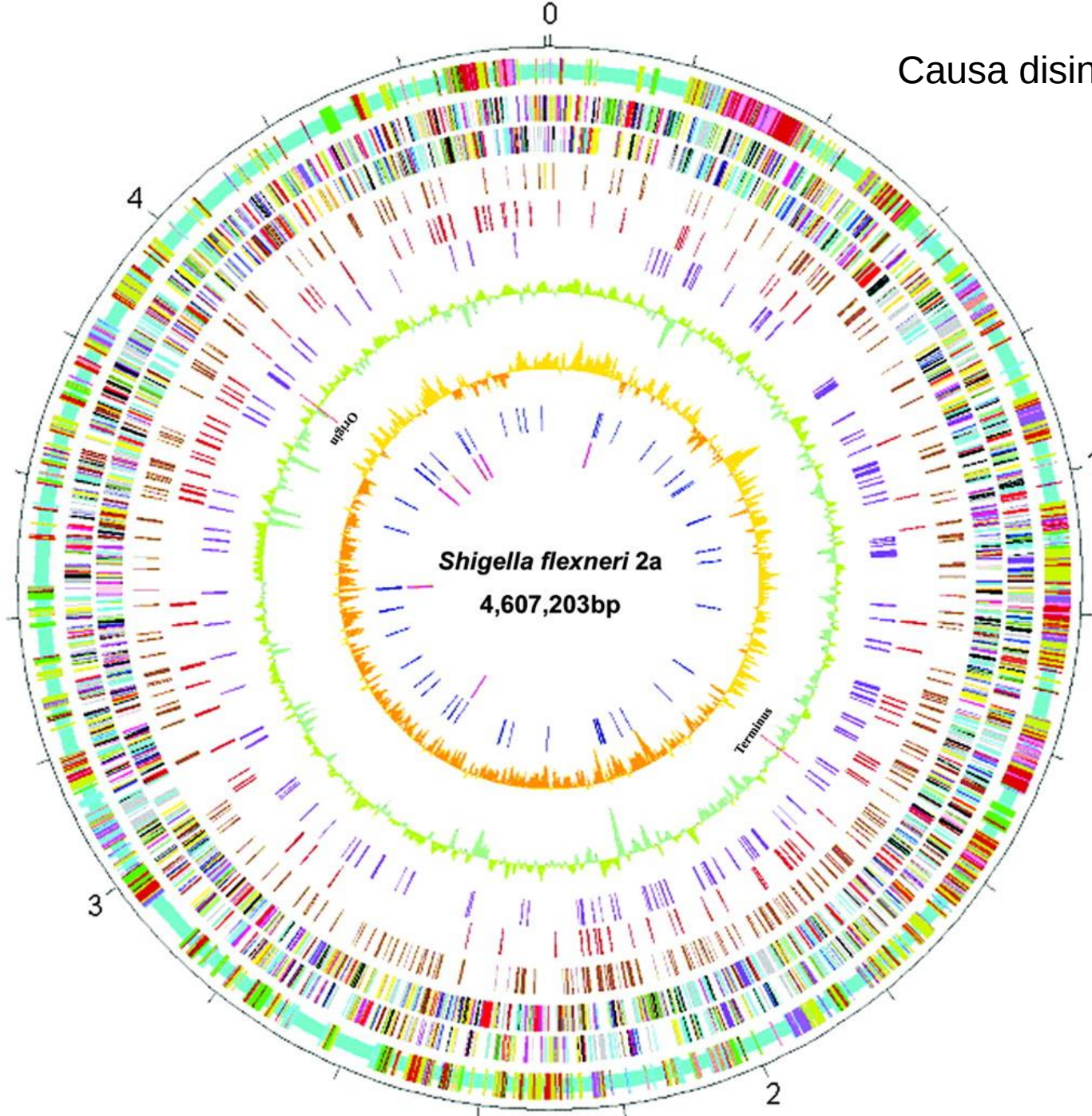
Choose a BLAST algorithm ?

>lcl|35099 t
Length=499

Score = 604 bits (1558), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 301/499 (60%), Positives = 365/499 (73%), Gaps = 25/499 (5%)

Query	21	DAACSEAAGDKSAMMHDALFERFSARLKAQVGPEVYASWFARLKLHTVSKSVVRFTVPTT	80
		DA C E ++ LF+ S++L+ QVG +VYASWF RLK +VS ++V +VPT	
Sbjct	23	DARCLETTCEE-----LFGNVSSKLEDQVGSDVYASWFQRLKFRSVSHNIVYLSVPTN	75
Query	81	FLKSWINNRYMDLITSLVQSEDPDVLKVEILVRSASRPVRPAQTEERAQPVQEVGAAPRN	140
		FLK+WI NRY+D IT L Q + VEI+VRS+ + P++T +	
Sbjct	76	FLKAWIKNRYIDTITKLFQESISSIQGVEIIVRSAA--LMPSETS-----S	119
Query	141	KSFIPSQSATAPAAQPMQAATLRQGGSGPLFGSPLDTRFTFDTFVEGSSNRVALAAAKT	200
		S I +A P + P+FGSPLD++F F F+EG SNRVALAAA T	
Sbjct	120	SSAIAHTTAKPPIINTGKISTIQGKQSINPVFGSPLDSKFVFSNFIEGPSNRVALAAAHT	179
Query	201	IAEAGAGA--VRFNPLFIHAGVGLGKTHLLQAIANAIAIDSPRNPRVVYLTAEYFMWRFAT	258
		IAE + + VRFNPLFIHA VGLGKTHLLQAIANAIAI N RVVYLTAEYFMWRFAT	
Sbjct	180	IAEENSSSCTVRFNPLFIHASVGLGKTHLLQAIANAIAIKKQNNLRVVYLTAEYFMWRFAT	239
Query	259	AIRDNDALTILKDTLRNIDLLVIDDMQFLOGKMIQHEFCHLLNMLLDSAKQVVVAADRAPW	318
		AIRDN AL KD LRNIDLL+IDDMQFLOGK+IQHEFCHLLN LLDSAKQ+V AADR P	
Sbjct	240	AIRDNYALNFKDCLRNIDLLVIDDMQFLOGKLIQHEFCHLLNSLLDSAKQIVAAADRPPS	299
Query	319	ELESLDPRVRSRLQGGMAIEIEGPDYDMRYEMLNRRMGSAQDDPSFEISDEILTHVAKS	378
		ELESLD R+RSRLQGG+A+ + D +MR +L R+ A++D+P IS+EIL VA++	
Sbjct	300	ELESLSRIRSRLQGGVAVPLGAHDIEMRLTILKNRLKMAKKNPKLYISEEILQVQAQT	359
Query	379	VTASGRELEGAFNQLMFRRSFEPNLSVDRVDELLSHLVSGEAKRVRIEDIQRIVARHYN	438
		VT SGREL+GAFNQL+FR SFEP L++ VDELLSHLV +GE K++RIEDIQR+V++HYN	
Sbjct	360	VTTSGRELDGAFNQLVFRNSFEPVLTIKMVDELLSHLVSAGETKKIRIEDIQRMVSKHYN	419
Query	439	VSRQELVSNRRTRVIVKPRQIAMYLAKMLTPRSFPEIGRRFGGRDHTTVLHAVRKIEDLI	498
		+SR +L+SNRR R IV+PRQIAMYL+K++TPRSFPEIGRRFG RDHTTVLHAVRKIE +	
Sbjct	420	ISRTDLLSNRRVRTIVRPRQIAMYLSKIMTPRSFPEIGRRFGDRDHTTVLHAVRKIEKSM	479
Query	499	SGDTKLGHEVELLKRLINE	517
		DT + EVELLKRLI+E	
Sbjct	480	EKDTVIKKEVELLKRLISE	498

Causa disinteria



Sumário

1. DNA: composição e estrutura
2. RNA: composição e estrutura
3. O código genético
4. Transcrição: DNA $\rightarrow$ mRNA
5. Tradução: mRNA $\rightarrow$ proteína
6. Replicação: DNA $\rightarrow$ 2 DNA
7. Expressão gênica
8. Sequenciamento de DNA
9. Bioinformática
10. Laboratórios