Aligning the unalignable: bacteriophage whole genome alignments

Sèverine Bérard, Annie Chateau, Nicolas Pompidor, Paul Guertin, Anne Bergeron and Krister M. Swenson

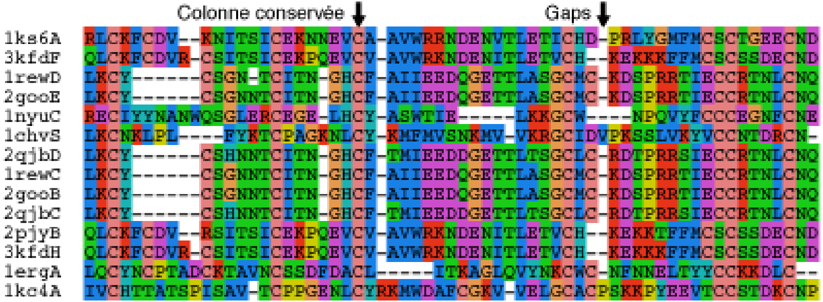
Colloque IPB 2015 Montpellier

Bacteriophages genomes: a mosaic structure

- Modular theory of phage genome organization [Botstein, 80]: biological functions are grouped into *modules* whose order is mostly conserved along the genomic sequence

- Each module has *variants* that perform the same functions, possibly encoded by dissimilar sequences

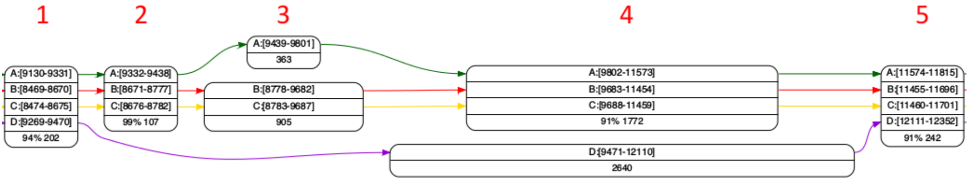
Classical multiple alignments



```
1ks6A  RLCKFCDV--KNITSICEKNNEVCA-AVWRRNDENVTLETICHD-PRLYGMFMCSCTGEECND
3kfdF  QLCKFCDVR-CSITSICEKPQEVCV-AVWRKNDENITLETVCH--KEKKKFFMCSCSSDECND
1rewD  LKCY------CSGN-TCITN-GHCF-AIIEEDQGETTLASGCMC-KDSPRRTIECCRTNLCNQ
2gooE  LKCY------CSGNNTCITN-GHCF-AIIEEDQGETTLASGCMC-KDSPRRTIECCRTNLCNQ
1nyuC  RECIYYNANWQSGLERCEGE-LHCY-ASWTIE-----LKKGCW----NPQVYFCCCEGNFCNE
1chvS  LKCNKLPL----FYKTCPAGKNLCY-KMFMVSNKMV-VKRGCIDVPKSSLVKYVCCNTDRCN-
2qjbD  LKCY------CSHNNTCITN-GHCF-TMIEEDDGETTLTSGCLC-RDTPRRSIECCRTNLCNQ
1rewC  LKCY------CSGNNTCITN-GHCF-AIIEEDQGETTLASGCMC-KDSPRRTIECCRTNLCNQ
2gooB  LKCY------CSGNNTCITN-GHCF-AIIEEDQGETTLASGCMC-KDSPRRTIECCRTNLCNQ
2qjbC  LKCY------CSHNNTCITN-GHCF-TMIEEDDGETTLTSGCLC-RDTPRRSIECCRTNLCNQ
2pjyB  QLCKFCDV--RSITSICEKPQEVCV-AVWRKNDENITLETVCH--KEKKTFFMCSCSSDECND
3kfdH  QLCKFCDVR-CSITSICEKPQEVCV-AVWRKNDENITLETVCH--KEKKKFFMCSCSSDECND
1ergA  LQCYNCPTADCKTAVNCSSDFDACL-----ITKAGLQVYNKCWC-NFNNELTYYCCKKDLC--
1kc4A  IVCHTTATSPISAV-TCPPGENLCYRKMWDAFCGKV-VELGCACPSKKPYEEVTCCSTDKCNP
```

Central dogma: sequence similarity ⇒ functional similarity

Phages multiple alignments



1

A:[9130-9331]
B:[8469-8670]
C:[8474-8675]
D:[9269-9470]
94% 202

2

A:[9332-9438]
B:[8671-8777]
C:[8676-8782]
99% 107

3

A:[9439-9801]
363

B:[8778-9682]
C:[8783-9687]
905

4

A:[9802-11573]
B:[9683-11454]
C:[9688-11459]
91% 1772

D:[9471-12110]
2640

5

A:[11574-11815]
B:[11455-11696]
C:[11460-11701]
D:[12111-12352]
91% 242

- The alignments rely on phages genomes features:
    - functional collinearity
    - a low duplication rate
    - the presence of long shared sequences

The Alpha aligner

Ideas

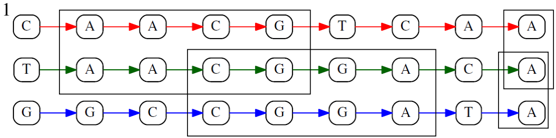
- to use exact matches (= anchor)
- to order them in a *graph* representing their succession in the genomes
- not to linearize the graph in an alignment

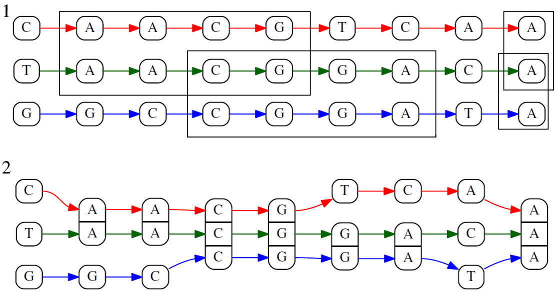
The Alpha aligner

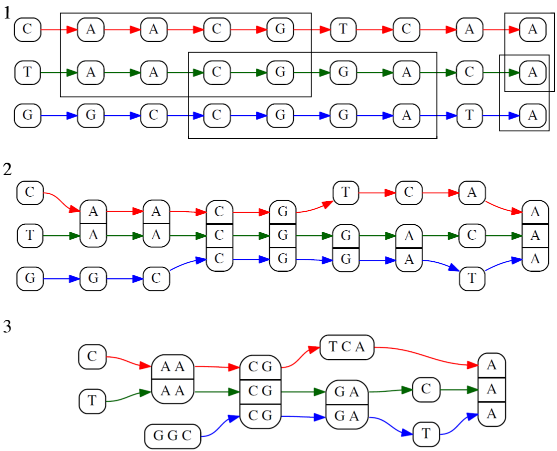
Ideas

- to use exact matches (= anchor)
- to order them in a *graph* representing their succession in the genomes
- not to linearize the graph in an alignment

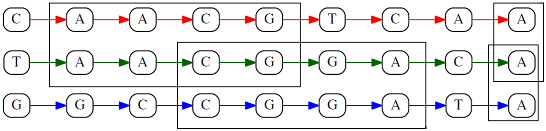
⇒ we work on graphs, representing partial order, to capture mosaicism of sets of genomes


1
C A A C G T C A A
T A A C G G A C A
G G C C G G A T A

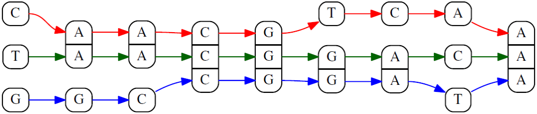

1
C A A C G T C A A
T A A C G G A C A
G G C C G G A T A
2
C T C A
A A C G A
T A A C G G A C A
C G G A
G G C T


1
C A A C G T C A A
T A A C G G A C A
G G C C G G A T A
2
C A A C G T C A A
T A A C G G A C A
G G C C G G A T A
3
C
T
AA
AA
GGC
CG
CG
CG
TCA
GA
GA
C
T
A
A
A

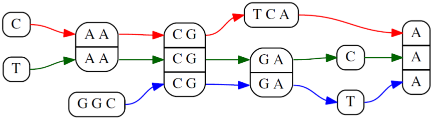
1


C A A C G T C A A
T A A C G G A C A
G G C C G G A T A

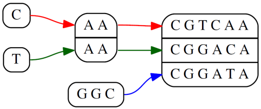
2


C
T
G
A A
A A
G
C C C
G G G
T C A
G G
A A
C
T
A A A

3


C
T
A A
A A
G G C
C G
C G
C G
T C A
G A
G A
C
T
A
A
A

4


C
T
A A
A A
G G C
C G T C A A
C G G A C A
C G G A T A

Alpha's characteristics

- One parameter m: minimal length of matches. Allow to adjust the alignment view

Alpha's characteristics

- One parameter m: minimal length of matches. Allow to adjust the alignment view

- Time complexity in $\mathcal{O}(\alpha(M_l)M_l + nN_l)$

  (where M_l is the total length of matches, α is the inverse of the Ackermann function, n is the number of genomes and N_l is the total length of genomes)

  = reasonable calculation time

Alpha's characteristics

- One parameter m: minimal length of matches. Allow to adjust the alignment view

- Time complexity in $\mathcal{O}(\alpha(M_l)M_l + nN_l)$
  (where M_l is the total length of matches, α is the inverse of the Ackermann function, n is the number of genomes and N_l is the total length of genomes)
  = reasonable calculation time

- Functional collinearity hypothesis: Alpha detects large rearrangements events that contradict this hypothesis and does not perform the alignment in such a case

Highlighting rearrangements (*e.g.* deletions)

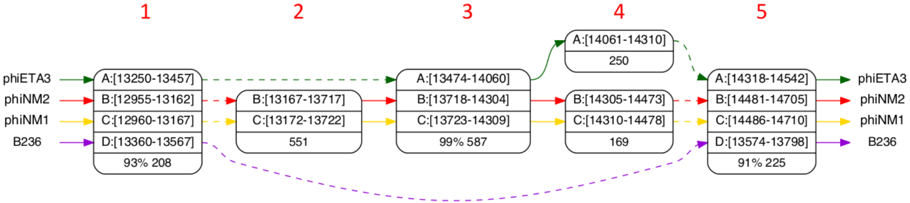


Figure 4 A local alignment of four S. aureus bacteriophages. In column 3, three phages are in a 587 bp exact alignment, with 99 % identical columns. A major deletion in phage B236 spans columns 2, 3, and 4, and the corresponding arrow is dotted to reflect the fact that some basepairs are not shown, 9 bp in this case.

Highlighting rearrangements (*e.g.* recombination)

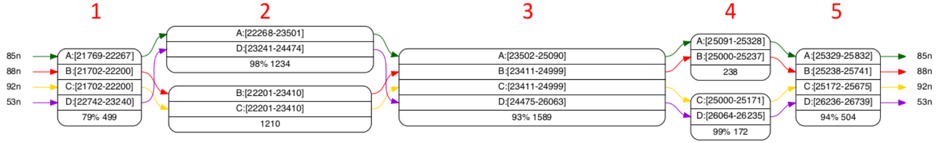
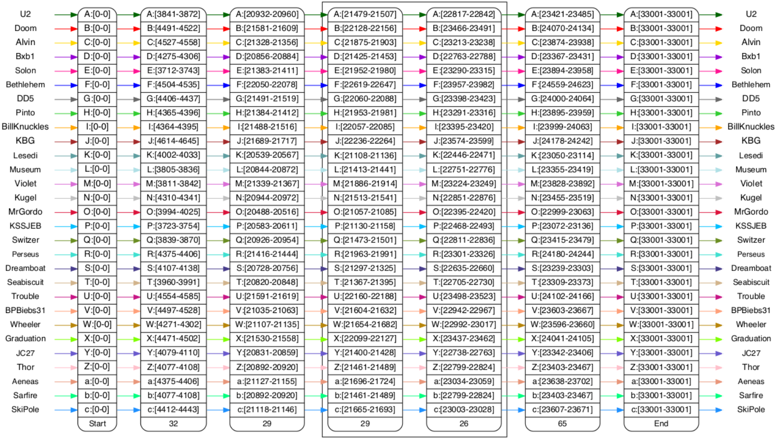


Figure 5 Modules and variants. Local alignment of four *S. aureus* bacteriophages clearly showing modules and variants. Notice that sequences with the same variant in the second column are switched in the fourth.

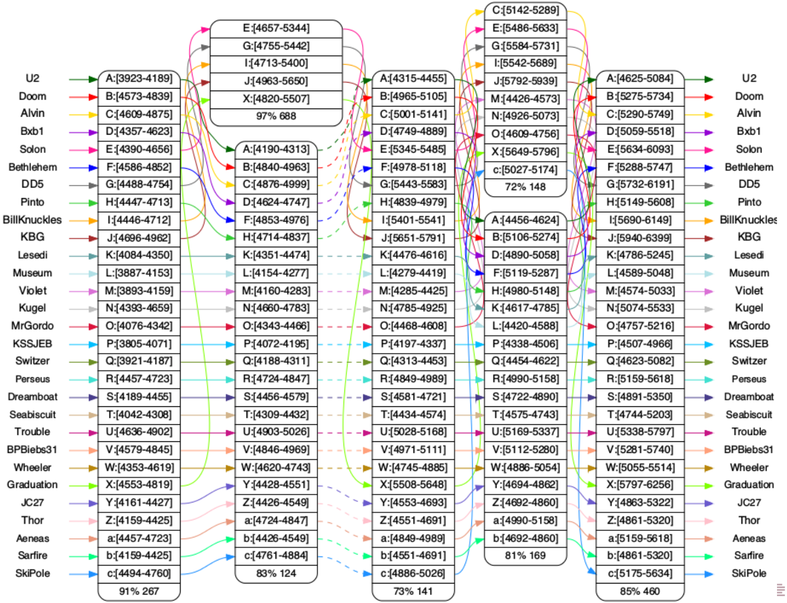
Anchors visualization, backbone



U2	A:[0-0]	A:[3841-3872]	A:[20932-20960]	A:[21479-21507]	A:[22817-22842]	A:[23421-23485]	A:[33001-33001]	U2
Doom	B:[0-0]	B:[4491-4522]	B:[21581-21609]	B:[22128-22156]	B:[23466-23491]	B:[24070-24134]	B:[33001-33001]	Doom
Alvin	C:[0-0]	C:[4527-4558]	C:[21328-21356]	C:[21875-21903]	C:[23213-23238]	C:[23874-23938]	C:[33001-33001]	Alvin
Bxb1	D:[0-0]	D:[4275-4306]	D:[20856-20884]	D:[21425-21453]	D:[22763-22788]	D:[23367-23431]	D:[33001-33001]	Bxb1
Solon	E:[0-0]	E:[3712-3743]	E:[21383-21411]	E:[21952-21980]	E:[23290-23315]	E:[23894-23958]	E:[33001-33001]	Solon
Bethlehem	F:[0-0]	F:[4504-4535]	F:[22050-22078]	F:[22619-22647]	F:[23957-23982]	F:[24559-24623]	F:[33001-33001]	Bethlehem
DD5	G:[0-0]	G:[4406-4437]	G:[21491-21519]	G:[22060-22088]	G:[23398-23423]	G:[24000-24064]	G:[33001-33001]	DD5
Pinto	H:[0-0]	H:[4365-4396]	H:[21384-21412]	H:[21953-21981]	H:[23291-23316]	H:[23895-23959]	H:[33001-33001]	Pinto
BillKnuckles	I:[0-0]	I:[4364-4395]	I:[21488-21516]	I:[22057-22085]	I:[23395-23420]	I:[23999-24063]	I:[33001-33001]	BillKnuckles
KBG	J:[0-0]	J:[4614-4645]	J:[21689-21717]	J:[22236-22264]	J:[23574-23599]	J:[24178-24242]	J:[33001-33001]	KBG
Lesedi	K:[0-0]	K:[4002-4033]	K:[20539-20567]	K:[21108-21136]	K:[22446-22471]	K:[23050-23114]	K:[33001-33001]	Lesedi
Museum	L:[0-0]	L:[3805-3836]	L:[20844-20872]	L:[21413-21441]	L:[22751-22776]	L:[23355-23419]	L:[33001-33001]	Museum
Violet	M:[0-0]	M:[3811-3842]	M:[21339-21367]	M:[21886-21914]	M:[23224-23249]	M:[23828-23892]	M:[33001-33001]	Violet
Kugel	N:[0-0]	N:[4310-4341]	N:[20944-20972]	N:[21513-21541]	N:[22851-22876]	N:[23455-23519]	N:[33001-33001]	Kugel
MrGordo	O:[0-0]	O:[3994-4025]	O:[20488-20516]	O:[21057-21085]	O:[22395-22420]	O:[22999-23063]	O:[33001-33001]	MrGordo
KSSJEB	P:[0-0]	P:[3723-3754]	P:[20583-20611]	P:[21130-21158]	P:[22468-22493]	P:[23072-23136]	P:[33001-33001]	KSSJEB
Switzer	Q:[0-0]	Q:[3839-3870]	Q:[20926-20954]	Q:[21473-21501]	Q:[22811-22836]	Q:[23415-23479]	Q:[33001-33001]	Switzer
Perseus	R:[0-0]	R:[4375-4406]	R:[21416-21444]	R:[21963-21991]	R:[23301-23326]	R:[24180-24244]	R:[33001-33001]	Perseus
Dreamboat	S:[0-0]	S:[4107-4138]	S:[20728-20756]	S:[21297-21325]	S:[22635-22660]	S:[23239-23303]	S:[33001-33001]	Dreamboat
Seabiscuit	T:[0-0]	T:[3960-3991]	T:[20820-20848]	T:[21367-21395]	T:[22705-22730]	T:[23309-23373]	T:[33001-33001]	Seabiscuit
Trouble	U:[0-0]	U:[4554-4585]	U:[21591-21619]	U:[22160-22188]	U:[23498-23523]	U:[24102-24166]	U:[33001-33001]	Trouble
BPBiebs31	V:[0-0]	V:[4497-4528]	V:[21035-21063]	V:[21604-21632]	V:[22942-22967]	V:[23603-23667]	V:[33001-33001]	BPBiebs31
Wheeler	W:[0-0]	W:[4271-4302]	W:[21107-21135]	W:[21654-21682]	W:[22992-23017]	W:[23596-23660]	W:[33001-33001]	Wheeler
Graduation	X:[0-0]	X:[4471-4502]	X:[21530-21558]	X:[22099-22127]	X:[23437-23462]	X:[24041-24105]	X:[33001-33001]	Graduation
JC27	Y:[0-0]	Y:[4079-4110]	Y:[20831-20859]	Y:[21400-21428]	Y:[22738-22763]	Y:[23342-23406]	Y:[33001-33001]	JC27
Thor	Z:[0-0]	Z:[4077-4108]	Z:[20892-20920]	Z:[21461-21489]	Z:[22799-22824]	Z:[23403-23467]	Z:[33001-33001]	Thor
Aeneas	a:[0-0]	a:[4375-4406]	a:[21127-21155]	a:[21696-21724]	a:[23034-23059]	a:[23638-23702]	a:[33001-33001]	Aeneas
Sarfire	b:[0-0]	b:[4077-4108]	b:[20892-20920]	b:[21461-21489]	b:[22799-22824]	b:[23403-23467]	b:[33001-33001]	Sarfire
SkiPole	c:[0-0]	c:[4412-4443]	c:[21118-21146]	c:[21665-21693]	c:[23003-23028]	c:[23607-23671]	c:[33001-33001]	SkiPole
	Start	32	29	29	26	65	End	

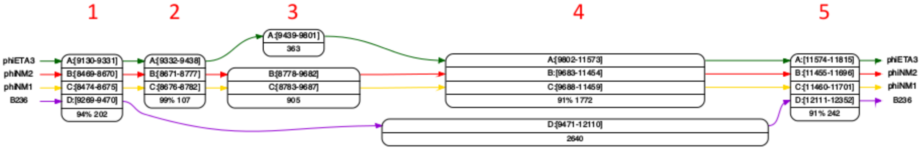
Figure 6 The anchor view. Aligning a set of 29 *Mycobacterium* bacteriophages using $m = 175$: the aligner can display the sequence of anchors, or backbone, allowing the user to zoom quickly to a specific region. Boxes group pairs of anchors bounding a gapless alignment.

Zooming from anchor view

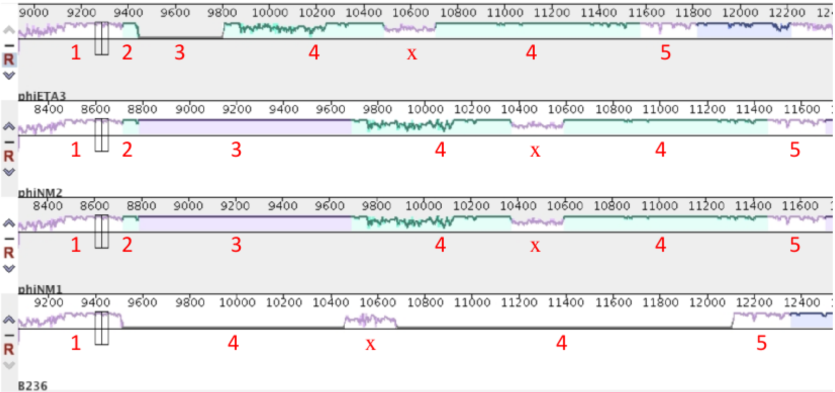


Comparison with Mauve

Alpha



Mauve



Comparison with Mauve

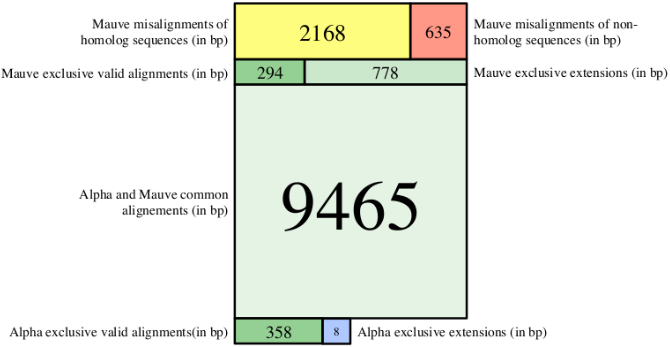

Mauve misalignments of homolog sequences (in bp)
2168
635
Mauve misalignments of non-homolog sequences (in bp)
Mauve exclusive valid alignments (in bp)
294
778
Mauve exclusive extensions (in bp)
Alpha and Mauve common alignements (in bp)
9465
Alpha exclusive valid alignments(in bp)
358
8
Alpha exclusive extensions (in bp)

Figure 9 Alignments comparison. Diagram representing the repartition of 13706 bp of 70 block alignments by Alpha or Mauve. The top 5 rectangles are from alignments produced by Mauve, and the bottom 3 are from Alpha alignments. Green and blue rectangles represent valid alignments, while yellow and red are misalignments.

Conclusion and perspectives

- Conclusion
  - Alpha produces biologically meaningful alignments

Conclusion and perspectives

- Conclusion
  - Alpha produces biologically meaningful alignments
  - Implemented in an interactive aligner

Conclusion and perspectives

- Conclusion
  - Alpha produces biologically meaningful alignments
  - Implemented in an interactive aligner
  - The model is almost parameter free

Conclusion and perspectives

- Conclusion
  - Alpha produces biologically meaningful alignments
  - Implemented in an interactive aligner
  - The model is almost parameter free
- Perspectives
  - Transfer of annotations

Conclusion and perspectives

- Conclusion
  - Alpha produces biologically meaningful alignments
  - Implemented in an interactive aligner
  - The model is almost parameter free
- Perspectives
  - Transfer of annotations
  - Extend our mathematical model to deal with rearrangements

Thank you for your attention

- References:
  - Article: “Aligning the unalignable: bacteriophage whole genome alignments”, *BMC Bioinformatics*, to appear
  - Tool: Alpha
    http://www.lirmm.fr/∼swenson/alpha/alpha.htm

Any questions ?

Alpha produces gapless alignment

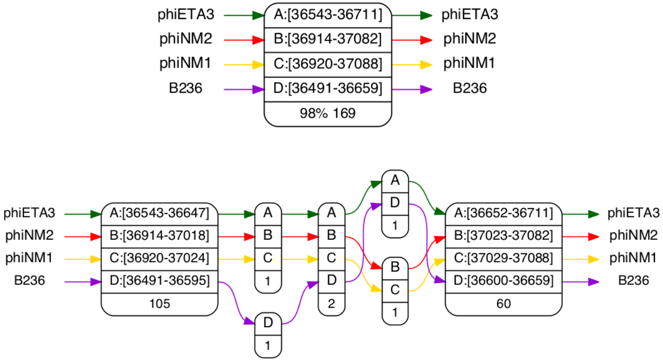


Figure 3 Contracted vertices. A example of a contracted vertex (top) and its expanded graph (bottom) in the alignment of four *S. aureus* bacteriophages. The length of the vertice is 169 nucleotides, and the percentage of identity is 98 %. In the expanded graph, there are two single nucleotide mutations that account for the 98% identity.

Quality of Alpha gapless alignments

- Comparison with 3 other alignment tools : Clustal, MUSCLE and T-Coffee
- 491 alignments compared over 3 datasets
- Only 8 regions of disagreement. For 7 regions, Alpha alignments are validated by translating in a.a. (tblastx). For the last region, the right solution is unknown.