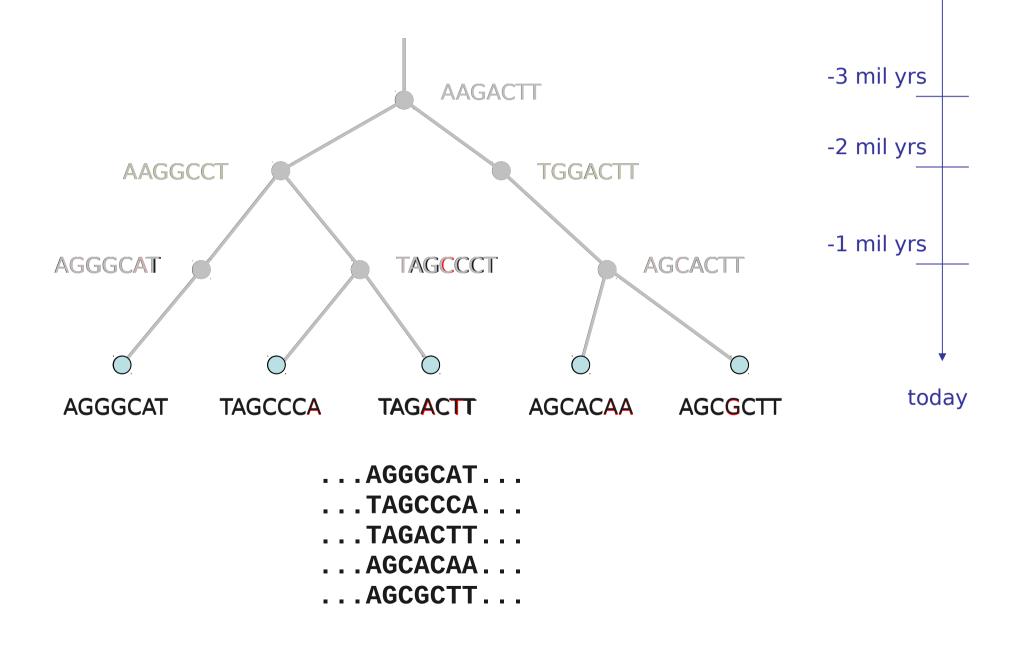
FASTSP: linear time calculation of alignment accuracy

Siavash Mir arabbaygi

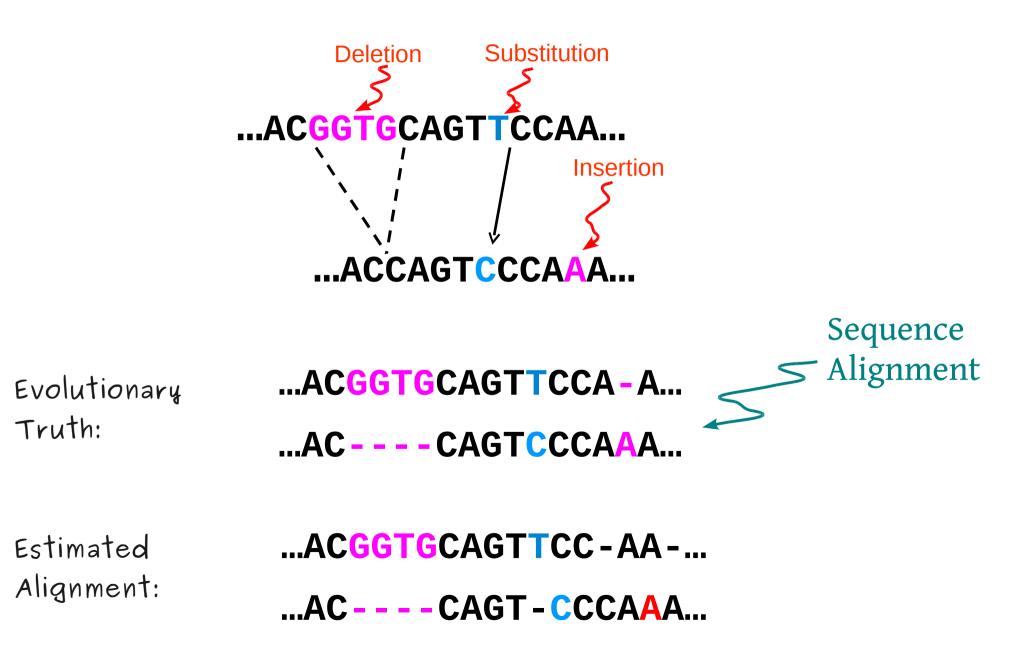
Research Preparation Exam

- Objective: Comparing very large Multiple Sequence Alignments efficiently (in linear time)
- Publication: Mirarab, S. and Warnow, T. (2011). Bioinformatics, 27(23), 3250–3258.
- Software: http://www.cs.utexas.edu/~phylo/software/fastsp/

DNA Sequence Evolution



Insertions and Deletions (indels)



Multiple Sequence Alignment (MSA)

	320	330	340	350	360	370	380	390	400	410	420	430
Acanthisitta_chloris_ <rifleman><mark>(</mark></rifleman>											A <mark>tgtaagtag</mark> a	
Alligator_mississippiensis/1-705												
Anas_platyrhynchos_ <mallard_(d<mark>l</mallard_(d<mark>												
Anolis_carolinensis_ <green_lizar< td=""><td>SCAA<mark>tt</mark>acaaa<mark>ggg</mark></td><td>TTATATGACGTO</td><td>CTTTCTCCA</td><td>A TCCA<mark>G</mark>A<mark>GG</mark>A</td><td>(GAACTT</td><td>TCA<mark>gg</mark>acaa<mark>gg</mark>cat</td><td>CA<mark>G</mark>A<mark>T</mark>ACC</td><td>CA<mark>g</mark>ca<mark>tg</mark>a<mark>tt</mark>cc</td><td>A<mark>gca<mark>gg</mark>aac(</mark></td><td>CTCATCAC<mark>G</mark>A)</td><td>A<mark>ggt</mark>cca<mark>gg</mark>a<mark>g</mark></td><td>B<mark>CTG</mark>ACC<mark>T</mark>C</td></green_lizar<>	SCAA <mark>tt</mark> acaaa <mark>ggg</mark>	TTATATGACGTO	CTTTCTCCA	A TCCA <mark>G</mark> A <mark>GG</mark> A	(GAACTT	TCA <mark>gg</mark> acaa <mark>gg</mark> cat	CA <mark>G</mark> A <mark>T</mark> ACC	CA <mark>g</mark> ca <mark>tg</mark> a <mark>tt</mark> cc	A <mark>gca<mark>gg</mark>aac(</mark>	CTCATCAC <mark>G</mark> A)	A <mark>ggt</mark> cca <mark>gg</mark> a <mark>g</mark>	B <mark>CTG</mark> ACC <mark>T</mark> C
Antrostomus_carolinensis_ <chud Apaloderma_vittatum_<bar-taile< td=""><td>SCAATTACAGAGGA</td><td>CTGCATGATGTT</td><td>тт<mark>е</mark>тстсст</td><td>TCTAGGGGC</td><td>GATGCT</td><td>TCAGGTAGGGCAAC</td><td>GAAGCTTC</td><td>a<mark>g</mark>aca</td><td>A<mark>ggg</mark>aacag</td><td>ATATTCACCA</td><td>A<mark>ggtgcagg</mark>a<i>a</i></td><td>ATTGACCTC</td></bar-taile<></chud 	SCAATTACAGAGGA	CTGCATGATGTT	тт <mark>е</mark> тстсст	TCTAGGGGC	GATGCT	TCAGGTAGGGCAAC	GAAGCTTC	a <mark>g</mark> aca	A <mark>ggg</mark> aacag	ATATTCACCA	A <mark>ggtgcagg</mark> a <i>a</i>	ATTGACCTC
Apaloderma_vittatum_ <bar-taile< td=""><td>SCAATTACAGAGGA</td><td>CTACATGATGTT</td><td>тт<mark>е</mark>тстсст</td><td>TCCAGGGGA</td><td>GACACT</td><td>TCAGGAAAAGGGA</td><td>CAGGTATC</td><td>CAACA<mark>TG</mark>ACA<mark>G</mark>C</td><td>AGCAGGAAT</td><td>CTCATCACCA</td><td>AGGTGCAGGAA</td><td>ATTGACCTC</td></bar-taile<>	SCAATTACAGAGGA	CTACATGATGTT	тт <mark>е</mark> тстсст	TCCAGGGGA	GACACT	TCAGGAAAAGGGA	CAGGTATC	CAACA <mark>TG</mark> ACA <mark>G</mark> C	AGCAGGAAT	CTCATCACCA	AGGTGCAGGAA	ATTGACCTC
Aptenodytes_forsteri_ <emperor_c< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></emperor_c<>												
Balearica_regulorum_ <grey_crot Buceros_rhinoceros_<rhinoceros< td=""><td>SCAATTACAGAGGAG</td><td>CTACATGATGTI</td><td></td><td></td><td>GAGACT</td><td></td><td>GAAGCTT</td><td> </td><td></td><td></td><td> CAGGAA</td><td>ATTGACCTC</td></rhinoceros<></grey_crot 	SCAATTACAGAGGAG	CTACATGATGTI			GAGACT		GAAGCTT				CAGGAA	ATTGACCTC
Buceros_minoceros_ <kninoceros Calypte_anna_<anna's_hummin< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></anna's_hummin<></kninoceros 												
Cariama cristata <red-legged s<="" td=""><td></td><td></td><td>TTETCTCCT</td><td></td><td>GATACT</td><td></td><td>CAGATATO</td><td></td><td>ANNINININIAT</td><td></td><td>ACCTOCACCA</td><td></td></red-legged>			TTETCTCCT		GATACT		CAGATATO		ANNINININIAT		ACCTOCACCA	
Canama_criscaca_ <ideo-leggeo_s Chaetura_pelagica_<chimney_st< td=""><td></td><td></td><td>TTETCTCCT</td><td></td><td></td><td></td><td></td><td>CAACATGACAGC</td><td></td><td></td><td>ABBIBCABBAA</td><td>ATTEACCTC</td></chimney_st<></ideo-leggeo_s 			TTETCTCCT					CAACATGACAGC			ABBIBCABBAA	ATTEACCTC
Charadrius_vociferus_ <killdeer></killdeer>	SCAATTACAGAGGAG	CTACATGATGTT	ттетстсст		GAGACT	TCAGG TAGGGCAA	GAAGCTT		ATCTGTCTT	T <mark>BB</mark> CCATCA	SOCTOGICACA	A GTTGTTGA
Chelonia mydas/1-705	CAATTATAGAGGA	CTGCATGATGTT					CAGATATO	ACATCACAAC	A B C A B A A A T	CTCATCACAA	AGGTTCAGGAG	TTGACCTC
Chlamydotis undulata <houbara< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></houbara<>												
Colius striatus <speckled mouse<="" td=""><td>GCAACTACAGAGGA</td><td>СТАСА<mark>ТВ</mark>АТВТТ</td><td>тт<mark>в</mark>тстсст</td><td>TCCA<mark>GGGG</mark>C</td><td>GATACT</td><td>TCA<mark>gg</mark>acaa<mark>ggg</mark>ad</td><td>CAGATATO</td><td>CAACA<mark>TG</mark>ACA<mark>G</mark>C</td><td>A<mark>g</mark>ca<mark>gg</mark></td><td></td><td> <mark>G</mark>AC</td><td>атт<mark>б</mark>асстс</td></speckled>	GCAACTACAGAGGA	СТАСА <mark>ТВ</mark> АТВТТ	тт <mark>в</mark> тстсст	TCCA <mark>GGGG</mark> C	GATACT	TCA <mark>gg</mark> acaa <mark>ggg</mark> ad	CAGATATO	CAACA <mark>TG</mark> ACA <mark>G</mark> C	A <mark>g</mark> ca <mark>gg</mark>		<mark>G</mark> AC	атт <mark>б</mark> асстс
Colius_striatus_ <speckled_mouse Columba_livia_<rock_pigeon_(de< td=""><td>SCAA<mark>tt</mark>aca<mark>g</mark>agga</td><td>СТАСА<mark>тб</mark>атбтт</td><td>тт<mark>атстсс</mark>т</td><td>ТССА<mark>бабб</mark>с</td><td>GATACT</td><td>TCA<mark>gg</mark>acaa<mark>ggg</mark>ad</td><td>CA<mark>G</mark>A</td><td></td><td></td><td> <mark>ACC</mark>A)</td><td>A<mark>ggtgcagg</mark>aa</td><td>ATT<mark>G</mark>ACCTC</td></rock_pigeon_(de<></speckled_mouse 	SCAA <mark>tt</mark> aca <mark>g</mark> agga	СТАСА <mark>тб</mark> атбтт	тт <mark>атстсс</mark> т	ТССА <mark>бабб</mark> с	GATACT	TCA <mark>gg</mark> acaa <mark>ggg</mark> ad	CA <mark>G</mark> A			<mark>ACC</mark> A)	A <mark>ggtgcagg</mark> aa	ATT <mark>G</mark> ACCTC
Corvus_brachyrhynchos_ <ameri<mark>(</ameri<mark>	<mark>bcaatt</mark> aca <mark>g</mark> aga	СТАСА <mark>ТБ</mark> АТ <mark>б</mark> ТТ	тт <mark>б</mark> тстсст	TCCA <mark>g</mark> a <mark>gg</mark> t	GATACT	TCA <mark>gg</mark> acaa <mark>ggg</mark> a(BAA <mark>g</mark> atato	CAACA <mark>tg</mark> aca <mark>g</mark> c	A <mark>gcagg</mark> aa <mark>t</mark> (СТСА <mark>Т</mark> САССАЛ	A <mark>ggt</mark> gca <mark>gg</mark> a <i>i</i>	A <mark>TT<mark>g</mark>acttc</mark>
Cuculus canorus <common cud<="" td=""><td>SCAATTACA<mark>g</mark>agga</td><td>СТАСА<mark>ТБ</mark>АТ<mark>б</mark>ТТ</td><td>тт<mark>в</mark>тстсст</td><td>ТССА<mark>бббб</mark>с</td><td>GAGACT</td><td>тса</td><td></td><td></td><td></td><td></td><td></td><td></td></common>	SCAATTACA <mark>g</mark> agga	СТАСА <mark>ТБ</mark> АТ <mark>б</mark> ТТ	тт <mark>в</mark> тстсст	ТССА <mark>бббб</mark> с	GAGACT	тса						
Egretta garzetta <little egret="">,</little>	SCAATTACA <mark>g</mark> agga	СТАСАТ <mark>б</mark> ат <mark>б</mark> тт	тт <mark>в</mark> тс <mark>в</mark> сст	ТССА <mark>бббб</mark> т	GATACT	TCA <mark>ggt</mark> a <mark>ggg</mark> caa	3 G A A <mark>G C T T T</mark>	G G	AA <mark>T</mark> (СТСА <mark>Т</mark> САССА	A <mark>ggt</mark> gca <mark>gg</mark> aa	A <mark>tt</mark> gacc <mark>t</mark> c
Eurypyga_helias_ <sunbittern>/1</sunbittern>	<mark>(</mark>	CA <mark>g</mark> aaacca <mark>gt</mark> t	CTT <mark>G</mark> ATCCT	Т <mark>СТ</mark>		G <mark>ca<mark>gg</mark>acaa<mark>ggg</mark>a(</mark>	SCA <mark>g</mark> a <mark>t</mark> ato	CAACA <mark>TG</mark> ACA <mark>G</mark> C	A <mark>g</mark> cc <mark>gg</mark> aac(C <mark>TCA<mark>TC</mark>ACC</mark> A,	A <mark>ggt</mark> gca <mark>gg</mark> a <mark>0</mark>	STT <mark>G</mark> ACCTC
Falco_peregrinus_ <peregrine_fa<mark>.</peregrine_fa<mark>	SCAA <mark>tt</mark> aca <mark>g</mark> aga	СТАСА <mark>ТБ</mark> АТ <mark>б</mark> ТТ	ТТ <mark>б</mark> ТСТССТ	TCCA <mark>gggg</mark> t	GATACT	TCA		<u>.</u>		ΤΤΑ <mark>σ</mark> ΤΤ <mark>σ</mark> ΤΤΑ <mark>(</mark>	<mark>g t g t</mark> a a <mark>g t</mark> a <mark>g</mark> a	A <mark>cacact</mark> a <mark>g</mark>
Fulmarus_glacialis_ <northern_fd< td=""><td>SCAA<mark>tt</mark>aca<mark>g</mark>agga(</td><td>CTACA<mark>TG</mark>A<mark>TG</mark>TT</td><td>тт<mark>в</mark>тстсст</td><td>TCCA<mark>ggggt</mark></td><td>G<mark>atact</mark></td><td>TCA<mark>ggt</mark>a<mark>ggg</mark>caat</td><td>GAA<mark>GCTTC</mark></td><td>CA<mark>G</mark></td><td> <mark>AA</mark>T(</td><td>С<mark>ТС</mark>А<mark>ТС</mark>АССА)</td><td>A<mark>ggtg</mark>ca<mark>gg</mark>a<i>i</i></td><td>A<mark>TTG</mark>ACC<mark>T</mark>C</td></northern_fd<>	SCAA <mark>tt</mark> aca <mark>g</mark> agga(CTACA <mark>TG</mark> A <mark>TG</mark> TT	тт <mark>в</mark> тстсст	TCCA <mark>ggggt</mark>	G <mark>atact</mark>	TCA <mark>ggt</mark> a <mark>ggg</mark> caat	GAA <mark>GCTTC</mark>	CA <mark>G</mark>	<mark>AA</mark> T(С <mark>ТС</mark> А <mark>ТС</mark> АССА)	A <mark>ggtg</mark> ca <mark>gg</mark> a <i>i</i>	A <mark>TTG</mark> ACC <mark>T</mark> C
Gallus_gallus_ <red_junglefowl_(< td=""><td>SCAA<mark>tt</mark>aca<mark>g</mark>agga(</td><td>CT<mark>G</mark>CAC<mark>G</mark>A<mark>TG</mark>TT</td><td>тт<mark>в</mark>тстсст</td><td>TCCAA<mark>ggg</mark>t</td><td>G A G A C T</td><td>TCA<mark>gg</mark>acaa<mark>ggg</mark>a(</td><td>CA<mark>GATATO</mark></td><td>CAACA<mark>TG</mark>AC</td><td>A<mark>gcaag</mark>aa<mark>t</mark>(</td><td>стта<mark>тс</mark>асса,</td><td>A<mark>ggt</mark>aca<mark>gg</mark>aa</td><td>A<mark>ctg</mark>acc<mark>t</mark>c</td></red_junglefowl_(<>	SCAA <mark>tt</mark> aca <mark>g</mark> agga(CT <mark>G</mark> CAC <mark>G</mark> A <mark>TG</mark> TT	тт <mark>в</mark> тстсст	TCCAA <mark>ggg</mark> t	G A G A C T	TCA <mark>gg</mark> acaa <mark>ggg</mark> a(CA <mark>GATATO</mark>	CAACA <mark>TG</mark> AC	A <mark>gcaag</mark> aa <mark>t</mark> (стта <mark>тс</mark> асса,	A <mark>ggt</mark> aca <mark>gg</mark> aa	A <mark>ctg</mark> acc <mark>t</mark> c
Gavia_stellata_ <red-throated_lo< td=""><td>SCAATTACA<mark>gagg</mark>a(</td><td>CTACAT<mark>G</mark>AT<mark>G</mark>TT</td><td>тт<mark>в</mark>тстсст</td><td>TCCA<mark>gggg</mark>c</td><td>GATACT</td><td>TCA<mark>gg</mark>acaa<mark>ggg</mark>a(</td><td>CA<mark>GATATO</mark></td><td>CAACA<mark>TG</mark>AC</td><td>A<mark>gc</mark>a<mark>gg</mark>aa<mark>t</mark>(</td><td>C<mark>TT</mark>A<mark>T</mark>CACCA</td><td>A<mark>ggtgcagg</mark>aa</td><td>A<mark>tt<mark>g</mark>acctc</mark></td></red-throated_lo<>	SCAATTACA <mark>gagg</mark> a(CTACAT <mark>G</mark> AT <mark>G</mark> TT	тт <mark>в</mark> тстсст	TCCA <mark>gggg</mark> c	GATACT	TCA <mark>gg</mark> acaa <mark>ggg</mark> a(CA <mark>GATATO</mark>	CAACA <mark>TG</mark> AC	A <mark>gc</mark> a <mark>gg</mark> aa <mark>t</mark> (C <mark>TT</mark> A <mark>T</mark> CACCA	A <mark>ggtgcagg</mark> aa	A <mark>tt<mark>g</mark>acctc</mark>
Geospiza_fortis_ <medium_groun<mark>(</medium_groun<mark>	BCAA <mark>tt</mark> aca <mark>g</mark> agga	CTACA <mark>TG</mark> A <mark>TG</mark> TT	тт <mark>б</mark> тс <u>т</u> сст	TCCA <mark>gggg</mark> c	GATGCT	TCA <mark>gg</mark> acaa <mark>ggg</mark> a(TAGATACC	CAACAC			<mark>GTGCAA</mark> GAA	A <mark>TTG</mark> ACTTC
Haliaeetus_albicilla_ <white-taile< td=""><td>SCAA<mark>tt</mark>aca<mark>g</mark>agga</td><td>CTACA<mark>TG</mark>ATGTT</td><td>тт<mark>в</mark>тсссст</td><td>TCCA<mark>gggg</mark>c</td><td>GATACT</td><td>TCA<mark>ggt</mark>a<mark>ggg</mark>cact</td><td>GAA<mark>GCTTC</mark></td><td>CA<mark>G</mark></td><td>A<mark>tctct</mark>t(</td><td>CT<mark>gg</mark>cca<mark>t</mark>ca</td><td>SGCTGGTGACA</td><td>A<mark>gtc</mark>aaaac</td></white-taile<>	SCAA <mark>tt</mark> aca <mark>g</mark> agga	CTACA <mark>TG</mark> ATGTT	тт <mark>в</mark> тсссст	TCCA <mark>gggg</mark> c	GATACT	TCA <mark>ggt</mark> a <mark>ggg</mark> cact	GAA <mark>GCTTC</mark>	CA <mark>G</mark>	A <mark>tctct</mark> t(CT <mark>gg</mark> cca <mark>t</mark> ca	SGCTGGTGACA	A <mark>gtc</mark> aaaac
Haliaeetus_leucocephalus_ <balc< td=""><td>SCAATTACAGAGGA</td><td>CTACATGATGTT</td><td>TTGTCCCCI</td><td>TCCAGGGGC</td><td>GATACT</td><td>TCAGGACAAGGGA</td><td>CAGATATO</td><td>CAGCA<mark>TG</mark>ACAGC</td><td>AGCAGGAAT</td><td>CTCATCACCA</td><td>AGGTGCAGGAA</td><td>ATTGACCTC</td></balc<>	SCAATTACAGAGGA	CTACATGATGTT	TTGTCCCCI	TCCAGGGGC	GATACT	TCAGGACAAGGGA	CAGATATO	CAGCA <mark>TG</mark> ACAGC	AGCAGGAAT	CTCATCACCA	AGGTGCAGGAA	ATTGACCTC
Homo_sapiens_ <human>/1-780 (</human>												
Leptosomus_discolor_ <cuckoo_k< td=""><td>SCAATTACAGAGGAG</td><td>CTUCATUATUT</td><td>TTGTCTCC1</td><td></td><td>GATACT</td><td></td><td></td><td>CAACATGACAGC</td><td>AGCAGGAAT</td><td>CTCATCACCA</td><td>AGGIGCAGGAG</td><td>ACCTC</td></cuckoo_k<>	SCAATTACAGAGGAG	CTUCATUATUT	TTGTCTCC1		GATACT			CAACATGACAGC	AGCAGGAAT	CTCATCACCA	AGGIGCAGGAG	ACCTC
Manacus_vitellinus_ <golden-coll Meleagris_gallopavo_<turkey>/10</turkey></golden-coll 			TTGTCTCCT		GATACT			CAACATGACAGC	AGCAGAAAT			
Melopsittacus_undulatus_ <budge< td=""><td></td><td></td><td>TTETCTCCT</td><td></td><td>GATACT</td><td></td><td>CAGATATO</td><td></td><td></td><td></td><td>ACCACCACCA</td><td>ACTO ACCTC</td></budge<>			TTETCTCCT		GATACT		CAGATATO				ACCACCACCA	ACTO ACCTC
Merops_nubicus_ <carmine_bee-4< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></carmine_bee-4<>												
Meriops_nabicas_ <carnine_bee-< td=""><td></td><td></td><td>TTETCTCCT</td><td></td><td>GATACT</td><td></td><td></td><td></td><td></td><td></td><td>A RETECASE A/</td><td>ATTEACCTC</td></carnine_bee-<>			TTETCTCCT		GATACT						A RETECASE A/	ATTEACCTC
Mesitornis_unicolor_ <brown_mesi Nestor_notabilis_<kea>/1-669</kea></brown_mesi 	SCAATTACAGAGGAG	CTGCATGATGTT	ттетстсст		GATACT	TCAGG TAGGGCAA	GAAGCTT	AT		• • • • • • • • • • • •		TTGACCTC
Wipponia nippon <crested bis=""></crested>	SCAATTACAGAGGAG	CTACATGATGTT	ттетстсст		GATACT		CAGATATO	AACATGACAGC	ABCABBAAT	TCATCACCA	A B B T B C A B B A A	ACCTC
Onhisthocomus hoazin <hoatzid< td=""><td>SCAATTACA<mark>G</mark>AGGA</td><td>СТАСА<mark>ТВ</mark>АТ<mark>В</mark>ТТ</td><td>тт<mark>в</mark>тстсст</td><td>тоса<mark>вава</mark>с</td><td>GATAGT</td><td>TCAGO TAGO CAAC</td><td>GAAGTGTI</td><td></td><td>A A <mark>B</mark>C A <mark>B</mark> A T A C</td><td>с<mark>тсатс</mark>асса.</td><td>А<mark>бот</mark>оса<mark>бо</mark>ая</td><td>ATT<mark>G</mark>ACCTC</td></hoatzid<>	SCAATTACA <mark>G</mark> AGGA	СТАСА <mark>ТВ</mark> АТ <mark>В</mark> ТТ	тт <mark>в</mark> тстсст	тоса <mark>вава</mark> с	GATAGT	TCAGO TAGO CAAC	GAAGTGTI		A A <mark>B</mark> C A <mark>B</mark> A T A C	с <mark>тсатс</mark> асса.	А <mark>бот</mark> оса <mark>бо</mark> ая	ATT <mark>G</mark> ACCTC
Pelecanus_crispus_ <dalmatian_i< td=""><td>CAATTACAGGGGA</td><td>TACATGATGTT</td><td>тт<mark>в</mark>тстсст</td><td>TCCAGGGG</td><td>GATACT</td><td>TCAGGACAAGGGA</td><td>CAGATATO</td><td>CAACATGACAGC</td><td>AGCAGGCAT</td><td>CTCATAACC -</td><td> NAA</td><td>ATTGACCTC</td></dalmatian_i<>	CAATTACAGGGGA	TACATGATGTT	тт <mark>в</mark> тстсст	TCCAGGGG	GATACT	TCAGGACAAGGGA	CAGATATO	CAACATGACAGC	AGCAGGCAT	CTCATAACC -	NAA	ATTGACCTC
Phaethon_lepturus_ <white-tailed< td=""><td>SCAATTACAGAGGA</td><td>СТАСА<mark>тб</mark>атбтт</td><td>тт<mark>в</mark>тстсст</td><td>тссавово</td><td>GATAAT</td><td>TCA<mark>gg</mark>acaa<mark>ggg</mark>ad</td><td>CA<mark>G</mark>ATACO</td><td>CAACA<mark>tg</mark>aca<mark>g</mark>c</td><td>ATCC<mark>GG</mark>AAT</td><td>CTCATCACCA</td><td>A<mark>ggtg</mark>ca<mark>gg</mark>aa</td><td>ATT<mark>G</mark>ACCTC</td></white-tailed<>	SCAATTACAGAGGA	СТАСА <mark>тб</mark> атбтт	тт <mark>в</mark> тстсст	тссавово	GATAAT	TCA <mark>gg</mark> acaa <mark>ggg</mark> ad	CA <mark>G</mark> ATACO	CAACA <mark>tg</mark> aca <mark>g</mark> c	ATCC <mark>GG</mark> AAT	CTCATCACCA	A <mark>ggtg</mark> ca <mark>gg</mark> aa	ATT <mark>G</mark> ACCTC
Phalacrocorax_carbo_ <cormoral< td=""><td>GCAA<mark>tt</mark>aca<mark>g</mark>agga</td><td>СТАСА<mark>ТВ</mark>АТВТТ</td><td>тт<mark>в</mark>тстсст</td><td>ТССА<mark>бббб</mark>т</td><td>GATACT</td><td>TCA<mark>gg</mark>acaa<mark>ggg</mark>a</td><td>ATCAACA</td><td>[NN NNN</td><td>NNNNNAAT</td><td>CTCATCACCA</td><td>A<mark>ggtgcagg</mark>a<i>a</i></td><td>А<mark>ТТ<mark>б</mark>асстс</mark></td></cormoral<>	GCAA <mark>tt</mark> aca <mark>g</mark> agga	СТАСА <mark>ТВ</mark> АТВТТ	тт <mark>в</mark> тстсст	ТССА <mark>бббб</mark> т	GATACT	TCA <mark>gg</mark> acaa <mark>ggg</mark> a	ATCAACA	[NN NNN	NNNNNAAT	CTCATCACCA	A <mark>ggtgcagg</mark> a <i>a</i>	А <mark>ТТ<mark>б</mark>асстс</mark>
Phoenicopterus ruber <america<mark>(</america<mark>	SCAA <mark>tt</mark> aca <mark>g</mark> aga(СТАСА <mark>ТВ</mark> АТ <mark>В</mark> ТТ	ТТА <mark>ТСТ</mark> ССТ	ТССС <mark>ввев</mark> с	GATACT	TCA <mark>gg</mark> acaa <mark>ggg</mark> a	CA <mark>GATATO</mark>	CAACA <mark>tg</mark> aca <mark>g</mark> c	A <mark>g</mark> ca <mark>gg</mark> ca <mark>t</mark> (СТСАТСАССАЛ	A <mark>ggtg</mark> ca <mark>gg</mark> a <i>m</i>	ATT <mark>G</mark> ACCTC
Picoides_pubescens_ <downy_wd< td=""><td>SCAA<mark>tt</mark>aca<mark>gggg</mark>a(</td><td>CTACA<mark>TG</mark>A<mark>TG</mark>TT</td><td>ТТ<mark>Б</mark>ТСТ<u>Т</u>СТ</td><td>TCCA<mark>gg</mark>a<mark>g</mark>t</td><td>GATACT</td><td>TCA<mark>gg</mark>acaa<mark>ggg</mark>a(</td><td>SCA<mark>G</mark>ATATO</td><td>CAACA<mark>tg</mark>aca<mark>g</mark>c</td><td>A<mark>gc</mark>a<mark>gg</mark>aac(</td><td>СТСА<mark>ТС</mark>АСТАЛ</td><td>A<mark>ggtgcagg</mark>aa</td><td>A<mark>TT<mark>g</mark>acctc</mark></td></downy_wd<>	SCAA <mark>tt</mark> aca <mark>gggg</mark> a(CTACA <mark>TG</mark> A <mark>TG</mark> TT	ТТ <mark>Б</mark> ТСТ <u>Т</u> СТ	TCCA <mark>gg</mark> a <mark>g</mark> t	GATACT	TCA <mark>gg</mark> acaa <mark>ggg</mark> a(SCA <mark>G</mark> ATATO	CAACA <mark>tg</mark> aca <mark>g</mark> c	A <mark>gc</mark> a <mark>gg</mark> aac(СТСА <mark>ТС</mark> АСТАЛ	A <mark>ggtgcagg</mark> aa	A <mark>TT<mark>g</mark>acctc</mark>
Podiceps_cristatus_ <great_crest< td=""><td>SCAA<mark>tt</mark>aca<mark>g</mark>aga</td><td>СТАСА<mark>ТБ</mark>А<mark>ТБ</mark>ТТ</td><td>ТТ<mark>А</mark>ТСТССТ</td><td>TCCA<mark>gggg</mark>c</td><td>GATACT</td><td>TCA<mark>gg</mark>acaa<mark>ggg</mark>a(</td><td>CA<mark>ggt</mark>ato</td><td>CAACA<mark>T</mark></td><td> <mark>GNN</mark>AA<mark>T</mark>(</td><td>СТСА<mark>ТС</mark>АССАЛ</td><td>A<mark>ggtg</mark>ca<mark>gg</mark>a<i>a</i></td><td>ATT<mark>G</mark>ACCTC</td></great_crest<>	SCAA <mark>tt</mark> aca <mark>g</mark> aga	СТАСА <mark>ТБ</mark> А <mark>ТБ</mark> ТТ	ТТ <mark>А</mark> ТСТССТ	TCCA <mark>gggg</mark> c	GATACT	TCA <mark>gg</mark> acaa <mark>ggg</mark> a(CA <mark>ggt</mark> ato	CAACA <mark>T</mark>	<mark>GNN</mark> AA <mark>T</mark> (СТСА <mark>ТС</mark> АССАЛ	A <mark>ggtg</mark> ca <mark>gg</mark> a <i>a</i>	ATT <mark>G</mark> ACCTC
Pterocles_gutturalis_ <yellow-thr<mark>(</yellow-thr<mark>												
Pygoscelis_adeliae_≺Adelie_Peng	GCAA <mark>tt</mark> aca <mark>gagg</mark> a	CT <mark>G</mark> CAT <mark>G</mark> AT <mark>G</mark> TT	TT <mark>GTCTCC</mark> T	TCCA <mark>gggg</mark> c	GATACT	TCA <mark>ggt</mark> a <mark>ggg</mark> caat	GAA <mark>GCTT</mark>	CA <mark>G</mark>	AA <mark>T</mark>	CTCA <mark>T</mark> CACCA)	A <mark>ggt</mark> aca <mark>gg</mark> a <i>i</i>	A <mark>ttg</mark> acc <mark>t</mark> c
Struthio_camelus_ <ostrich>/1-7</ostrich>	BCAA <mark>tt</mark> aca <mark>gggg</mark> a	CTACAT <mark>G</mark> AT <mark>G</mark> TT	CT <mark>GTCT</mark> CCT	CCCAA <mark>gg</mark> cc	G A G A C T	TCA <mark>gg</mark> acaa <mark>ggg</mark> a	CA <mark>GATATO</mark>	CAACA <mark>tg</mark> aca <mark>g</mark> c	A <mark>gt</mark> aggaa <mark>t</mark>	CTCA <mark>T</mark> CACCA)	A <mark>ggtg</mark> ca <mark>gg</mark> a <i>i</i>	A <mark>ttg</mark> acctc
Taeniopygia_guttata_ <zebra_fin< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></zebra_fin<>												
Tauraco_erythrolophus_ <red-cre< td=""><td>GCAA<mark>tt</mark>aca<mark>gagg</mark>a(</td><td>CTACATGATGTT</td><td>TTGTCTCC1</td><td>TCCA<mark>GGGG</mark>C</td><td>GATGCT</td><td>TCA<mark>gg</mark>acaa<mark>gg</mark>aat</td><td>CA<mark>GATATO</mark></td><td>CAÁCA<mark>TG</mark>AC</td><td> A<mark>NN</mark></td><td>C<mark>T</mark>CA<mark>T</mark>CACCA,</td><td>A<mark>ggtg</mark>ca<mark>gg</mark>a<i>i</i></td><td>ATTGACCTC</td></red-cre<>	GCAA <mark>tt</mark> aca <mark>gagg</mark> a(CTACATGATGTT	TTGTCTCC1	TCCA <mark>GGGG</mark> C	GATGCT	TCA <mark>gg</mark> acaa <mark>gg</mark> aat	CA <mark>GATATO</mark>	CAÁCA <mark>TG</mark> AC	A <mark>NN</mark>	C <mark>T</mark> CA <mark>T</mark> CACCA,	A <mark>ggtg</mark> ca <mark>gg</mark> a <i>i</i>	ATTGACCTC
Tinamus_major_ <great_tinamol< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></great_tinamol<>												
Tyto_alba_≺Barn_Owl≻/1-708 【	SCAA <mark>tt</mark> aca <mark>g</mark> a <mark>gg</mark> a(CTACATGAT <mark>G</mark> TT	TT <mark>GT</mark> CACCT	TCCA <mark>gggg</mark> C	GATACT	TCAGGACAA <mark>ggg</mark> a(CA <mark>G</mark> ATATO	CAACA <mark>TG</mark> ACA <mark>g</mark> c	A <mark>gcagg</mark> aa <mark>t</mark>	CTCATCACCA)	A <mark>ggtgcagg</mark> a <i>i</i>	ATT <mark>G</mark> ACCTC
i					1 2							
Consensus												
	GCAATTACAGAGGA	CTACATOATOT	TTGTOTOOT	TCCAREES	GATACT	TCARGACAARROAM			AGCA00447	CTCATCACCA	AGGTGCAGGA	ATTGACCTC
ľ	GCAATTACAGAGGA	UTACATGATGT1	- IGICICCI	IICCAGGGGG4	-GATACT	ICAGGACAAGGGAI	JUAGATAT(LAACAIGACAGC		UTUATCACCA.	REGISCAGGA	ATTGACCTC
									A 54%			
1												
		<										

MSA Estimation Methods

Basis: score alignments based on a similarity matrix and gap penalties

Most formulations of the problem are NP-complete.

Polynomial for two sequences (dynamic programming)

There are plenty of methods to estimate alignments:

- Progressive methods: use a guide tree to align sequences two at a time, from most similar to more distantly related.
- Iterative methods: similar to progressive, but allow updating pairwise alignments if scores are improved
- Hidden Markov models: model "current" alignment as a Markov model, and use Viterbi algorithm to successively add new sequences to the current alignment

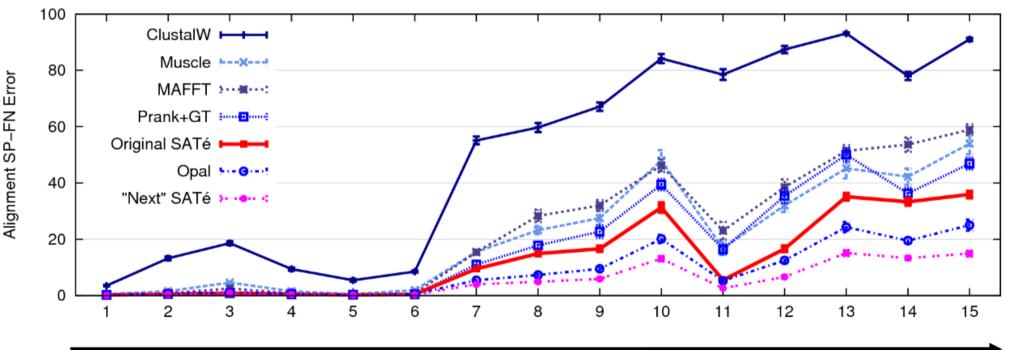
Alignment Comparison

Many ways to estimate alignments

• Alignments need to be compared

Alignment Comparison: performance Study

- Assessing accuracy in performance studies
- Example:



1000 taxon models ranked by difficulty

From: Liu,K. et al. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science, 324, 1561–1564.

Alignment Comparison: Phylogenetic Uncertainty

• Different MSA methods produce alignments that differ enough to introduce phylogenetic uncertainty (Wong et al., 2008)

• Alignment error increases with the size of the dataset (Liu et al., 2009, 2010)

Using several alignments, and comparisons of these alignments

Alignments Comparison Metrics

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Homology: Any pair of characters in the same column of a MSA

012345678 0123456789

- AGTGCTTC- AGTGCTTC--
- **1** A---CTCCA **1** A---CT-CCA
- 2 AC-CGTCCA 2 ACC-GT-CCA

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Homology: Any pair of characters in the same column of a MSA

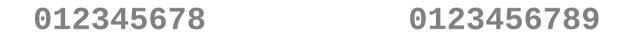
012345678 0123456789

- **O** AGTGCTTC-- **O** AGTGCTTC--
- **1** A---CTCCA **1** A---CT-CCA
- 2 AC-CGTCCA 2 ACC-GT-CCA

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Homology: Any pair of characters in the same column of a MSA



- AGTGCTTC- AGTGCTTC--
- **1 A**---CTCCA **1** A---CT-CCA
- 2 AC-CGTCCA 2 ACC-GT-CCA

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Homology: Any pair of characters in the same column of a MSA

012345678 0123456789

- **O** AGTGCTTC-- **O** AGTGCTTC--
- **1 A**---CTCCA **1** A---CT-CCA
- 2 AC-CGTCCA 2 ACC-GT-CCA

Homologies (count)

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Number of Homologies: two chose number of characters per column

012345678 0123456789

- AGTGCTTC- AGTGCTTC--
- **1** A---CTCCA **1** A---CT-CCA
- **2** AC-CGTCCA

310133331 total=18

2 ACC-GT-CCA

3110330311

total=16

Representing Characters

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Character Representation: a pair (a,b) where a indicates the row in the alignment matrix b indicates the position of the character in *unaligned* sequence

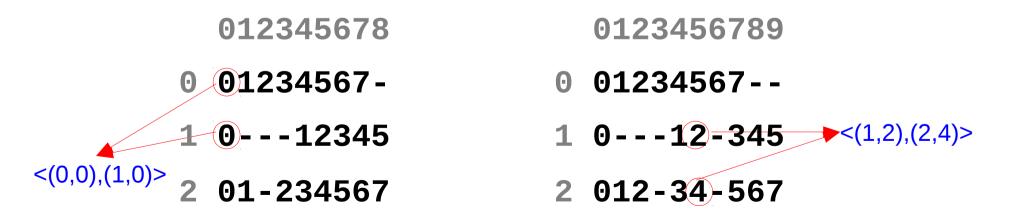


Representing Homologies (homology)

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Homology Representation: a pair <(a,b),(c,d)> where (a,b) each represent a character in the alignment, and (a,b) and (c,d) are in the same column of the alignment.



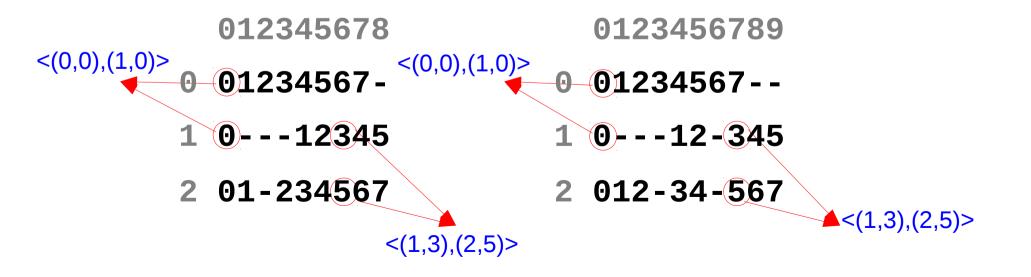
Note: the order doesn't matter: $\langle (a,b), (c,d) \rangle = \langle (c,d), (a,b) \rangle$

Shared Homology

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Shared Homologies: two homologies are shared between the two alignments if they have the exact same representation.



Shared Homology

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

Shared Homologies: two homologies are shared between the two alignments if they have the exact same representation.



SP-Score

• The Developer score = SP-score (sum-of-pairs):

Percentage of Homologies in Reference Alignment that are found in the estimated alignment (shared homologies).

SP-Score: find all homologies in both alignments, find those that are shared, and divide by the number of homologies in the reference alignment.

Reference: 012345678 Estimated: 0123456789

- 0 01234567- 0 01234567--
- **1** 0---12345 **1** 0---12-345
- **2** 01-234567 **2** 012-34-567

ALL: **310133331** =18 SHARED: **310033111** =13

SP-Score=13/18=72%

Modeler Score

• The Modeler score:

Percentage of Homologies in the estimated Alignment that are found in the reference alignment (shared homologies).

SP-Score: find all homologies in both alignments, find those that are shared, and divide by the number of homologies in the reference alignment.

Reference: 012345678 Estimated: 0123456789

- **0** 01234567- **0** 01234567--
- **1** 0---12345 **1** 0---12-345 **Modeler Score=**
- 2 01-234567 2 012-34-567 ^{13/16=81%} 310133331 _{ALL:} 3110330311 ₌₁₆

SHARED: **3100330111** =13

Total Column Score

• Total Column (TC) score:

Percentage of *aligned* columns in the reference alignment that are found in the estimated alignment.

Reference: 012345678 Estimated: 0123456789

- 0 01234567- 0 01234567--
- **1** 0---12345 **1** 0---12-345
- **2** 01-234567 **2** 012-34-567
- ALIGNED: **YYNYYYYY** =8
- SHARED: **YYNNYYNNY** =6

TC Score= 6/8=75%

Definitions

k = number of characters in the longest sequence
k1 = number of sites in the reference alignment
k2 = number of sites in the estimated alignment
n = number of sequences

Reference: 012345678 Estimated: 0123456789

- 0 01234567- 0 01234567--
- **1** 0---12345 **1** 0---12-345
- **2** 01-234567 **2** 012-34-567

k=7 k1=9

n=3 k2=10

Brute Force Calculation

 Homologies in each alignment can be represented as a presence/absence matrix with n.k rows and columns

• O(n²k²) time and memory.

FastSP: Objectives

Show that all three scores can be calculated in linear time (with respect to k.n)

Implement an efficient algorithm to calculate alignment scores

FastSP: Idea

- Characters in each column (x) of the reference alignment are dispersed in one or more columns in the estimated alignment.
- Divide characters in column x into equivalence classes, such that all characters in the same equivalence class are in the same column in the estimated alignment

 $\binom{1}{2} + \binom{2}{2} = 1$

- Number of shared homologies contributed by column x is
 - sum (for all equivalence classes S of x) |S| choose 2

```
      Reference:
      012345678
      Estimated:
      012345679

      0
      01234567-
      0
      01234567--

      1
      0---12345
      1
      0---12-345

      2
      01-234567
      2
      012-34-567
```

FastSP: Algorithm



1- Read reference alignment and save it with this character representation

• (also find k and n).

FastSP: Algorithm

Reference:	
Keterence.	
ROTOTOTIOO.	

012345678

- 0 01234567-
- 1 0---12345
- 2 01-234567

Estimated:

0123456789

- 0 01234567--
- 1 0---12-345
- 2 012-34-567

Matrix S:

01234567

- 0 01234567
- 1 045789--
- 2 01245789

2- Read estimated alignment and create a n by k matrix S such that

• S[i,j]=x iff Estimated_Alignment[i,x]=j.

Reference:

012345678

- 01234567-
- $1 \quad 0 - 12345$
- 2 01-234567

Estimated:

0123456789

- 0 01234567--
- $1 \quad 0 - 12 345$
- 2 012-34-567

Matrix S:

01234567

- 01234567 Θ
- 1 045789--
- 2 01245789

- Mu = An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x] = $\sum \binom{Mu_j}{2}$

Reference:

012345678

- 0 1234567-
- 1 0---12345
- 2 01-234567

Estimated:

0123456789

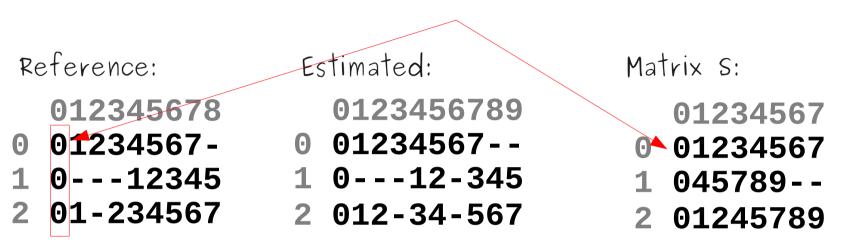
- 0 01234567--
- 1 0---12-345
- 2 012-34-567

Matrix S:

- 01234567
- 0 01234567
- 1 045789--
- 2 01245789

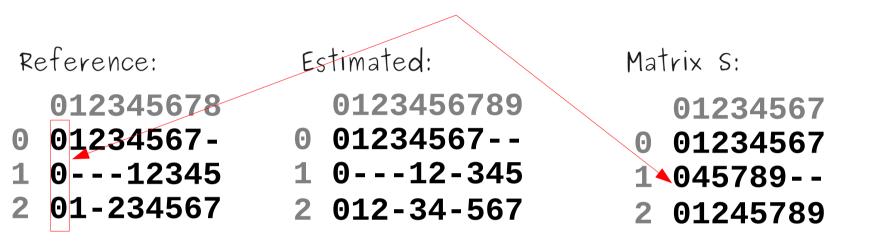
Mu=[0 0 0 0 0 0 0 0 0 0 0 0]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x] = $\sum \binom{Mu_j}{2}$



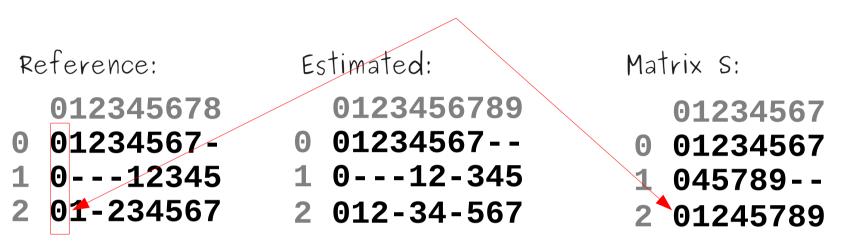
Mu=[1000000000]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x] = $\sum \binom{Mu_j}{2}$



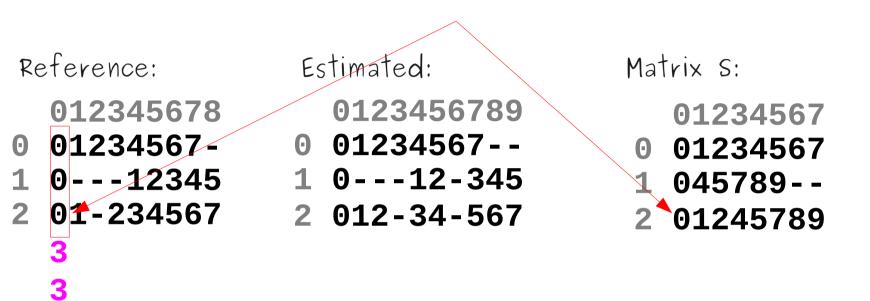
Mu=[2000000000]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x] = $\sum \binom{Mu_j}{2}$



Mu=[3000000000]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x] = $\sum \binom{Mu_j}{2}$



Shared= $\binom{3}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} + \ldots = 3$ Mu=[3 0 0 0 0 0 0 0 0 0]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]

- Shared [x] =
$$\sum {Mu_j \choose 2}$$

	•	
Dot	OVO	nce:
PC I	CIC	nce.

012345678

- 01234567-Θ
- 2 01-234567 31

31

Estimated:

0123456789

- 0 01234567--
- $1 \ 0 - 12345 \ 1 \ 0 - 12 345$
 - 2 012-34-567

Matrix S:

- 01234567
- 0 01234567
- 1 045789--
- 2 01245789

Shared = $\binom{0}{2} + \binom{2}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} = 1$ Mu=[0 2 0 0 0 0 0 0 0 0]

- Mu = An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$

ſ	`	
Dot	OKO	MOQ.
NEI	ere	NUC.

012345678

- 01234567-Θ
- 2 01-234567 310 310

Estimated:

0123456789

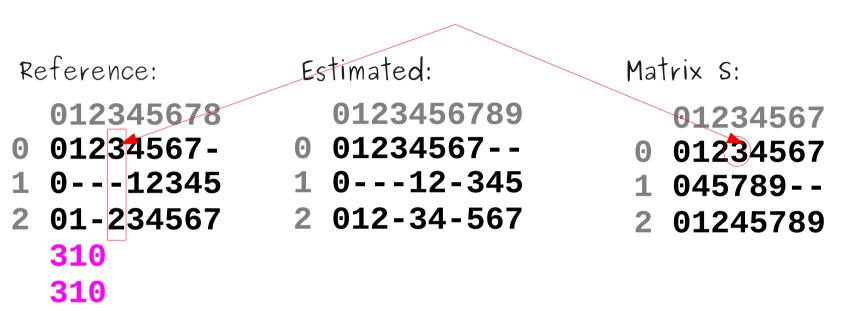
- 0 01234567--
- $1 \ 0 - 12345 \ 1 \ 0 - 12 345$
 - 2 012-34-567

Matrix S:

- 01234567
- 0 01234567
- 1 045789--
- 2 01245789

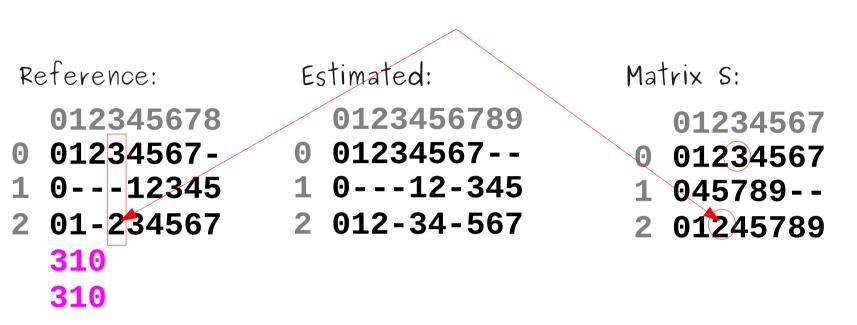
Shared = $\binom{0}{2} + \binom{0}{2} + \binom{1}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} + \binom{0}{2} = 0$ Mu=[001000000]

- Mu = An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$



Mu=[000100000]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$



Mu=[0 0 1 1 0 0 0 0 0 0]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$

Reference:	

012345678

- 01234567-Θ
- 2 01-234567 3100 3101

Estimated:

0123456789

- 0 01234567--
- 1 0 - 12345 1 0 - 12 345
 - 2 012-34-567

Matrix S:

- 01234567
- 0 01234567
- 1 045789--
- 2 01245789

Shared = $\binom{0}{2} + \binom{0}{2} + \binom{1}{2} + \binom{1}{2} + \binom{0}{2} + \binom{0}{2} + \dots = 0$ Mu=[0 0 1 1 0 0 0 0 0]

- Mu = An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$

Reference:

012345678

- 0 01234567-
- **1** 0---**1**2345
- 2 01-2<mark>3</mark>4567 31003 31013

Estimated:

0123456789

- 0 01234567--
- 1 0---12-345
- 2 012-34-567

Matrix S:

- 01234567
- 0 01234567
- 1 045789--
- 2 01245789

Mu=[0 0 0 0 3 0 0 0 0 0]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$

Reference:

012345678

- 0 01234567-
- **1** 0---**1**2345
- 2 01-23<mark>4</mark>567 310033 310133

Estimated:

0123456789

- 0 01234567--
- 1 0---12-345
- 2 012-34-567

Matrix S:

- 01234567
- 0 01234567
- 1 045789--
- 2 01245789

Mu=[0 0 0 0 0 3 0 0 0 0]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$

Reference:

012345678

- 0 01234567-
- $1 \ 0 - 12 \ 345 \ 1 \ 0 - 12 345$
- 2 01-234567 3100331 3101333

Estimated:

0123456789

- 0 01234567--
- 2 012-34-567

Matrix S:

- 01234567
- 0 01234567
- 1 045789--
- 2 01245789

Shared=...+ $\binom{0}{2}$ + $\binom{0}{2}$ + $\binom{1}{2}$ + $\binom{1}{2}$ + $\binom{2}{2}$ +... = 1 Mu=[0 0 0 0 0 1 2 0 0]

- Mu = An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum {\binom{Mu_j}{2}}$

Reference:

012345678

- 01234567- $\mathbf{\Theta}$
- $1 \ 0 - 12345 \ 1 \ 0 - 12 345$
- 2 01-234567 31003311 31013333

Estimated:

0123456789

- 0 01234567--

 - 2 012-34-567

Matrix S:

01234567

- 0 01234567
- 1 045789--
- 2 01245789

Shared=...+ $\binom{0}{2}$ + $\binom{0}{2}$ + $\binom{1}{2}$ + $\binom{1}{2}$ + $\binom{2}{2}$ +... = 1 Mu=[0 0 0 0 0 0 1 2 0]

- Mu = An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$

- 012345678
- 0 01234567-
- **1** 0---12345
- 2 01-234567 310033111 310133331

Estimated: 0123456789 0 01234567--

- 1 0---12-345
- 2 012-34-567

Matrix S:

- 01234567
- 0 01234567
- 1 045789--
- 2 01245789

Shared=...+ $\binom{0}{2}$ + $\binom{0}{2}$ + $\binom{0}{2}$ + $\binom{0}{2}$ + $\binom{2}{2}$ = 1 Mu=[0 0 0 0 0 0 0 0 2]

- Mu= An array of length k2 initialized by 0 (or a dictionary)
- For character M in row r
 - Increment Mu[S[r][M]]
- Shared [x]= $\sum \binom{Mu_j}{2}$

Reference:

012345678

- 0 01234567 -
- 1 0---12345
- 2 01-234567 310033111=13 310133331=18

Estimated:

0123456789

- 0 01234567--
- 1 0---12-345
- 2 012-34-567

Matrix S:

01234567

- 0 01234567
- 1 045789--
- 2 01245789

SP-Score=13/18

4- Report (sum of shared)/(sum of reference) as SP-Score

Running Time Analysis

1- read reference alignment and save it with our character representation

O(n.k1)

- 2- read estimated alignment and create a n by k matrix S such that
- S[i,j]=x iff Estimated_Alignment[i,x]=j.
 O(n.k2)
- 3- For each column of reference alignment (k1)
 - Mu= An array of length k2 initialized by 0 (or a dictionary)
 - For character M in row r (n)
 - Increment Mu[S[r][M]]
 - Shared [x]= $\sum \binom{Mu_j}{2}$
 - O(n.k1)
- 4- Report (sum of shared)/(sum of reference)

O(k1)

Overall=O(max(k1,k2).n)

Memory Analysis

1- read reference alignment and save it with our character representation

O(n.k1)

- 2- read estimated alignment and create a n by k matrix S such that
- S[i,j]=x iff Estimated_Alignment[i,x]=j.
 O(n.k)
- 3- For each column of reference alignment
 - Mu= An array of length k2 initialized by 0 (or a dictionary) O(k2) or O(n)
 - For character M in row r
 - Increment Mu[S[r][M]]
 - Shared [x]= $\sum {\binom{Mu_j}{2}}$
- 4- Report (sum of shared)/(sum of reference)

Overall=O((k1+k).n)

Memory Analysis

 Trick: choose smallest of k1 and k2 as reference alignment and the other as estimated alignment. Number of shared homologies will be the same either way.

Overall=O((min(k1,k2)+k).n)

Modeler and TC scores

- Both Modeler and TC scores can be calculated with FastSP algorithm without any sacrifice to running time and memory
- Modeler: we just need to calculate total number of homologies in the estimated alignment
- TC: As we go through column of reference alignment, if we get only one equivalence class, we have a correct column.

Implementation

- Implemented in Java
 - Computes SP, Modeler, and TC in one run

• 420 LOC

 Available publicly at http://www.cs.utexas.edu/~phylo/software/fastsp/

Performance Study: datasets

S.Mirarab and T.Warnow

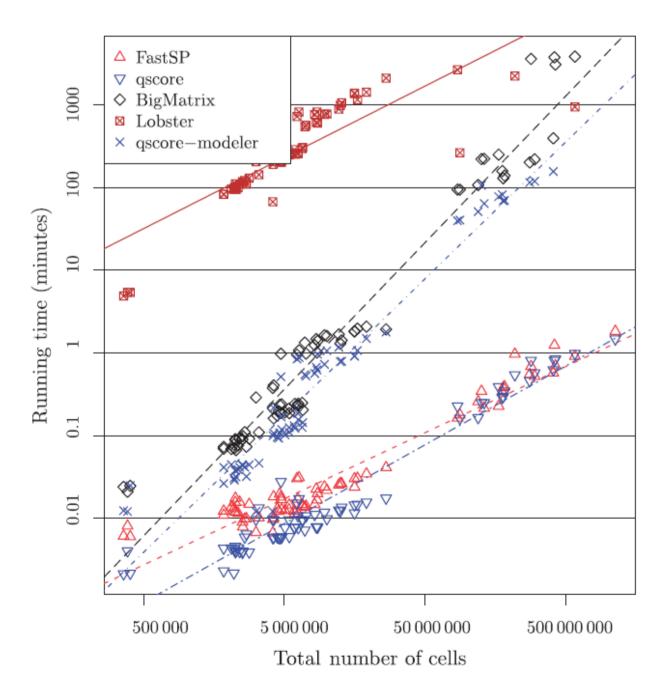
Table 1. Datasets and their sizes

Dataset	n ^a	L	Ref.	MAFFT	OPAL	PART	PRANK	QUICK	SATé	SATé-II
100L1-R0	100	1089	2287	1563						1737
500L1-R0	500	1110	4992	3307						3421
1000L1-R0	1000	1079	3517	907						2856
Price-78K	78132	1286	1287			1504				
23S.E	117	5317	9079	8929	9860	11018	13941	6796	10352	
23S.E.aa_ag	144	1079	8619	8123	9576	10956	14343	7017	9029	
23S.M.aa_ag	263	4483	10305	7353	13625	12320	13471	5522	7815	
23S.M	278	4216	10738	7478	10447	13384	13639	5311	8746	
16S.M	901	2023	4722	4418	12812	9496	12826	3216	4776	
16S.M.aa_ag	1028	2672	4907	4493	13785	11225	20856	3317	48881	
16S.3	6323	4066	8716			19775		5310	10186	20414
16S.T	7350	4066	11856	10891	43797	25951		6109	12301	29156
16S.B.ALL	27643	1851	6857			14217		3413		8463
16S.GG-50K	50000	1701	7682			14877				

 a_n is the number of sequences. *L* is the maximum number of nucleotides in any of the sequences. *Ref.* is the length of the reference alignment (i.e. including gaps). The rest of the columns show the length of the alignment for each estimated alignment. An empty cell indicates that the respective alignment method was not run on that particular dataset. The first four datasets are simulated datasets, while the rest are all real biological datasets.

We have included results from two different runs of SATé-II on 16S.B.ALL dataset. The second version had a length of 8209.

Performance Study: All Techniques



Performance Study: Q-Score

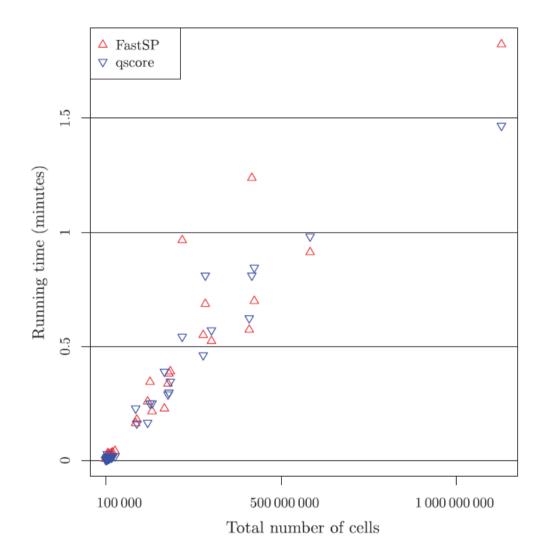


Fig. 4. Comparison of FASTSP and QSCORE-default with respect to running time on machines with 'at least' 8 GB of main memory.

Performance Study: Memory

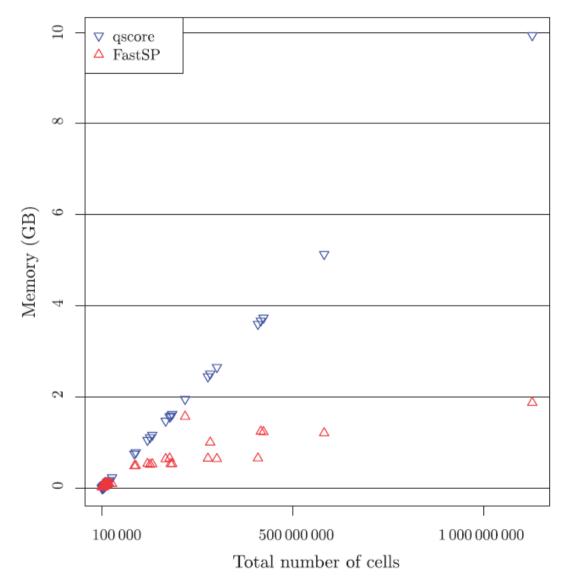


Fig. 5. Comparison of FASTSP and QSCORE-default with respect to peak memory usage on machines with 'at least' 8 GB of main memory.

Performance Study: Limited Memory

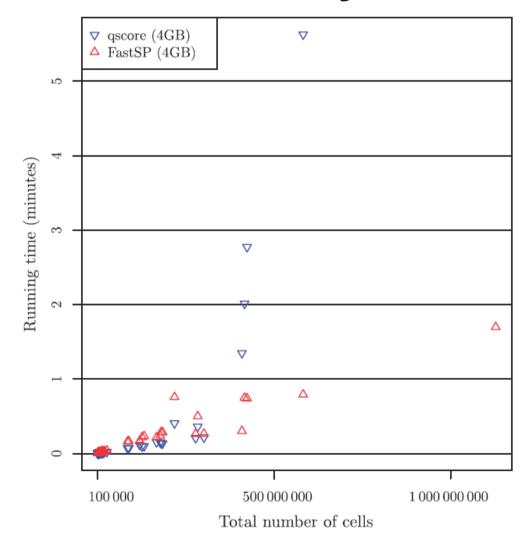


Fig. 1. Log-scaled running time on machines with 4 GB of main memory. QSCORE is run only in default setting, and so computes only SP- and TC-scores. Note that QSCORE fails to analyze the largest dataset.

Summary

 Two alignments can be compared in terms of SP-Score, Modeler Score, and TC in linear time (linear with respect to k.n)

• FastSP provides a memory-efficient tool for comparing alignments