Correction of Genome-MSAs around Exons

Background

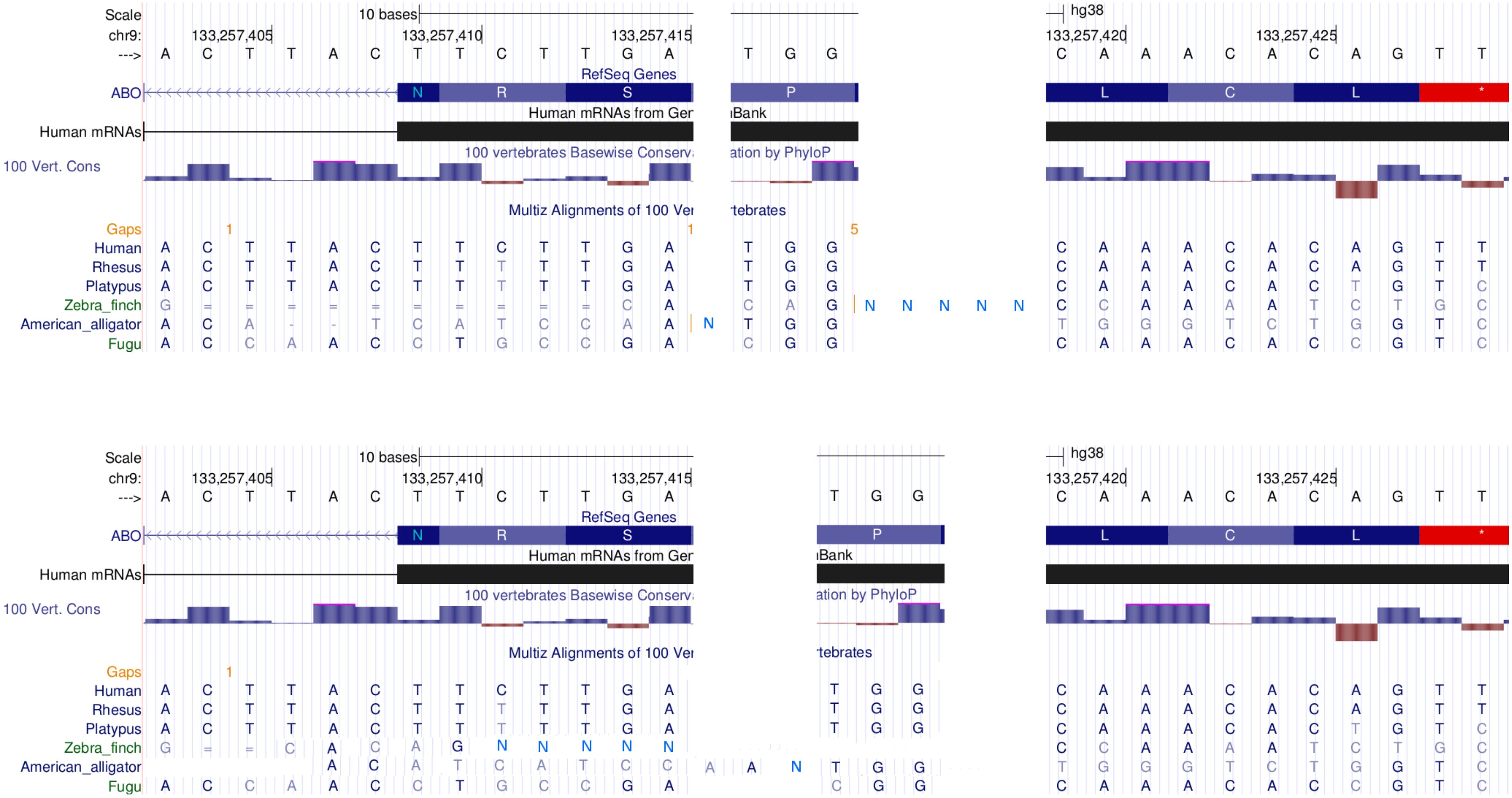
The tree of life is denser and denser populated with sequenced genomes. Genomes of related species are compared, for example to analyse evolutionary events that changed the protein sequence or even the gene structure. For closely related species a multiple alignment of the genomes can be used for that purpose. For example, the UCSC Genome browser group has already aligned the genomes of more than 100 vertebrate species – and an extension to 10 000 species is planned.
In constructing a multiple genome alignment, the alignment program often can choose between several similarly plausible local variants of an alignment, e.g.

```
AAAGTACTGTCGA        AAAGTACTGTCGA
                or
AGAGT-----CGA        AGA-----GTCGA
```

This happens more often if the species are not very closely related, e.g. mammals and fish. The multiple genome alignment is usually constructed using iterative pairwise alignments, with a scoring scheme that uses a 4x4 scoring matrix and affine gap costs. The alignment program is unaware of the gene structure, in particular, it does not exploit the fact that in genes

- bases that are aligned with each other should almost always be in the same frame (0, 1 or 2 depending on the position in the codon)
- alignments in which the exon boundaries are aligned with each other should be preferred if they are not clearly worse.

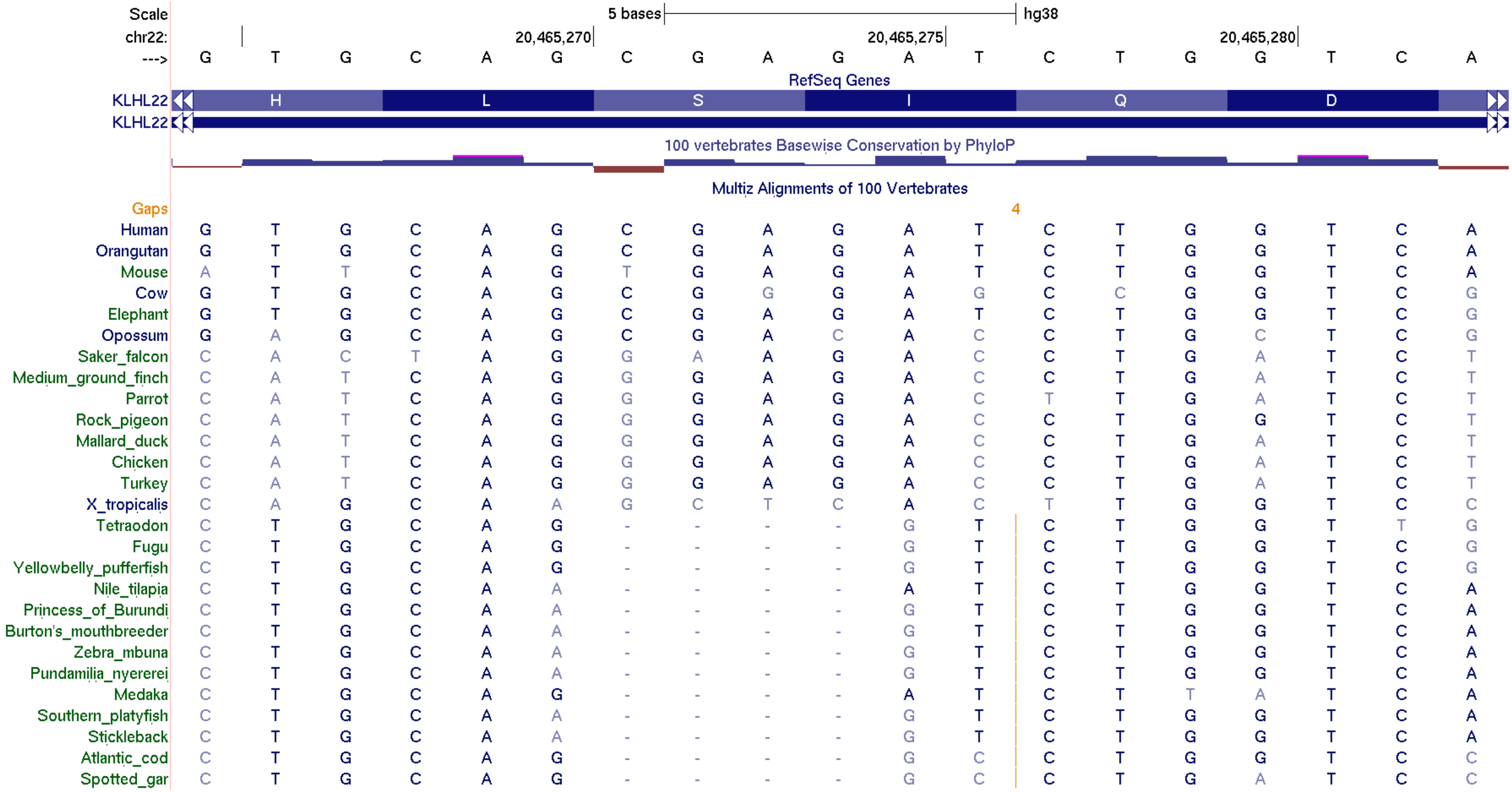
The next two pages show examples of likely alignments errors, as the aligments can be corrected using above two criteria by changing a few gaps only.



```
Scale                           10 bases                                      hg38
chr9:        133,257,405       133,257,410       133,257,415              133,257,420       133,257,425
--->      A  C  T  T  A  C  T  T  C  T  T  G  A     T  G  G               C  A  A  A  C  A  C  A  G  T  T
                                        RefSeq Genes
ABO                       N     R        S          P                       L        C        L      *
                              Human mRNAs from GenBank
Human mRNAs
                       100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
                           Multiz Alignments of 100 Vertebrates
Gaps            1                                1         5
Human            A  C  T  T  A  C  T  T  C  T  T  G  A     T  G  G                C  A  A  A  C  A  C  A  G  T  T
Rhesus           A  C  T  T  A  C  T  T  T  T  T  G  A     T  G  G                C  A  A  A  C  A  C  A  G  T  T
Platypus         A  C  T  T  A  C  T  T  T  T  T  G  A     T  G  G                C  A  A  A  C  A  C  T  G  T  C
Zebra_finch      G  =  =  =  =  =  =  =  =  =  =  C  A     C  A  G | N N N N N    C  C  A  A  A  A  T  C  T  G  C
American_alligator A C  A  -  -  T  C  A  T  C  C  A  A | N  T  G  G                T  G  G  G  T  C  T  G  G  T  C
Fugu             A  C  C  A  A  C  C  T  G  C  C  G  A     C  G  G                C  A  A  A  C  A  C  C  G  T  C
```

```
Scale                           10 bases                                      hg38
chr9:        133,257,405       133,257,410       133,257,415              133,257,420       133,257,425
--->      A  C  T  T  A  C  T  T  C  T  T  G  A           T  G  G         C  A  A  A  C  A  C  A  G  T  T
                                        RefSeq Genes
ABO                       N     R        S                P                 L        C        L      *
                              Human mRNAs from GenBank
Human mRNAs
                       100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
                           Multiz Alignments of 100 Vertebrates
Gaps            1
Human            A  C  T  T  A  C  T  T  C  T  T  G  A           T  G  G          C  A  A  A  C  A  C  A  G  T  T
Rhesus           A  C  T  T  A  C  T  T  T  T  T  G  A           T  G  G          C  A  A  A  C  A  C  A  G  T  T
Platypus         A  C  T  T  A  C  T  T  T  T  T  G  A           T  G  G          C  A  A  A  C  A  C  T  G  T  C
Zebra_finch      G  =  =  C  A  C  A  G  N  N  N  N  N                            C  C  A  A  A  A  T  C  T  G  C
American_alligator           A  C  A  T  C  A  T  C  C  A  A  N  T  G  G          T  G  G  G  T  C  T  G  G  T  C
Fugu             A  C  C  A  A  C  C  T  G  C  C  G  A           C  G  G          C  A  A  A  C  A  C  C  G  T  C
```

Figure 1: *Top:* UCSC-Browser shot of a MULTIZ alignment in the range of a donor splice site of a human (hg38 assembly) gene on the reverse strand. The inserions (1bp in american alligator, 5bp in zebra finch) are shown as N only but the sequence is available. *Bottom:* By realigning a part of the zebrafinch and alligator sequence the splice site consensus at the rightmost two intronic bases (`AC` = reverse complement of `GT`) can be conserved. Further, in the new alignment the relative frame of all aligned bases is the same, in contrast to the original alignment at the top.


Scale
5 bases
hg38
chr22:
20,465,270
20,465,275
20,465,280
RefSeq Genes
KLHL22
KLHL22
H
L
S
I
Q
D
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates

--->	G	T	G	C	A	G	C	G	A	G	A	T		C	T	G	G	T	C	A
Gaps													4							
Human	G	T	G	C	A	G	C	G	A	G	A	T		C	T	G	G	T	C	A
Orangutan	G	T	G	C	A	G	C	G	A	G	A	T		C	T	G	G	T	C	A
Mouse	A	T	T	C	A	G	T	G	A	G	A	T		C	T	G	G	T	C	A
Cow	G	T	G	C	A	G	C	G	G	G	A	G		C	C	G	G	T	C	G
Elephant	G	T	G	C	A	G	C	G	A	G	A	T		C	T	G	G	T	C	G
Opossum	G	A	G	C	A	G	C	G	A	C	A	C		C	T	G	C	T	C	G
Saker_falcon	C	A	C	T	A	G	G	A	A	G	A	C		C	T	G	A	T	C	T
Medium_ground_finch	C	A	T	C	A	G	G	G	A	G	A	C		C	T	G	A	T	C	T
Parrot	C	A	T	C	A	G	G	G	A	G	A	C		T	T	G	A	T	C	T
Rock_pigeon	C	A	T	C	A	G	G	G	A	G	A	C		C	T	G	G	T	C	T
Mallard_duck	C	A	T	C	A	G	G	G	A	G	A	C		C	T	G	A	T	C	T
Chicken	C	A	T	C	A	G	G	G	A	G	A	C		C	T	G	A	T	C	T
Turkey	C	A	T	C	A	G	G	G	A	G	A	C		C	T	G	A	T	C	T
X_tropicalis	C	A	G	C	A	A	G	C	T	C	A	C		T	T	G	G	T	C	C
Tetraodon	C	T	G	C	A	G	-	-	-	-	G	T	|	C	T	G	G	T	T	G
Fugu	C	T	G	C	A	G	-	-	-	-	G	T	|	C	T	G	G	T	C	G
Yellowbelly_pufferfish	C	T	G	C	A	G	-	-	-	-	G	T	|	C	T	G	G	T	C	G
Nile_tilapia	C	T	G	C	A	A	-	-	-	-	A	T	|	C	T	G	G	T	C	A
Princess_of_Burundi	C	T	G	C	A	A	-	-	-	-	G	T	|	C	T	G	G	T	C	A
Burton's_mouthbreeder	C	T	G	C	A	A	-	-	-	-	G	T	|	C	T	G	G	T	C	A
Zebra_mbuna	C	T	G	C	A	A	-	-	-	-	G	T	|	C	T	G	G	T	C	A
Pundamilia_nyererei	C	T	G	C	A	A	-	-	-	-	G	T	|	C	T	G	G	T	C	A
Medaka	C	T	G	C	A	G	-	-	-	-	A	T	|	C	T	T	A	T	C	A
Southern_platyfish	C	T	G	C	A	A	-	-	-	-	G	T	|	C	T	G	G	T	C	A
Stickleback	C	T	G	C	A	A	-	-	-	-	G	T	|	C	T	G	G	T	C	A
Atlantic_cod	C	T	G	C	A	G	-	-	-	-	G	C	|	C	T	G	G	T	C	C
Spotted_gar	C	T	G	C	A	G	-	-	-	-	G	C	|	C	T	G	A	T	C	C

Figure 2: A deletion of 4bp in the fish with respect to the other vertebrates is followed only 2bp later by an insertion of 4bp as indicated by the orange 4 and the vertical lines. An obvious correction – by removing any indels – would conserve the frame throughout this alignment section. (hg38, MULTIZ)

Task

Given a section of an input multiple genome alignment around a likely exon, write a program that improves the input alignment: Firstly, develop a scoring model that considers above two criteria besides the classical scoring scheme mentioned above (4x4, gap open, gap extension). Secondly, realign sequences in areas with possible errors such that the new score is optimized.

Afterwards, apply the algorithm to large numbers of example regions and analyse the changes that the program introduced. E.g. how often have the corrections a better or worse classical score? How many base pairs had to be realigned? Estimate the number of errors in the input alignment.

A possible approach

A possible approach is based on the assumption that the majority of the input alignment is correct. Each alignment row can be corrected individually by realigning it to the MSA. Every input alignment column can be assigned a nucleotide profile. A second profile (f_{ij}) could be constructed that contains for each alignment column i and each relative frame $j \in \{0, 1, 2, \text{NULL}\}$ the relative frequency of frame j in column i (NULL if missing, i.e. outside of exon). Thirdly, an extra score is introduced for alignments that place the required consensus dinucleotide for splice sites (or a start or stop codon in cases of initial and terminal coding exons) at the alignment columns which correspond to the exon boundaries given in the input.

Programming

Writing the code in C/C++ is preferred, as data structures can be reused and – if successful – an efficient implementation of this project will be used in our code base and applied to whole genome vertebrate alignments.

Refereces

- Stefanie König, Lars Romoth, Lizzy Gerischer, and Mario Stanke (2015), "Simultaneous gene finding in multiple genomes. PeerJ PrePrints", e1594, `http://peerj.com/preprints/1296.pdf`
- `http://genome.ucsc.edu`
- Schwartz et al., 2003, "Human-Mouse Alignments with BLASTZ", `http://genome.cshlp.org/content/13/1/103.full.pdf+html`. This is the pairwise alignment program used inside of MULTIZ. The paper contains the 4x4 matrix and the affine gap cost model.
- Blanchette et al. 2004, "Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner", `http://genome.cshlp.org/content/14/4/708.full.pdf+html`. This program was used to create the alignments in above examples.