Additional Supplemental Material For: A Novel And Well-Defined Benchmarking Method For Second Generation Read Mapping

Manuel Holtgrewe Anne-Katrin Emde David Weese
Knut Reinert

Version 1
June 2011

Abstract

This manuscript contains extended explanations and proofs of well-formedness for the equivalence classes defined in

M. Holtgrewe, A.-K. Emde, D. Weese and K. Reinert. "A Novel And Well-Defined Benchmarking Method For Second Generation Read Mapping", *BMC Bioinformatics*, Vol. 12, May, 2011.

1 Equivalence Class Definitions

In the following, we will state the complete Definitions 1–3 from the paper again and then refine the definition of match equivalency $\equiv$ in Definition 4.

For numbers a, b, c in the following, we assume w.l.o.g. that $a \leq b \leq c$. All other cases follow by a simple role change argument on these variables. Thus, symmetry is not explicitely considered in the following.

Definition 1 (Neighbour Equivalence) *Two feasible matches (identified by their end positions) a, b are* neighbour equivalent *($a \overset{N}{\equiv} b$) if for all x, $a \leq x \leq b$ the following holds: $\delta(x) \leq k$.*

Well-formedness of $\overset{N}{\equiv}$. The reflexivity of $\overset{N}{\equiv}$ follow directly from the definition. Now, let $a \overset{N}{\equiv} b$ and $b \overset{N}{\equiv} c$. Since b is feasible as well, $\delta(x) \leq k$ holds for all $a \leq x \leq b$ and $\overset{N}{\equiv}$ is an equivalence relation.

Definition 2 (Trace Equivalence) *We define two matches a, b to be* trace equivalent *($a \overset{T}{\equiv} b$) if their traces share a part. This is the case if their canonical start positions are equal.*

Well-formedness of $\overset{T}{\equiv}$. Reflexivity and transitivity follow directly from the definition through the trace.

Observation 1 *When two position a, b are trace equivalent, then they are also trace equivalent to all x with $a \leq x \leq b$.*

Any canonical trace from x must cross the traces from a or b. Therefore, it must have the same canonical start position s.

Definition 3 (k-Trace Equivalence) *Two matches a, b are* k-trace equivalent *($a \overset{kT}{\equiv} b$) if one of the following holds: (1) They are feasible, neighbour equivalent, and trace equivalent. (2) There exist feasible, trace equivalent matches α, β and a separating match ζ such that $\alpha \leq a \leq \zeta \leq b \leq \beta$.*

A separating match *ζ is a match with $\delta(\zeta) > k$ and there exists α, β, $\alpha < \zeta < \beta$ such that $\delta(\alpha), \delta(\beta) \leq k$.*

Well-formedness of $\overset{kT}{\equiv}$. Reflexivity is easy to see in both cases (1) and (2).

For transitivity, we perform a case distinction. Given $a \overset{kT}{\equiv} b$ and $b \overset{kT}{\equiv} c$. There now are four cases, depending which case (1, or 2) holds for the left and right relation.

Case (1, 1). The transitivity follows since $a \overset{N}{\equiv} b$ and $b \overset{N}{\equiv} c$ holds.

Cases (1, 2) and (2, 1). (We show the proof for case (2, 1), case (1, 2) follows analogously.) a, b are feasible, neighbour equivalent, and trace equivalent. For b, c, there exist feasible, trace equivalent matches α, β and a separating match ζ such that $\alpha \leq b \leq \zeta \leq c \leq \beta$. Because of Observation 1 and $\alpha \leq b$ and $a \overset{T}{\equiv} b$, a and α have to be trace equivalent. Transitively, a and β have to be trace equivalent. a can take the role of α from before: $a \leq a \leq \zeta \leq b \leq \beta$ and thus transitivity is shown.

Case (2, 2). For a, b, c exist feasible matches $\alpha, \beta, \alpha', \beta'$ and separating matches ζ, ζ', such that $\alpha \leq a \leq \zeta \leq b \leq \beta$, and $\alpha' \leq b \leq \zeta' \leq c \leq \beta'$. It follows that $\alpha \leq a \leq \zeta \leq b \leq \zeta' \leq c \leq \beta'$ and thus $\alpha \leq a \leq \zeta \leq c \leq \beta'$. α and β are trace equivalent; the same hold for α' and β'. Because of Observation 1, and $\alpha' \leq \beta$ it follows that α and β' have to be trace equivalent, too. Thus, we now have shown that case (2) of k-trace equivalency holds for a and c.

Definition 4 (Match Equivalence) *We say that two matches a, b are equivalent ($a \equiv b$) if there exist $\ell \geq 0$ feasible* connecting *matches $a \leq m_1 \leq \ldots \leq m_\ell \leq b$ such that:*

$$(a \overset{kT}{\equiv} m_1 \vee a \overset{N}{\equiv} m_1) \wedge \ldots \wedge (m_{i-1} \overset{kT}{\equiv} m_i \vee m_{i-1} \overset{N}{\equiv} m_i) \wedge \ldots \wedge (m_\ell \overset{kT}{\equiv} b \vee m_\ell \overset{N}{\equiv} b).$$

If $\ell = 0$ then two matches are equivalent if $a \overset{kT}{\equiv} b$ or $a \overset{N}{\equiv} b$.

The motivation for the disjunction in the definition of match equivalence is to join neighbouring k-trace equivalence classes/intervals when being directly adjacent to other k-trace equivalence classes or overlapping with neighbour equivalence classes. Such intervals/classes I, J (w.l.o.g. I left of J) are (in some sense) minimally joined by the rightmost match of I being neighbour equivalent to J or by the overlap.

Well-formedness of $\equiv$. Reflexivity follows from the case where $k = 0$. Transitivity for $a \equiv b$, $b \equiv c$ follows from using b as a connecting match if it is feasible. If b is not feasible then there exist feasible matches α, β that are trace-equivalent to b and they can be used as connecting matches.