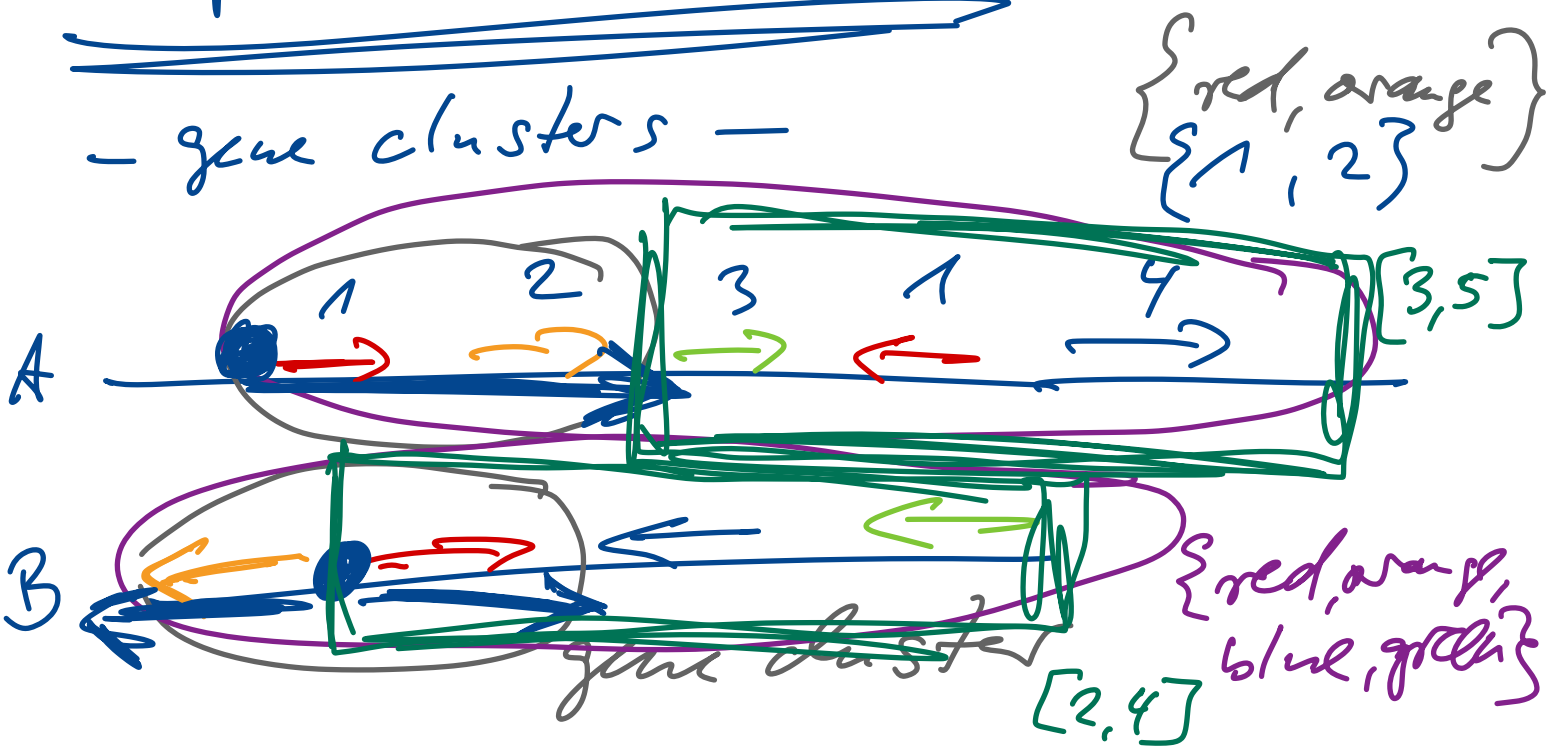
Comparative Genomics II

5. Feb. 2021

— gene clusters —



formally: common interval

S : sequence of "genes" (unsigned) $N = \{1, \dots, n\}$

$S[i]$: the i-th element of S

$[x,y]$: the set of indices $\{x, x+1, x+2, \dots, y\}$

$S[x,y]$: the substring $S[x]\,S[x+1] \dots S[y]$

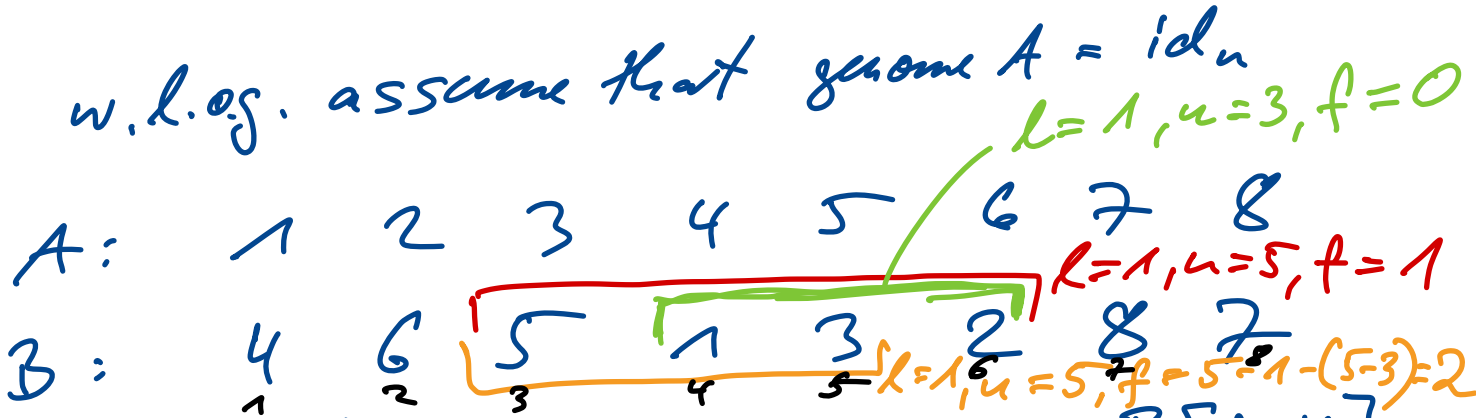
$\mathcal{Z}(S) = \{ S[i] \mid 1 \le i \le |S| \}$, "character set of S"

Def: Given a family of k sequences $\mathcal{S} = (S_1, \dots, S_k)$, then a common interval of $\mathcal{S}$ is a k-tuple $C = ([l_1, u_1], [l_2, u_2], \dots, [l_k, u_k])$ if and only if $\mathcal{Z}(S_1[l_1, u_1]) = \dots = \mathcal{Z}(S_k[l_k, u_k])$.

genome models

① genome = permutation of $\{1, \dots, n\}$

w.l.o.g. assume that genome A = id_n



test functions: $\ell(x,y) = \min B[x,y]$

$u(x,y) = \max B[x,y]$

$f(x,y) = u(x,y) - \ell(x,y) - (y-x)$

observation: $f(x,y)$ = number of "missing" genes in $B[x,y]$.

Lemma: $f(x,y) = 0 \iff A[x,y]$ is a common interval of A and B.

Algorithms

k=2 genomes

1. straightforward — $O(n^3)$ time
   → build all $\frac{n(n-1)}{2}$ intervals and compute f.
2. a bit more clever, using running min and running max — $O(n^2)$ time
3. Uno & Yagiura (2000) [no details] — $O(n+K)$ time where K = output size

4. Extension to Uno & Yagiura for $k \geq 2$ genomes – $O(kn+K)$ time

Heber, Mayr, Stoye (2011)

(2) genome = sequence over $N = \{1, \dots, n\}$

(Schmidt & Stoye, 2004/07)

[$O(n^2)$ time and space (for $k=2$ sequences)
$O(kn^2)$ time for k sequences

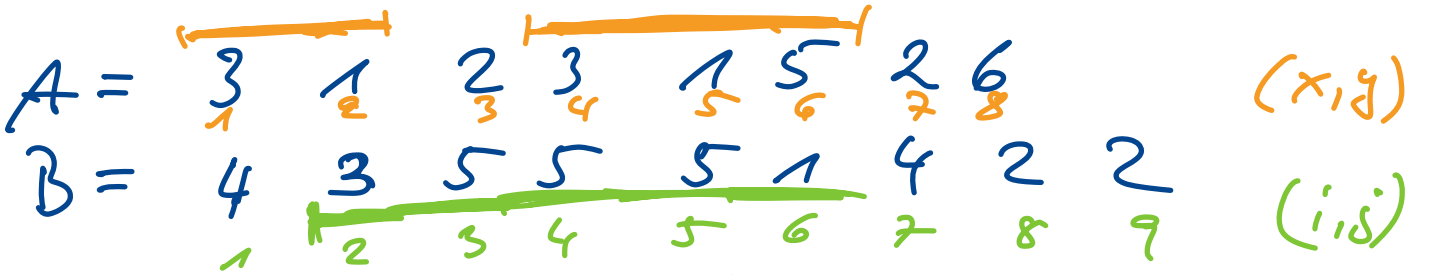
Preprocessing

$POS[c]$ = list of all occurrences of c in A

$NUM[x,y] = |\mathcal{Z}(A[x,y])|$, $x < y$

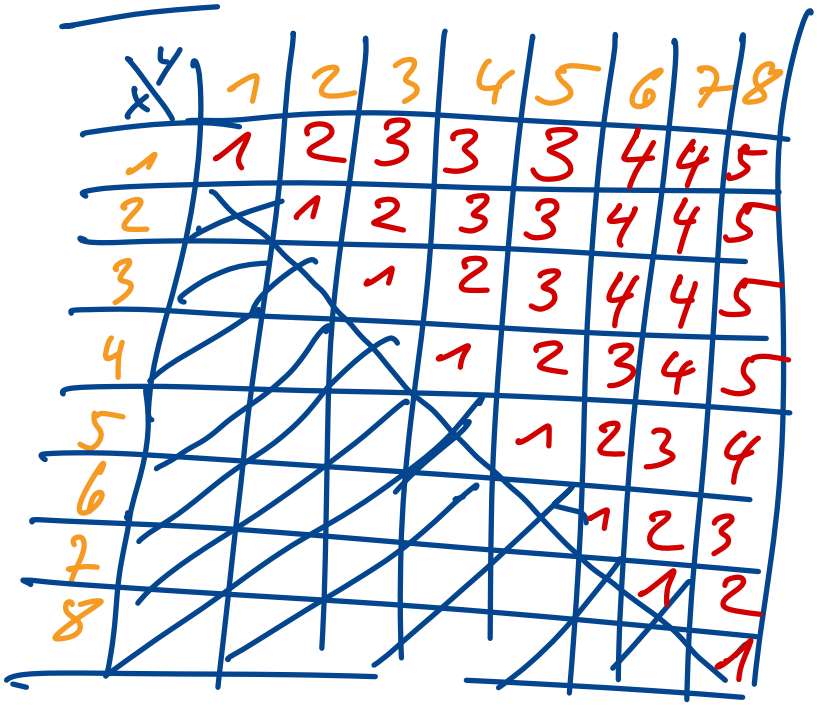
Algorithm: enumerate all intervals $[i,j]$ of B, while marking maximal intervals $[x,y]$ of A.

→ whenever such an interval fulfills the test that $|\mathcal{Z}(B[i,j])| = NUM[x,y]$ then output a common interval.



POS[1] = 2, 5
POS[2] = 3, 7
POS[3] = 1, 4
POS[4] = /
POS[5] = 6
POS[6] = 8

NUM



output:

([4,6], [2,6])

....

---

variants of the gene cluster model

(A) above: common interval of A and B.

(B) more general: approximate common interval

→ add some error tolerance to the common interval model

common interval: $Z(A[x,y]) = Z(B[i,j])$

approx. c. i.: $|Z(A[x,y]) - Z(B[i,j])| \leq t$

where

$\triangle$ is the symmetric set difference:

$$A \triangle B = (A \cup B) \setminus (A \cap B)$$

and t is a threshold value ≥ 0

(c) weak common intervals in indeterminate strings

example: A: 1 2 (3/4) 2 (3/4) (1/2/4) 3 ....

B: 2 3 (1/4) 3 (2/5) 3 ...

dfs w.c.i. if by removing some alternatives of indeterminate positions, we get a c.i.