Fast Python sequence classification

Piotr Styczyński

MIM UW

2024

Testing

Assuming we have match (r_1, r_2) and expect (e_1, e_2) LOL We define the following:

- $d_1 := |e_1 - r_1|$
- $d_2 := |e_2 - r_2|$
- $d_s := d_1 + d_2$
- $d_m := max(d_1, d_2)$
- $score((r_1, r_2)) = AA$ iff $d_s < 10$
- $score((r_1, r_2)) = AB$ iff $d_s < 20$
- $score((r_1, r_2)) = C$ iff $d_m < 20$
- $score((r_1, r_2)) = D$ iff $d_m \geq 20$

case name	ok	unmapped	AA	AB	C	D (bad)
reads20Mb	1000	0	936	64	0	0
reads20Ma	1000	0	975	25	0	0
reads2	100	0	90	10	0	0

References

Walther Wombat and Klaus Koala. "The true meaning of 42". In: *Journal of modern skepticism* (2016).

Thank you