#flo #inclass

---

1 | **Bio informatics!**

guinea pig time

1.1 | **bio overview**

about genomic information transfer and also, viruses!

recall: exons, introns, coding regions, non-coding regions, all that good stuff it's hard to find these exons and introns! not obvious also the concept of `consensus sequence`

real world gene time! :: 11k base pairs for a single gene in what is considered a simply organism worm genes are defined as `##-L` good test cases, as well documented! poor c elegans tho.. (but who cares, they are worms.)

c elegans advantage: - dvides quickly - has a development map (knows where each section leads to in the development) - and also, available? i mean, they're worms man

`https://docs.google.com/presentation/d/1Cj1jMeNIUOh3GchMhSkgA53OebWP1gdicwsPKuPHcps/edit?usp=sharing` ## some ideas - ml to identify certain aspects of base pair sequences? exrons/introns, etc → different protein versions from one gene, called transcripts. this is done with diff combinations of exrons - the ones that appear are experimentally verified

- semantic similarity graph? some time of fdg?
  - context-dependant similarty?
- some type of word vectors? genomic embedding space
- compression??
  -
- predict *folding/function* similarity with sequence?

F → E

(E_1 vs. E_2) → "true" similarity → model (L_1, L_2) → predict "true" similarity → similarity metric

metric vs. edit distance

"true" vs. nlp metrics

ACD =>1, 2, 9

[1, 9, 1, …, 2 2 4 4]