Description of seqenv's main internal data structures

Lucas Sinclair, lucas.sinclair@me.com

Last modified:

Thu Jul 23 2015

All numbers are fake
This document just describes a plan for processing data.

seq_to_gis
C0 -> [343206452, 343204350, 199999999]
C1 -> [234523451, 343204354]
C2 -> []

gi_to_text (For all NCBI GIs)
343206452 -> "the deepest sea mud of the Mariana Trench"
343204354 -> "isolated from terrestorial samples near ..."
199999999 -> ""
343204350 -> "the deepest sea mud of the Mariana Trench"
... etc etc

gi_to_matches (For only GIs found)
343206452 -> [ENVO:123123, ENVO:123777, ENVO:123666]
343204354 -> [ENVO:23452345]
199999999 -> []
343204350 -> [ENVO:123123, ENVO:123777, ENVO:123666]

gi_to_counts (For all NCBI gis)
343204354 -> []
199999999 ->
343204350 ->

seq_to_counts (For all NCBI gis)
343206452 ->
343204354 ->
199999999 ->
343204350 ->