SemanticGenomeGraphs

Simon Heumos[1][0000−0002−7449−1266] and Jerven Bollemann[2][0000−0002−0875−1680]

[1]Quantitative Biologiy Center Tuebingen
simon.heumos@qbic.uni-tuebingen.de
[2]Swiss Institute of Bioinformatics
jerven.bolleman@sib.swiss

Abstract. The current linear reference based methods of representing genomic variation are limiting our insights into the variation between genomes. Genome graphs are a set of techniques that can accurately represent large structural variation as well as single nucleotide polymorphism. As any graph can be serialized as an RDF (Resource Description Framework) one, we show some advantages and disadvantages of making a Genome Graph available on the Semantic Web in a FAIR (Findable Accessible Interoperable Reusable) way. Demonstrating how we can use SPARQL to drive visualizations and integrate with non genome graph knowledge.

1 RDF in VG

The most prominent variation graph toolkit, vg[1] (variation graph) encodes nucleotide sequences in nodes which are connected by directed edges. Genomes are defined as paths through the nodes. Being a graph data structure itself, vg can be serialized into a Turtle[1] document. This textual representation of an RDF graph is structured in semantic N-Triples: subject, predicate, object. Turtle's syntax is very similar to SPARQL, an RDF query language. This allows accessing variation graphs using SPARQL[2], presenting an accessible and interoperable way in order to open Genome Graphs to the Semantic Web.

We extended the variant graph visualization Javascript module Sequence Tube Map[2] with the possibility to select and query a vg SPARQL endpoint[3]. One can dynamically browse through a whole genome graph. A second advantage of the RDF serialization is that one can combine data from very distinct sources annotating the current visualized genome graph. This allows visualizations to include more than just graph topology information.

However, our current approach only allows us to query a static version of a genome graph. This is great for visualization, but limits its reusability. In the future, we want to be able to translate any SPARQL queries directly to genome graph calls leading to a FAIR data management concept of genome graphs.

[1] https://github.com/vgteam/vg
[2] https://github.com/vgteam/vg/wiki/RDF:-for-VG
[3] https://github.com/graph-genome/MatrixTubeMap/tree/splitsparql2

2 SPARQL Against VG

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vg:<http://biohackathon.org/resource/vg#>
SELECT
 DISTINCT ?node ?sequence
WHERE {
  {
    SELECT ?otherpath (MIN(?otherposition) AS ?moreoffset)
    WHERE {
?step vg:node ?sharednode ;
vg:position ?position ;
                        vg:path <${path}> .
        ?step2 vg:node ?sharednode ;
                        vg:position ?otherposition ;
                        vg:path ?otherpath .
        FILTER(!sameTerm(?otherpath, ?path ))

        FILTER(?position >= ?offset && ?position <= ?upto))
        } GROUP BY ?otherpath
  }
  ?step3 vg:node ?node ;
        vg:position ?position3 ;
        vg:path ?otherpath .
  ?node rdf:value ?sequence .
  FILTER(?position3 >= ?moreoffset && ?position3 <= ?moreoffset < ?distance)
} VALUES ?path ?offset ?upto ?distance
```

Example 1: Find the nodes that are in the same linear area of differing paths. Where they are selected by a path section of reference e.g. Reference genome chromosome X from it's 1000 upto 2000 nucleotide. This query will find nodes in the non reference genome also in that area of their chromosome Xs.

References

1. Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G.: RDF 1.1 turtle. W3C recommendation, W3C (Feb 2014), https://www.w3.org/TR/2014/REC-turtle-20140225/
2. Beyer, W., Novak, A.M., Hickey, G., Chan, J., Tan, V., Paten, B., Zerbino, D.R.: Sequence tube maps: making graph genomes intuitive to commuters. Oxford Bioinformatics btz597 (2019), https://doi.org/10.1093/bioinformatics/btz597
3. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., Paten, B., Durbin, R.: Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature Biotechnology 36, 875–879 (2018), https://doi.org/10.1038/nbt.4227