**Utility of a Pathogen G-Quadruplex (G4) - Host Stress Protein (SP) Interactome Knowledge Graph**

| KEN4256 Maastricht University Period 4 Spring 2023  Written by: Farida Hanna Campbell (i6344578) | Date due: 3 April 2023  Date submitted: 30 March 2023 |
| --- | --- |

# Abstract:

There is a need for information that links heat shock protein (HSP) of human cells (hosts) infected by pandemic-type(“spread” related) virus genomes at the g-quadruplex points of intelligence (G4s). The proposal for this project aimed to satisfy the link with the creation of a knowledge graph that would present such pathogenic interactomes based on literature findings for the HIV-1 virus G-quadruplex (G4s) with human host cellular epigenetic stress proteins (HSP70). The results of the project were complex. The initial method proposed modifying an existing machine-language RDF disease-virus knowledge graph ([PathoPhenoDB](http://patho.phenomebrowser.net/#/search) ) and relevantSPARQL. It required access to G4, virus genomic and disease ontologies that in fact were either unavailable, deficient or inaccurate. Within this experience, the decision to deliver an RDF (JSON-LD) knowledge graph standard began based on manual annotations of knowledge from two critical literary publications into an TAO ontology available through *BIOLINK,* <https://collaboratory.semanticscience.org/annotate>. The G4-HSP relationship accuracy through calls to appropriate URL references and namespaces was verified. It is the goal of the author to bring the project forward into a larger scope beyond HIV-1 and to ensure adaptations that fulfil an urgent need in mutagenesis research subdomains of pathogen ecology. This will be hosted at a dedicated website <http://g4hsp.bio/> for public access as well as on the github [G4HSP](https://github.com/Cartesian1671/G4HSP) repository.

# Results:

The results of this project effort provide broad relation extraction from genomic domains of biomedicine together with finer-defined domains of Grich/G4/Gquartet/Gquadruplex named entity recognition and stress proteins such as HSP27, 40, 70,90 and 110. While the results of this project created a G4HSP knowledge graph in RDF/TTL and SPARQL query,they are no longer based on delivering a modified version of the [PathoPhenoDB](http://patho.phenomebrowser.net/#/search) database.

The results are listed as follows:

1. Prototype ontology: Annex I, Ontology.png
2. Prototype descriptions: Annex II, HSP-G4 descriptions, Descriptions.xlsx
3. G4HSP Rudimentary Prototype and csv-datafiles: G4HSP.Prototype.ipynb *plus* csvs
4. Extraction-model /Manually modified RDF: RDF Files.zip
5. SparQL: SparQL.txt
6. Shapl: Shapl.txt

# 

|  |
| --- |
| Figure 1: Schematic Ontology. *Top*, the prototype for development. *Bottom,* the scope (pink) and the overall future project (to be built with Biolink, https://collaboratory.semanticscience.org/annotate) |

# 

# Discussion:

The “G4-SP Interactome” combines HIV-1 pathogenic virus G-quadruplexes in the virus genome (Figure 2, G-Quadruplex structure) and its relationships to human host cell stress proteins during infection, Figure 3. The model acquired new ontologies that represent characteristics of the virus life cycle, and named stress proteins that are signatures of cellular infection and resistance in the literature.

|  |
| --- |
| **Figure 1.G-quadruplex structure.**  (**a**) Genomic structure of the HIV-1 provirus  (**b**) LOGO representation of the G-rich region for the HIV-1 promoter generated using the weblogo software[(Qiu et al. 2013)](https://paperpile.com/c/gGXwFO/QE8u) and based on an alignment of 1684 HIV-1 sequences provided by the HIV-1 database ([www.hiv.lanl.gov](http://www.hiv.lanl.gov/)).  (**c**) The U3 region of the LTR from the HXB2\_LAI;[NC\_001802](https://www.ncbi.nlm.nih.gov/nuccore/NC_001802) HIV-1 representative strain.  (**d**, **e** and **g**) Topologies of the LTR G4s determined using Clerocidin and DMS-mediated footprinting assays [(Perrone et al. 2013)](https://paperpile.com/c/gGXwFO/9CQp) for the (d) 1–32 nt segment,  (e) 12–44 nt segment and  (**f**) 26–44 nt segment. (g) Topology of the 12–37 nt segments determined using nuclear magnetic resonance (NMR) [(Amrane et al. 2014)](https://paperpile.com/c/gGXwFO/CPDJ).  Source: Métifiot M, Amrane S, Litvak S, Andreola ML. G-quadruplexes in viruses: function and potential therapeutic applications. Nucleic Acids Res. 2014 Nov 10;42(20):12352-66. doi: 10.1093/nar/gku999. Epub 2014 Oct 20. PMID: 25332402; PMCID: PMC4227801. Image and text reproduced with permission under Creative Commons 4.0 |

|  |
| --- |
| **Figure 2. HSPs interaction with HIV-1 life cycle, representing the following stages:**  (1) attachment  (2) entry  (3) uncoating  (4) nuclear import  (5) replication/transcription  (6) translation and protein folding  (7) assembly  (8) virion release from the human host cell  Source; Schematic is an adaptation of Iyer, K., Chand, K., Mitra, A. et al. Diversity in heat shock protein families: functional implications in virus infection with a comprehensive insight of their role in the HIV-1 life cycle. Cell Stress and Chaperones 26, 743–768 (2021). <https://doi.org/10.1007/s12192-021-01223-3>. Reproduced with permission by RightsLink application 28.03.2023. |

HSP are vaccine adjuvants[(Iyer et al. 2021)](https://paperpile.com/c/gGXwFO/LQQe). Until now however, there is no knowledge graph that links literature to do with the HSP genomic G4 interactome. This is unfortunate because virus chief “intelligence” for its life cycle multiplication properties depends on its G4 entities. Likewise, virus hijacking of host cell infrastructure depends on HSP signaling and yet, HSP properties are expressions of environmental abiotic resistance, including infectious (biotic) resistance and cancer related apoptosis. This means that research in HSP units combined with knowledge graph lies is an important forecasting resource for pathogen ecology.

# List of challenges faced:

The following databases were considered crucially necessary to the successful and simple [PathoPhenoDB](http://patho.phenomebrowser.net/#/search) database modification objectives cited in the original proposal. However, they were found to harbor limitations such as inconsistencies or incompleteness, or no longer working and available.

They include:

<http://patho.phenomebrowser.net/#/>

<http://scottgroup.med.usherbrooke.ca/G4RNA/> : No longer working/No longer available

<https://bioinformatics.ramapo.edu/GRSDB2/> : No longer working/No longer available

<http://athena.bioc.uvic.ca/genomes/index.html>. :No longer working/No longer available

<https://viralzone.expasy.org/5183> : Incomplete/insufficient/inaccurate

<https://www.genome.jp/virushostdb/note.htm> : Incomplete/insufficient/inaccurate

<https://github.com/AnimaTardeb/G4Hunter/blob/master/Mitochondria_NC_012920_1.fasta> : Incomplete/insufficient/inaccurate

Specific entities to do with HIV-1 genome and HSP interaction during infection were limited but still useful to the development process.

They include the following: [https://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=http%3A%2F%2F](https://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FSNOMEDCT%2F444356002)

[purl.bioontology.org%2Fontology%2FSNOMEDCT%2F444356002](https://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=http%3A%2F%2Fpurl.bioontology.org%2Fontology%2FSNOMEDCT%2F444356002)

<https://www.ncbi.nlm.nih.gov/gds/?term=human+immunodeficiency+virus+2>

<https://bioportal.bioontology.org/ontologies>

[https://bioportal.bioontology.org/ontologies/PR/?p=classes&conceptid=http%3A%2F%2Fpurl.oboli](https://bioportal.bioontology.org/ontologies/PR/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FSO_0000704&jump_to_nav=true)

[brary.org%2Fobo%2FSO\_0000704&jump\_to\_nav=true](https://bioportal.bioontology.org/ontologies/PR/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FSO_0000704&jump_to_nav=true)

<https://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes&conceptid=36115006>

<https://bioportal.bioontology.org/ontologies/PR/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FGO_0016485>

<https://bioportal.bioontology.org/ontologies/PR/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FPR_P07900&jump_to_nav=true>

<https://bioportal.bioontology.org/ontologies/PR/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FPR_P04792&jump_to_nav=true>

<https://bioportal.bioontology.org/ontologies/GO/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FGO_0031072&jump_to_nav=true>

<https://disease-ontology.org/?id=DOID:526>

<http://icd9cm.chrisendres.com/index.php?action=search&srchtext=042-042.99>

A creative attempt to use the xml tree for [PathoPhenoDB](http://patho.phenomebrowser.net/#/search) database was rejected:

<https://www.proteinatlas.org/ENSG00000134057.xml>

Additionally, there is a lack of URI references such as for the following entities and their G4 relationships:

HSPB10, HSPB11, DNAJA4, DNAJB1, DNAJB2, DNAJB3, DNAJB4, DNAJB5, DNAJB6, DNAJB7, DNAJB8, DNAJB14, DNAPJC1, DNAPJC2, DNAPJC3, DNAPJC4,DNAPJC30, HSPD1, CCT1, CCT2, CCT3, CCT4, CCT5, CCT6A, CCT6B, CCT7,CCT8,HSPA1A,HSPA1B,HSPA1C, HSPA2,HSPA6,HSPA8,HSPA9,HSPA12A,HSPA12B,HSPA13,HSPA14,HSPA90AA1,HSPA90AA2,HSPA90B1, TRAP1, HSPH1,HSPH2,HSPH3,HSPH4.

# Solutions to Challenges:

A “Plan B” solution scenario began with contemplation of the knowledge graph development “from scratch”:

1. To develop a rudimentary csv-dependent knowledge graph, including with example.com namespace references and python RDF libraries and stardog.

2. From this prototype, selection of key documents for curation, and the machine-powered development of an RDF knowledge graph. Step 2 includes the deployment of an extraction model (Biolink <https://collaboratory.semanticscience.org/annotate>) to annotate much richer reference and namespace requirements than possible at a manual scripting level. The Knowledge Collaboratory is a biomedical web service to query and publish Nanopublications for the NCATS Biomedical Data Translator project, that allows nanopublications to be queried using the Translator Reasoner API (TRAPI) specifications, and uploading via authentication keys to the server; developed and hosted by the Institute of Data Science at Maastricht University.

### Class development:

The virus-biomedical text with genome and proteomic statements were manually curated from two selected documents,*G-quadruplexes in viruses: function and potential therapeutic applications*[(Métifiot et al. 2014)](https://paperpile.com/c/gGXwFO/WWDL) and *Diversity in heat shock protein families: functional implications in virus infection with a comprehensive insight of their role in the HIV-1 life cycle*[(Iyer et al. 2021)](https://paperpile.com/c/gGXwFO/LQQe) *.* The curations were then deployed in the BioLink application to generate RDF using standard identifiers associated with the US NIH NCATS Translator SRI Name Resolution API. Each of the rendered annotations had to be inspected for accuracy and URI, and then, either saved or corrected, or deleted. The result includes RDF (JSON-LD) knowledge graph standard based on TAO ontology and both G4-HSP relationship accuracy through calls to appropriate URL references and namespaces. The project will be continued as part of a larger objective in pathogen ecology, and will be hosted at a dedicated website <http://g4hsp.bio/> for public access as well as on the github [G4HSP](https://github.com/Cartesian1671/G4HSP) repository.

### Adaptability:

This project supports new (virus) pathogen classifications based on RNA virus in G4-genomic relations to human host cells, in particular. This represents a multi-host, poly pandemic and epidemic species update rapidly. Future users can use this knowledge graph to generate important medical research investigations in the precise nature of G4-HSP interaction.

### Novel insights that add to the body of scientific knowledge

The scope of the G4HSP concept has powerful implications for pathogen-genome interception and cellular pathology. This is because the knowledge graph queries support cellular stress response relationships at the core of G4 metabolics in applications such as cellular senescence and oncology. Such a knowledge graph is important to the next generation of pharmaceutical intervention strategy.

Additional questions remain to be answered. This includes the form of entity and relation detection, prediction, and linking:

1. To verify the G4HSP only contains mention of a single entity and single relation once and in the correct form (entity/relation detection/linking).
2. To allow the union of two queries where appropriate to answer a question where is the answer to the query.
3. Where there is noise or incompleteness of information in the knowledge graph reactions, that the tool Biolink tool can compensate the query with recall from similar publications to those used in our model generation.
4. Where there may be multiple entities/relations leading to duplicate names for the same entity (such as Grich, G4, Gquadruplex, and Gquartet which all represent different or duplicate names for the same entity), that the predicted entity/relation will still yield an accurate match for the response to the query.

# Conclusions:

Knowledge domains of Grich/G4/Gquartet/Gquadruplex genetic compounds in the literature for HIV-1 pathogenic virus were linked to HSP27, 40, 70,90 and 110 protein descriptions in the interactome of a human host cell. The process of development included a fundamental prototype that was instrumental to the selection of literature and manual annotation of statements in a Tao Ontology using Biolink. In the next phase of development, this project will be adapted for multiple virus.

# Selected References

[1] [J. Acunzo, M. Katsogiannou, P. Rocchi, Small heat shock proteins HSP27 (HspB1), αB-crystallin (HspB5) and HSP22 (HspB8) as regulators of cell death, Int. J. Biochem. Cell Biol. 44 (2012) 1622–1631.](http://paperpile.com/b/5SneVw/EVafD)

[2] [F.H. Campbell, Trajectories of RNA Virus Mutation Hidden by Evolutionary Alternate Reality Thermodynamic Endpoints in Transformations in Response to Abiotic Habitat Stresses, in: Environmental Management-Pollution, Habitat, Ecology, and Sustainability, IntechOpen, 2021.](http://paperpile.com/b/5SneVw/R0e3)

[3] [L. Gottschlich, M. Bortfeld-Miller, C. Gäbelein, S. Dintner, J.A. Vorholt, Phosphorelay through the bifunctional phosphotransferase PhyT controls the general stress response in an alphaproteobacterium, PLoS Genet. 14 (2018) e1007294.](http://paperpile.com/b/5SneVw/66M1)

[4] [L. Gottschlich, P. Geiser, M. Bortfeld-Miller, C.M. Field, J.A. Vorholt, Complex general stress response regulation in Sphingomonas melonis Fr1 revealed by transcriptional analyses, Sci. Rep. 9 (2019) 1–13.](http://paperpile.com/b/5SneVw/Ac7u)

[5] [Y. He, R.R. Vines, A.R. Wattam, G.V. Abramochkin, A.W. Dickerman, J.D. Eckart, B.W.S. Sobral, PIML: the Pathogen Information Markup Language, Bioinformatics. 21 (2005) 116–121.](http://paperpile.com/b/5SneVw/ZEpqR)

[6] [S. Kumar, C. Ramamurthy, D. Choudhary, A. Sekar, A. Patra, N.S. Bhavesh, P. Vivekanandan, Contrasting roles for G-quadruplexes in regulating human Bcl-2 and virus homologues KSHV KS-Bcl-2 and EBV BHRF1, Sci. Rep. 12 (2022) 5019.](http://paperpile.com/b/5SneVw/CRpW7)

[7] [N. Kurbatova, R. Swiers, Disease ontologies for knowledge graphs, BMC Bioinformatics. 22 (2021) 377.](http://paperpile.com/b/5SneVw/pZBZ)

[8] [Q. Li, J. Nie, J. Wu, L. Zhang, R. Ding, H. Wang, Y. Zhang, T. Li, S. Liu, M. Zhang, C. Zhao, H. Liu, L. Nie, H. Qin, M. Wang, Q. Lu, X. Li, J. Liu, H. Liang, Y. Shi, Y. Shen, L. Xie, L. Zhang, X. Qu, W. Xu, W. Huang, Y. Wang, SARS-CoV-2 501Y.V2 variants lack higher infectivity but do have immune escape, Cell. 184 (2021) 2362–2371.e9.](http://paperpile.com/b/5SneVw/cKHV)

[9] [M. Métifiot, S. Amrane, S. Litvak, M.-L. Andreola, G-quadruplexes in viruses: function and potential therapeutic applications, Nucleic Acids Res. 42 (2014) 12352–12366.](http://paperpile.com/b/5SneVw/OX2Qy)

[10] [J. Miskiewicz, J. Sarzynska, M. Szachniuk, How bioinformatics resources work with G4 RNAs, Brief. Bioinform. 22 (2021).](http://paperpile.com/b/5SneVw/iibt)

[11] [M.D. Perng, R.A. Quinlan, The Dynamic Duo of Small Heat Proteins and IFs Maintain Cell Homeostasis, Resist Cellular Stress and Enable Evolution in Cells and Tissues, Heat Shock Proteins. (2015) 401–434.](http://paperpile.com/b/5SneVw/iCBK)

[12] [G.G. Rao, Risk factors for the spread of antibiotic-resistant bacteria, Drugs. 55 (1998) 323–330.](http://paperpile.com/b/5SneVw/rCg2)

[13] [E. Ruggiero, I. Zanin, M. Terreri, S.N. Richter, G-Quadruplex Targeting in the Fight against Viruses: An Update, Int. J. Mol. Sci. 22 (2021).](http://paperpile.com/b/5SneVw/aSgWA)

[14] [S. Sarrou, Evaluation of the molecular assay Genspeed® Superbug CR as screening test in a tertiary hospital with high prevalence of multidrug resistant Gram-negative bacteria, in: 26th European Congress of Clinical Microbiology and Infectious Diseases Apr 9-12, 2016, Medical School, University of Thessaly, n.d.](http://paperpile.com/b/5SneVw/ew8g)

[15] [M.D. Wilkinson, M. Dumontier, I.J.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Sci Data. 3 (2016) 160018.](http://paperpile.com/b/5SneVw/Itcg)

[16] [World Health Organization, Prioritizing diseases for research and development in emergency contexts, World Health Organisation. (2022).](http://paperpile.com/b/5SneVw/trvMh)

[17] [Z. Xiang, Y. Tian, Y. He, PHIDIAS: a pathogen-host interaction data integration and analysis system, Genome Biol. 8 (2007) R150.](http://paperpile.com/b/5SneVw/iBdd)

[18] [A. Zappa, A. Splendiani, P. Romano, Towards linked open gene mutations data, BMC Bioinformatics. 13 Suppl 4 (2012) S7.](http://paperpile.com/b/5SneVw/XVPH)