Wellness Knowledge Provider for NCATS' Biomediacal Data Translator

Derivation of knowledge from wellness data

- We have analyzed the ISB Wellness dataset, which has been phenotyped extensively, affording many types of correlations and connections to be uncovered. Expanding on our original wellness study of 108 individuals [5], this deep phenotyping data set integrates many data types including WGS and/or SNP genotyping, clinical blood tests, salivary cortisol, weight and BMI, blood pressure, health assessments, provider notes, gut microbiome, blood metabolomics, blood proteomics, activity tracking, sleep tracking, and heart rate. The cohort includes 4,879 individuals with at least one blood draw via Arivale. Integrative analysis of this multidimensional data set is already leading to significant novel findings, e.g., on the connection between blood metabolites and the microbiome [6] and how this reflects aging [7]. The data were collected in longitudinal 'snapshots' that enable a more detailed analysis of data accrual and stability than a single 'final' data set view can.
- We have created and deployed multiple versions of the Multiomics Wellness KG. We computed correlations among attributes in the chemistries, metabolomics and proteomics tables in the ISB Wellness dataset. These attributes are clinical labs, metabolites, and proteins, respectively. Each attribute can either be a blood analyte or an index computed from one or more analytes. For example, the chemistries table has Albumin, Globulin and also the ratio of the two as three different attributes. The resulting KG includes statistically significant correlations from the inner join of the clinical labs, protein panels and metabolites, thus extending on the original version that included only correlations within each table.
- We performed a detailed curation of LOINC codes for the analytes that have significant correlations with other analytes for the attributes in the chemistries table, and modified the biolink concepts of a subset of nodes to ClinicalFinding to retain LOINC codes that best preserve the identity of a node.
- We transformed the resulting set of attributes and correlations into a KG (**Figure 1**) and expressed them in the KGX format. This format expresses KGs as two TSV files: one for the nodes with necessary columns for curies (a compressed URI that uniquely identifies a node) and a Biolink model concept; and the other representing edges between these nodes, requiring the curies for the two end nodes, a Biolink concept for the relationship between the two nodes and similarly another relationship concept expressed in some standard ontology. In collaboration with Kevin Xin (Su Lab, Service Provider) we deployed this KG via BioThings API.

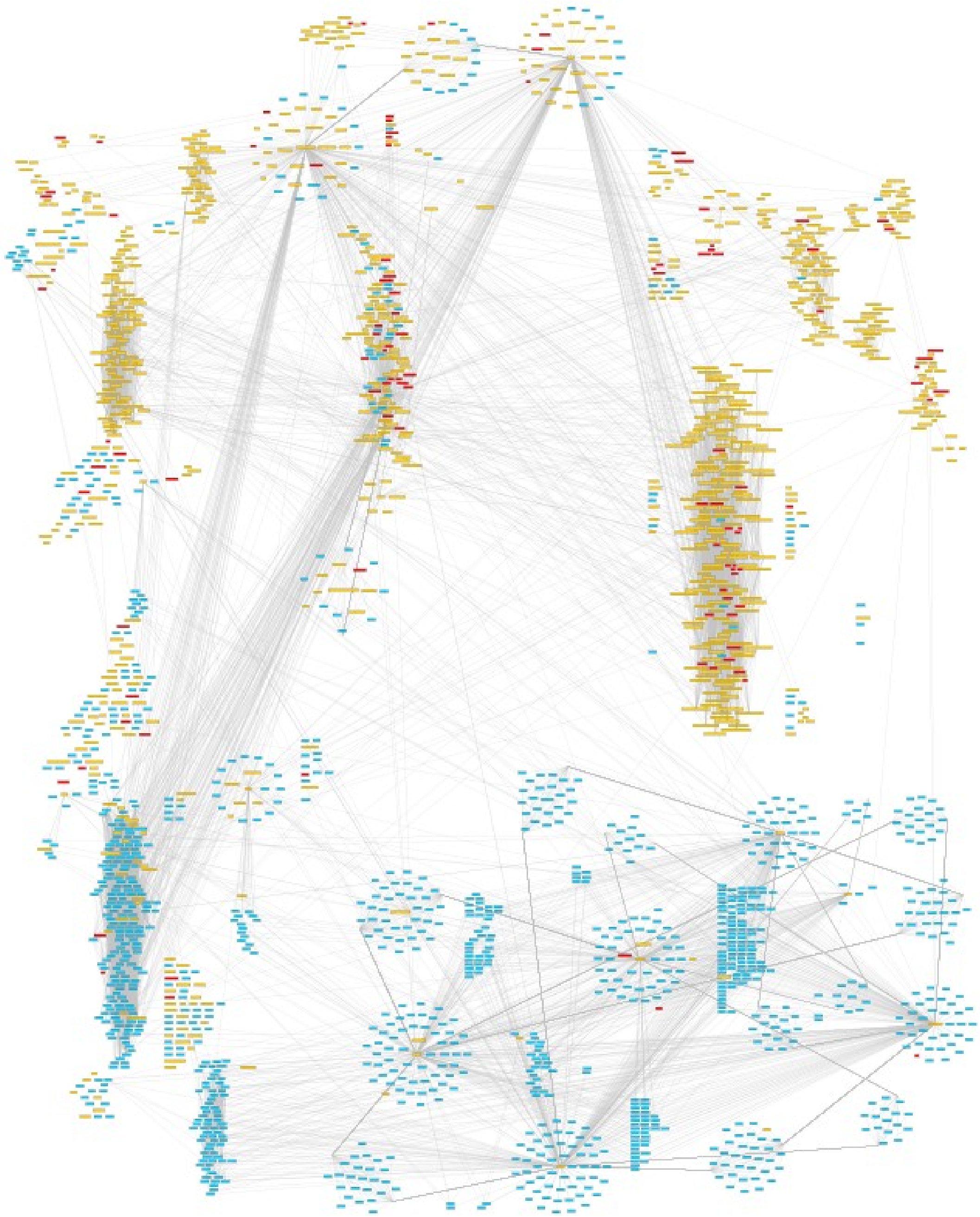

Figure 1. *Visualization of the core of the ISB Wellness Multiomics KG, v.1.3, integrating clinical chemistries (red), proteins (cyan) and metabolites (yellow), which includes drugs. The edges of the network are correlations between various analytes, retaining only edges with p-value under 10-100. The complete network is much more extensive and complex than depicted, and subsequent versions (e.g., current v.1.6) even more so. Several metabolites are 'hubs' connected to multiple proteins; most of these hub metabolites are drugs.*

- Based on the analysis of snapshots (described below), we implemented metrics of knowledge confidence by taking into account the p-values of correlations between analytes as observed in former data snapshots. The information is presented in the form of a new column in the edges TSV called 'weighted p-value', the weight of the p-value in each snapshot is scaled by the factor of the number of observations in the corresponding snapshot. We will further refine this method by making significantly different snapshots (as assessed via data fingerprint comparisons) contribute more to the confidence in the derived knowledge.

- We have created the first version of a predictive model by creating a regression model for the analytes that are correlated with the most other analytes. As a result, we have a number of new edges and connections between analytes that are actually not significantly correlated with an analyte but are still a regressor of theirs. We used ridge regression for this purpose, so as to have a reliable prediction model for the number of baseline observations, which are fewer than the number of analytes. For example, the protein CVD3_O00300 (tumor necrosis factor receptor superfamily, member 11b) is not significantly correlated with adiponectin in the Wellness graph, but the regression analysis shows that this protein is one of the most important regressors (predictors) of adiponectin, consistent with the literature [8].
- The new edges so created have the predicate ‘related_to’ and a relation from the RO ontology that more closely describes the relationship of statistical prediction. Biolink is being updated to create a suitable predicate, possibly called ‘regressor_of’ as a hierarchical descendant of ‘predicts’ - see issue GitHub Biolink:731.
- We incorporated the results of stratification of correlations with additional context information on the individuals and the observations, like sex, age group, ancestry, seasonality, and proclivity to having extreme values for specific analytes.
- We updated the Wellness graph to adhere to the latest Biolink version to support results for the December demo.
- We created a new version of the Wellness graph with recomputed correlations between all analytes with a much higher concept pair count. We integrated results for the gut microbiome represented as molecular activity, in the form of its correlations with blood analytes and each other. There are ~5000 unique kegg orthologs that represent the gut microbiome analytes. We expect these results to be soon deployed by the Service Provider.
- We computed the *interactions* between various blood analytes and gut microbiome. For the analytes that were significantly correlated, we modeled them pairwise with other analytes to obtain a predictive relationship. We used *generalized linear model*, to obtain the interaction term (the coefficient of interaction) and its significance using the model:analyte1 ~analyte2 * analyte3 The interaction term analyte2:analyte3 gives a measure of the change in relationship between analyte1and analyte2(or analyte3) in the presence of analyte3(or analyte2). Representation of such knowledge requires an association between three nodes - a limitation in the current Biolink Model.
- For example, among the interactions between metabolites and gut microbiome, one significant interaction found was between 1-methylnicotinamide, pyridoxate and glutamate carboxypeptidase [EC:3.4.17.11]. The interaction between pyridoxate and glutamate carboxypeptidase changes both in direction and magnitude of coefficient of relation between the three analytes:

analyte1	1-methylnicotinamide
analyte2	pyridoxate
analyte3	glutamate carboxypeptidase [EC:3.4.17.11]
Coefficient analyte3	-2.474

Coefficient analyte2	-0.028
Coefficient interaction	1.78
P-value analyte3	2.65e-05
P-value analyte2	1.39e-16
P-value interaction	2.63e-91

- We encountered an extension of the ‘interactions’ paradigm when we explored the application of Differential Rank Conservation (DIRAC) [9] to ISB’s wellness data. DIRAC requires two-fold separation within data. It provides quantitative measures of how network rankings differ either among networks for a selected phenotype or among phenotypes for a selected network. While the phenotypic separation can be extracted trivially from the Wellness dataset, for example sex or ethnicity based stratification, the network modules within data are hard to construe. The method is designed to analyze relative regulation of modules that are involved in biologically known processes among phenotypes.
- We applied DIRAC to the metabolomics data of ISB’s wellness cohort. We divided the cohort into Males and Females, and identified five network modules to work with: ‘Cofactors and Vitamins’, ‘Carbohydrates’, ‘Nucleotides’, ‘Energy’ and ‘Peptides’. The normalized rank indices of the metabolites that form these modules, among males and females are:

Network	Females	Males
Cofactors and Vitamins	243.8	200.9
Energy	33.6	34.4
Carbohydrate	97.9	85.8
Nucleotide	150.5	99.9
Peptide	544.3	411.6

These results indicate a biologically significant general trend of tighter regulation in the relative abundance of metabolites in most metabolic networks in females. As the Biolink model is adapted to support knowledge generated by graphs that contain N-ary relationships, we can expand the phenotypes for such analyses to contain rank conservation among network modules of diseases.

Analysis of snapshots in ISB wellness data to derive a knowledge stability metric

- Datasets that grow over time can have ‘snapshots’ (or ‘freezes’) stored periodically, for example to ensure that analyses done on the dataset are stable and reproducible, and don’t change with every small change to the dataset. The ISB Wellness dataset is such a dataset, with 96 snapshots spanning slightly over two years of data collection, and multiple data tables per snapshot. With each newer snapshot, the complexity of the dataset increased by adding timepoints for the same individuals, by adding new individuals, by adding (and sometimes, removing) attributes to existing tables, and even by adding new tables. The dataset could also change by removal of individuals, e.g., those that withdrew consent to participate in research.

- We sought to leverage the availability of the many dataset snapshots to derive metrics of stability or reliability of the knowledge derived from the data. Does confidence in knowledge derived from data increase as more data are added, or does it become weaker? Our preliminary results from correlations in the chemistries table suggests this can go both ways, and thus that there is value in performing this analysis.
- In some cases, consecutive snapshots may be very similar; some of the tables may be entirely unchanged. It would therefore be a mistake to take these to be independent when doing any statistical computation over the different snapshots, to assess the stability of edges in the knowledge graphs created from similar snapshots. This calls for an algorithm to assess the information content offered by a snapshot, and to prune [10] redundant snapshots, so as to increase the reliability of the stability metric.
- There is therefore a need for metrics of similarity between snapshots. Our data fingerprinting method [11] can efficiently yield such a metric. We computed data fingerprints for the ‘chemistries’ table in each of 96 snapshots of the ISB wellness data (gray-filled circles), and visualized using PCA (**Figure 2**). We observed a significant transition point followed by a faster rate of change at the beginning of 2019; a less pronounced shift in early 2018 corresponded to the addition of two columns to the table. Subsequent analyses explained the main (2019) effect as a result of changes in ‘reflexive’ measurements; excluding such measurements canceled the effect, and revealed a further minor shift in mid-2018 caused by the removal of two columns from the table.
- Insights from this analysis led to an improvement in the quality of the knowledge presented in the Wellness KG. Removal of the ‘reflexive’ measurements from the ‘chemistries’ table affected the derived knowledge. In one example, the statistically significant positive correlation between zinc_plasma_or_serum and triglycerides (which contradicts reports in the literature) was canceled by removal of the ‘reflexive’ measurements.
- We have a manuscript in preparation presenting these analyses, aiming to include similar analyses on other datasets with snapshot structure, to validate the generalizability of the method.
- We are currently working on adapting the code for the data fingerprinting algorithm [11] to be able to run it in the environment for EHR data. We have identified over 30,000 patients in the EHR data who have been diagnosed with IBD (Inflammatory Bowel Disease) and have lab results for them over a period of over 10 years. This information can be organized into snapshots similar to ISB’s wellness dataset. We sought to use the algorithms in [10] and [11] to derive similar metrics of stability for risk factors for IBD.
- We ported to Python the instance reduction algorithm [10] to make it platform independent. The new code is in Python which saves compute time and is easily refactorable.
- We evaluated additional datasets in snapshot format (e.g., PDB) for performing similar analyses.
- We expect this methodology to be applicable also to performing QC on versions of KGs stored in the Knowledge Graph Exchange (KGE) Archive.

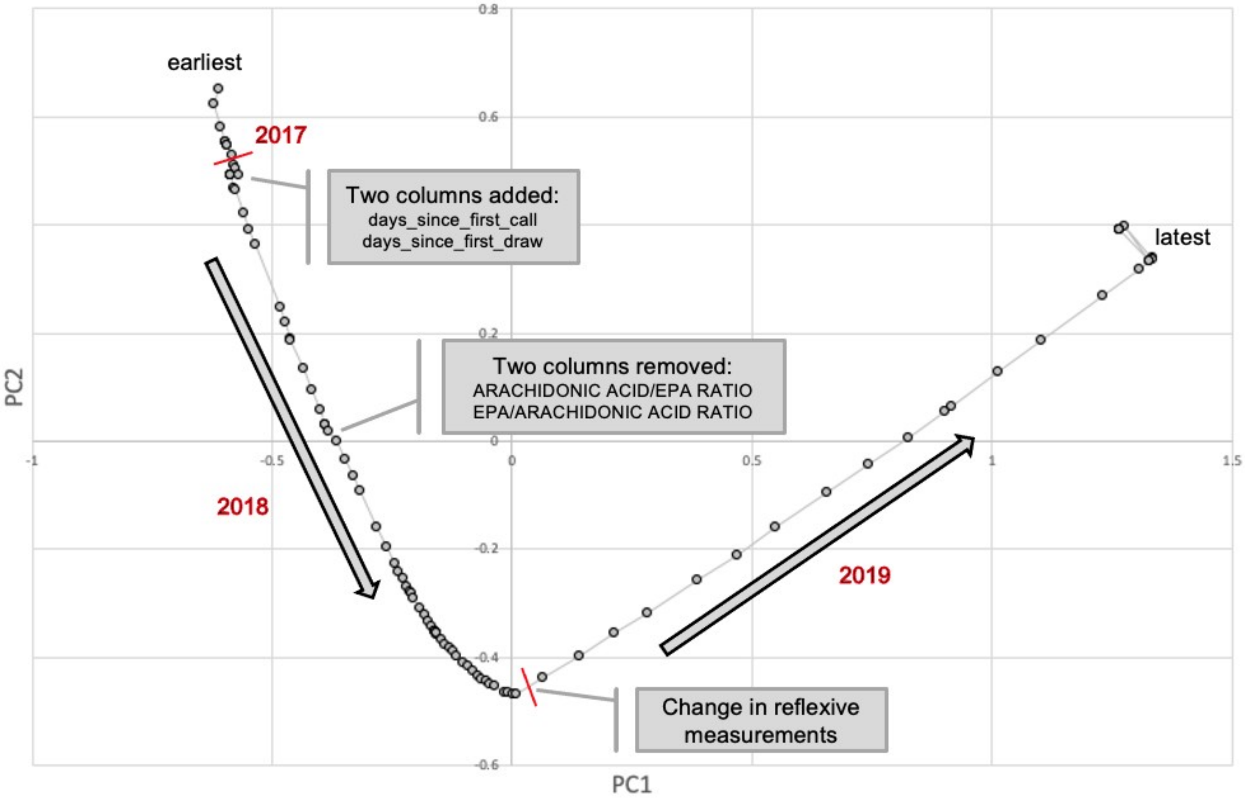

Figure 2. *Visualization of the trajectory of the ISB Wellness dataset (clinical chemistries table) over time.*

Challenges encountered and potential solutions

- A challenge was mapping concepts to ontologies that can be used (e.g., while LOINC codes are available for our chemistries table, they are not readily usable by other Translator components). Finding a way to convert LOINC codes to equivalent alternate ontology: e.g., three different nodes with different LOINC codes involve zinc, do these nodes get combined into one node in an alternate ontology? What about measurements that don't exist in any other alternate ontology?
- The actual Wellness graph is more complex and diverse than what is served via the KP because of naming difficulties for certain nodes. We expect these to be resolved gradually over time through ongoing curation.
- Identifying, tracing and rationalizing multiple arbitrary encoding decisions that were made while generating and organizing the data, due to real-world complexity.
- Representation of interactions knowledge (as in Multiomics Wellness KG) requires an association between three nodes (**Figure 3**). This poses a limitation for the current Biolink model, which expects an association to only exist between two nodes (a subject and an object). This is a specific case of an N-ary relationship (see GitHub issue: 566) for which we are currently working on with the Data Modeling and SRI teams to find a suitable solution.

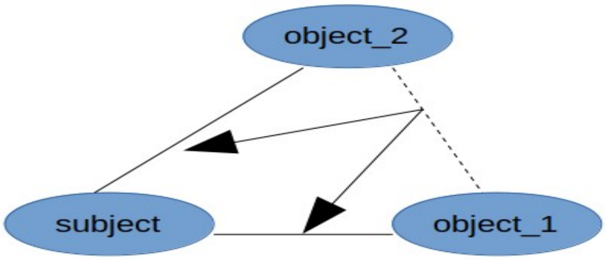

***Figure 3**. Interaction between nodes object_1 and object_2 would affect the relationship (residualized relationship) between subject & object_1 and subject & object_2*

References

1. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46:D1074–82.

2. Ursu O, Holmes J, Bologa CG, Yang JJ, Mathias SL, Stathias V, et al. DrugCentral 2018: an update. Nucleic Acids Res. 2019;47:D963–70.

3. Wang Y, Zhang S, Li F, Zhou Y, Zhang Y, Wang Z, et al. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Res. 2020;48:D1031–41.

4. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017;16:19–34.

5. Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat Biotechnol. Nature Publishing Group; 2017;35:747–56.

6. Wilmanski T, Rappaport N, Earls JC, Magis AT, Manor O, Lovejoy J, et al. Blood metabolome predicts gut microbiome α-diversity in humans. Nat Biotechnol. Nature Publishing Group; 2019;37:1217–28.

7. Wilmanski T, Diener C, Rappaport N, Patwardhan S, Wiedrick J, Lapidus J, et al. Gut microbiome pattern reflects healthy ageing and predicts survival in humans. Nat Metab. 2021;3:274–86.

8. Alikaşifoğlu A, Gönç N, Özön ZA, Sen Y, Kandemir N. The relationship between serum adiponectin, tumor necrosis factor-alpha, leptin levels and insulin sensitivity in childhood and adolescent obesity: adiponectin is a marker of metabolic syndrome. J Clin Res Pediatr Endocrinol. 2009;1:233–9.

9. Eddy JA, Hood L, Price ND, Geman D. Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). PLoS Comput Biol. 2010;6:e1000792.

10. Joshi A, Haspel N. A Novel Data Instance Reduction Technique using Linear Feature Reduction. AIS. 2020;191–206.

11. Robinson M, Hadlock J, Yu J, Khatamian A, Aravkin AY, Deutsch EW, et al. Fast and simple comparison of semi-structured data, with emphasis on electronic health records [Internet]. bioRxiv. 2018 [cited 2018 Apr 14]. p. 293183. Available from: https://www.biorxiv.org/content/early/2018/04/02/293183

12. Su Y, Chen D, Yuan D, Lausted C, Choi J, Dai CL, et al. Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. Cell. 2020;183:1479–95.e20.

13. Dai CL, Kornilov SA, Roper RT, Cohen-Cline H, Jade K, Smith B, et al. Characteristics and Factors Associated with COVID-19 Infection, Hospitalization, and Mortality Across Race and Ethnicity. Clin Infect Dis [Internet]. 2021; Available from: http://dx.doi.org/10.1093/cid/ciab154

14. Desai RA, Davies AL, Del Rossi N, Tachrount M, Dyson A, Gustavson B, et al. Nimodipine Reduces Dysfunction and Demyelination in Models of Multiple Sclerosis. Ann Neurol. 2020;88:123–36.