**NBD(NIST Big Data) Requirements WG Use Case Template Aug 11 2013**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Comparative analysis for metagenomes and genomes | |
| **Vertical (area)** | | Scientific Research: Genomics | |
| **Author/Company/Email** | | Ernest Szeto / LBNL / [eszeto@lbl.gov](mailto:eszeto@lbl.gov) | |
| **Actors/Stakeholders and their roles and responsibilities** | | Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) project. Heads: Victor M. Markowitz, and Nikos C. Kyrpides. User community: JGI, bioinformaticians and biologists worldwide. | |
| **Goals** | | Provide an integrated comparative analysis system for metagenomes and genomes. This includes interactive Web UI with core data, backend precomputations, batch job computation submission from the UI. | |
| **Use Case Description** | | Given a metagenomic sample, (1) determine the community composition in terms of other reference isolate genomes, (2) characterize the function of its genes, (3) begin to infer possible functional pathways, (4) characterize similarity or dissimilarity with other metagenomic samples, (5) begin to characterize changes in community composition and function due to changes in environmental pressures, (6) isolate sub-sections of data based on quality measures and community composition. | |
| **Current**  **Solutions** | **Compute(System)** | | Linux cluster, Oracle RDBMS server, large memory machines, standard Linux interactive hosts |
| **Storage** | | Oracle RDBMS, SQLite files, flat text files, Lucy (a version of Lucene) for keyword searches, BLAST databases, USEARCH databases |
| **Networking** | | Provided by NERSC |
| **Software** | | Standard bioinformatics tools (BLAST, HMMER, multiple alignment and phylogenetic tools, gene callers, sequence feature predictors…), Perl/Python wrapper scripts, Linux Cluster scheduling |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Centralized. |
| **Volume (size)** | | 50tb |
| **Velocity**  **(e.g. real time)** | | Front end web UI must be real time interactive. Back end data loading processing must keep up with exponential growth of sequence data due to the rapid drop in cost of sequencing technology. |
| **Variety**  **(multiple datasets, mashup)** | | Biological data is inherently heterogeneous, complex, structural, and hierarchical. One begins with sequences, followed by features on sequences, such as genes, motifs, regulatory regions, followed by organization of genes in neighborhoods (operons), to proteins and their structural features, to coordination and expression of genes in pathways. Besides core genomic data, new types of “Omics” data such as transcriptomics, methylomics, and proteomics describing gene expression under a variety of conditions must be incorporated into the comparative analysis system. |
| **Variability (rate of change)** | | The sizes of metagenomic samples can vary by several orders of magnitude, such as several hundred thousand genes to a billion genes (e.g., latter in a complex soil sample). |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | Metagenomic sampling science is currently preliminary and exploratory. Procedures for evaluating assembly of highly fragmented data in raw reads is better defined, but still an open research area. |
| **Visualization** | | Interactive speed of web UI on very large data sets is an ongoing challenge. Web UI’s still seem to be the preferred interface for most biologists. It is use for basic querying and browsing of data. More specialized tools may be launched from them, e.g. for viewing multiple alignments. Ability to download large amounts of data for offline analysis is another requirement of the system. |
| **Data Quality** | | Improving quality of metagenomic assembly is still a fundamental challenge. Improving the quality of reference isolate genomes, both in terms of the coverage in the phylogenetic tree, improved gene calling and functional annotation is a more mature process, but an ongoing project. |
| **Data Types** | | Cf. above on “Variety” |
| **Data Analytics** | | Descriptive statistics, statistical significance in hypothesis testing, discovering new relationships, data clustering and classification is a standard part of the analytics. The less quantitative part includes the ability to visualize structural details at different levels of resolution. Data reduction, removing redundancies through clustering, more abstract representations such as representing a group of highly similar genomes in a pangenome are all strategies for both data management as well as analytics. |
| **Big Data Specific Challenges (Gaps)** | | The biggest friend for dealing with the heterogeneity of biological data is still the relational database management system (RDBMS). Unfortunately, it does not scale for the current volume of data. NoSQL solutions aim at providing an alternative. Unfortunately, NoSQL solutions do not always lend themselves to real time interactive use, rapid and parallel bulk loading, and sometimes have issues regarding robustness. Our current approach is currently ad hoc, custom, relying mainly on the Linux cluster and the file system to supplement the Oracle RDBMS. The custom solution oftentimes rely in knowledge of the peculiarities of the data allowing us to devise horizontal partitioning schemes as well as inversion of data organization when applicable. | |
| **Big Data Specific Challenges in Mobility** | | No special challenges. Just world wide web access. | |
| **Security & Privacy**  **Requirements** | | No special challenges. Data is either public or requires standard login with password. | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | A replacement for the RDBMS in big data would be of benefit to everyone. Many NoSQL solutions attempt to fill this role, but have their limitations. | |
| **More Information (URLs)** | | <http://img.jgi.doe.gov> | |
| **Note:** <additional comments> | | | |