Data Like This: Ranked Search of Genomic Data

Vision Paper

V.M. Megler[1] David Maier[1] Daniel Bottomly[2]
[1]Portland State University
Department of Computer Science
Portland, Oregon
vmegler,maier@cs.pdx.edu

Libbey White[2] Shannon McWeeney[2] Beth Wilmot[2]
[2]Oregon Health Sciences University
Oregon Clinical and Translational Research Institute
Portland, Oregon
bottomly,whiteli,mcweeney,wilmotb@ohsu.edu

ABSTRACT

High-throughput genetic sequencing produces the ultimate "big data": a human genome sequence contains more than 3B base pairs, and more and more characteristics, or annotations, are being recorded at the base-pair level. Locating areas of interest within the genome is a challenge for researchers, limiting their investigations. We describe our vision of adapting "big data" ranked search to the problem of searching the genome. Our goal is to make searching for data as easy for scientists as searching the Internet.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods.* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models, search process.* H.2.8 [**Information Systems**]: Database Applications – *scientific databases, spatial databases & GIS.*

General Terms

Design

Keywords

Genome search, scientific data, ranked data search, data exploration.

1. INTRODUCTION

The last decade has seen massive growth in the amount of scientific data collected. This growth is particularly pronounced in the field of bioinformatics and genomic research, with gene sequencing being applied in more and more clinical research settings. While the general attitude seems to be "more data is better," growth in holdings can actually introduce impediments to science [2]. In an archive consisting of thousands of datasets, or a growing collection of whole genome sequences, a scientist can find it daunting to identify the subset of data relevant to her research interests. The harder it is to find relevant data, the fewer research questions are asked and studied [14]. Growth in data must be accompanied by improvements in tools that help scientists easily find the data they need [2].

Despite much progress in providing data access through data browsers, visualizers, portals and gateways, the problem of how a scientist efficiently finds the data she feels is worth accessing has not been solved [8, 14]. This challenge crosses industries and disciplines in scientific research, and is seen as a constraint on discovery [2, 8]. The cost of time spent in searching for data, and the potential failure to gain value from relevant data that was collected but cannot be located, provides the motivation for our research. Existing data access and search tools primarily focus on returning results that exactly match the user's request, and return nothing when no exact match is found – or a vast number of results when there are many matches. In a large collection, such searches can take hours.

In the world of genomics, tools such as Basic Local Alignment Search Tool (BLAST) [3] provide a form of similarity search for gene-sequence patterns. However, scientists also wish to search for physical properties or annotations, such as areas with high conservation and low methylation. Existing tools focus on browsing or visualizing annotations for the exact area the scientist has requested (e.g., [5, 11, 12]). Here, the scientist must first identify and then view the sequence segment of interest. But scientists wish to manipulate the whole genome, and collections of genomes. Browsing through the genome looking at each segment is not practical. Once a scientist has found one sequence of interest, she would like to quickly identify other parts of the genome that resemble it – perhaps not in the sequence itself, but in terms of its characteristics, as reflected in the annotations.

Our vision is to let scientists work with these massive data collections in a non-exhaustive way. Wide acceptance and use of interactive Internet search engines, such as Microsoft's Bing and Google, have made interactive, ranked search results over a huge, summarized collection (e.g., the Internet) a familiar paradigm for users of text search, including scientists in their non-science activities. Our vision is to apply this paradigm to the area of "big data" genomic search. We build on prior research that has provided one proof-point – called Data Near Here – with the very different discipline and data of oceanography.

2. RELATED WORK

Many researchers propose novel visualizations as a method for scientists to deal with the high volumes of data [5, 12]. However, these methods assume that the scientist knows which subsets of

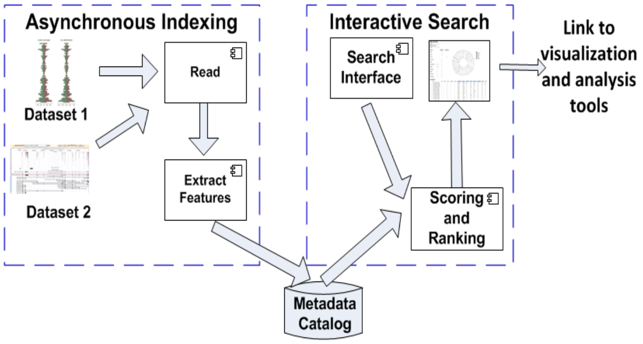
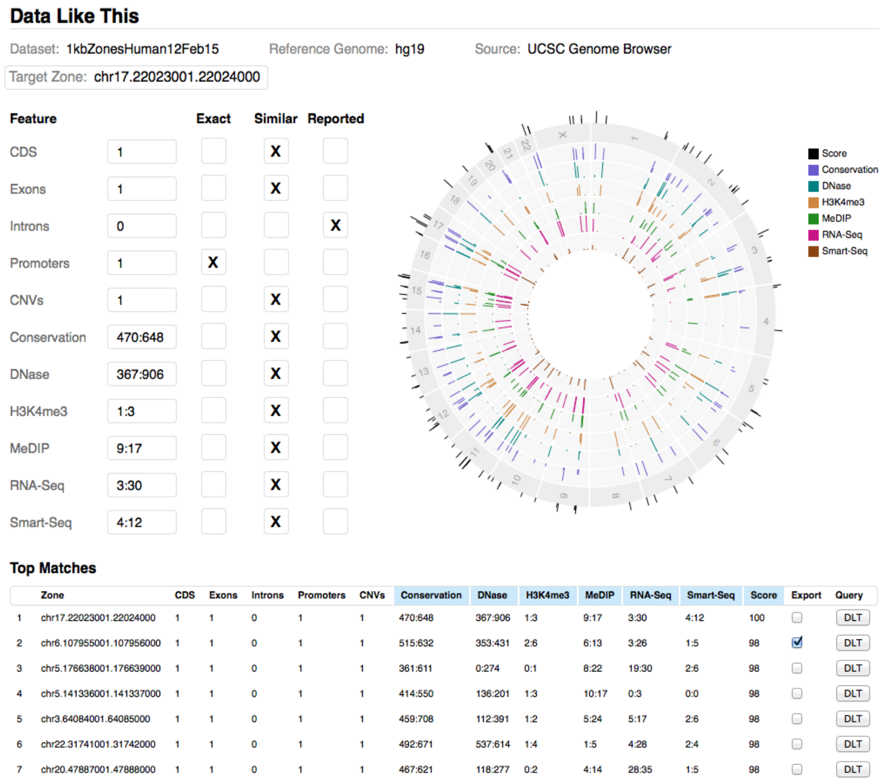


Figure 1. High-level dataset search architecture.

ExploreDB'15, May 31, 2015, Melbourne, Vic, Australia.

the data he wishes to visualize. With data the size of the genome, it is time-consuming to locate those areas. We directly address this challenge and propose an approach – ranked similarity search over summarized numeric data, using ideas from Internet search – for genome scientists to quickly identify regions of the genome that are similar in some characteristics while allowing flexibility in other characteristics.

In prior research, we successfully applied this approach to searching a large archive of numeric data stored in diverse formats in an ocean observatory ("Data Near Here" (DNH), at CMOP) [9, 10]. Figure 1 shows our high-level architecture, adapted from Internet search architectures. The Asynchronous Indexing component performs a one-time scan of each dataset in an archive to construct a summary, or "footprint," for each; we summarize its spatial and temporal extent, and record the ranges of physical variables. The summaries are stored in a Metadata Catalog. The Interactive Search component compares a search request to each dataset summary using a similarity function, allowing interactive search over an extensive and diverse archive. Datasets are ranked by similarity score, and presented to the user along with a "thumbnail" description and a map location. By providing the scientists with a ranked list of datasets "near" their search, we allow them to quickly explore the available data, allowing them to serendipitously discover nearby data and then narrow in on the most interesting subset. DNH was implemented at CMOP in 2012 and integrated into their suite of tools [7].



Data Like This

Dataset: 1kbZonesHuman12Feb15   Reference Genome: hg19   Source: UCSC Genome Browser

Target Zone: chr17.22023001.22024000

Feature		Exact	Similar	Reported
CDS	1	□	☒	□
Exons	1	□	☒	□
Introns	0	□	□	☒
Promoters	1	☒	□	□
CNVs	1	□	☒	□
Conservation	470:648	□	☒	□
DNase	367:906	□	☒	□
H3K4me3	1:3	□	☒	□
MeDIP	9:17	□	☒	□
RNA-Seq	3:30	□	☒	□
Smart-Seq	4:12	□	☒	□

Top Matches

	Zone	CDS	Exons	Introns	Promoters	CNVs	Conservation	DNase	H3K4me3	MeDIP	RNA-Seq	Smart-Seq	Score	Export	Query
1	chr17.22023001.22024000	1	1	0	1	1	470:648	367:906	1:3	9:17	3:30	4:12	100	□	DLT
2	chr6.107955001.107956000	1	1	0	1	1	515:632	353:431	2:6	6:13	3:26	1:5	98	☑	DLT
3	chr5.176638001.176639000	1	1	0	1	1	361:611	0:274	0:1	8:22	19:30	2:6	98	□	DLT
4	chr5.141336001.141337000	1	1	0	1	1	414:550	136:201	1:3	10:17	0:3	0:0	98	□	DLT
5	chr3.64084001.64085000	1	1	0	1	1	459:708	112:391	1:2	5:24	5:17	2:6	98	□	DLT
6	chr22.31741001.31742000	1	1	0	1	1	492:671	537:614	1:4	1:5	4:28	2:4	98	□	DLT
7	chr20.47887001.47888000	1	1	0	1	1	467:621	118:277	0:2	4:14	28:35	1:5	98	□	DLT

Figure 2. Prototype search interface for "Data Like This", showing a sample search for zones "like" a target zone for a subset of annotations: 5 Boolean (CDS, …, CNVS), one ordinal (conservation) and 5 numeric features. Result "zones" are shown in the bottom panel and in a Circos plot. In the ranked list of answers, one complete match for the search conditions (the target zone) was found; 8 partial matches are shown in the part of the text panel segment shown here. Partial matches are also shown on the Circos plot.

While there has been work to extend ranked search to handle numeric data [1, 4, 13], those efforts target web documents and embedded tables, rather than large collections of scientific data.

3. THE NEED FOR "DATA LIKE THIS"

We now apply these ideas to genomic research being performed at Oregon Health Sciences University. Here the "big data" is a collection of genome sequences of one or more cohorts of interest, with a large quantity of fine-grained annotations (such as levels of expression and regulation) for each dataset. These annotations are a mix of numeric, ordinal and binary data types, and apply to specific positions in the sequence data. Our researchers have methods for identifying the most relevant subsets of the genome using sequence similarity. Now they wish to also explore similarities in annotations, using approximate matching methods and exploratory search. The researchers wish to find and compare different regions in the genome that have some combination of features of several different data types in common. They are looking for "Data Like This" (DLT).

4. SEARCHING GENOME ANNOTATIONS

We are separately experimenting with indexing, scoring-and-ranking, and user interface components to work with genomic data. During asynchronous indexing, we segment a large dataset, in this case the genome, into smaller sections and create (at least) one summary for each segment. For our first test we arbitrarily divided each chromosome into contiguous "zones" of 1,000 base pairs, resulting in 3.0M zones for our initial genome of interest. Any segment summary can be returned for a search, allowing the most relevant subset of a large dataset to be found. We do not ascribe meaning to the size or location of the zone, but purely to the content.

We create a zone summary per zone from the base-pair annotations for 11 "annotations of interest" of our researchers (e.g., level of conservation, or existence of a promoter). Each of these annotations is summarized in one of three ways: a. into a Boolean value (e.g., this zone contains an intron); b. into an ordinal ranking of this zone as compared to others (e.g., conservation); or c. into a numeric range, representing the low, median and high value of a measured value across all base-pair positions in the zone (e.g., RNA-seq). While hundreds of variables could be extracted for this genomic dataset, our researchers selected these 11 as representative of the variety of annotations and as primary targets for research, thus forming a rich starting point for experimentation. Ordinal and Boolean data types are new for us, adding to our previous work with numerics.

We wish to understand if summarizing a genome in this way has meaning to genome researchers, and whether these summaries allow researchers to quickly explore research ideas, clarify their thinking, and reduce the data they need to analyze in detail. We plan to experiment with zones of other sizes and potentially with semantic meaning; for example, a larger section of the genome that has little variation could be treated as a single zone, while a section with great variation could be split into several smaller zones. It is also possible to summarize a single annotation in several different ways and make all versions available for search simultaneously, allowing multiple views of the data. We can also include measures of zone sequence similarities from such tools as BLAST, providing even richer search options.

We search over the summaries, returning results ranked by a measure of similarity [9]. We currently use an existing similarity measure from DNH, for which we have experimental evidence that it is a good proxy for how a user population ranks similarity of data ranges [10]. We believe this similarity measure is a good starting point. However, we wish to explore whether another similarity measure may provide a better model for these data types while retaining many of the qualities of our original measure (simple, fast to compute, and "good enough" – that is, returns a set of results ranked in an order that users generally agree with). How minimally can we modify our similarity measures across different data types and still get "good enough" results?

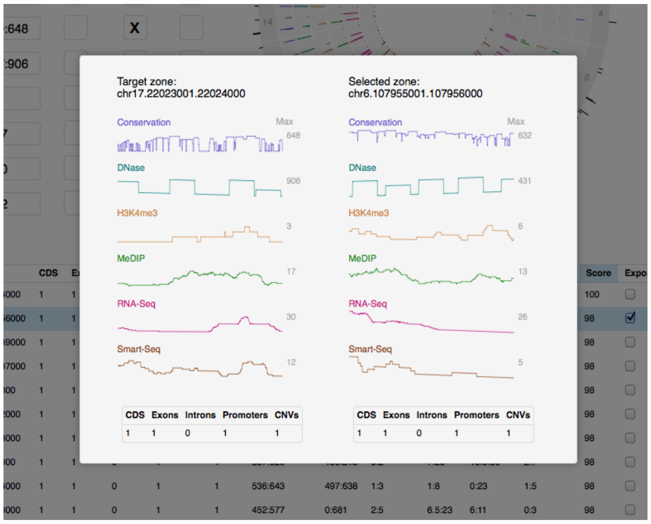
Each scientific discipline is accustomed to interacting with their data in certain ways. We show our current prototype user interface in Figures 2 and 3 (with real data). A Circos plot [6] (see top right of Figure 2) replaces DNH's Google map as a method to display the "locations" of search results. The "export" button in the search results will allow one or more identified zones to be migrated to analytic tools for further analysis; each zone summary identifies its start and end location, and many tools can be scripted or parameterized to load a specific genome region. Figure 3 shows a mock-up of a way a scientist can view details of a particular zone in the search results, in order to quickly evaluate its relevance.

We see some differences in the way these scientists wish to search for data as compared to the oceanographers and microbiologists who use DNH. Our colleagues sometimes wish to specify certain "required" annotation values as part of the search; for example, she may only wish to see results that have a promoter (see Figure 2, "Exact"). A search may now contain an arbitrary mix of required search terms (acting as a hard filter on the results) and similarity search terms (the default). Our collaborators are also exploring the idea of locating a zone of interest, and then using that zone as a "seed" for a search ("Target Zone"); that is, specifying a subset of the annotations for which they are looking for "like" zones. The proposed user interface in Figure 2 shows all these features.

We ran a small set of test searches over this set of "zone summaries" to test interactivity. Some of our test searches return in a few seconds, while others take several minutes. (The example search in Figure 2 took 3 seconds). In all these cases, these response times are significantly faster than searches using current approaches, where these same searches may take hours, or require significant manual effort or significant technical skill to construct. Even so, we believe that there is much opportunity for us to further improve response times, allowing scientists to rapidly explore many more candidates before narrowing in on a few.

5. CONCLUSION

This paper offers an approach for providing interactive, ranked search over annotations to bioinformatics data, such as over a whole genome or collections of genomes. We describe an approach that on builds on ideas successfully implemented in another scientific discipline, extending them to genomic data. Unlike existing tools aimed at fast data visualization, our approach allows scientists to search for, identify and reduce the data they wish to further visualize or analyze. These characteristics allow scientists to spend less of their time on data manipulation and more on high-value research and discovery.


Target zone:
chr17.22023001.22024000

Selected zone:
chr6.107955001.107956000

CDS	Exons	Introns	Promoters	CNVs
1	1	0	1	1

CDS	Exons	Introns	Promoters	CNVs
1	1	0	1	1

Figure 3. Proposed details view: Mousing over an entry in the results list will cause an overlay to appear comparing profiles of the data values for the target and selected zones.

6. ACKNOWLEDGMENTS

This work is supported by NSF award OCE-0424602, NIH/NCATS (5UL1RR024140), NIH/NCI (5P30CA06533), and Leukemia and Lymphoma Society (7005-11).

7. REFERENCES

[1] Agrawal, R. and Srikant, R. 2003. Searching with numbers. *IEEE TKDE*. 15, 4 (Aug. 2003), 855 – 870.

[2] Ahrens, J.P. et al. 2011. Data-intensive science in the US DOE. *CISE*. 13, 6 (Dec. 2011), 14 –24.

[3] Altschul, S.F. et al. 1997. Gapped BLAST and PSI-BLAST. *Nucleic acids res*. 25, 17 (1997), 3389–3402.

[4] Cafarella, M.J. et al. 2008. Webtables: exploring the power of tables on the web. *VLDB*. 1, 1 (2008), 538–549.

[5] CURSOR: *http://cursor.businesscatalyst.com/index.html*. Accessed: 2015-02-23.

[6] Krzywinski, M. et al. 2009. Circos: An information aesthetic for comparative genomics. *Genome Research*. 19, 9 (Sep. 2009), 1639–1645.

[7] Maier, D. et al. 2012. Navigating oceans of data. *Scientific and Statistical Database Management* (2012), 1–19.

[8] Martin Sanchez, F. et al. 2013. Exposome informatics. *J. of Am. Medical Informatics Ass*. 21, 3 (Nov. 2013), 386–390.

[9] Megler, V.M. 2014. *Ranked Similarity Search of Scientific Datasets (PhD Dissertation)*. Portland State University.

[10] Megler, V.M. and Maier, D. 2015. Are Datasets Like Documents?. *IEEE TKDE*. 27, 1 (Jan. 2015), 32–45.

[11] Robinson, J.T. et al. 2011. Integrative Genomics Viewer. *Nature Biotechnology*. 29, (2011), 24–26.

[12] UCSC Genome Browser: *http://genome.ucsc.edu/*. Accessed: 2015-02-23.

[13] Venetis, P. et al. 2011. Recovering semantics of tables on the web. *Proceedings of VLDB*. 4, 9 (2011), 528–538.

[14] Weidman, S. and Arrison, T. 2009. *Steps toward large-scale data integration in the sciences*. National Research Council of the National Academies.