RESEARCH

GEMINI: A Search Engine for Genomic Data

Timothy DeFreitas[1,2] and Patrick Flaherty[2,3,4*]

*Correspondence:
pjflaherty@wpi.edu
[3]Biomedical Engineering Department, Worcester Polytechnic Institute, 100 Institute Rd, 01609 Worcester, USA
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: Low-cost DNA sequencing allows organizations to accumulate massive amounts of genomic data and use that data to answer a diverse range of research questions. Presently, users typically extract data from these online resources using keyword or accession number searches. However, this search paradigm is slow and error-prone because the form of the query – a text-based string – is mismatched with the form of the target – a genomic profile.

Results: To improve access to massive genomic data resources, we have developed a fast search engine, GEMINI, that uses a genomic profile as a query to search for similar genomic profiles. GEMINI uses a vantage-point tree to store a database of n profiles and in certain circumstances achieves an $\mathcal{O}(\log n)$ expected query time in the limit. We show that GEMINI accurately identifies nearest-neighbor samples when applied to breast and ovarian cancer gene expression data from The Cancer Genome Atlas project and achieves a query time that scales as the logarithm of the number of records in practice on genomic data. In a database with 10^6 samples, GEMINI identifies the nearest neighbor in 0.05 sec compared to a brute force search time of 15.9 sec.

Conclusions: GEMINI is a fast search engine that uses a query genomic profile to search for similar profiles in a very large genomic database. It enables users to identify similar profiles independent of sample label, data origin or other meta-data information.

Keywords: genomic search; vantage-point tree; Cancer Genome Atlas

Background

Organizations such as individual research labs, sequencing core facilities, hospitals and research consortiums are accumulating large databases of gene expression and other genomic data from primary patient samples. Currently, the GEO database for microarray data contains more than 800,000 samples [1], the International HapMap 3 Project contains 1.6 million common SNPs for 1,184 individuals [2], and the Cancer Genome Atlas Project has 1.059 petabytes of genomic data on more than 20 types of cancer [3]. While these databases are already massive, the low-cost of next-generation sequencing is making it easier to add more data to these repositories and to build massive private data repositories [4]. Samples in these databases are often lightly annotated with clinical information or deidentified entirely for patient privacy. The question we address here is: "When a new patient sample arrives, what other samples, among those that we have seen, is most similar to this new one?" The solution we describe here, GEMINI, is a search engine that provides fast access to relevant samples in a database based only on similarity of gene expression profile, much like the PageRank algorithm provides access to internet web pages based on similarity between query terms and terms used in the web page content [5].

Previous work on search engines for gene expression data largely falls into two categories those that use a gene set query and those that use an expression profile query. ExpressionBlast takes as input a species type, a gene list, an output species and a distance metric and uses text analysis methods and uses an algorithm to standardize based on the query gene set to output labeled relevant experiments [6]. SEEK uses a novel cross-validaton-based algorithm to prioritize ranking and network information to identify relevant neighbors based on a query gene set for human data [7]. GeneChaser is an earlier effort that identifies all experiments where a single gene is differentially expressed [8]. In contrast to gene-set-based query search tools, ProfileChaser uses a GEO accession number to experiments that are similar to query [9]. The focus of that work is on choosing good data representation, dimensionality reduction, and similarity/distance metrics. However, they do not evaluate the computational performance or scalability of their approach. We build on the work in ProfileChaser by focusing on speed and scalability while allowing for interchangeability in dimensionality reduction methods and distance metrics.

The application domain of GEMINI is different and in some ways complementary to gene-set based methods such as SEEK and ExpressionBlast. Our focus is on developing a method that is amenable to different data representations, dimensionality reduction methods, and distance metrics, and, importantly, is also scalable so that it can be used by many concurrent users. To address the problem of sample similarity directly, GEMINI does not use gene-sets to select database records, and instead uses the similarity of expression profiles in the database to the query profile. In order to be practically useful such a search engine must be fast.

Tree data structures are common in search applications where optimizing query time is important [10]. By structuring the data records into a tree, a suitable algorithm is able to exclude irrelevant records from consideration and reduce search time to less than the brute force complexity of $\mathcal{O}(n)$, where n is the number of records in the database. Some binary tree data structures used for search include kd-trees [11], SR-trees [12], R*-trees [13]. Hash tables have very good search time for finding an exact match, but there is no good way to locate a record that is a nearest neighbor to a query. So, while hash tables are often used in applications where exact matches are needed, they are rarely used in application where near matches are needed.

GEMINI uses a vantage-point tree (vp-tree) data structure to store genomic data records [14, 15]. The vantage-point tree is a special case of a binary search tree where the left subtree of a node contains records that are closer than some distance, μ, and the right subtree contains records that are further than μ. The tree gets its name because the subtree nodes are partitioned from the vantage point of the current node. The advantage of the vp-tree in genomic search applications lies in the fact that is does not impose a particular coordinate structure on the data and instead employs a user-definable metric to measure distance. The construction and search algorithms for the vp-tree are described in the Implementation section.

Implementation

We describe the algorithms for the construction and search for the vantage-point tree here. GEMINI is implemented in python as a stand-alone command-line program

and as a public web site; we describe those implementations in the availability and requirements section.

Data Organization

A record in GEMINI is constructed a normalized gene-expression profile. In the cancer genome atlas project, this profile is a level 3 processed gene expression tab-delimited file. These records are converted to a HDF5 file format for compatibility and then preprocessed into a vantage-point tree. Internally, each tab-delimited file from the TCGA project consists of a vector of gene identifiers (e.g. "BRCA1"), and a vector of sample identifiers (e.g. "TCGA-59-2349..."), along with a matrix of the log2 normalized expression value for each gene-sample pair. A query is likewise an HDF5 file with the same attributes but with only one sample. A search therefore returns the most similar expression profiles in a dataset to the profile in the query.

Each vantage point tree is implemented as a python heapq object, and therefore must be entirely loaded into RAM to search a dataset. For datasets with thousands or millions of samples of complete gene expression profiles, the object requires several gigabytes of memory. Though memory performance is somewhat system-dependent, in our tests a database of 1 million records required 4GB. By reducing the complexity of the profiles using principal component analysis (PCA), the memory footprint can be reduced by more than 3 orders of magnitude. Datasets of with millions of profiles are currently size are rare – those from the TCGA contain roughly 17,000 genes and hundreds of samples, but PCA enables GEMINI to scale to accommodate increasing availability of data.

Vantage-point Tree Construction

Construction of the vp-tree takes $\mathcal{O}(n \log n)$ time for records with constant dimension where n is the number of records in the dataset. We briefly summarize the simplest version of the recursive construction algorithm here and refer to the original article for further details and extensions [14].

function `MakeVPTree(`$\mathcal{S}$`):`
- **Data**: a set of records, $\mathcal{S}$
- **Result**: a pointer to the root of the vp-tree
- **if** $\mathcal{S} = \emptyset$ **then return** $\emptyset$;
- node $\leftarrow$ a pointer to a new node;
- node.p $\leftarrow$ random element of $\mathcal{S}$;
- node.mu $\leftarrow$ median $d(\mathrm{p}, s)$ over all $s \in \mathcal{S}$;
- L $\leftarrow \{s \in \mathcal{S} - \{p\} | d(\mathrm{p}, s) < \mathrm{mu}\}$;
- R $\leftarrow \{s \in \mathcal{S} - \{p\} | d(\mathrm{p}, s) \geq \mathrm{mu}\}$;
- node.left $\leftarrow$ `MakeVPTree(`L`)`;
- node.right $\leftarrow$ `MakeVPTree(`R`)`;
- **return** node;

Algorithm 1: Vantage-point tree construction algorithm

This binary search tree construction works by taking a set $\mathcal{S}$ of records. If the $\mathcal{S}$ is not empty, we create a new node and store a random element, p, in the node. We store the median distance between p and all the other elements in $\mathcal{S}$ in mu in the node using any distance metric that satisfies the triangle inequality. We partition

the set $\mathcal{S}$ into two roughly equal size sets L and R, where L contains all of the elements of $\mathcal{S}$ that are closer to p than the median distance, mu and R contains all of the elements of $\mathcal{S}$ that are further than mu. The function recurses by calling itself with arguments L and R for the left (closer) and right (further) subtrees. The recursion ends when the subtree sets are empty and the algorithm returns the pointer to the root node. Clearly, because the size of the set in each subtree is half the original set, due to the use of the median distance, the time to construct the tree is $\mathcal{O}(n \log n)$.

Vantage-point Tree Search

Search in the vantage-point tree proceeds by recursive depth-first search. The left subtree of a node contains records that are closer than μ from the vantage point of the current node's records. Symmetrically, the right subtree contains records further than μ.

If we have a query profile, q and a vantage-point node, p, by symmetry and the triangle inequality of a distance metric $d(\cdot, \cdot)$, we have

$$d(q, s) \geq |d(q, p) - d(p, s)| = d_p(q, s), \tag{1}$$

where s is any other record in the database and $d_p(\cdot, \cdot)$ is defined as the vantage-point distance. Since the vantage-point distance shrinks the true distance between q and s, if $d_p(q, s) \geq \tau$, then $d(q, s) \geq \tau$ [14].

Suppose that we have found a record at distance τ from the query and we are at vantage-point node p in the tree. If $d(p, q) \geq \tau + \mu$, then the nearest-neighbor is not closer than μ and we can fathom (remove from further consideration) the left subtree as shown in Figure 1A. Conversely, if $d(p, q) + \tau \leq \mu$, then the nearest-neighbor is certainly closer than μ and we can fathom the right subtree (Figure 1B). Thus, the vantage-point tree data structure allows us to exclude records from examination and we achieve super-linear search time. As shown by Yianilos, the average-case querying time scales as $\mathcal{O}(\log n)$ when the data is low-dimensional [14].

The search algorithm can be written as a recursive depth-first search algorithm as described previously [14]. The algorithm holds the node of the nearest neighbor in the global variable `best` and is initialized with $\tau \leftarrow 0$. If d(q, node) $< \mu + \tau$, only the left subtree is traversed and if d(q, node) $> \mu - \tau$ then only the right subtree is traversed. If $\mu - \tau \leq$ d(q, node) $\geq \mu + \tau$ then both subtrees are traversed.

function `SearchVPTree(`*`node`*`):`
  Data: a vantage point tree root node, `root`
  Result: a pointer to the root of the vp-tree
  if *node* $= \emptyset$ **then return**;
  if *d(q, node)* $< \tau$ **then**
    $\tau \leftarrow d(q,$ `node`$)$;
    `best` $\leftarrow$ `node`
  end
  if *d(q, node)* $< \mu + \tau$ **then** `SearchVPTree(`*node.left*`)`;
  if *d(q, node)* $> \mu - \tau$ **then** `SearchVPTree(`*node.right*`)`;
return

Algorithm 2: Vantage-point tree search algorithm

Vantage point tree structures support any distance metric that satisfies the triangle inequality, but the optimal distance function is not yet known. For simplicity, GEMINI currently uses the euclidean distance between samples after principal component transformation. However, weighted distance functions utilizing genomic knowledge could better facilitate a particular search. For example, for a cancer dataset, one could limit the genes compared to known oncogenes, thereby finding which sample showed the most similar oncogenic profile to the query. This search should be more sensitive to small changes in particular genes, and and therefore result in less statistical noise, though we do not attempt to prove this in this paper.

Results

Comparison to Other Search Methods

We compare the vp-tree to the related KD-Tree as well as a brute-force approach in Figure 2. The brute force algorithm simply compares the query to every record in the database. As expected, the brute force approach scales linearly in the size of the database. However, the tree structure approaches scale as the log of the size of the database because of the savings achieved by being able to exclude distance samples from consideration based on their position in the vp-tree.

Though both of the tree based query algorithms scale similarly with the log of the database size in the average case for low-dimensional data, and their end structures are related, they differ in their construction algorithms and use of distance metrics. KD-Trees use non-leaf nodes to divide the dataset using a hyperplane whose normal vector is equivalent to one of the dimensions of the data. Splits continue recursively until the number of instances in each node is smaller than some threshold. [14] showed that query time for both the kd-tree and vp-tree scales exponentially with the dimension of the data set (Figure 6 in that paper). Therefore, for both methods, it is important to perform some form of dimensionality reduction prior to storing the data in the data structure.

Both kd and vp-trees are constructed in $\mathcal{O}(n \log n)$ with a linear time median-finding algorithm. Other trees achieve similar complexity and differ in the use of split heuristics and amount of reinsertion during construction.

Search in Cancer Genome Atlas Database

We tested GEMINI on a database of gene expression data from the cancer genome atlas (TCGA) comprised of 559 ovarian (OV) and 599 breast cancer (BRCA) samples. First, we projected the probe-level data onto the first 10 principal components to reduce the dimensionality of the data from 17,813 features to 10. The choice of 10 was selected due to the diminishing returns associated with each subsequent dimension. Using the BRCA and OV dataset, this projection preserved 40% of the variance in the data, while 4 dimensions preserved just 25% and 100 more were required to achieve 70%. A plot of the first two principal components for the BRCA and OV samples is shown in Figure 3. Clearly, the two cancer types differ in their gene expression patterns and cluster. However, there are two ovarian samples that do not cluster with the rest. One falls within a group of BRCA samples and the other falls outside of either cluster.

We tested GEMINI using four queries against the database of combined OV and BRCA samples. The four queries (shown circled in Figure 3) are: (A) a prototypical

BRCA sample, (B) a prototypical OV sample, (C) a BRCA-like OC sample, and (D) an outlier OV sample. The prototypical OV and BRCA sample is the nearest Euclidean neighbor to the average OV and BRCA expression respectively. The top 10 hits by similarity to the prototypical OV sample are all OV samples and the top 10 hits for the prototypical BRCA are all BRCA samples as expected (Figure 4). The BRCA-like OV sample (59-2349) has 4 BRCA samples and 5 OV samples in the top 9 hits. This result indicates that the BRCA-like OV sample is genomically similar to both types of cancer. The OV outlier, surprisingly, shows the most similarity to 9 BRCA samples. This result indicates that though the sample was isolated as an ovarian type cancer, it appear to most resemble breast cancer. Indeed, the genomic similarity between ovarian and breast cancer has been noted, with clear therapeutic implications [3]. Therapeutics are generally approved for types of cancer categorized by anatomical site of presentation. However, if the genomic driver of tumorigenesis of an ovarian tumor is identical to that of a breast tumor and there is an approved therapeutic for the breast cancer, the therapeutic may also be effective for the ovarian tumor.

Discussion

GEMINI forms the basis of an open-source platform for machine learning optimization of search result relevance in genomic data repositories. A clinician may be interested in different types of profiles than a patient or basic researcher. By observing the click-through behavior of an individual or group of users, the platform may learn and re-rank results based on individualized probabilistic assessments of relevance.

This search engine fits in the context of a large database of profiles that are centrally located as well as with distributed databases. While it may be impossible to store all of the public genomic data in one repository, autonomous software that crawls the web identifying genomic data resources can temporarily store the profile long enough to identify the insertion location in the tree. Only the url of the root source of the data would then be needed in the vp-tree. Then, if the record is identified as a near-neighbor, the profile can be retrieved on-demand.

Our capability to generate genomic data is outpacing our capability to analyze and re-use that data. A fast, accurate search engine for genomic data may enable researchers to make discoveries using community-collected data more effectively. GEMINI uses a vp-tree to enable us to make effective use of the massive genomic data repositories that we have created.

Conclusions

Current genomic data search engines use text-based queries to search for numerical (e.g. gene expression) genomic data profiles. But this paradigm represents a mismatch between the subject and object of the query. Our genomic data search engine, GEMINI, matches the query and database record forms and leverages a vp-tree data structure to deliver relevant results in worst-case $\mathcal{O}(\log n)$ time.

Availability and requirements

We have implemented GEMINI as a python module, a standalone command-line program and as a website. Our code extends an implementation of the vp-tree originally written by Paul Harrison, whose code is available in the public domain [16].

The KD-tree was implemented used a scipy library written by Anne Archibald [17]. Usage documentation for the python module is provided with the source code.

The standalone command-line program has two sub-commands: `build` and `search`. The build sub-command takes a HDF5 format file with three datasets: "Sample", "Feature", "Data" and returns a pickled vp-tree data structure. The source data contains sample names in "Sample", genomic features names (genes) in "Features" and the data matrix (features x samples) in "Data". The `search` sub-command loads the vp-tree structure created in the `build` step and a HDF5 file in the same format as the source data except with a single column for the "Data" vector as the query. GEMINI prints the top K matches in the source data matrix where K is 10 by default but can be modified in command-line options.

The web interface at genomics.wpi.edu/gemini has only one entry box for the user to specify the query HDF5 file. The vp-tree is built off-line and loaded using a separate administrative tool and associated with a specific query page for the data source. This design choice provides a robust and simple interface and minimizes the user-effort to search. After submitting the query, the user is directed to a results page that shows a heatmap representation of the top 10 matches to the query.

Project name: GEMINI

Project home page: http://genomics2.wpi.edu/gemini

Operating system: platform independent

Other requirements: python modules listed in requirements.txt on website. None for website.

License: in process

Competing interests

The authors declare that they have no competing interests.

Author's contributions

PF and TD implemented the algorithm and website. PF conceived of the project and TD performed the experiments. PF and TD both contributed to writing the manuscript. All authors read and approved the final manuscript.

Author details

[1]Computer Science Department, Worcester Polytechnic Institute, 100 Institute Rd, 01609 Worcester, USA. [2]Program in Bioinformatics and Computational Biology, 100 Institute Rd, 01609 Worcester, USA. [3]Biomedical Engineering Department, Worcester Polytechnic Institute, 100 Institute Rd, 01609 Worcester, USA. [4]Department of Mathematics and Statistics, University of Massachusetts, Amherst, 710 N. Pleasant St, 01003 Amherst, USA.

References

1. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N., Holko, M., Ayanbule, O., Yefanov, A., Soboleva, A.: NCBI GEO: archive for functional genomics data sets—10 years on. Nucl. Acids Res. **39**(suppl 1), 1005–1010 (2011)
2. International HapMap 3 Consortium: Integrating common and rare genetic variation in diverse human populations. Nature **467**(7311), 52–58 (2010)
3. Network, T.C.G.A.: Comprehensive molecular portraits of human breast tumours. Nature **490**(7418), 61–70 (2012)
4. Rung, J., Brazma, A.: Reuse of public genome-wide gene expression data. Nat Rev Genet **14**(2), 89–99 (2013)
5. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report (November 1999)
6. Zinman, G.E., Naiman, S., Kanfi, Y., Cohen, H., Bar-Joseph, Z.: ExpressionBlast: mining large, unstructured expression databases. Nature Methods **10**(10), 925–926 (2013)
7. Zhu, Q., Wong, A.K., Krishnan, A., Aure, M.R., Tadych, A., Zhang, R., Corney, D.C., Greene, C.S., Bongo, L.A., Kristensen, V.N., Charikar, M., Li, K., Troyanskaya, O.G.: Targeted exploration and analysis of large cross-platform human transcriptomic compendia. Nature Methods **12**(3), 211–43214 (2015)
8. Chen, R., Mallelwar, R., Thosar, A., Venkatasubrahmanyam, S., Butte, A.J.: GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed. BMC Bioinformatics **9**(1), 548 (2008)

9. Engreitz, J.M., Morgan, A.A., Dudley, J.T., Chen, R., Thathoo, R., Altman, R.B., Butte, A.J.: Content-based microarray search using differential expression profiles. BMC Bioinformatics **11**(1), 603 (2010)
10. Knuth, D.E.: Optimum binary search trees. Acta Informatica **1**(1), 14–25 (1971)
11. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 881–892 (2002)
12. : The SR-tree An Index Structure for High-dimensional Nearest Neighbor Queries. ACM, New York, NY, USA (1997)
13. : The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. ACM, New York, NY, USA (1990)
14. : Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (1993)
15. : Bregman Vantage Point Trees for Efficient Nearest Neighbor Queries. IEEE (2009)
16. Harrison, P.: VP Tree. Technical report (2006)
17. Archibald, A.: KDTree. Technical report (2008)

Figures

Figure 1 Vantage-point tree structure allows search algorithm to fathom subtrees. Given we have already found a record at distance τ from the query node q and we are at vantage point p with a right subtree containing records further than μ from p and a left subtree with records closer than μ. (A) If $d(p,q) \geq \tau + \mu$, then the nearest-neighbor is not closer than μ and we can fathom the left subtree containing records closer than μ. (B) If $d(p,q) + \tau \leq \mu$, then the nearest-neighbor is certainly closer than μ and we can fathom the right subtree.

Figure 2 Timing comparison of GEMINI and other search methods. The search time in seconds is shown for a typical query in databases ranging in size from 100 samples to 1,000,000 samples. The brute force search time scales linearly with the size of the database, while GEMINI search time scales as the log of the size of the database.

Figure 3 Differential gene expression for ovarian and breast cancer samples from TCGA. (left) Principal component analysis is used to project the 17,813 dimension gene expression data to two dimensions for visualization. The ovarian samples and breast samples clearly cluster. One ovarian sample (C) has an expression pattern similar to breast cancer samples and one (D) shows an expression pattern outside of both the ovarian and breast clusters. Representative breast (A) and ovarian (B) samples are circled. (right) A boxplot of all non-zero pairwise distances in the joint breast and ovarian cancer data sets. The nearest neighbors for the four queries are shown as symbols in the legend. We find that the nearest neighbors all fall closer than the lower quartile of all of the distances.

Figure 4 GEMINI heat map results showing 9 nearest neighbors to the query (top row) for four samples. Four query profiles were used to search for nearest-neighbor profiles in a database containing both ovarian and breast cancer samples. The nearest neighbors of the prototypical breast cancer profile are all breast cancer samples and the nearest neighbors of the prototypical ovarian cancer profile are all ovarian cancer samples as expected. The ovarian cancer sample that falls in the breast cancer cluster is nearest neighbors with both ovarian and breast cancer samples. The ovarian cancer outlier has all breast cancer samples as nearest neighbors indicating that the differential gene expression patterns for that sample most closely resemble breast cancer.