BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 2 2012, pages 160–166
doi:10.1093/bioinformatics/btr652

Genome analysis Advance Access publication November 24, 2011

Using Sybil for interactive comparative genomics of microbes on the web

David R. Riley[1,*], Samuel V. Angiuoli[1], Jonathan Crabtree[1], Julie C. Dunning Hotopp[1,2] and Hervé Tettelin[1,2]

[1]Institute for Genome Sciences and [2]Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Analysis of multiple genomes requires sophisticated tools that provide search, visualization, interactivity and data export. Comparative genomics datasets tend to be large and complex, making development of these tools difficult. In addition to scalability, comparative genomics tools must also provide user-friendly interfaces such that the research scientist can explore complex data with minimal technical expertise.

Results: We describe a new version of the Sybil software package and its application to the important human pathogen *Streptococcus pneumoniae.* This new software provides a feature-rich set of comparative genomics tools for inspection of multiple genome structures, mining of orthologous gene families and identification of potential vaccine candidates.

Availability: The *S.pneumoniae* resource is online at http://strepneumo-sybil.igs.umaryland.edu. The software, database and website are available for download as a portable virtual machine and from http://sourceforge.net/projects/sybil.

Contact: driley@som.umaryland.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 4, 2011; revised on November 10, 2011; accepted on November 20, 2011

1 INTRODUCTION

The increasing throughput and reduced cost of second- and third-generation DNA sequencing technologies has resulted in a deluge of genome sequence data in the public domain. These require efficient bioinformatics tools enabling visualization and mining of multigenome comparisons. Such tools are difficult to develop because of the size and complexity of such large-scale comparisons. The challenge is to develop a tool that can be flexible, extensible and usable while providing the following four basic features: search, visualization, interactivity and export.

1.1 Search

The Internet and the ubiquity of Google, Yahoo and Bing have proven that when datasets become extremely large, search becomes a critical entry point. Internet searching is typically done based on unstructured keyword searches with some structured filters such as date or country of origin. In the case of biological data, structured and unstructured searches are both important features. Unstructured keyword searching gives users the power to find items of interest based on free text that matches against fields like gene name or gene function in the same way one might search the Internet. Structured search parameters include filtering for source organism, feature length, presence of a particular motif and overlap with another feature of interest. Both these search strategies combined can form the basis of a powerful tool with the ability to find a reasonable number of features of interest in an otherwise vast and inaccessible dataset. BioMart (Smedley *et al.*, 2009) is a widely used tool that integrates large datasets and provides advanced structured and unstructured search capability. While not specifically designed for comparative genomics, BioMart's power includes its flexibility in handling data of many different types and formats. It is in this spirit that search capability should be implemented into a full-featured comparative genomics tool.

1.2 Visualization

Summarizing large datasets in pictures can reveal important patterns and results in the data and can serve as a valuable entry point for data querying. Illustrating a small-scale genomic rearrangement or deletion is typically the most effective way to describe the event. Existing comparative genomics visualization tools include linear genome browsers, dot-plot viewers and circular genome viewers (Nielsen *et al.*, 2010). Linear browsers draw pictograms of the genes in the genome with links between related genes or regions. For example, SynView (Wang *et al.*, 2006), Gbrowse-syn (McKay *et al.*, 2010) and SynBrowse (Pan *et al.*, 2005) are linear browsers useful for comparing genomic regions between a few genomes. The UCSC genome browser (Kent *et al.*, 2002) and VISTA (Frazer *et al.*, 2004) are linear representations providing conservation plots that indicate regions where the genomes share conserved primary sequence. Such conservation plots can provide a high-level visualization of whole genome alignments. This type of view is most useful when looking at large, distantly related genomes. Dot-plot viewers are an alternative way of visualizing whole genome alignment data with a particular emphasis on large-scale changes in synteny and indel detection. Dot-plots are limited in that they can only compare a limited number of genomes at once and typically do not display annotated feature information. The whole genome aligner Mauve has a built-in viewer that color-codes locally syntenic regions across several genomes (Darling *et al.*, 2004). GMAJ is another viewer that specializes in displaying whole genome alignment data as percent

*To whom correspondence should be addressed.

identity plots and dot-plots (Blanchette *et al.*, 2004). These tools are particularly useful when looking for large-scale rearrangements, insertions and deletions. Circular representations like those from the Circos package (Krzywinski *et al.*, 2009) draw genomes in a circle and draw arcs between regions of conservation. These displays can capture much of the information of the linear browser with the added benefit of being compact and having fewer line crossings. Circular browsers can also work at the whole-genome scale. The challenge for comparative genomics visualization methods is including the ability to navigate seamlessly from a whole genome view that accurately depicts feature content and conservation information to a localized view that provides detailed imagery of genomic features and comparisons.

1.3 Interactivity

While search and visualization can provide valuable entry points to large datasets, the ability to interact with these results turns a simple tool into a powerful application. The power to sort results by any field, further refine results and navigate to visualizations makes search a jump-off point to more in-depth data mining. A more difficult feature to implement is interactivity of visualizations. Giving the user the power to obtain detailed information about drawn features, download sequence data and navigate to other views for a different perspective make an otherwise static graphic into another powerful entry point. Gbrowse-based (Stein *et al.*, 2002) tools have the ability to leverage mouseover events, which give these tools the power to hide detailed information from the visualization until requested. Clicking on gene features in these tools can bring the user to a report page with additional details about the gene. Similarly, the UCSC genome browser provides several links to more detailed information about the different features it can display. These features allow the user to begin at a particular search result or graphic and investigate features of interest without having the impression of reaching a dead end.

1.4 Data export

Export is essential in allowing users the flexibility to take data and results out of the system for publication and further inspection. In large comparative genomics datasets, interactive search and visualization capabilities are used to identify a set of genes, gene clusters or genomic regions of particular interest. Export of these data allows for additional analyses outside of the tool. Data export can include the ability to capture a vector graphics-based publication quality image, a table or a FASTA file for use in downstream bioinformatics analyses.

1.5 Sybil

With Sybil, we sought to combine interactive search, visualization and data export in a dynamic web-based comparative genomics software package that provides the freedom to analyze many data types in a comparative genomics context (Crabtree *et al.*, 2007). The Sybil suite of feature-rich web tools allows users to search and visualize several genomes based on clusters of orthologous genes. The views in Sybil are interactive allowing users to click on each drawn feature to retrieve more detailed information. Navigation between views is supported and provides Sybil with the unique ability to explore from the whole genome level to the protein and nucleotide multiple sequence alignment level without leaving the application. Export of tab-delimited tables, sequence data, and SVG and PDF publication-quality figures is possible from many of the views.

The Strepneumo Sybil-based website http://strepneumo-sybil.igs.umaryland.edu was developed as a comparative genomics resource to house all published, annotated genome sequences of the human pathogen *Streptococcus pneumoniae.* The aim of the project focused primarily on combining annotated genomics data in a comparative context with several additional data types related to vaccine candidate identification. Infections caused by *S.pneumoniae* continue to be a major cause of human morbidity and mortality worldwide (O'Brien *et al.*, 2009). This is a result of the pathogen's extreme adaptability that leads to antibiotic resistance and escape from the current polysaccharide capsule-based vaccines (Mera *et al.*, 2008). There is a crucial need for novel candidates for the development of efficacious protein-based vaccines. Efficient mining of the wealth of genome sequence data available for *S.pneumoniae* will open novel avenues for vaccine development.

2 METHODS

2.1 Clustering

Multigenome homologous gene clusters are the primary comparative data type in Sybil. These clusters must be pre-computed and loaded into a relational database that supports the Sybil software. Gene clusters are treated as generic objects and alternative clustering methods can be used. For the Strepneumo site, the clusters were generated using a modified form of reciprocal best BLAST (Altschul *et al.*, 1990) match corrected for paralogs as described in Crabtree *et al.* (2007). In brief, an all-vs.-all BLASTP with an *e*-value cutoff of 1e-05, a percent identity cutoff of 80% and a percent coverage cutoff of 70% results in a hit graph. The hit graph is used to first identify homologs within a single genome, termed paralogs. Paralogs are grouped by computing a Jaccard similarity coefficient (Jaccard, 1908) for each pair of proteins and applying a cutoff of 0.6. The resulting groups of paralogs and singleton genes are used as input to identify clusters of genes across genomes or orthologs. In the most basic example, two genes are grouped in an ortholog cluster if they are from different genomes and are each other's best hit, or reciprocal best BLAST match. When orthology involves Jaccard paralog clusters, all the genes in the cluster are treated together such that any member that is a best hit can cluster a paralog cluster with a singleton gene or another paralog cluster. For example, if genes A + B are paralogs in genome X and genes C + D are paralogs in genome Y. If gene A's best hit in genome Y is C but C's best hit in X is B these four genes will end up in an ortholog cluster together despite not having an individual reciprocal best hit. For Strepneumo, this pipeline is implemented in an instance of Ergatis (Orvis *et al.*, 2010) and results are loaded into a relational database that supports Sybil.

2.2 Data types

In addition to comparative computes, Sybil supports a number of additional data types that can be visualized and used as search criteria. These data types fall into two major categories: genome data and protein/gene data. Genome data is the result of a search or analysis that finds features located on the raw genome sequence and therefore has genome coordinates. In the case of Strepneumo, a simple repeat library was used along with Repeatmasker (Smit *et al.*, 1996) to predict simple repeats, Phobos (Mayer, 2006–2011) was used to identify tandem repeats, HMMER (Eddy, 1998) was used to identify several genomic repeats particular to *S.pneumoniae* and genomic islands were identified using IslandPath (Hsiao *et al.*, 2003). Protein/gene-based data are associated or located on genes and are the result of an analysis

Table 1. Feature types loaded into the Strepneumo Sybil system

Feature name	Tool	No. of features found
Genome features		
Simple repeat	Repeatmasker	3470
Homopolymeric tract	Custom Perl	7964
Tandem repeat	Phobos	17223
BoxA repeat	HMMER	4368
BoxB repeat	HMMER	6111
BoxC repeat	HMMER	4057
RUP repeat	HMMER	3555
Genomic island	IslandPath	162
Protein features		
Signal peptide	SignalP	13899
B-cell epitope	BepiPred	10240
Antigenic region	EMBOSS antigenic	9668
Lipoprotein	Custom Perl	1647
LPxTG	HMMER	1264
Membrane/surface motif	HMMER	6398
Bacteriocin motif	HMMER	195
Fibronectin motif	HMMER	34
Transmembrane region	tmhmm	79543
Subcellular localization	psortb	*NA*

These features are divided into Genome features, found and located on genomic coordinates, and Protein features, found and located on polypeptide sequences. Detailed information on each can be found in the Supplementary Table S1.

that operates on protein or nucleotide gene sequences. In this case, the analysis can yield either a feature with gene coordinates or it can predict or identify a particular property of the target gene. Strepneumo contains signal peptides from SignalP (Bendtsen *et al.*, 2004), B-cell epitope predictions from BepiPred (Larsen *et al.*, 2006), transmembrane regions from tmhmm (Krogh *et al.*, 2001) and subcellular localizations from psortb (Yu *et al.*, 2010). The psortb subcellular localizations exemplify an analysis that gives a gene a particular property but does not predict features with coordinates. The EMBOSS tool antigenic (Kolaskar and Tongaonkar, 1990) was used to predict antigenic regions. In addition, several protein motifs were predicted using hmmpfam (Eddy, 1998) including LPxTG Gram-positive cell wall anchors, as well as bacteriocin and fibronectin binding sites. Any data type with either genomic or gene coordinates can be used in the system. New data types must be loaded into the underlying database and configured in the Sybil site configuration file. A complete list of the data types provided in the Strepneumo Sybil site is found in Table 1 with more details in Supplementary Table S1.

2.3 Data storage

All data are stored in a Chado (Mungall and Emmert, 2007) relational database using PostgreSQL. The results of the protein and genomic feature searches are localized to the relevant protein and genome sequences, respectively. Clusters are stored as 'match' features with member proteins linking to this central match. The Chado schema is designed with a high degree of abstraction and normalization resulting in reduced performance with very large datasets. To address this, a data mart was developed called ChadoMart. This mart consists of three tables: cm_proteins, cm_clusters and cm_cluster_members. The cm_proteins table contains de-normalized information about each protein while the cm_clusters table includes summary statistics on all protein clusters. The cm_cluster_members table links the cm_clusters table with the cm_proteins table, mapping all the proteins to their respective clusters and vice versa. These three tables contain the data most frequently queried by Sybil. Therefore, using ChadoMart significantly improves the website performance. In addition to the ChadoMart, server-side query caching was implemented to further improve performance. For overlap or ranged queries an interval tree is used. Using an interval tree was found to be much faster than doing large table joins in the relational database.

The dataset can be quite large with the all vs. all BLAST results constituting a majority of the data. In the most recent version of the Strepneumo data, the BLAST results accounted for ∼10.5 million of the 11 million rows in the central 'feature' table. Therefore, for distribution of the data for local use, the BLAST results can be left out of the database, allowing more modest hardware to run the system successfully.

2.4 Visualization/user interface

The Sybil website is composed of a combination of traditional standard-based HTML/CSS and JavaScript along with elements from the JavaScript framework ExtJS. HTML/CSS and JavaScript follow the HTML 4.01 standard and the implementation of the site is designed to work in most modern web browsers. The use of asynchronous JavaScript server requests (AJAX) gives Sybil the feel of an application rather than a website. Complementing the traditional HTML/CSS/JavaScript is the use of ExtJS, a JavaScript library that speeds development of feature-rich web applications. ExtJS is used exclusively on the Protein/Cluster Search Pages and the Genomic Comparative View. The ExtJS grid on the search pages provides useful functionality like sorting, column reordering and pagination. ExtJS windows are used as informational popups when rendered features are clicked and ExtJS ToolTips are used to display help information when a user mouses over a box with a question mark (?) as shown in Figure 1.

Visualization is central to the Sybil system and is accomplished using two different methods. The first leverages the BioPerl (Stajich *et al.*, 2002) Bio::Graphics Perl package. An additional Sybil-specific glyph allows visualization of the cluster relationships across genomes (Fig. 2). Bio::Graphics is used for the Genomic Comparative View, Cluster Report and Protein Report pages. In general, this package is good at displaying comparative data for up to ∼100 kb of bacterial sequence. The second method is a custom module called Sybil::Graphics::GenomeImage developed as part of the new implementation of Sybil. GenomeImage currently leverages SVG natively. This package is used for drawing whole-genome scale images such as the Gradient View (Fig. 3) and the Whole-Genome Display (Fig. 4). GenomeImage allows for the drawing of large amounts of data with less detail than Bio::Graphics while still retaining an interactive image map for web display.

3 RESULTS

3.1 Data mining and visualization

A number of interactive visualization and querying tools were developed to aid interactive analysis of the *S.pneumoniae* data. Effort was primarily focused on providing tools that are responsive and informative for comparisons of many genomes using a web browser. Particular attention was paid to supporting navigation between the various Sybil views. Providing this capability gives Sybil the feel of a fully integrated application rather than a set of independent web pages.

3.1.1 Search Sybil provides extensive search capability, and the results are interactive and exportable. Gene queries can be initiated using locus identifiers, database accessions, gene names and protein functions. Results can be filtered by taxonomy, length, subcellular location, protein motifs, overlap with genomic features, membership in orthologous clusters and BLAST matches. The orthologous gene clusters are also searchable based on the same criteria as gene searches with the addition of taxonomic profile and cluster size. The search results are displayed in a paginated table in the web interface as shown in Figure 1. The data can be sorted by any column,

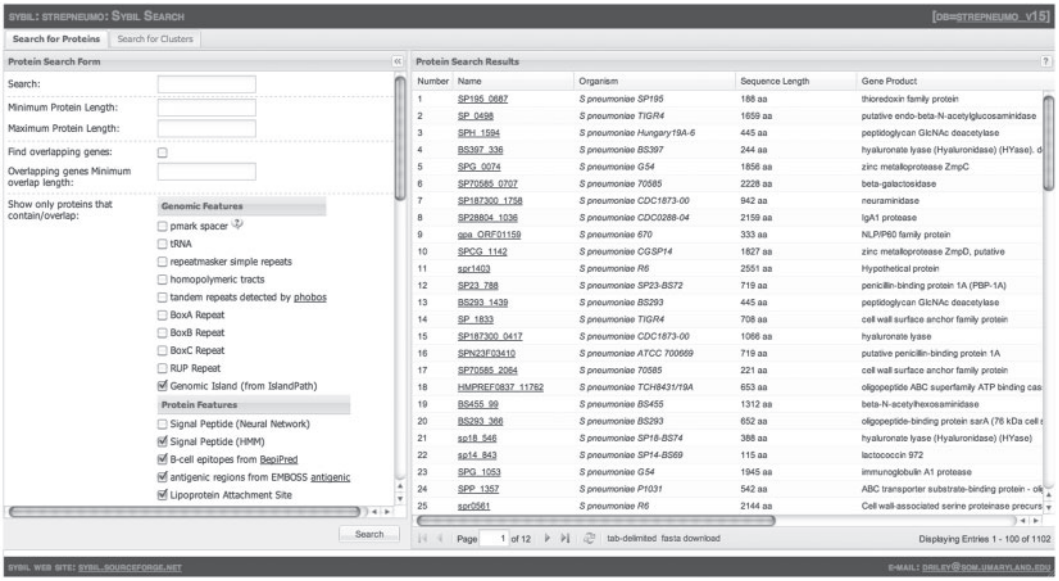

Fig. 1. The Sybil search page is made up of two panels: a search form on the left and a table of results on the right. The form on the left provides a variety of search options including free text, length filters, overlapping feature filters and more. The table on the right shows the results of the search in a paginated form. The table includes links to other Sybil pages for more detailed information and visualization. The data can be exported as a tab-delimited or multi-FASTA file.

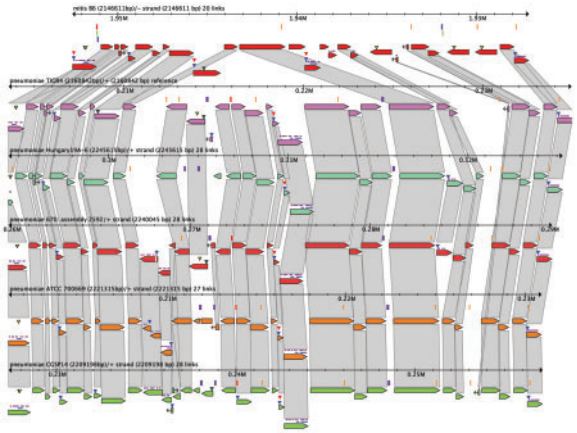

Fig. 2. The Genomic Comparative View provides the ability to visually compare genomic regions. Genes in each region of each genome are drawn with arrowheads indicating their orientation. Genes that share a common cluster are connected via a shaded polygon. Additional genome/protein feature types are drawn as different shapes in different colors in their respective positions. This particular image shows the deletion of several genes in the top genome.

and links are provided to gene and cluster report pages. Results are also downloadable as a tab-delimited text file or FASTA file for downstream analysis.

3.1.2 Cluster report page/genomic comparative display Visualizing genomic comparisons is central to Sybil's power. Both the cluster report page and the genomic comparative page allow the user to view genes and other genomic features drawn in a logical linear fashion (Fig. 2). Gene cluster relationships are easily visible providing clear illustration of insertions, deletions, rearrangements and annotation/sequencing anomalies. On the cluster report page, the cluster of interest is drawn in the center with pink shading. The genomic comparative display supports two drawing modes: search mode and edit mode. Search mode allows the user to specify a reference region and search query genomes for similar regions. Edit mode allows the user to manually define all viewable regions and can also be used to perfect a view to create publication ready figures. Clicking on genes, gene clusters or any protein or genomic feature will generate a popup with data pertinent to the clicked feature and links to other views/pages. This allows Sybil displays to forego displaying labels for all features, decluttering complex images. Clicking on a gene or gene cluster will provide several pieces of data and additionally provide a link to center on this gene or cluster in the Genomic Comparative View. Several links are provided allowing the user to center on the gene or cluster in the Genomic Comparative View with a variety of base pair pads and search other genomes in the dataset for homologous regions. These links give the user the power to explore regions of interest more fully without leaving the system. Multi-FASTA files of protein or nucleotide sequences can be exported for external analyses and publication quality figures can be exported in SVG or PDF format using buttons along the bottom of the page.

3.1.3 Whole genome display/synteny gradient display The synteny gradient display enables the visualization of changes in synteny relative to a reference (Fig. 3). In this view, a reference genome's genes are colored from yellow to blue on a gradient from left to right. If a query genome shares a cluster with a reference gene, then it is drawn above the matching reference gene but using a color that corresponds to the query gene's position in its native genome. The resulting figure reveals conservation of the color gradient in syntenic regions while shared genes located in rearrangements will

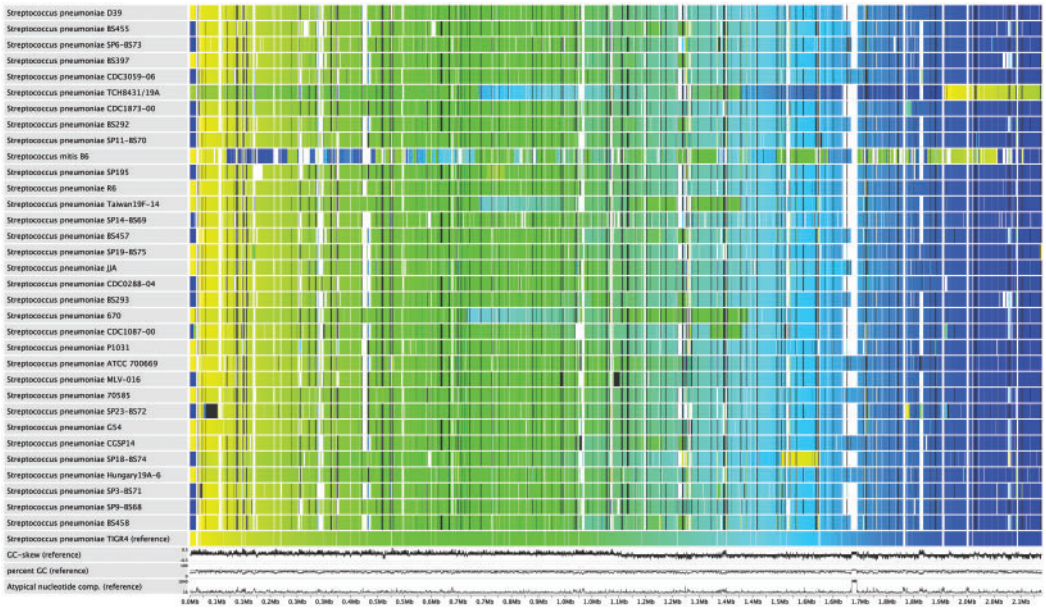

Fig. 3. The synteny gradient display is a unique representation of conserved gene content/order between several genomes that can be used to identify rearrangements as well as large-scale insertions/deletions. The view is based on a reference sequence (in this case *S.pneumoniae* TIGR4) that is drawn in the bottom panel. Genes in the reference genome are colored yellow to blue from left to right. White in the reference denotes a region with no gene annotation. Matching genes in each of the 33 query sequences are drawn atop their reference match, i.e. not in the order in which they appear in their respective genomes. To highlight rearrangements and conserved synteny, the matching genes are colored based on the relative position in their respective genomes (yellow for the beginning and blue for the end). Genes shown in black are part of a paralogous cluster in their respective genome and therefore do not have a single native location. GC-skew, %GC and atypical nucleotide composition are plotted for the reference genome.

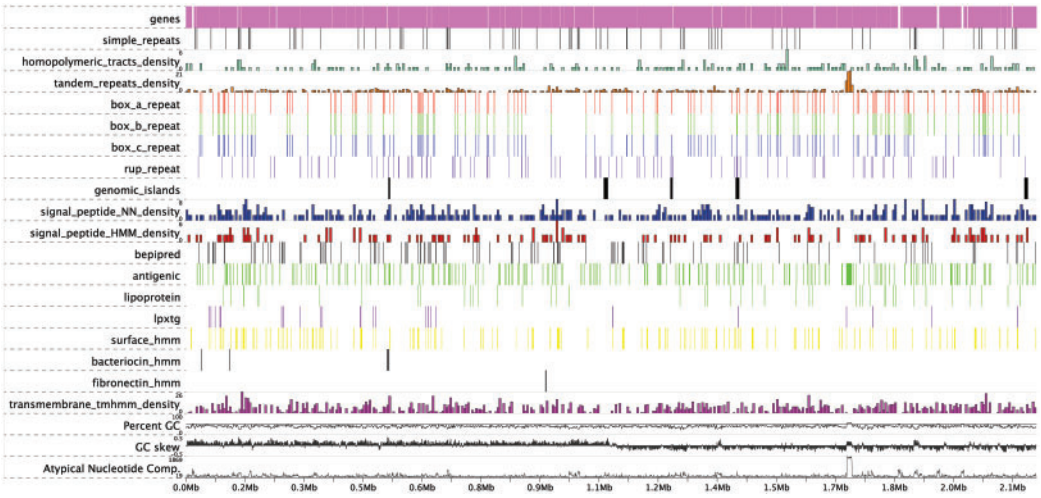

Fig. 4. This whole genome display illustrates Sybil's data type flexibility. In this figure, all the genes from an entire *S.pneumoniae* genome are drawn at the top. Below the gene track are several additional annotations listed in Table 1. Any feature that has a genomic or gene coordinate can be drawn. For signal peptides and *trans*-membrane regions, the density of features is drawn. Lastly, % GC, GC-skew and atypical nucleotide composition plots are drawn.

show up as color mismatches between the reference and query genomes. Gaps in query genomes represent genes that are present in the reference but not the query. Query genes with paralogs will show up as either colorless or black depending on the user's preference. Multiple drawings can be made at once with each genome serving as the reference if the user chooses. Additionally, the genes in the reference are clickable. This generates a popup with information about the gene, its clusters and links to other Sybil views or pages. Images can be exported as high-quality SVG or PDF files for publication or other uses.

As the name implies, the whole genome display provides visualization of an entire genome or chromosome (Fig. 4). This

display draws an entire genome in a linear and compact fashion with genes drawn as vertical bars. The bars are clickable and more information is provided in popups including links to the genomic comparative display, and cluster and protein report pages. The power of this display is the ability to draw additional genomic data types and graphs of feature densities. This view also allows filtering the displayed genes by gene function keyword. Like the synteny display, images are exportable as SVG or PDF.

3.2 Strepneumo Sybil

The Sybil system has been leveraged to explore the genomes of the human pathogen *S.pneumoniae*. Currently, the Strepneumo public website houses 34 streptococcal genomes (33 *S.pneumoniae* and 1 close relative, *S.mitis*) comprising 76 539 genes with a total of 3407 ortholog clusters. In addition to the cluster data, several other data types are present, including genome features (e.g. genomic islands, repeats) and protein features (trans-membrane spans, cell wall anchors) (Table 1). These data types can be used to search for genes and gene clusters and can also be visualized in all the Sybil views.

3.2.1 Potential vaccine candidates Many of the data types used in Strepneumo were inspired by the need to identify potential vaccine candidate genes for the purpose of reverse vaccinology (Sette and Rappuoli, 2010). These features and properties have been used in combination to find genes that are exposed on the cell surface and could therefore be accessible to the immune system. Antigenicity and epitope prediction are also used to identify protein regions that have the greatest chance of interacting with the immune system. The goal is to improve reverse vaccinology by reducing the number of predicted vaccine candidates such that downstream experimental confirmation is less burdensome. No single analysis is a good predictor for whether a gene will make a good vaccine candidate. However, combining a variety of properties and visualizing them across the different strains can identify a reduced set of likely candidate genes.

3.2.2 Strepneumo-Sybil virtual machine While Strepneumo is freely available to everyone with access to an Internet connection, some situations like travel or unreliable Internet connections may necessitate a local copy of the tool. The Sybil package requires considerable expertise to install, therefore distribution of mirrors and local copies of the Strepneumo Sybil system was accomplished via virtual machines (VM). The Strepneumo VM leverages the CloVR (Angiuoli *et al.*, 2011b) architecture, contains an install of Ubuntu Linux and comes pre-loaded with all the dependencies required for running Sybil and the Strepneumo website. The virtual machine is ∼1.7 GB in size when compressed and is available for download from the Strepneumo website. Any computer that can run the 64-bit VMWare player, has at least 15 GB of disk space and 2 GB of RAM can run the Strepneumo Sybil website locally.

3.3 Sybil screencasts

One often-overlooked aspect of Bioinformatics software development is documentation and training. For Sybil it was decided that documentation would be part of the tool. To accomplish this goal, descriptions of the various panels and their use are available by hovering the cursor over the '?' boxes at the top of each panel. In the case of complex features like the protein/cluster search page, cluster and protein reports, and Genomic Comparative View, video screencasts were recorded and uploaded to YouTube (www.youtube.com/SybilScreencasts). Each screencast is ∼10 min in length and describes each of the basic features of the tool. The screencasts are narrated and the watcher observes the tool in use. These screencasts can be accessed by going directly to the SybilScreencasts YouTube page or by hovering over the '?' icon in one of the tools where a screencast is available.

4 DISCUSSION

4.1 Sybil VM

In order to provide a system that is portable, deployable and scalable, the Strepneumo Sybil website was deployed using the CloVR virtual machine. The VM proved to be invaluable in setting up the system in two remote sites: the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in India and the International Livestock Research Institute (ILRI) in Kenya. These two sites provided mirrors of the Strepneumo Sybil instance by running a VM based on the CloVR project. These installations did not require travel and were frequently updated in a matter of a few days with minimal effort.

The computes that form the basis of the Sybil Strepneumo system are complex, long running and require significant computational resources. For example, the computes for Strepneumo data version 15 required ∼136 CPU hours. Planned integration of these compute pipelines into the CloVR VM will allow users with limited resources to set up Sybil websites with their own data by allowing them to run their computes in a cloud environment with minimal expertise.

4.2 Genomic data

The Sybil system is designed to work with genomic data from any type of organism. In addition to numerous instances involving bacterial genomes, including the public Strepneumo site described here, the system has been employed for the eukaryotic *Aspergillus* database AspDB at www.aspgd.org (Arnaud *et al.*, 2010) to provide comparative genomics capabilities. Internally, the system has been used to compare several protozoan parasites. These capabilities can be as easily leveraged for archaea, other eukaryotes, or viruses.

Whole genome shotgun second- and third-generation sequencing approaches result in large numbers of draft or unfinished genome sequences being used as the final product. Sybil accepts unfinished genome data, but it does not attempt to connect together any separate contigs. This can cause issues when looking for regions of synteny as in the gradient display and the genomic comparative display. Currently, this can be addressed by building pseudomolecules, where contigs are ordered and oriented based on a reference genome and made into a single FASTA entry representing one prediction of the genome. However, this can incorrectly be perceived as the true order of the genome when the information is actually lacking. One way to address this issue is incorporation of whole genome alignment data. Generation of gene clusters based on data from the whole genome aligner Mugsy (Angiuoli and Salzberg, 2011) has already been accomplished and extension of this work to dynamically order previously unordered contigs will give Sybil

the power to better visualize draft genome data (Angiuoli *et al.*, 2011a).

4.3 Optimizations and scalability

As datasets grow, the Sybil system will need to implement more aggressive optimizations to maintain acceptable performance. The Strepneumo Sybil system contains 11 million rows in the feature table with 22 million feature locations and 43 million feature properties. The normalized structure of the Chado database can require numerous table joins, making some queries slow. Optimizations like the ChadoMart and query caching have allowed the system to scale to near 50 bacterial genomes, but other strategies are needed for larger data. A prototype NoSQL document-based database was benchmarked using MongoDB. This method of data storage does away with the relational schema and opts for a schema-less data model where every gene is a conceptual 'document'. Using MongoDB as a primary data store allowed the system to scale to >100 $\sim$5–6 MB bacterial genomes. The bottleneck at this level becomes the visualization strategies used since graphic drawing in Sybil is done on demand. Future iterations of the system will likely require a refactored drawing strategy to provide the flexibility and interactivity that has become standard in the current Sybil system.

5 CONCLUSION

As is typical in bioinformatics, the biological problem defines the needed software. Development of Sybil has been focused on manual data mining efforts, which give the user the power to navigate countable numbers of genomes in search of potentially important biological features. As datasets become larger and less manageable, the comparative genomics tools of the future will need to adapt to provide meaningful information without overwhelming the user. Sybil has addressed this problem by transitioning from a system capable of drawing a handful of genomes to one that is able to handle >100 genomes.

ACKNOWLEDGEMENTS

We thank Etienne de Villiers at ILRI and Trushar Shah at ICRISAT for their efforts in setting up the remote mirrors. We also thank Dave Kemeza and IGS IT for their work maintaining the IGS grid compute, database and web-hosting infrastructure.

Funding: Platform for Appropriate Technology in Health (PATH) Contracts (GAT.0782-01990-COA and 003024).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Angiuoli,S.V. and Salzberg,S.L. (2011) Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, **27**, 334–342.

Angiuoli,S.V. *et al.* (2011a) Improving pan-genome annotation using whole genome multiple alignment. *BMC Bioinformatics*, **12**, 272.

Angiuoli,S.V. *et al.* (2011b) CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics*, **12**, 356.

Arnaud,M.B. *et al.* (2010) The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community. *Nucleic Acids Res.*, **38**, D420–D427.

Bendtsen,J.D. *et al.* (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

Blanchette,M. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.

Crabtree,J. *et al.* (2007) Sybil: methods and software for multiple genome comparison and visualization. *Methods Mol. Biol.*, **408**, 93–108.

Darling,A.C. *et al.* (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Frazer,K.A. *et al.* (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, W273–W279.

Hsiao,W. *et al.* (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, **19**, 418–420.

Jaccard,P. (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudense Sci. Naturel.*, **44**, 223–270.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kolaskar,A.S. and Tongaonkar,P.C. (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett.*, **276**, 172–174.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Krzywinski,M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.

Larsen,J.E. *et al.* (2006) Improved method for predicting linear B-cell epitopes. *Immunome Res.*, **2**, 2.

Mayer,C. (2006–2011) Phobos 3.3.11. Available at http://www.rub.de/spezzoo/cm/cmphobos.htm.

McKay,S.J. *et al.* (2010) Using the Generic Synteny Browser (GBrowse_syn). *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9 12.

Mera,R. *et al.* (2008) Serotype replacement and multiple resistance in *Streptococcus pneumoniae* after the introduction of the conjugate pneumococcal vaccine. *Microb. Drug. Resist.*, **14**, 101–107.

Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.

Nielsen,C.B. *et al.* (2010) Visualizing genomes: techniques and challenges. *Nat. Methods*, **7**, S5–S15.

O'Brien,K.L. *et al.* (2009) Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet*, **374**, 893–902.

Orvis,J. *et al.* (2010) Ergatis: a web interface and scalable software system for bioinformatics workflows. *Bioinformatics*, **26**, 1488–1492.

Pan,X. *et al.* (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, **21**, 3461–3468.

Sette,A. and Rappuoli,R. (2010) Reverse vaccinology: developing vaccines in the era of genomics. *Immunity*, **33**, 530–541.

Smedley,D. *et al.* (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.

Smit,A.F.A. *et al.* (1996) RepeatMasker Open-3.0. Available at http://www.repeatmasker.org.

Stajich,J.E. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Stein,L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Wang,H. *et al.* (2006) SynView: a GBrowse-compatible approach to visualizing comparative genome data. *Bioinformatics*, **22**, 2308–2309.

Yu,N.Y. *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.