AnnotateGenomicRegions: a web application

Heiko Muller✉, Luca Zammataro, Gabriele Bucci
Computational Research, Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia (IIT), Genova, Italy

Motivation and Objectives

A common denominator for all applications of New Generation Sequencing technology is the need to annotate genomic regions of interest. Tools such as Galaxy (Giardine et al., 2005), CisGenome (Ji et al., 2008), or the Bioconductor ChIPpeakAnno package (Zhu et al., 2010) have been published to perform this task. However, using these tools often requires a significant amount of bioinformatics skills and/or downloading and installing dedicated software. A widely accepted, web-based annotation tool available to bioinformaticians and biologists with widely varying skill levels is not available. Here we present AnnotateGenomicRegions, a web application that accepts genomic regions as input and outputs overlapping and/or neighboring genome annotations chosen on a simple web-form.

Genomic data sets are diverse. However, a common denominator of all studies is the possibility to represent the data as a set of genomic regions identified by "chromosome name : start base - end base", followed by some quantitative or qualitative measure characteristic of the data set. This data format is also used by genome browsers to display known genome features and is called browser embedded format (.bed). Therefore, the most straight-forward way of annotating a genomic data set is based on using genomic regions of interest as genome browser queries.

Tools performing this task have been developed in the past. For example, a bioinformatician with programming skills may use the EnsEMBL core API or the Bioconductor ChIPpeakAnno package. Slightly less demanding is the use CisGenome or Galaxy. All of these options require considerable programming skills, the download of dedicated software, or both. A simple web tool that accepts genomic regions as input and outputs annotations in a format ready to be pasted into an Excel sheet is, to the best of our knowledge, not available. Here we address this need by presenting AnnotateGenomicRegions.

AnnotateGenomicRegions is an open-source web application that can be installed on any computer running the Glassfish web server. This might be a personal laptop or an institute's Linux cluster. AnnotateGenomicRegions is available at: http://bioserver.iit.ieo.eu/AnnotateGenomicRegions

Methods

AnnotateGenomicRegions uses a set of simple Java servlets to process the annotation queries and returns the annotations as zipped, tab-delimited tables. It has been developed using Java Enterprise technology on the NetBeans 6.9 Integrated Development Environment and the Glassfish version 3 web server. This choice is motivated by the better scalability and portability of Java Enterprise as opposed to common gateway interface based web applications. AnnotateGenomicRegions is a Sourceforge project and can be downloaded from http://sourceforge.net/projects/annotatelocus/ along with detailed descriptions of input and output formats.

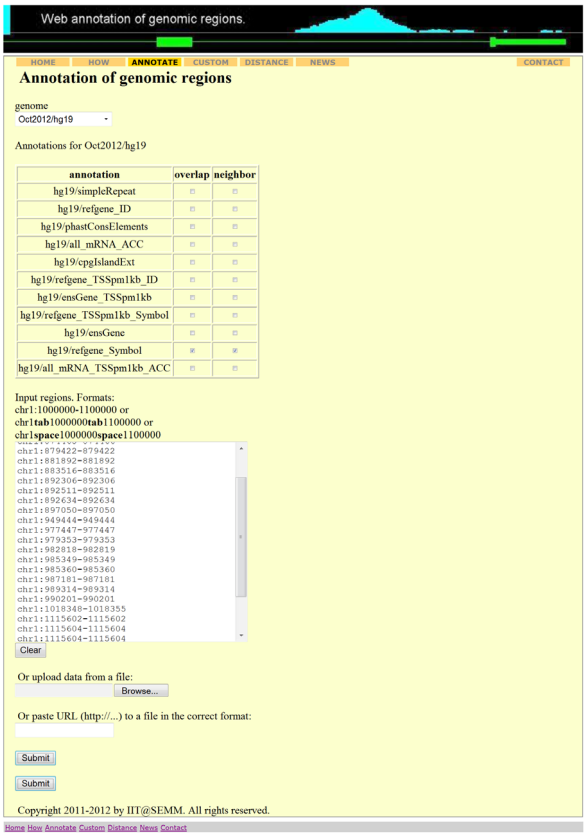
Results and Discussion

The design of AnnotateGenomicRegions is based a few simple requirements:

1. Genomic regions shall be used as input query.
2. The output shall be pastable into an Excel table.
3. The application shall be web-based.
4. No programming skills required to use the application.
5. It must be fast enough to annotate hundreds of thousands of genomic regions within seconds

The steps to be followed by the user to annotate his/her data are: on the "Annotate" pane (Figure 1 A) choose the genome, choose the desired features for annotation and whether the feature shall be overlapping and/or neighboring the query regions, paste or upload the query regions, and finally submit the query. The results of an annotation query are displayed in tabular form (Figure 1 B). The results can be downloaded in zip format and pasted into an Excel spreadsheet.

For non-standard annotations, a "CUSTOM" menu option has been provided. Here, the user



download

region	hg19/refgene_Symbol_ol	hg19/refgene_Symbol_ln	hg19/refgene_Symbol_rn
chr1:69538-69538	OR4F5	FAM138A	LOC729737
chr1:874447-874447	SAMD11	LOC100130417	NOC2L
chr1:874456-874456	SAMD11	LOC100130417	NOC2L
chr1:874465-874466	SAMD11	LOC100130417	NOC2L
chr1:879422-879422	SAMD11	LOC100130417	NOC2L
chr1:881892-881892	NOC2L	SAMD11	KLHL17
chr1:883516-883516	NOC2L	SAMD11	KLHL17
chr1:892306-892306	NOC2L	SAMD11	KLHL17
chr1:892511-892511	NOC2L	SAMD11	KLHL17
chr1:892634-892634	NOC2L	SAMD11	KLHL17
chr1:897050-897050	KLHL17	NOC2L	PLEKHN1
chr1:949444-949444	ISG15	HES4	AGRN
chr1:977447-977447	AGRN	ISG15	RNF223
chr1:979353-979353	AGRN	ISG15	RNF223
chr1:982818-982819	AGRN	ISG15	RNF223
chr1:985349-985349	AGRN	ISG15	RNF223
chr1:985360-985360	AGRN	ISG15	RNF223
chr1:987181-987181	AGRN	ISG15	RNF223
chr1:989314-989314	AGRN	ISG15	RNF223
chr1:990201-990201	AGRN	ISG15	RNF223
chr1:1018348-1018355	C1orf159	RNF223	LOC254099
chr1:1115602-1115602	TTLL10	MIR429	TNFRSF18
chr1:1115604-1115604	TTLL10	MIR429	TNFRSF18
chr1:1115604-1115604	TTLL10	MIR429	TNFRSF18
chr1:1115604-1115604	TTLL10	MIR429	TNFRSF18
chr1:1159233-1159233	SDF4	TNFRSF4	B3GALT6
chr1:1164118-1164118	SDF4	TNFRSF4	B3GALT6
chr1:1192497-1192497	UBE2J2	FAM132A	SCNN1D

Figure 1: Screenshot of AnnotateGenomicRegions. A) Annotation pane. B) output example

can upload an annotation file in bed format along with the queries. The user chooses the number of desired annotation files, browses to the local files with the annotations, specifies the column numbers for chromosome, start, end, and annotation name, and chooses whether overlap or neighbors queries are desired. When submitting the queries, the annotations will be uploaded to the server, processed for fast annotation, and annotations will be provided as a zipped output file. Distances can be calculated using the "DISTANCE" pane. The annotations used for distance calculations must be provided by the user including strand information.

Design criterion 5 regards the speed and the scaling of the application. Without going into too much detail, the core of the application is located in a Java class called Query. This class ensures that both the query regions and the annotations of interest are sorted first by chromosome and then by start position. For each chromosome, a separate Hashtable object is created that holds the query regions sorted by start position in an ArrayList. Similar Hashtables are created for each annotation. Then, auxiliary Hashtables are generated that make sure that querying a chromosomal region in the vicinity of a previous query does not result in searching a region that has already been searched by the previous query, which is guaranteed to have a start position smaller than or equal to the start position of the current query. The Query class performs searches for hundreds of thousands of query regions and tens of annotations in a matter of seconds and the scaling with the number of query regions or the size of annotation files is linear.

ChIP-Seq analysis tools have been developed that comprise functional annotation, for example CisGenome, W-ChIPeaks, Sole-Search, or CASSys (Ji et al., 2008; Blahnik et al., 2010; Lan et al., 2011; Alawi et al., 2011). These tools are focusing on the identification of enriched regions in ChIP-Seq experiments and annotation of genomic regions is provided as a side-aspect. Therefore, using these tools for annotation purposes only is cumbersome. Command-line tools such as BEDtools (Quinlan and Hall, 2010) are extremely powerful at identifying overlapping regions in two bed formatted files. But being command-line tools, they are off-limits for most biologists. The same is true for the BioConductor ChIPpeakAnno package (Zhu, 2010). Tools such as the EnsEMBL Ruby API (Strozzi and Aerts, 2011) require considerable programming skills, which precludes widespread use by biologists.

Galaxy (Giardine et al., 2005) is a sophisticated web-based suite of genome analysis tools

that can also perform annotation of genomic regions as part of the "Operate on Genomic Intervals" menu option. It is an expert tool that requires some familiarity. The option "Fetch closest non-overlapping feature" will find annotations that have been defined as "neighbors" in this work. The file defining the neighbors must be uploaded along with the query regions. No default annotations for neighbor fetching are provided. Only one annotation can be fetched at the time. Identification of overlapping features requires the use of a different menu option ("Intersect"). In contrast to AnnotateGenomicRegions, none of the above mentioned tools can be used easily by non-experts..

Acknowledgements

We thank Dr. Davide Cittaro for helpful discussions on application design and implementation.

References

6. Alawi M, Kurtz S, Beckstette, M (2011) CASSys: an integrated software-system for the interactive analysis of ChIP-seq data. J Integr Bioinform. 8, 155. doi:10.2390/biecoll-jib-2011-155
7. Blahnik KR, Dou L, et al. (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. Nucleic Acids Res. 38, e13. doi: doi:10.1093/nar/gkp1012
8. Giardine B, Riemer C, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 15, 1451-5. doi:10.1101/gr.4086505
9. Ji H, Jiang H, Ma W, Johnson DS, Myers RM et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol. 26, 1293-300. doi:10.1038/nbt.1505
10. Lan X, Bonneville R, et al. (2011) W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. Bioinformatics 27, 428-30. doi:10.1093/bioinformatics/btq669
11. Quinlan AR, Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-2. doi: 10.1093/bioinformatics/btq033
12. Strozzi F, Aerts J (2011) A Ruby API to query the Ensembl database for genomic features. Bioinformatics 27, 1013-4. doi:10.1093/bioinformatics/btr050
13. Zhu LJ, Gazin C, et al. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 11, 237. doi:10.1186/1471-2105-11-237