DE GRUYTER OPEN

JACSM 2014, Vol. 6, No. 2, pp. 101 - 110

10.1515/jacsm-2015-0002

ANNOGENE: RESTFUL WEB SERVICE FOR ANNOTATING GENOMIC FEATURES

Andrzej Tomski[1,2], Marcin Piechota[2], Ryszard Przewłocki[2]

[1]Institute of Mathematics, Jagiellonian University, Kraków, Poland
Andrzej.Tomski@im.uj.edu.pl

[2]Department of Molecular Neuropharmacology, Institute of Pharmacology of the Polish Academy of Sciences, Kraków, Poland
marpiech@if-pan.krakow.pl
nfprzewl@cyf-kr.edu.pl

Abstract

Modern high-throughput sequencing techniques generate a constantly increasing amount of genomic data from eukaryotes. The main problem is quickly identifying the data that may provide information about the nature of intracellular processes, such as the targeting of transcription factor-binding sites. Typically, thousands of peaks or signals are found across the genome and the nearby genes must be annotated. We introduce AnnoGene - a web service for annotating genomic features. AnnoGene was implemented in a representational state transfer (REST) architectural style. The program searches for the gene nearest to the center of a genomic position. Subsequently, the location and annotationsof the gene are shown. The tool can be downloaded and run on a local computer, but it was designed to be a web service. AnnoGene is freely available through a web browser. Moreover, our paper covers examples of the REST clients written in the Python, R and Java programming languages. AnnoGene only requires genomic positions from the user. Even when annotating several thousand positions, the output is typically ready in a few seconds. Moreover, this tool supports Seqinspector – a web tool for finding regulators of the genes.

Key words:BED annotation, RESTful web service, ChIP-seq peaks

1 Introduction

A large amount of data is currently obtained from high-throughput sequencing experiments, such as ChIP-seq [1] or RNA-seq [2]. One of the important challenges is finding the nearest genes for a given set of genomic positions and assigning the gene symbols, such as MGI [3] or HGNC [4] Gene

Symbol and Ensembl ID [5]. Usually, these data are ChIP-seq peaks and they are presented in the Browser Extensible Data format (BED)[6]. In the last few years, many gene annotation tools have been developed, but they still have some disadvantages. These applications should be as simple as possible. Our goal was to deliver a tool in the form of a very simple online web service that can be called from Python, R, Java or any other environment. This tool only requires data entry into a form. Using thehuman (hg19) and mouse (mm9, mm10) genome assemblies, the tool provides a list of the genes that correspond to the queried data.

To meet these demands, we introduce AnnoGene. The idea of such a program is clear: annotation of the genes for most peaks. AnnoGene is a web service designed in representional state transfer (REST) **[7]**. We also provide a few REST clients, written in the Python, R and Java programming languages. Each clientcan be run from an operating system command line. AnnoGene differs from similar programs in the following aspects: AnnoGene has been implemented differently, has an alternative searching criteria, was noticeably faster during tests than the other examined tools and it works with the data, which the user submits to Seqinspector, an online tool to investigate interactions betweenthe proteins. This case will be described in details later. In a situation when many data points are produced from one experiment, an application with a faster execution time can provide an advantage. The final result with hundreds or even thousands of annotations should be obtained rapidly but accurately. Parsing the data or cutting the data into smaller parts should not be required, especially for extraordinarily large data. We tested the accuracy and the query response times of the most popular tools for annotating genes for comparison with AnnoGene. AnnoGene fills the gap between the need to easily annotate thousands of genomic positions and the desire to significantly shorten the long analysis time.

2 Implementation

AnnoGene is a RESTful web service on an Apache server. AnnoGene annotates the genes nearest to the centers of queried genomic positions with Ensembl ID and MGI (HGNC) symbols. In addition, a user sets the desired search accuracy. The entire application was designed in Python (v2.7.1), using the web.py (v0.37) [8] framework. The resource is freely available at http://bedanno.cremag.org through a web browser or through one of the REST clients written in the following three programming languages: Java, Python and R. The urllib and urllib2 libraries [9] for Python and the R Curl (v1.95-4.1) [10] package for R were used. The source code is freely available at http://github.com/andrewtom/AnnoGene. To address the early obstacles

that arose when using the application, we prepared a proper Web Application Description Language (WADL) [11] file.

3 Results

3.1 AnnoGene - web interface

AnnoGene facilitates the search for the genes nearest to the centers of the queried genomic positions and annotates these genes with Ensembl ID and MGI (HGNC) symbol. The interface requires three simple steps (Fig. 1). First, the user enters the data into the 'Data' form. Then, the user specifies an accuracy for the analysis in the 'Accuracy' text box (the accuracy indicates the distance from the center of the region the operation is performed). Finally, the user chooses one of the available genomes. AnnoGene finishes by presenting all of the lines together with their annotations. If the accuracy is limited to a very small number of base nucleotides, there may be no genes sufficiently close to the entered position. However, when no accuracy is specified, the nearest gene will be found.

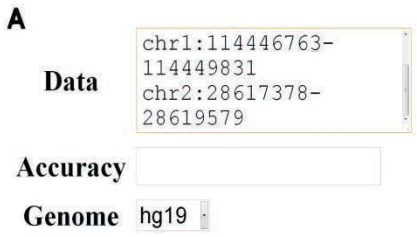

Figure 1. (A) The window that appears when AnnoGene is accessed. There are three inputs: 'Data', 'Accuracy' and 'Genome'. (B) A sample output where all the lines are tab-delimited and contain a query line with the genome assemble, position of the nearest gene, its Ensembl ID and the HGNC symbol

Genomic positions can be entered in two different formats:
- divided by whitespace (i.e. tab delimited): chrName Start End
- the format used in most of genomic databases: chrName:Start-End

3.2 Client Programs

AnnoGene can be accessed through any web browser or through one of the provided clients (Python, R and Java). Below we present the client code listings.

(I) Python code:

```
#loading the libraries

import urllib
import sys
url="http:////bedanno.cremag.org"
pa-
rams=urllib.urlencode({"Data":sys.argv[1],"Accuracy":sys.
argv[2],"Genome":sys.argv[3]})}
if sys.argv[3] in ["mm9","mm10","hg19"]:
   #retrieve a URL containing parameters
response=urllib.urlopen(url,params).read()
 if "negative" in response:
print "Accuracy must be positive!"
  elif "Incorrect" in response:
     print "Wrong match: line 1"
else:
print response
else:
print "Genome not available. Try again!"
```

(II) R code:

```
  args← commandArgs(TRUE)
  #RCurl loading
  library("RCurl")
  #fetch a URL and submit forms
  url ← "http:////bedanno.cremag.org"
  if (args[2] !=" "){
     if (as.integer(args[2])<0){
      cat("Accuracy must be positive!\n");quit(save="no")
  }
  }
  tryCatch({
  result ←postForm(url,Data=args[1],Accuracy=args[2],
  Genome=args[3])
  }
  error=function(e){
cat("Genome not available!$n");quit(save="no")}
  })
  tryCatch({
```

```
cat(rawToChar(result))}
  }
  error=function(e){
cat("Wrong match: line 1 \n")
  })
```

(III) Java code:

```
String bedString = "";
for(BedItem item: bedItems)
bedString += item.toString() + "\n";
 bedString = bedString.trim();
String parameters = "Data=" + bedString +"&Accuracy=" +
 "0" + "&Genome=" + genome;
URL url = new URL("http:////bedanno.cremag.org");
URLConnection conn = url.openConnection();
conn.setDoOutput(true);
OutputStreamWriter writer = new OutputStreamWriter
 (conn.getOutputStream());
 ////write parameters
writer.write(parameters);
 writer.flush();
 ////get the response
BufferedReader reader = new BufferedReader
(new InputStreamReader(conn.getInputStream()));
 String line;
while ((line = reader.readLine()) != null){
    System.out.println(line);
}

writer.close();
reader.close();
```

By default, the three most popular and most frequently used genome assemblies are available: mm9 and mm10 for mouse and hg19 for human.

New assemblies will be added in the future. The results can be useful for determining the expression level of nearby genes when large amounts of data are obtained from high-throughput sequencing experiments.

3.3 Example of the usage

We have examined the usage of AnnoGene for annotation of Nfkb1 ChIP-seq peaks in bone marrow-derived dendritic cells 30 min post LPS stimulation [12]. ChIP-seq peaks were obtained from Gene Expression Omnibus (GEO) under accession number GSM881153. Nfkb1 is a subunit of the NF-kappa B protein complex, which is activated by extracellular stimuli related to inflama-

tion such as cytokines, oxidant-free radicals and bacterial products [13]. The top 100 of peaks in the BED format were submitted to AnnoGene. Afterwards, obtained gene symbols were submitted to DAVID Bioinformatic Resources [14], [15] to identify overrepresented gene ontology terms. The overrepresented terms are associated with immune response, cytokine activity or chemotaxis (see Table 1).

Table 1. Annotation of Nfkb1 ChIP-seq peaks in bone marrow-derived dendritic cells 30 min post LPS stimulation [12], followed by identifying overrepresented gene ontology terms in DAVID [14], [15]

Category	Term	Genes	P-value	Benjamini
GOTERM BPFAT	responsetowounding	11	1.7E-8	1.1E-5
GOTERM BPFAT	immuneresponse	10	1.9E-8	5.9E-6
SPPIR	cytokine	10	2.9E-8	3.2E-6
GOTERM MFFAT	cytokineactivity	8	2.9E-8	2.0E-6
GOTERM BPFAT	defenseresponse	10	1.7E-7	3.5E-5
GOTERM BPFAT	inflammatory response	8	2.7E-7	4.3E-5
INTERPRO	smallchemokine	5	4.8E-7	5.0E-5
GOTERM MFFAT	chemokineactivity	5	6.9E-7	2.3E-5
GOTERM MFFAT	chemokinereceptor	5	7.7E-7	1.75E-5
GOTERM BFFAT	chemotaxis	6	6.6E-6	3.2E-4
GOTERM BPFAT	taxis	6	2.6E-6	3.2E-4

AnnoGene has been integrated with Seqinspector (available at http://seqinspector.cremag.org), a web tool for finding putative regulators of genes and studying protein-protein interactions. A typical application involves obtaining peaks from a ChIP-seq experiment, then calculating the average coverage for all tracks, and performing a two-sample t-test with a comparison to a reference. P-values significant under Bonferroni correction lead to the identification of the associated transcription factors.

3 Discussion

PinkThing [16] calculates the distance to the transcription start sites and the histogram peaks near the 3' and 5' ends of genes. However, the user must prepare a BED file or Wiggle track and include a few extra parameters, such as the 'file description'. The output does not include any information about the coordinates of the approximate gene locations. The default value setting is for the human genome and no mistake is reported when data from other genomes are entered. Moreover, the response time for Pink Thing is long for large datasets (~ 30 minutes for about 100 000 positions). AnnoGene works faster and does not require the preparation of any specific files.

GPAT [17] is a web interface that can label genomic positions with several types of annotations. Along with the necessity of preparing the data in the BED file format, other limitations are imposed. It is impossible to upload any file larger than 10 MB or any longer input than 10 000 lines. Consequently, pasting tab-separated values in the text box fails.

ChIPBase's annotationTool [18] returns the annotations for ChIP-seq regions and a distribution of the distance between the center of the peaks and the transcript start sites (TSS). However, the output information is incomplete: there is no information about the coordinates of the nearest gene and not all of the genes are annotated. In addition, AnnoGene clearly produces results more rapidly.

GREAT [19] predicts some functions of cis-regulatory regions. At the cost of generating a highly detailed overview of the position, the query response time for a small portion of the data is quite long. Moreover, the API requires an upload of a BED file to a URL, and the server can only accept 5 batch queries at a time from all of the users.

PyCogent [20] is an integrated Python framework for analyzing the evolution of biological sequences and for querying databases. PyCogent is mainly meant for use by Python programmers. Nevertheless, the framework still appears to be unfinished - many options, such as switching between different genome data releases, are unavailable.

We compared the average query response times for these programs. For this comparison, we prepared input data of 100, 1000 and 100 000 lines in a manner approved by all of the tools. Afterwards, we connected to the programs through an anonymous proxy server. We observed significant differences, especially for the data of 1 000 or more genomic positions (see Table 2). AnnoGene produces results much faster than most of the similar applications. Additionally, AnnoGene catches typical mistakes, such as swapping the start and end positions of the gene. Entering incorrect positions can lead to an incorrect match or to no match at all.

Table 2. Query response times. A comparison of the query response times for the most popular gene annotation tools. 100, 1 000 and 100 000 peaks were prepared. A publicly available proxy server was used. Each measurement was repeated five times, and the average value was calculated (s – seconds)

Input (lines)	**100**	**1**	**100000**
AnnoGene	0.86s	2.91s	170s
GPAT	6.8s	8.15s	notpossiblewithatext box
ChIPBase	6.1s	10.22s	6minutes
PinkThing	11.7s	31.45s	26minutes
GREAT	13.2s	14.15s	generatesusererror

Early growth response gene 1 (known as EGR1) is a transcription factor that plays an important role in cell apoptosis. There has been interest in EGR1 for several years due to its possible influence on carcinogenesis. EGR1 overexpression was observed in the most common types of prostate tumor cells, whereas EGR1 showed little activity in brain gliomas [21]. However, the direct regulator its expression level remains unknown. Tracing distant peaks linked with EGR1 may help to show other molecular mechanisms that regulate its activity. AnnoGene even works for data points that are far from the beginning of the gene. This function can provide a new point of view on the causes of the diverse behavior of genes in various types of tumors.

3 Conclusions

Many bioinformatics tools are available for analyzing and annotating large numbers of data points from high-throughput sequencing technologies. However, most of the existing implementations are slower than our service, have significant data entry limits and often require the preparation of files specific formats. AnnoGene has been designed to work with large amounts of data. This RESTful web service annotates ChIP-seq peaks or data points from other experiments with acceptable precision. All the genes are annotated with Ensembl ID and MGI (HGNC) symbols. We hope that AnnoGene will play a role in those biological experiments that produce large amounts of data and that require a quick identification of the genes whose expression levels changed.

Competing interests

The authors declare that they have no competing interests.

Author Contributions

AT wrote the paper, designed and implemented AnnoGene, wrote the Python & R REST clients and prepared the WADL file. MP conceived the idea, revised the algorithm, wrote the Java client and provided the experimental data. RP coordinated the study and approbated the manuscript.

Acknowledgments

We thank Michał Korostyński for testing the application.This work was supported by the grants NCN 2011\01\N\NZ204827 and POIG DeMeTer 3.1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation the manuscript.

References

1. Park P.J.,2009,ChIP-seq: advantages andchallengesofamaturing technology,NatRevGenet10, pp. 669-680.
2. WangZ., GersteinM.,SnyderM., 2009, RNA-Seq:arevolutionary toolfortranscriptomics,NatRevGenet10, pp. 57-80.
3. Bult C.J., BlakeJ.,KadinJ.,EppigG.,RingwaldM.,Richardson J.etal., 2007,P4-SThe MouseGenomeInformaticsDatabase: AnIntegratedResourcefor MouseGenetics andGenomics,JBiomolTech.
4. GrayK.A.,Daugherty L.C.,GordonS.M.,SealR.L.,WrightM.W.,BrufordE.A., 2013,ChIP-seq: advantages and challenges ofa maturing technology. Genenames.org: the HGNCresourcesin2013,NucleicAcidsRes41.
5. Hubbard T.,BarkerD.,BirneyE.,CameronG.,ChenY.,ClarkL.etal., 2002,TheEnsembl genomedatabase project,NucleicAcidsRes30(1), pp. 38-41.
6. KentW.J., ZweigA.S.,Barber G.,HinrichsA.S.,Karolchik D., 2010,BigWigand BigBed:enabling browsingoflargedistributeddatasets,Bioinformatics 26(17), pp. 2204-2207.
7. Fielding R.,Taylor R.N., 2002,Principled Design ofModern Web Architecture, ACM TransactionsonInternetTechnology, NewYork: AssociationforComputing Machinery 2(2), pp. 115-150.
8. web.pyv0.37,Availableat: http://webpy.org.,Accessed in2013 and 2014.
9. Python HOWTOFetchInternet Resources Using urllib2, Available at:http://www.w3.org/Submission/wadl.,Accessedin 2013.
10. Package'RCurl'(v1.95-4.1),Generalnetwork(HTTP/FTP/...)clientinterfacefor R.,Availableat: http://www.omegahat.org/RCurl,Accessed in2013 and 2014.
11. Web Application Description Language (WADL) W3C Member Submission 31 August 2009,Availableat: http://www.w3.org/Submission/wadl., Accessed in 2013.
12. Garber M.,YosefM.,GorenA.,Costa A.M.,Raychowdhury R.,WuJ.etal., 2012,Ahigh- throughputchromatinimmunoprecipitationapproach revealsprinciplesofdynamic generegulationinmammals,MolCellSep 10, pp. 669-80.
13. Barnes P.J., KarinM., 2009,Nuclear factor-BA pivotal transcriptionfactor inchronic inflammatorydiseases 10, pp. 669-680.
14. Huang D.W., Sherman B.T., Lempicki R.A., 2009, Systematic and integrativeanalysis oflarge genelists using DAVID Bioinformatic Resources, Nature Protoc Genet 4(1), pp. 44-57.
15. Huang D.W.,Sherman B.T., LempickiR.A., 2009,Bioinformaticsenrichmenttools:paths towardthecomprehensivefunctionalanalysisof largegenelists.NucleicAcidsRes Genet37(1), pp. 1-13.
16. NielsenF.G., Kooyman M., Kensche P., Hendrik M., Stunnenberg H., Huynen M., 2013, ThePinkThing foranalysingChIP profilingdata intheir genomiccontext.,BMC ResearchNotes 6, pp. 133.
17. KrebsA.,Frontini M.,Tora L., 2008,GPAT: Retrieval ofgenomicannotationfromlarge genomicpositiondatasets,BMCBioinformatics 9, pp. 533.
18. YangJ.H.,LiJ.H.,JiangS.,ZhouH.,QuL.H., 2013,ChIPBase: Adatabase fordecoding thetranscriptionalregulation of longnon-codingRNAandmicroRNAgenesfrom ChIP-Seqdata,NucleicAcidsRes 41, pp. 177-187.

19. McLean C.Y., Bristor D.,HillerM., Clarke S., Schaar B.T., Lowe, C.B.et al., 2010,GREATimprovesfunctionalinterpreationofcis-regulatoryregions,NatBiotechnol 28(5), pp. 495-501.
20. KnightR.,MaxwellP.,Birmingham A.,Carnes J.,Caporaso J.G., Easton B.C.,etal., 2007, PyCogent:atoolkitformakingsense fromthesequence.GenomeBiol 10, pp. 669-680.
21. YangS.Z,Abdulkadir SA., 2010,Earlygrowthresponsegene1modulatesandrogenreceptorsignalinginprostate carcinomacells.JBiolChem 10(39), pp. 906-911.