31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Visual Annotation of the Gene Database

Jing Wen[1], Xishu Wang[1], Warren Kibbe[2], Simon Lin[2], Hui Lu[1]*

Abstract— **The genes in NCBI databases are currently annotated with itemized text (Gene Reference Into Function, or GeneRIF). A previous work suggests that the visual presentation can be more effective when time and space are under heavy constraints. Here we report a novel annotation of the genome information using Web 2.0 technologies: GeneGIF (Gene Graphics Into Function). The users can quickly scan through important functions of each gene from a graph, and then go to detailed pages when they find interesting annotations. The modular implementation makes it easily pluggable into other widely used databases without reprogramming. Similar approaches are being developed to incorporate information to other types of genomics and proteomics databases.**

I. INTRODUCTION

GeneRIF [1] (Gene Reference Into Function) provides a simple way to gene functional annotation: each GeneRIF is a textual statement up to 255 characters to document the function of a gene. One can scan the functions of a gene through each GeneRIF quickly. However, when a gene has dozens or even hundreds of GeneRIFs, it will be time-consuming to go through all of them. The situation is getting worse when users need to get some ideas of a long list of genes, such as over-expressed genes from microarray experiments. The users will get lost in the information and miss the message of interests. To get a general idea of what functions a gene has, a faster and more intuitive way is in demand.

The first observation is that for genes with many GeneRIFs, there are quite some overlaps among each of them. Thus if we emphasis the keywords it will be easier for users to grasp the functions. One solution to this problem is the so called in-line html tag cloud which appeared firstly in Douglas Coupland's book[2] in 1995. Tag cloud drawing[3] has become a popular way to display data with its frequency used in many website such as New York Times[4] (`http://www.nytimes.com/gst/mostsearched.html?format=tagcloud&period=1`) and Flickr[5]. The advantage of this method is that it delivers the important information by a bunch of key words according to its popularity and avoid going through several paragraphs of sentences. Usually, in this method, to check what a key word is referring to, one has to navigate to a new page. However, one needs to go back to the tag cloud homepage to check the details of another word. A recent solution is to use a nice and convenient word clouds graph called Wordle[6] (Fig.1) which is developed by Jonathan Feinberg. The graph is also generated base on the word frequency. The font size is proportional to its word counts. Compared with in-line html tag, Wordle-like graphs utilize space more effectively by meshing up the words. To enhance readability, colors and directions are addede to each word.

Fig. 1. Word clouds for gene annotation of KLF4 using Wordle

To compare different representations of gene annotation, a survey[7] was conducted among experts who work on genomic data analysis using microarrays. In this survey, a graph (similar to Fig.1) is generated by Wordle based on GeneRIF of gene KLF4 followed by 8 questions related with usage of GeneGIF and participant characteristics (such as gender, age, education level, study field and native language). 53 valid responses were collected in the end. The result[7] showed that in terms of usage, 64% of the users were either positive or neutral toward using GeneGIF in their daily work; in terms of preference, 51% of the users preferred visual (GeneGIF) information than textual (GeneRIF) information.

Some of the participants gave very useful comments on both the advantages and drawbacks of these two types of gene annotation methods: GeneRIF vs GeneGIF. Most participants think GeneGIF is more convenient when one needs to check functions of many genes at a time or get an outline of the functions. It gives a quick idea of the functional annotation for a gene. But traditional GeneRIF provides a more precise description of gene function, it is necessary when one needs to study the gene function in detail. Many of the participants suggested to make GeneGIF clickable so that the GeneRIF could be displayed when needed.

In this work, we are presenting an upgrade version of GeneGIF which combines Tag cloud, Wordle-like graphic as a pluggable web 2.0 component to biologists.

1. Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607
2. The Biomedical Informatics Center and The Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL 60611
*Hui Lu, corresponding author (`huilu@uic.edu`)

II. SYSTEM AND METHODS

A. Data Set:

The genes used in this project cover the entire human genome whose Taxonomy ID is 9606 in the NCBI[8] database. By March 2009, there are 40765 human genes and 24493 GeneRIFs. 12101 genes have at least one GeneRIF related to it, so we implemented our graphic annotation to these genes.

B. Text Mining:

Before generating the graphs, first we need to pre-process the sentences in GeneRIFs and derive a word frequency list from the raw GeneRIFs. In this stage, some meaningless words (stop words) will be removed while phrases and terms related to genomic data will be identified.

1) Stopwords removing: After carefully checking the graph in Fig. 1, we found some words in the graph with high frequency but useless in terms of gene function. We divided them into three domains: common English stopwords (eg. ”the”, ”of”), biology domain-specific stopwords (eg. ”active”, ”protein”) and experts suggested stopwords (eg. ”paper”, ”review”). (table 1 in [7]).These words should be filtered before we generate our graphs. Also the self-referring words, such as the gene symbol (”KLF4” in Fig.1) and the gene name are removed from the graph.

2) Morphological unification: We unify the different forms of nouns and verbs into their prototype. For example, the word ”cells” is considered the same as its singular form ”cell”. Also ”regulates”, ”regulated”, ”regulating” and ”regulation” will be counted into the frequency of ”regulate”. Porter Stemming Algorithm[9] is used to implement this function.

3) Phrases recognition: In gene annotation, sometimes, phrases contain much more information than single words and should be parsed as a whole term. For example, ”cell cycle” is recognized as one term, because neither the individual word ”cell” nor ”cycle” can express its meaning. And for the phase recognition part, we are going to use the Gene Ontology Terms[10] as our glossary to define the meaningful phrases.

C. Graph Generating

Given the processed word frequency list, we can generate a colorful annotation graph using Thomas Boutell’s open source GD-library [11].

With GD-library, it is very easy to generate the bounding box (the smallest rectangle containing the word) of the word with arbitrary font, color, size and rotation.

The font size is based on our text mining result, proportional to its frequency and scaled for best viewing. The position and rotation of each word is currently set to random which means we randomly throw the word on to the screen until it fits. See Fig.3 for an example.

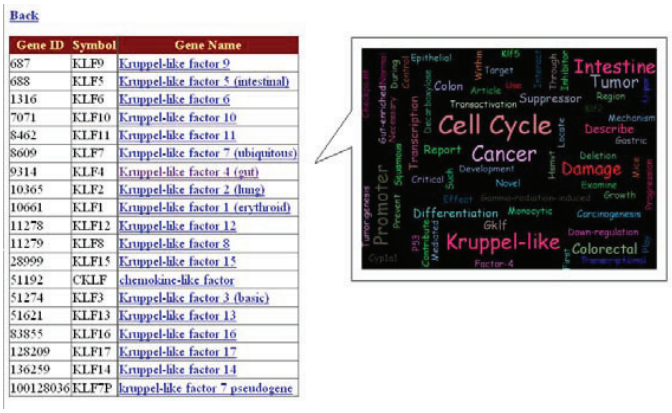

Fig. 2. Search result page with thumbnail of GeneGIF when mouse over a gene name

D. Advanced Web Features

Based on our graph, we implement advanced features to enhance the usability. After user searching for the specific gene according to its ID or symbol, in the list of search result, an thumbnail of the GeneGIF (Fig.2) will display when mouse moves over the gene name. The user can look through each image to get a quick idea of genes’ function. If one is interested in a specific gene, the full size annotation graph will be presented after clicking the gene name (Fig.3). On this full size graph, when user clicks on each word, a panel of all the GeneRIFs containing this word will be displayed with key words highlighted. The panel can be dragged anywhere on the page. With this implementation, the detailed information of GeneRIF is included in the new version of GeneGIF.

E. Conclusions

In this work we have implemented an improved version of GeneGIF for gene annotation which used of tag cloud and Wordle-like graphic techniques. To our knowledge, this is the first report of annotating the functions of each gene in the human genome visually. In the mean time, we employ new web tools to make the website more user-friendly. The project can be accessed here http://proteomics.bioengr.uic.edu/genegif/. This graphic annotation can be easily embedded in other database and resources.

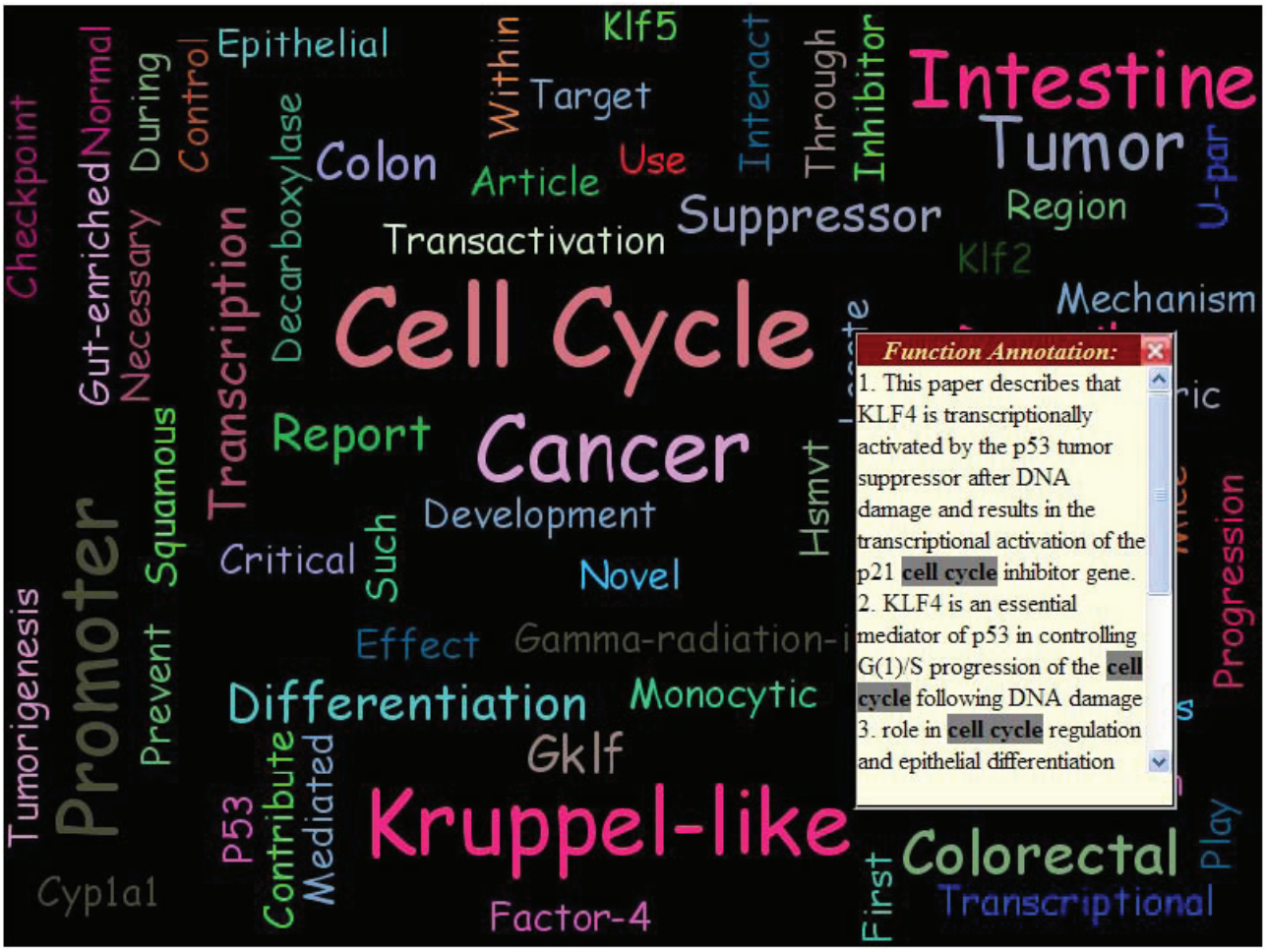

Fig. 3. The GeneGIF of gene "KLF4" and the related GeneRIFs when clicking a specific word. The words related to the important gene functions like "Cell Cycle", "Cancer" etc. stand out in the graph.

III. ACKNOWLEDGMENTS

The authors would like to thank Jairav Desai for the proof-of-concept implementation and Rhett Sutphin at NUCATS for the discussion of Wordle. The authors would also like to thank Martin Wattenberg, Matthew M Mckeon, and Jonathan Feinberg at IBM Research for helpful discussion and comments on this manuscript. This project was supported in part by Award Number UL1RR025741 from the National Center for Research Resources. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health.

REFERENCES

[1] http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html
[2] D. Coupland. Microserfs. [Book], 1995.
[3] Owen Kaser and Daniel Lemire, Tag-Cloud Drawing: Algorithms for Cloud Visualization, WWW2007
[4] http://www.nytimes.com/
[5] Yahho! Inc.Flickr, 2007
[6] http://www.wordle.net/
[7] Jairav Desai et.al. Visual versus Textual Preferences in Presenting Gene Annotation information
[8] http://www.ncbi.nlm.nih.gov/
[9] C.J. van Rijsbergen, S.E. Robertson and M.F. Porter, 1980. New models in probabilistic information retrieval.
[10] Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet 25(1): 25-9
[11] Thomas Boutell, http://www.libgd.org