BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2534–2536
doi:10.1093/bioinformatics/btu241

Data and text mining Advance Access publication April 23, 2014

GenCLiP 2.0: a web server for functional clustering of genes and construction of molecular networks based on free terms

Jia-Hong Wang[1,†], Ling-Feng Zhao[2,†], Pei Lin[1], Xiao-Rong Su[1], Shi-Jun Chen[1], Li-Qiang Huang[3], Hua-Feng Wang[4], Hai Zhang[5], Zhen-Fu Hu[6,*], Kai-Tai Yao[1,*] and Zhong-Xi Huang[1,*]

[1]Cancer Institute, [2]Key Laboratory of Zebrafish Modeling and Drug Screening for Human Diseases of Guangdong Higher Education Institutes, Department of Cell Biology, Southern Medical University, Guangzhou 510515, [3]Guangzhou Biotechnology Center, Guangzhou, 510630, [4]School of Basic Medical Sciences, [5]Network Center and [6]Department of Plastic Surgery, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China.

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Identifying biological functions and molecular networks in a gene list and how the genes may relate to various topics is of considerable value to biomedical researchers. Here, we present a web-based text-mining server, GenCLiP 2.0, which can analyze human genes with enriched keywords and molecular interactions. Compared with other similar tools, GenCLiP 2.0 offers two unique features: (i) analysis of gene functions with free terms (i.e. any terms in the literature) generated by literature mining or provided by the user and (ii) accurate identification and integration of comprehensive molecular interactions from Medline abstracts, to construct molecular networks and subnetworks related to the free terms.

Availability and implementation: http://ci.smu.edu.cn.

Contact: zxhuang@smu.edu.cn

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

Received on January 27, 2014; revised on April 3, 2014; accepted on April 17, 2014

1 INTRODUCTION

Given a set of genes, for example, from high-throughput experiments, it can be helpful to know which biological functions and molecular networks may be involved, or whether genes from a given list or all human genes are related to certain topics, such as various biological and pathological processes. Some pre-defined annotation databases, such as Gene Ontology (GO), or pathway databases, such as KEGG, or Protein–Protein interaction (PPI) databases, such as HPRD (Keshava Prasad *et al*., 2009) and IntAct (Aranda *et al*., 2010), can be used as a gold-standard description. Some annotation tools that integrate these manually curated databases, such as DAVID (Huang da *et al*., 2009) and EGAN (Paquette and Tokuyasu 2010), provide convenient and practical application. However, owing to structured vocabularies and manual curations, pre-defined annotations are inevitably limited in scope, quantity and flexibility.

Some text-mining tools can compensate for these deficiencies. Martini (Soldatos *et al*., 2010) and CoPub 5.0 (Fleuren *et al*., 2011) adopted a keyword-based approach to annotate gene function; however, the keywords were still limited in a pre-defined thesauri. iHOP (Hoffmann and Valencia, 2005) and STRING (Franceschini *et al*., 2013) generate gene networks based on genes co-occurrence in the literature. However, even though gene pairs co-occur in the same sentences, only 30% of pairs have an actual interaction (Cohen *et al*., 2008). FACTA+ (Tsuruoka *et al*., 2011) finds hidden associations between concepts and extracts genes related to a topic, but it does not search phase or select genes from an input list. Previously, we developed stand-alone software called GenCLiP (Huang *et al*., 2008) that annotated gene functions with free terms and generated gene co-occurrence networks related to free terms. However, GenCLiP had three distinct disadvantages: (i) the free terms were limited to single words, (ii) the gene network was constructed based on genes co-mentioned in the same abstracts, often leading to high false positives and (iii) the analysis period generally took 2 weeks, and most of the time was spent on literature download. Thus, we have developed GenCLiP 2.0, a web server that inherits the advantages of GenCLiP, and extends it by incorporating five new features: (i) good performance in gene recognition, with F-measure rising from 0.72 to 0.828, (ii) expansion of free terms to phrases, (iii) molecular interaction extraction accuracy of nearly 90%, (iv) search genes related to free terms and (v) complete analysis in minutes.

2 METHODS

For details about methods, please see Supplementary Data S1.

2.1 Retrieving gene-related literature

The human gene thesaurus was compiled from the HUGO Nomenclature Committee database and the Entrez Gene. We used dictionary-based and rule-based approaches to identify gene names in Medline abstracts. Furthermore, we recognized genes in sentences and built indexes of words and phrases with corresponding genes, sentences and abstracts, to support the search of genes related to any topic.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

2.2 Gene annotation with keywords

As in GenCLiP (Huang *et al*., 2008), terms (including a single word, GO terms and phrases followed with acronyms) that appeared frequently in certain gene-related literatures were considered as keywords for these genes. A fuzzy cluster algorithm (Huang da *et al*., 2009) was used to group statistically overrepresented keywords to annotate input genes. A user can add or remove keywords manually. To explore gene–gene and term–term relationships graphically, a heat map can be created after average linkage hierarchical clustering analysis.

2.3 Construction of gene network

For *de novo* extraction of molecular interaction, a rule-based approach that considered words surrounding gene names and interaction words, and distance between two genes or between interaction word and gene, etc., was used to search sentences. Gene pairs from four manually curated databases [HPRD, BioGRID (Stark *et al*., 2006), CORUM (Ruepp *et al*., 2008) and IntAct] that were co-mentioned in sentences were additionally considered as molecular interactions. The interactive gene network was implemented through a customized Cytoscape Web (Lopes *et al*., 2010). Meanwhile, a subnetwork can be constructed based on the free terms specified by a user. When the free terms appear in the sentence (or an abstract containing the sentence) of a gene pair, the connection will be created. Simultaneously, random simulation was performed to determine whether a gene network was specific for the input genes.

3 RESULTS

3.1 Gene name recognition

Our gene recognition procedure achieved an F-measure of 82.8% (recall: 83.8%, precision: 81.8%) on BioCreative II (GN) test set, which compared favorably with other tested methods (Morgan *et al*., 2008). Moreover, we evaluated our procedure on the test set of iHOP (Hoffmann and Valencia, 2005). The F-measure was 0.86, which was better than iHOP. From whole Medline abstracts, we identified 19 764 genes that occurred in ~3 540 000 abstracts and 13 370 000 sentences.

3.2 Recognition of keyword and molecular interaction

We identified 16 448 keywords for 19 691 of 19 764 genes, where 3395 keywords were phrases with an acronym and 2053 were GO terms. The *de novo* approach recognized 10 545 genes forming 76 437 pairs of molecular interactions, where 62 806 pairs were not collected by the four PPI databases. In our manually defined and other test sets, the precision of molecular interactions was nearly 90%. Details and comparison with other tools is available in Supplementary Data S1. After integrating the four databases, molecular interactions increased to 94 058 pairs, which appeared in ~2 440 000 sentences and 960 000 abstracts.

3.3 Application

GenCLiP 2.0 has three modules for text mining: ‘Gene Cluster With Literature Profiles’, ‘Literature Mining Gene Networks’ and ‘Word Related Gene Search’. As an example, we took 65 upregulated and 53 downregulated genes (Sample 2 on the main page) of keloid to compare with hypertrophic scar to illustrate the application of our server. Unlike hypertrophic scars, keloids are disfiguring scars that extend beyond the original wound borders and resist treatment.

In our analysis, enriched keywords were mostly related with cell growth, extracellular matrix, epithelial mesenchymal transition (EMT), cell migration, cell adhesion, mesenchymal stem cell and wound healing. ‘Collagen’ was manually input as a search term, and found that 10 upregulated genes (P = 7.366e-11) were closely associated with collagen. These keywords are mostly concordant with well-known characteristics of keloid. Interestingly, keratnocyte and keratinocyte differentiation were also annotated as keywords, and associated genes were all downregulated genes except for one (Supplementary Fig. S2A). This reminded us that we should pay more attention to keratnocyte. Recent study has also shown the important role of keloid keratinocytes in keloid scarring, and it was also reported that there were a substantial number of upregulated genes involved in EMT in keloid keratinocytes (Hahn *et al*., 2013). In the ‘Word Related Gene Search’ module, we obtained 28 of the 118 input genes that co-occurred with ‘epithelial mesenchymal transition’ in sentences. We confirmed that 12 upregulated and 5 downregulated genes were related to EMT. In our experience, in general, at least 50% of the genes will be related to search terms. Resulting gene networks (Supplementary Fig. S2B) showed that upregulated MMP2 played an important role in the network. Interestingly, THBS2, CST3 and GLB1, as activators of MMP2, were upregulated, whereas three inhibitors, IL1RN, S100A8 and S100A9, were downregulated. Most of these genes had not been investigated in keloid; however, keywords and the related gene search provided strong evidence that they were closely associated with extracellular matrix, EMT, cell migration and cell growth. Consequently, we proposed that abnormal expression of these genes can cause upregulation of MMP2 and may impact keloid progress.

4 CONCLUSIONS

GenCLiP 2.0 is a web-based tool that can analyze human genes through three functions: (i) generation of enriched and clustered keywords, which are generated based on occurrence frequencies of free terms in gene-related literature or provided by a user, (ii) construction of a gene-network using accurate molecular interactions and generation of subnetworks based on user-defined query terms and (iii) querying of genes cooccurring with search terms in a sentence or abstract. Our testing showed that GenCLiP 2.0 is a practical tool for the analysis of high-throughput experimental results. The databases will be updated every 6 months.

Funding: Guangzhou Condition Construction Projects (2009-7-24) and by National High Technology Research and Development Program of China (Program 863) (grant numbers: 2012AA02A501).

Conflict of Interest: none declared.

REFERENCES

Aranda,B. *et al*. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res*., **38**, D525–D531.

Cohen,K.B. *et al*. (2008) Nominalization and alternations in biomedical language. *PLoS One*, **3**, e3158.

Fleuren,W.W. *et al*. (2011) CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic Acids Res*., **39**, W450–W454.

Franceschini,A. *et al*. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*., **41**, D808–D815.

Hahn,J.M. *et al*. (2013) Keloid-derived keratinocytes exhibit an abnormal gene expression profile consistent with a distinct causal role in keloid pathology. *Wound Repair Regen*., **21**, 530–544.

Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21** (**Suppl. 2**), ii252–ii258.

Huang da,W. *et al*. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc*., **4**, 44–57.

Huang,Z.X. *et al*. (2008) GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. *BMC Bioinformatics*, **9**, 308.

Keshava Prasad,T.S. *et al*. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res*., **37**, D767–D772.

Lopes,C.T. *et al*. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.

Morgan,A.A. *et al*. (2008) Overview of BioCreative II gene normalization. *Genome Biol*., **9** (**Suppl. 2**), S3.

Paquette,J. and Tokuyasu,T. (2010) EGAN: exploratory gene association networks. *Bioinformatics*, **26**, 285–286.

Ruepp,A. *et al*. (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*., **36**, D646–D650.

Soldatos,T.G. *et al*. (2010) Martini: using literature keywords to compare gene sets. *Nucleic Acids Res*., **38**, 26–38.

Stark,C. *et al*. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*., **34**, D535–D539.

Tsuruoka,Y. *et al*. (2011) Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics*, **27**, i111–i119.