Journal of
Data Mining in Genomics & Proteomics

Jere et al., J Data Mining Genomics Proteomics 2014, 6:1
DOI: 10.4172/2153-0602.1000164

Research article Open Access

BioGyan: A Tool to Identify Gene Functions from Literature

Shiva Kumar, Vijaykumar Ghadage, Indhupriya Subramanian, Aarti Desai, Vivek K Singh and Abhay Jere*

LABS, Persistent Systems Limited, Aryabhata – Pingala, 9A/12 Erandwane, Pune-411004, India

Abstract

Background: The primary objective of life science research is to understand complex cellular mechanisms and the interplay of various genes/proteins in multiple cellular processes. For this, PubMed is still the primary source of biomedical information even though multiple other databases such as UniProt, Protein Data Bank (PDB) and Reactome exist.

Objective: With the available large volume data from high-throughput technologies and multiple databases, finding relevant information for gene-process-phenotype has now become extremely challenging and tedious. No tool is currently available to simultaneously search PubMed and multiple other databases to get holistic information. Moreover, a typical PubMed search returns large number of articles, which need to be manually screened for identifying relevant literature. Hence, we developed BioGyan, a literature mining tool to simplify the combinatorial search for genes, cell-types and cellular processes in PubMed and other relevant databases.

Methods: BioGyan uses a robust scoring method to rank articles relevant to user search terms. The scoring method is based on the weighted sum of co-occurrence of gene, process and interactions terms in an abstract.

Results: BioGyan retrieves PubMed articles supporting association between queried genes and processes, relevant pathways from pathway databases and 3-dimensional structures from PDB. For easy viewing, all information to the user is available in single window. BioGyan showed an accuracy of 85.46% in predicting relevance of articles to a gene-process association, and performed better than PESCADOR.

Conclusion: BioGyan has several key features such as batch query of genes as well as processes, offline reading of articles, export of list of articles as bibliography and flexibility for user to revise the article relevance, making it a vital tool for literature search. Thus, BioGyan is a unique tool that offers holistic search across multiple databases while greatly automating the entire process.

Keywords: Protein Data Bank; BioGyan; Genes; Proteins

Introduction

Biologists and life science researchers are primarily interested in understanding the complex cellular mechanism and the interplay of these mechanisms at cellular, tissue and organ level. The underlying quest is to decipher the correlation between genotype-phenotype and a disease state. Currently, with availability of data from high-throughput technologies and multiple information databases dealing with genotypic and phenotypic information, it has become extremely challenging and time-consuming for any researcher to find relevant information and make sense from large volume of available data. This becomes critical while designing high-throughput experiments and interpreting their results, more so when one considers that genotypic-phenotypic correlations alter with cell or tissue types or organism. Hence, experimental data in one cell, tissue or organism may not be directly extrapolated to other cell, tissue types or organisms.

NCBI PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) is one of the most widely used databases, and is generally considered as the primary source for biomedical literature with about 23 million citations. Usually, a typical PubMed search returns large number of articles and reading through all the retrieved articles is very time-consuming and may not be an efficient way for accessing the required information. To ease PubMed search, some tools which mine PubMed to either rank articles [1], cluster the articles [2], or enrich the results [3] are already available. Also, Signaling Gateway Molecule Pages (SGMP) database provides information on the functional state of proteins and the biological processes associated with each state [4]. However, to our knowledge, no tool exist that can help mine PubMed based on genotypic-biological process (phenotypes) correlations directly, especially in a particular cell-type, or tissue or organism. Apart from PubMed, biologist routinely use databases such as Gene Ontology (GO) [5], BioCyc [6], KEGG [7] and Reactome [8,9] to identify gene-process associations. However, these databases suffer from one or both of the following limitations: (1) Not all the gene to process associations are supplemented by relevant literature evidence; and (2) Information regarding relevance of gene to process association for a particular cell-type is missing [10,11].

To address these current multiple limitations and to offer a unique integrated gateway, we developed a tool 'BioGyan' (*Bio*: Biology and *Gyan*: Knowledge) (http://biogyan.com/). BioGyan is a unique single window search tool that mines multiple databases including PubMed, PDB, GO and Reactome simultaneously. It uniquely allows searching directly by gene-process association and helps retrieve multidimensional information which includes relevant articles, associated pathways, processes and 3-dimensional (3D) structures. BioGyan greatly automates literature and database searches and their interpretation by: (1) supporting combinatorial queries of list of genes and processes; and (2) ranking research articles as per 'relevance' to

***Corresponding author:** Abhay Jere, LABS, Persistent Systems Limited, Aryabhata – Pingala, 9A/12 Erandwane, Pune-411004, India, Tel: +91 (20) 670-34562; E-mail: abhay_jere@persistent.co.in

Received November 4, 2014; **Accepted** November 28, 2014; **Published** December 07, 2014

Citation: Kumar S, Ghadage Vk, Subramanian I, Desai A, Singh VK, et al. (2014) BioGyan: A Tool to Identify Gene Functions from Literature. J Data Mining Genomics Proteomics 6: 164. doi:10.4172/2153-0602.1000164

queried gene to process associations. The relevance scoring and ranking of articles is done on the basis of *in-house* scoring algorithm, which uses text-mining and heuristic rules. When tested on a set of 6889 unique articles from PubMed, BioGyan accurately identified 85.46% articles for their relevance to queried genes-process associations. Furthermore, for any search query, BioGyan retrieves data from multiple databases and stores in XML format on the system, thus allowing user to work even in offline mode.

In summary, we believe that BioGyan is a unique platform which not only offers holistic search across multiple databases but also helps substantially reduce the time required to find the relevant biological information by automating the entire process.

Materials and Methods

The methodologies, components and computations employed in BioGyan are schematically represented in Figure 1.

Technologies

BioGyan is a Java based desktop application with user-interface developed using Swing. The 3D structure of proteins is displayed using Jmol [12] and the charts using JFreeChart (http://www.jfree.org/jfreechart/), a Java chart library. Data from NCBI database and UniProt [13] are fetched using E-utilities and REST API (http://www.uniprot.org/faq/28) respectively.

Fetching and collation of data

BioGyan has two search options: BioGyan search and PubMed search. In case of BioGyan search, PubMed, NCBI Gene database, UniProt [13], pathway databases and PDB [14] are mined for information related to the user query. NCBI Gene database and UniProt [13] are queried for gene symbols and aliases. Pathway databases queried are NCBI BioSystems [15] which has data from BioCyc [6], GO [5], KEGG [7], Pathway Interaction Database [16], Reactome [8,9] and WikiPathways [17]. BioGyan search accepts standard NCBI gene names in the query. In case of PubMed search, only PubMed is queried for the keywords. For this search, the results are identical to routine PubMed search.

In both, BioGyan as well as PubMed search: (1) Sentences with search terms are highlighted and this might help in expediting, albeit manually, the screening process for abstracts; and (2) The data and articles fetched against a query are combined and stored as a single XML file. This gives user an advantage to work even in offline mode or share the results with others.

Process list and connection terms

As stated earlier, the primary function of BioGyan is to retrieve articles highlighting association between genes and processes/phenotypes specified by the user. To enable quick search, BioGyan has included a pre-compiled list of 202 biological processes (e.g. Differentiation) obtained from KEGG, GO and literature (Table 1 in Appendix 1). The pre-compiled list includes generic as well as cell specific processes.

Moreover, the pre-compiled list of processes is stored as keywords and used for finding all verb forms of a process in the abstract. For example, the process "Differentiation" is searched using keyword: differentiat*. This ensures that BioGyan captures words such as

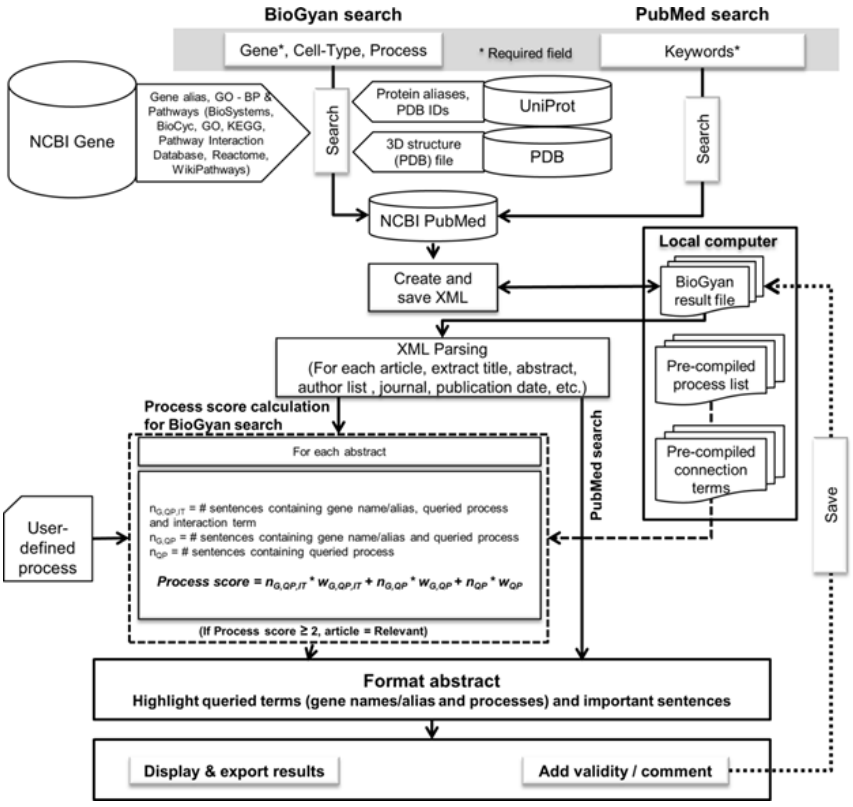

Figure 1: BioGyan: Application work-flow. BP: Biological Process; 3D: 3-dimensional

Gene	TP	TN	FP	FN	# Total abstracts
AQP3	145	233	10	10	398
ARNT	178	1088	85	31	1382
FADD	855	5106	562	134	6657
HDAC1	478	2125	221	152	2976
IRF3	155	1123	109	58	1445
PERP	62	201	21	2	286
TP63	84	1074	30	770	1958
Total	1957	10950	1038	1157	15102
Sensitivity	[1957/(1957 +1157)]*100 = 62.84%				
Specificity	[10950/(10950+1038)]*100 = 91.34%				
Accuracy	[(1957+10950)/(1957+10950+1038+1157)] *100 = 85.46%				

TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative

Table 1: Prediction performance of BioGyan

differentiation, differentiated and differentiating. For certain processes such as "Cornification", multiple keywords "cornificat", "keratini" and "cornified" are used.

BioGyan also uses a set of interaction terms defined as words/phrases that provide evidence for a relationship between two entities (e.g. induces, activates, etc.). The interaction terms searched for in the abstracts are a combination of those obtained from http://www2.informatik.hu-berlin.de/~hakenber/corpora/interactors.html [18] and manual reading of research articles (Table 2 in Appendix 1). A total of 126 keywords corresponding to these interaction terms are currently available in BioGyan.

Scoring of research articles

Ranking of articles for their relevance to the search terms will certainly facilitate rapid screening of literature. For predicting the relevance of articles retrieved from PubMed for a particular query, we have devised a unique scoring mechanism called Process Score.

The Process Score is calculated as represented in Equation 1

$$\text{Process Score} = n_{G,QP,IT} * w_{G,QP,IT} + n_{G,QP} * w_{G,QP} + n_{QP} * w_{QP}$$

… Equation 1

where,

$n_{G,QP,IT}$ = number of sentences in an abstract containing at least one gene name or its alias, one of the queried processes and one or more interaction term

$w_{G,QP,IT}$ = 2; weight assigned to each sentence in the abstract that has at least one gene name or its alias, one of the queried processes and one or more interaction term in the sentence. Sentences containing the search as well as interaction terms provide strong support for the gene-process association and hence are given high weightage.

$n_{G,QP}$ = number of sentences in an abstract containing at least one gene name or its alias and one of the queried processes, but no interaction term

$w_{G,QP}$ = 1; weight assigned to each sentence in the abstract that has one gene name or its alias and one of the queried processes, but no interaction term. Sentences containing the search terms, but not the interaction terms may still be supporting the gene-process association, albeit with less certainty and hence are given lower weightage.

n_{QP} = number of sentences containing at least one of the queried processes, but no gene name or its alias and interaction term

w_{QP} = 0.2; weight assigned to each sentence containing at least one of the queried processes, but no gene name or its alias and interaction term. w_{QP} was assigned a weight of 0.2, as a higher weight resulted in many irrelevant articles being picked up as relevant (Table 3 in Appendix 1). During the computation of Process Score, the gene nomenclatures from NCBI Gene database are used for screening the abstracts.

A Process Score threshold is used to segregate 'relevant' articles (Process Score≥ 2.0) from the 'irrelevant' ones (Process Score<2.0). A Process Score threshold of 2.0 was chosen because during the optimization stage, Process Score threshold of ≥ 2.0 indicated an article relevant to the queried terms and showed best prediction performance (Table 4 in Appendix 1). Process Score threshold lower than 2.0 led to higher number of false positives (i.e. irrelevant articles predicted as relevant), while threshold higher than 2.0 missed many true positives (Table 4 in Appendix 1).

Abstracts that do not contain any of the queried/pre-compiled processes, but have the gene information are grouped under 'Miscellaneous' and the scoring system is not applied on them.

Prediction performance

BioGyan was evaluated for sensitivity, specificity and accuracy as described in Equations 2, 3 and 4 respectively. In the equations, True Positive (TP) is defined as the number of relevant abstracts correctly predicted as relevant, False Positive (FP) is number of irrelevant abstracts wrongly predicted as relevant, True Negative (TN) is number of irrelevant abstracts correctly predicted as irrelevant and False Negative (FN) is number of relevant abstracts wrongly predicted as irrelevant.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$ … Equation 2

$$\text{Specificity} = \frac{TN}{TN + FP}$$ … Equation 3

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad \ldots \text{Equation 4}$$

Results and Discussion

Features and user-interface

BioGyan is a literature-mining tool specially designed for biologists to search, annotate and rank scientific literature from PubMed, and other important biological databases (Figure 2). It is freely available as a desktop application for Windows, Linux and Macintosh at: http://biogyan.com/. Key features of BioGyan include:

- Single window gateway allowing batch querying for list of genes against list of cellular processes and cell-types
- Ranking of articles based on relevance to the search terms
- Fetching of data from multiple databases simultaneously
- Retrieval of biological pathway information and 3D structures for queried genes/proteins
- Easy visualization of gene-process associations as bar charts and pie charts
- Manual intervention for user to annotate and validate results
- Highlighting of the search terms and relevant sentences facilitating easy manual screening
- Exporting citations to conventional bibliographic file formats (EndNote and BibTeX)
- Offline review of search results

BioGyan search

'BioGyan search', a key feature, empowers user to search biomedical literature directly indicating associations between genes of interest and multiple biological processes. Standard NCBI gene names are required for BioGyan as input. Users have the option to specify process name along with the gene name for query or use the pre-compiled process list. In the latter scenario, only the gene name is queried in PubMed while the gene name and pre-compiled processes are used in scoring of articles. Cell-type/tissue can optionally be used as keywords to refine the search.

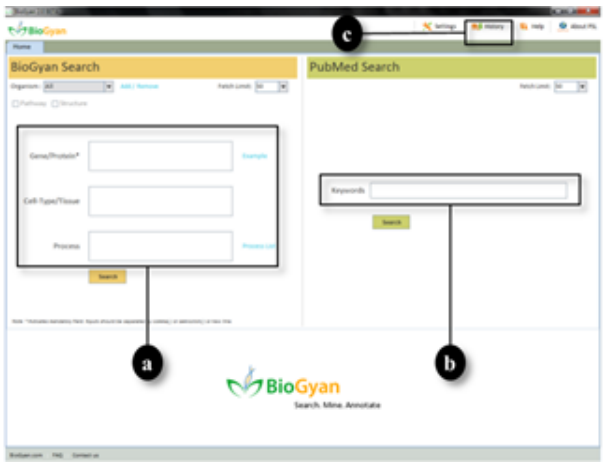

Figure 2: Home screen of BioGyan. (a): Key parameters for BioGyan search; (b): Generic PubMed search utility; (c): History - Utility to view previously executed searches.

A typical result dashboard for any query is shown in Figure 3. The key features include: number of articles identified for each gene to process association in chronological order (starting with the latest) and list of articles indicated as relevant/irrelevant. 'Process Score' for the abstracts are represented and the queried genes, their aliases and important sentences are highlighted. As stated in Methods section, Process Score determines the relevance of an article.

BioGyan also facilitates easy visualization of results through bar and pie charts. This enables user to broadly identify the processes affected by genes without having to go through abstracts (Figure 4). If specified by the user, BioGyan also fetches biological pathways and 3D structures of queried genes from public databases. Jmol [12] command window is offered alongside to issue commands for customizing the display of structures (Figure 5).

PubMed search

BioGyan provides 'PubMed search' utility to allow search functionality similar to that offered by NCBI PubMed. Here results section is different compared to typical BioGyan search, where the gene to process association table is replaced with table of just search terms. Although, results for PubMed search through BioGyan tool is identical to NCBI PubMed website, BioGyan offers an enhancement by highlighting the search terms, processes from the pre-compiled list and sentences containing the search terms. This helps in faster screening of abstracts manually. For PubMed search, the Process Score is not calculated and the articles are not ranked for relevance.

Validation of articles by user

BioGyan empowers user to mark an article valid or invalid after manual assessment. This allows user to override any incorrect prediction by BioGyan. Furthermore, user can annotate an article (Figure 3 (e)) with comments. BioGyan facilitates storing these annotations for future reference.

Storage and export of results

BioGyan stores the executed searches under "History". This facilitates phase-wise or offline reading. The results can also be exported in EndNote and BibTeX, Spreadsheet, and XML format.

Assessment of prediction capability

Evaluation of prediction performance: We evaluated the prediction performance of BioGyan on the basis of abstracts fetched by querying 7 human genes against the 202 pre-compiled processes. These genes are known to participate in multiple cellular processes in diverse roles such as activator, inhibitor and transcriptional regulator. This query predicted 424 gene-process associations and fetched 15102 abstracts. An abstract can contain information for more than one gene-process association and hence is counted multiple times when calculating the total number of abstracts fetched in BioGyan. After accounting for this, there were 6889 unique abstracts in the query results. To calculate the prediction accuracy and specificity of the returned results, manual verification of the articles classified as relevant and irrelevant was performed.

Our analysis (Table 1) showed that for BioGyan, the accuracy of prediction for relevance or irrelevance for fetched abstracts was high (=85.46%) indicating that the underlying algorithm is robust enough to correctly identify relevant abstracts for a gene-process association. Moreover, the high specificity (=91.34%) substantiates BioGyan's

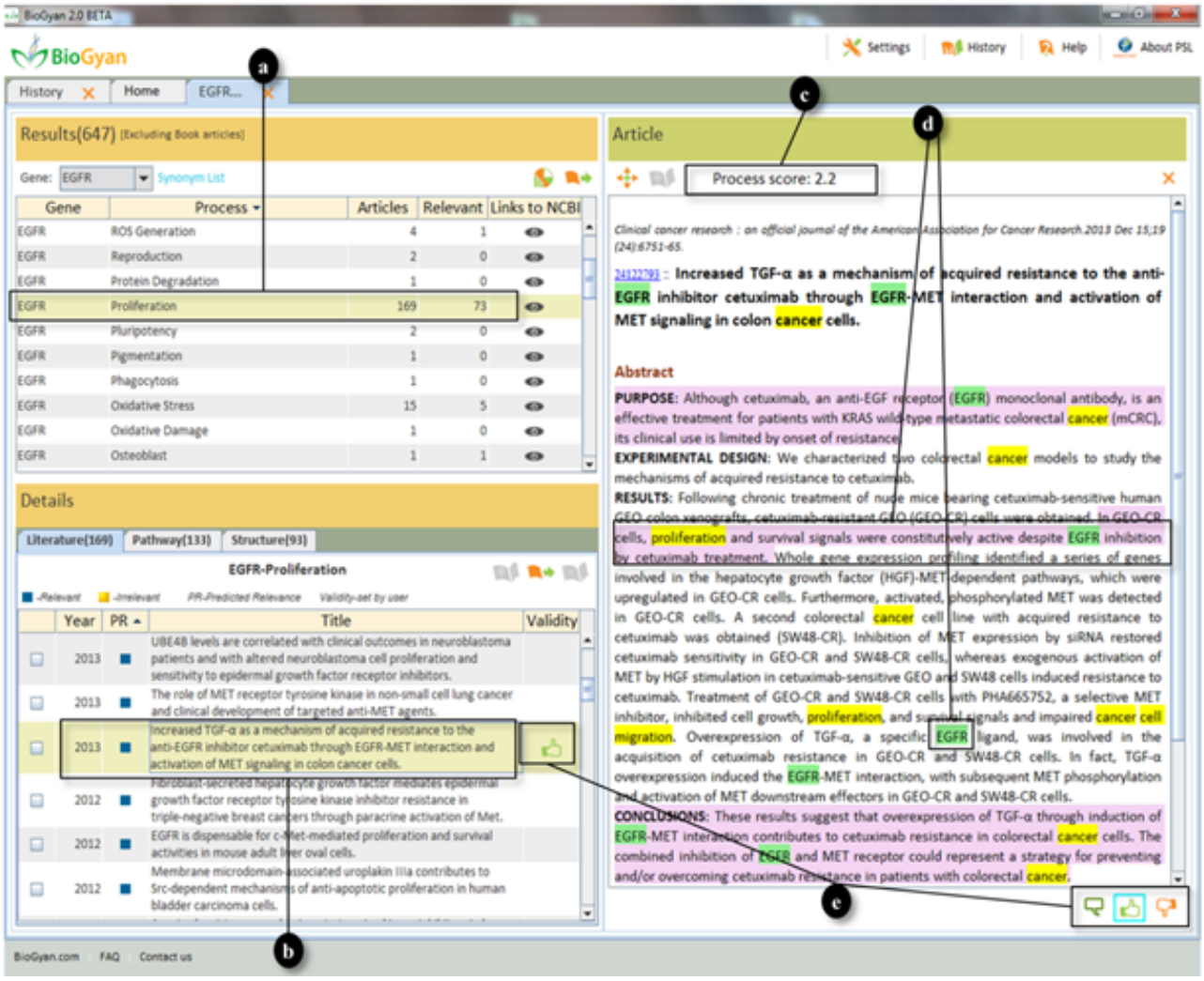

Figure 3: Typical result from a BioGyan search. (a): Gene to process associations, total number of articles found for each gene to process association and number of relevant articles; (b): Details of the articles for a gene to process association; (c): Process Score for a selected abstract; (d): Selected abstract in which the queried terms and relevant sentences are highlighted; (e): Utilities for user's annotation.

ability of rejecting irrelevant literature, thus significantly helps in reducing the number of articles to be screened manually and increases the efficiency.

Undoubtedly, BioGyan, is the only tool available which assists in quickly determining genes associated with cellular processes or phenotype with good accuracy and high specificity. Furthermore, BioGyan ranks the articles, thus, enabling easy identification of important articles. A similar search in PubMed requires multiple queries (7 genes x 202 processes = 1414 possible combinations) and manual curation of thousands of articles in order to validate the association of genes with these processes. As BioGyan allows for batch querying of a list of genes against a list of processes, the above search can be done in a single query.

Comparative assessment: We have compared BioGyan's prediction performance with an established tool PESCADOR [19] that helps to determine protein-protein interactions. PESCADOR also helps generate protein networks using LAITOR [20] text-mining method [19]. Although, the primary objective of PESCADOR is not to establish gene to process associations, it allows customized search to provide this information as its output.

PESCADOR was challenged with the same set of 7 human genes and 202 processes that were used for evaluating BioGyan's performance.

As shown in Table 2, BioGyan identified equal or higher number of processes for each queried gene as compared to PESCADOR. Moreover, upon manual verification, the average prediction efficiency of gene to process associations was 79.0% for BioGyan while only 53.6% for PESCADOR. Further, for the gene PERP, which is known to play a key role in cell adhesion and apoptosis, BioGyan identified 15 associated processes of which, 12 were found to be relevant on manual verification. PESCADOR, on the other hand, found no associated processes for PERP. Thus, BioGyan performs better in identifying gene to process associations, when compared with an established tool, PESCADOR.

Case-study: Identifying literature evidence for the role of gene(s) in a particular process

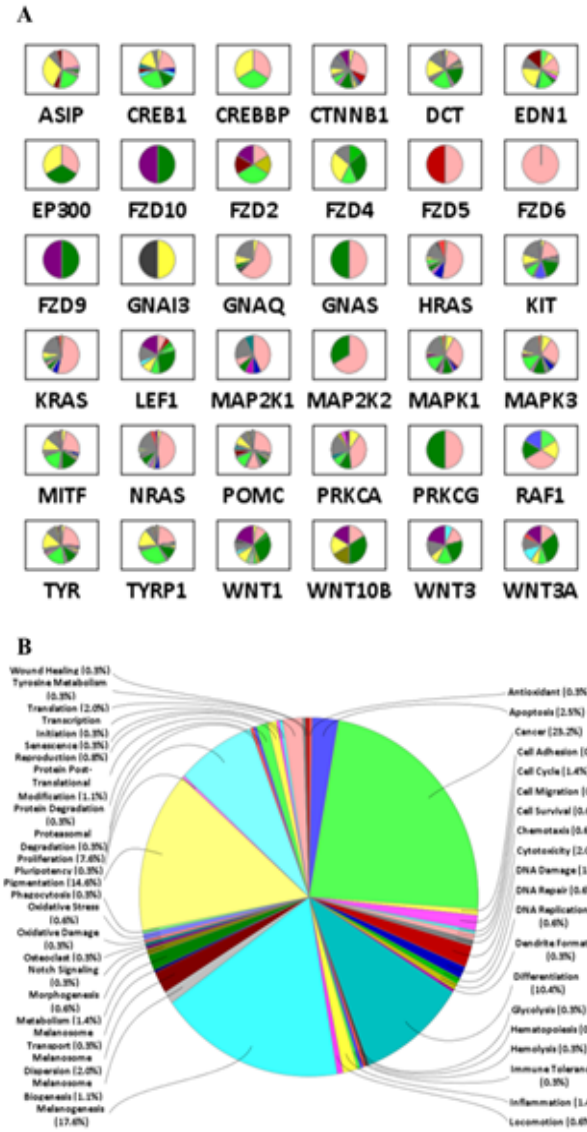

Figure 4: Consolidated representation of the association of queried genes to processes. Panel A: Pie charts showing queried genes and their participation in processes. Each section in a chart refers to a unique process. Panel B: Pie chart corresponding to TYR gene. Processes and percentage of the total articles found by BioGyan are shown.

Gene	# Processes predicted by PESCADOR	# Processes predicted by BioGyan	% Correct prediction by PESCADOR	% Correct prediction by BioGyan
AQP3	25	30	68.0	86.7
ARNT	38	36	84.2	83.3
FADD	25	44	32.0	72.7
HDAC1	45	53	80.0	86.8
IRF3	29	31	37.9	67.7
PERP	0	15	0.0	80.0
TP63	22	21	72.7	76.2

Table 2: Comparison of BioGyan with PESCADOR with respect to prediction of gene-process associations

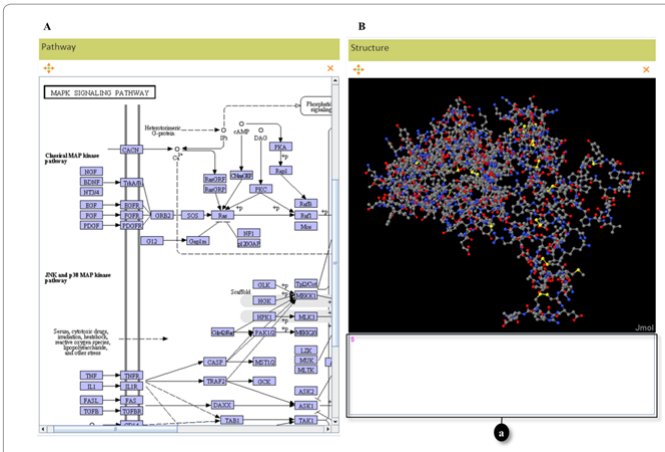

Figure 5: Biological pathway and 3-dimensional structure of a queried gene as displayed by BioGyan. Panel A: MAPK signaling pathway displayed for the queried gene, EGFR. Panel B: EGFR structure corresponding to PDB ID: 1DNQ; (a): Jmol command window to facilitate customization of display of structures.

Life Science researchers are frequently interested in identifying genes associated with a biological process in a specific cell-type. Conventionally this is done by querying public databases for the biological process and identifying the genes associated with it. However, the genes so obtained may not be supplemented with cross-references to published literature[10], [11], and, may not be relevant in a specific cell-type/tissue. In such cases, BioGyan could be extremely relevant for identifying published scientific articles to supplement gene-process associations.

As our group works extensively in the field of skin biology, we used BioGyan to identify genes involved in 'keratinocyte differentiation' process in skin epidermis. Keratinocyte differentiation is a key biological process responsible for maintaining human skin homeostasis, and its imbalance may result in psoriasis [21-24]. We queried GO (database release 2013-11-30) for genes involved in keratinocyte differentiation and retrieved 108 genes. Out of these 108 genes, 38 genes have experimental evidence assigned by a curator in GO. The following "Evidence Code" parameters were used for identifying these genes: EXP, IGI, IEP, IC, IMP, IDA, TAS and IPI (See http://www.geneontology.org/GO.evidence.shtml# for details). Out of the remaining 70 genes, 63 were obtained from "Electronic Annotation" that were not assigned by a curator in GO whereas 7 have "Evidence Code" assigned through computational methods or based on authors' statements.

To ascertain the involvement of these 108 genes in keratinocyte differentiation through literature evidence, following query was used in BioGyan.

"Gene/Protein" – List of 108 genes obtained from GO

"Cell-Type/Tissue" – Keratinocyte

"Organism" – Human (*Homo sapiens*)

"Fetch Limit" – All

As keratinocyte differentiation was the process of our interest, differentiation and its associated keywords were only retained in the pre-compiled process list. BioGyan results were further verified manually to ascertain the gene-process associations predicted.

On analyzing the results, we found that for 58 genes BioGyan

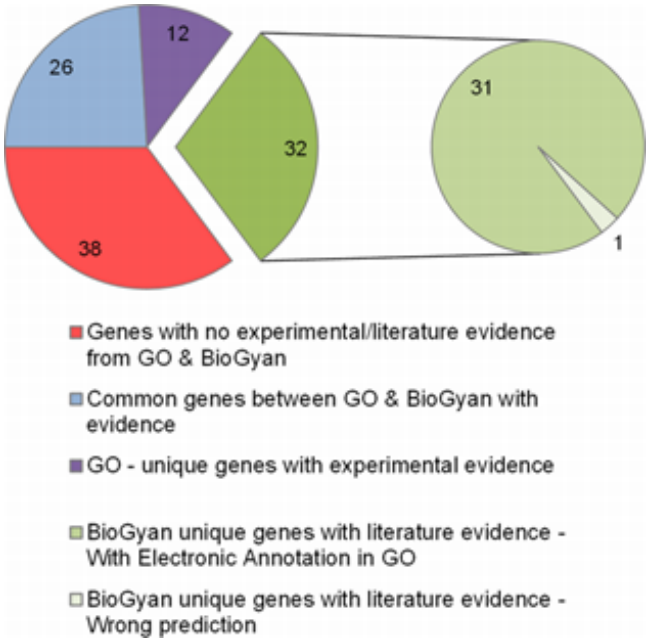

Figure 6: Comparison of genes with literature evidence obtained from BioGyan and GO.

identified literature supporting an association with the keratinocyte differentiation process. It is noteworthy that out of these 58 genes, 32 genes were those for which there was no direct experimental evidence in GO. On manual curation of the articles retrieved in support of association between these 32 genes and keratinocyte differentiation, we found that BioGyan's predictions were correct for 31 genes. For the remaining one gene KRT17, the literature retrieved by BioGyan did not support an association. GO also had experimental evidence for the remaining 26 genes correctly predicted by BioGyan (Figure 6).

Additionally, for 38 out of 108 genes, neither there is experimental evidence in GO nor BioGyan was able to retrieve articles supporting an association with keratinocyte differentiation. This further emphasizes the importance of BioGyan as it can be used to identify potentially incorrect gene-process associations in the public databases. BioGyan missed 12 genes that had experimental evidence in GO due to the following key reasons: (1) the gene names used in the abstract were not the standard gene names/aliases as provided by NCBI [25]; and (2) the association to keratinocyte differentiation is mentioned in the full-text of the article but not in the abstract. The first issue is a result of occurrence of description/full form of gene in the abstract. This can be handled by maintaining a map of description of a gene and its standard name in BioGyan and we plan to address this issue in a future version. The second issue can be resolved by running the computation on the full-text instead of the abstracts as BioGyan currently does. However, this will not always work because the full-text of all articles are not freely available and is dependent on institutional or personal subscription to the journals.

Present limitations and future plan

BioGyan fails to predict gene-process associations for some genes that have names similar to common English words (e.g. WAS and SET) and gene symbols that are also used as abbreviations for other biological terms (e.g. PHB gene refers to prohibitin while PHB can also mean poly 3- hydroxybutyrate).

Presently, BioGyan can provide only qualitative indication for cell-type/tissue specific gene to process associations. We plan to soon incorporate computation for pathway enrichment in BioGyan such that a quantitative insight can be obtained. Furthermore, BioGyan currently relies on PubMed heuristics for identifying cell- or tissue specificity. We plan to include the heuristics for identifying cell- or tissue specificity in BioGyan algorithm. We also plan to include a functionality that will enable researchers to identify genes involved in biological processes of interest.

Conclusion

Identifying gene to process to phenotype associations in a cell/tissue specific manner is a key aspect of integrative biology. Our biomedical literature mining tool BioGyan is currently the most powerful solution available for easily identifying the role(s) played by genes in cellular processes in cell/tissue specific manner. Various features of BioGyan like batch query, user annotation of articles, collation of data from multiple databases and ability to work in offline mode makes it one of the most user-friendly literature mining application currently available for biologists.

Acknowledgements

We thank Janhavi Kodilkar, Nitisha Danve, Sharmilli Chettiar and Kenaz Siddiqui for manually reading the articles and verifying BioGyan predictions, and Mayank Solanki and Priya Kulkarni for valuable inputs on conceptualization and optimization of the tool. The work was funded by Persistent Systems Limited, the employer of all authors.

References

1. Yu H, Kim T, Oh J, Ko I, Kim S, et al. (2010) Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. BMC Bioinformatics 11 Suppl 2: S6.
2. Smalheiser NR, Zhou W, Torvik VI (2008) Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. J Biomed Discov Collab 3: 2.
3. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, et al. (2007) EBIMed--text crunching to gather facts for proteins from Medline. Bioinformatics 23: e237-244.
4. Dinasarapu AR, Saunders B, Ozerlat I, Azam K, Subramaniam S (2011) Signaling gateway molecule pages--a data model perspective. Bioinformatics 27: 1736-1738.
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29.
6. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, et al. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 38: D473-479.
7. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.
8. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, et al. (2013) The Reactome pathway knowledgebase. Nucleic Acids Res.
9. Milacic M, Haw R, Rothfels K, Wu G, Croft D, et al. (2012) Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers (Basel) 4: 1180-1211.
10. Bauer-Mehren A, Furlong LI, Sanz F (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol Syst Biol 5: 290.
11. Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8: e1002375.
12. Jmol: an open-source Java viewer for chemical structures in 3D.
13. UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res 41: D43-47.
14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.

15. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, et al. (2010) The NCBI BioSystems database. Nucleic Acids Res 38: D492-496.

16. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) PID: the Pathway Interaction Database. Nucleic Acids Res 37: D674-679.

17. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, et al. (2012) WikiPathways: building research communities on biological pathways. Nucleic Acids Res 40: D1301-1307.

18. Temkin JM, Gilder MR (2003) Extraction of protein interaction information from unstructured text using a context-free grammar. Bioinformatics 19: 2046-2053.

19. Barbosa-Silva A, Fontaine JF, Donnard ER, Stussi F, Ortega JM, et al. (2011) PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinformatics 12: 435.

20. Barbosa-Silva A, Soldatos TG, Magalhães IL, Pavlopoulos GA, Fontaine JF, et al. (2010) LAITOR--Literature Assistant for Identification of Terms co-Occurrences and Relationships. BMC Bioinformatics 11: 70.

21. Evans ND, Oreffo RO, Healy E, Thurner PJ, Man YH (2013) Epithelial mechanobiology, skin wound healing, and the stem cell niche. J Mech Behav Biomed Mater 28: 397-409.

22. Jensen PJ, Wheelock MJ (1996) The relationships among adhesion, stratification and differentiation in keratinocytes. Cell Death Differ 3: 357-371.

23. Hoss E, Austin HR, Batie SF, Jurutka PW, Haussler MR, et al. (2013) Control of late cornified envelope genes relevant to psoriasis risk: upregulation by 1,25-dihydroxyvitamin D3 and plant-derived delphinidin. Arch Dermatol Res 305: 867-878.

24. Proksch E, Brandner JM, Jensen JM (2008) The skin: an indispensable barrier. Exp Dermatol 17: 1063-1072.

25. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2010) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 39: D38–D51.