MedCrave
Step into the World of Research

MOJ Proteomics & Bioinformatics

Research Article

Open Access CrossMark

Text mining and network analysis for discovery of novel genes associated with *staphylococcus* biofilms

Abstract

Text mining is the process of extracting relevant information from the unstructured data obtained from the published corpora in biomedical literature. A vast amount of literature is available in context of biofilms in PubMed database. It is a very daunting task to extract useful information in the form of genes and proteins from text manually; therefore, text mining approaches are used for this purpose. The extracted relevant information can be appropriately visualized using network biology approaches. Therefore, here we present a framework that would help in discovering novel biomarkers (genes) and gene-drug associations, involved in *Staphylococcus* biofilm using the available published corpora. We initially utilized COREMINE tool to select biological processes (i.e. biofilm formation, growth, quorum sensing and pathogenesis) and extract genes from PubMed literature database for biofilms. We selected the co-occurring genes for four biofilm processes through network biology approach and validated them by using GO enrichment analysis. *AgrB* gene was validated as the gene involved in two biological processes i.e. biofilm quorum sensing and pathogenesis. Therefore, structure of AgrB encoded protein (AgrB) was predicted using bioinformatics tools and analysis of ligand binding was done using molecular docking which was further used in blocking quorum sensing and pathogenesis processes. The validation of results obtained from text mining suggests that this approach can be extended to reveal interesting trends and associations among different biological entities related to biofilms and other diseases.

Keywords: text mining, biofilms, network analysis, GO enrichment analysis, AgrB, biomedicine

Volume 4 Issue 5 - 2016

Mansi Jain, Kanika Gupta, Ashok Kumar
Centre for Systems Biology and Bioinformatics, Panjab University, India

Correspondence: Ashok Kumar, Centre for Systems Biology and Bioinformatics, Block-III, Sector-25, Panjab University, Chandigarh-160014, India,
Email ashokkumar@pu.ac.in, ashokbiotech@gmail.com

Received: August 24, 2016 | **Published:** December 08, 2016

Introduction

Advances in technologies related to biomedicine lead to the tremendous data generation through respective researches and discoveries. A wealth of knowledge obtained through these researches is available in the form of published texts. But it is a challenging task to extract relevant information like genes, proteins, diseases, drugs, biological mechanisms etc. from the available unstructured data.[1] Integration and analysis of the information obtained from the literature can lead to the new discoveries in medical health, medicine, therapeutic and diagnostic processes.[2] Therefore, this information could be easily extracted from the publications by using text mining approaches[3] and could be visualized using network biology approaches. Text mining is the branch of data mining which is used to find hidden relevant information from the available published corpora.[4] The purpose of text mining here is to process the unstructured textual information related to biofilms in order to extract knowledge like set of biofilm genes, their interactions with biofilm biological processes and associated drug relationship.[5] The application of text mining has earlier been seen on the literature of biomedical sciences like respiratory disease, breast cancer, heart diseases etc.[6]

In the present study we have used the text mining application on the PubMed literature available for *Staphylococcus* biofilms. Biofilms are the basically the assembly of various living and reproducing microorganisms that adhere to the surface via glue-like slime and extracellular polymeric substances (EPS).[6–8] These microorganisms are various species of bacteria, fungi, algae, yeast, protozoans and other microorganisms.[9] *Staphylococci* are recognized as the most frequent causes of biofilm-associated infections.[10] This exceptional status among biofilm-associated pathogens is due to the fact that staphylococci are frequent commensal bacteria on the human skin and mucous surfaces (and those of many other mammals).[11] Thus, staphylococci are among the most likely germs to infect any medical device that penetrates those surfaces, such as when being inserted during surgery.[12] Therefore, there is need to explore the genes and proteins of *Staphylococcus* involving in biofilm formation and other biological processes and their related pathways, drugs and other processes associated with these genes. The networks help in appropriate visualization of genetic disorders and all their known gene associations,[13] or of drugs and all their known protein targets,[14] enabling worthwhile insights into disease and disease therapy.[15]

Methodology

Data retrieval

For text mining purpose, the availability of text on the relevant topic is a fundamental necessity.[16] This text is made available in the form of published corpora in biomedical literature of biofilms obtained through major resource i.e. PubMed database.[17] MeSH major topic “Biofilms” and MeSH term “*Staphylococcus*” was searched from MeSH database.[18,19] Then the literature containing both terms and published during last 10years was obtained.

Exploring Connections from Biomedical Literature

Many tools are available for text mining the biomedical literature. Here COREMINE tool[19] was used which is a biomedical text mining tool and extracts all the information like list of genes and proteins, related MeSH terms, processes, diseases, drugs etc. from the literature. It gives the results in an interactive graphical interface. The corpora for biofilms were used to search for genes. The results were downloaded in the form of excel sheet. The results are sorted on the basis of significance score i.e. P-value. This significance score is calculated on the basis of the occurrence of two terms in the same title/abstract of the article. Higher the number of articles lower will be the P-value and thus more significant the connection.[20] Four processes i.e. biofilm formation, growth, quorum sensing and pathogenesis was also selected using the same tool.

Co-occurrence of genes

Co-occurrence is a term used in semantics[21] that means the occurrence of two terms in a text. If two terms are occurring in same article for the same topic they are more likely to be related to each other. Here we found the co-occurring genes by using network analysis approach. Co-occurrence networks are generally used to provide a graphic visualization of potential relationships. The generation and visualization of co-occurrence networks has become practical with the advent of electronically stored text amenable to text mining.

Network building

All the processes along with their corresponding genes were shown in a network using Cytoscape version 3.3.3.[22] Network analysis was done to differentiate the genes on the basis of their significance score. In the network nodes represent the genes and edges represent the significance score. Also all the networks were merged into one and the genes that were co-occurring for multiple processes were selected. Cystoscope is an open source bioinformatics software platform for visualizing molecular interaction networks. Additional features are available as plugins in cystoscope here merged network plug-in was used to merge networks.

GO Enrichment analysis

The genes that were co-occurring for all of the four processes were selected for further enrichment by using GO enrichment analysis.[23] This step provides the validation to the genes for their classification to particular processes. These validated genes were further shown in a network corresponding to their validated processes. Again using network analysis genes common for more than one process were found.

Molecular modeling and docking

In this step various bioinformatics tools were used for the structure prediction of selected protein. FALCON@ home tool[24] was used for structure prediction. In the Critical Assessment of protein Structure Prediction (CASP11) in the year of 2014, the FALCON@ home based prediction was ranked the 12th in the template-based modeling.[17] The ligands that correspond to the selected proteins were found using text mining tool like Coremine and curated database like Stitch4.0. The ligands common from both searches were taken and their binding with corresponding genes/proteins was further validated using molecular docking. Fire Dock (Fast Interaction Refinement in molecular Docking) tool[25] was used to refine the results obtained from Patch Dock.[26]

Results and discussion

Data retrieval

There were total 13907 and 87198 PubMed articles published for "Biofilms [MeSH major topic]" and "*Staphylococcus* [MeSH terms]" respectively and a total of 1568 articles were published in PubMed for both the terms in the previous 10years. These terms were co-occurring in the abstracts and full text of the published articles (Figure 1).

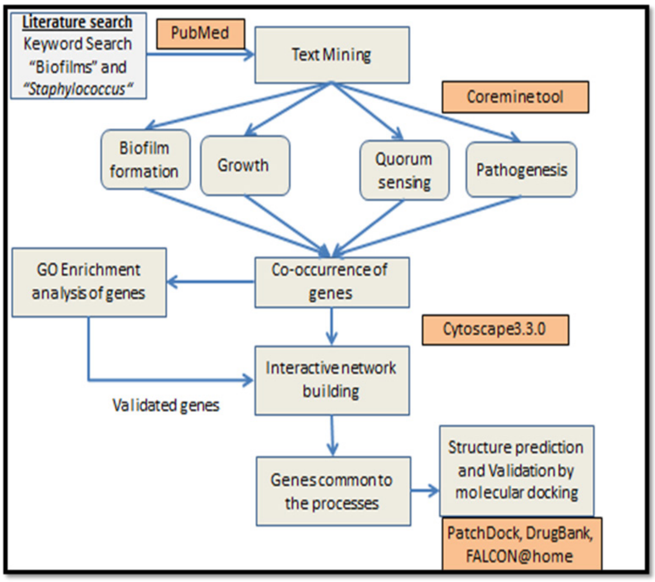

Literature search
Keyword Search "Biofilms" and "Staphylococcus"
PubMed
Text Mining
Coremine tool
Biofilm formation
Growth
Quorum sensing
Pathogenesis
GO Enrichment analysis of genes
Co-occurrence of genes
Cytoscape3.3.0
Interactive network building
Validated genes
Genes common to the processes
Structure prediction and Validation by molecular docking
PatchDock, DrugBank, FALCON@home

Figure 1 Review of the procedure and instruments utilized as a part of the procedure.

Exploring connections from biomedical literature

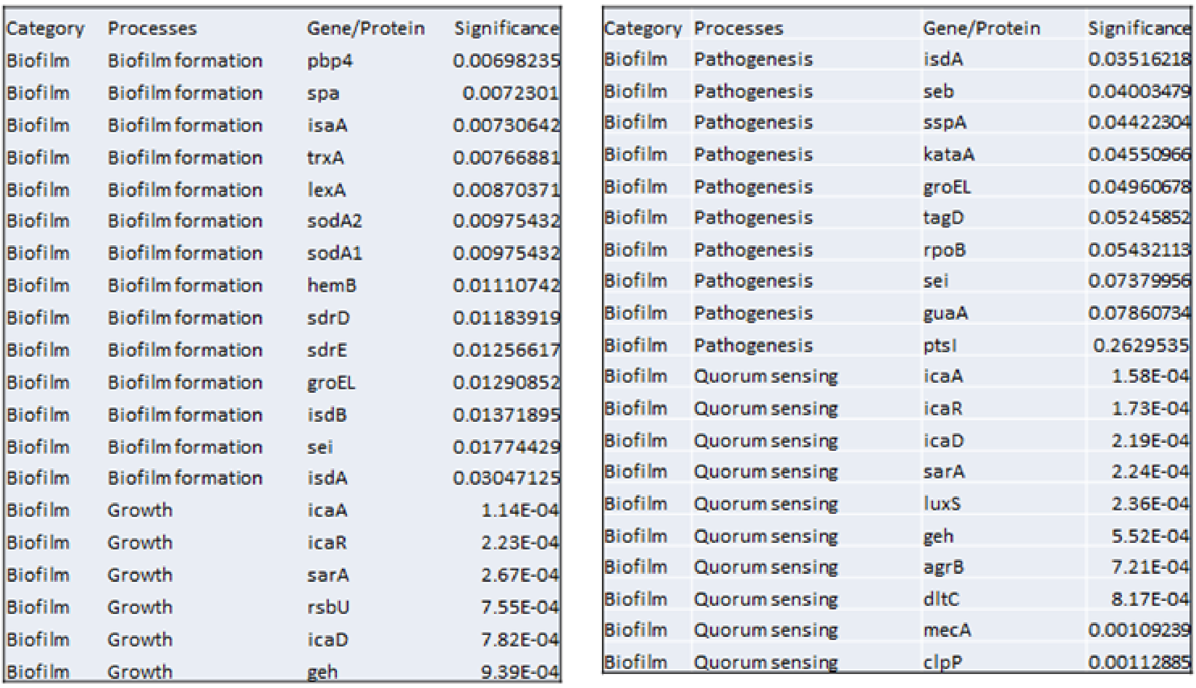
Biofilm term was searched and *Staphylococcus* genes were selected after using text minig approach (Figure 2). There were 25062 connections available for this term in Coremine. Only biological processes and Gene/protein option was selected as shown in Table 1.

Table 1 Total number of genes for each process and their range of significance scores

Process	Number of genes	Range of significance score	Highly cited gene	Number of articles	Less Cited gene	Number of articles
Formation (P1)	72	1.31E-05-0.03047125	*ica A*	31	*isdA*	1
Growth (P2)	82	1.14E-04-0.15396536	*ica A*	30	*norA*	1
Pathogenesis (P3)	63	1.21E-04-0.04513	*ica A*	14	*arlR*	2
Quorum Sensing (P4)	21	1.58E-04-0.02311349	*ica A*	10	*kataA*	2

Four Biofilm processes were selected in Coremine i.e. Biofilm formation, Growth, Pathogenesis and Quorum sensing. There were a total of only 238 genes corresponding to these four processes. Each connection was associated with a significance score. Thus genes are sorted on the basis of significance score, which is P-value, where low value represents higher significance and thus higher rank as shown in Figure 2.

Citation: Jain M, Gupta K, Kumar A. Text mining and network analysis for discovery of novel genes associated with *staphylococcus* biofilms. *MOJ Proteomics Bioinform.* 2016;4(5):311–316. DOI: 10.15406/mojpb.2016.04.00135



Category	Processes	Gene/Protein	Significance
Biofilm	Biofilm formation	pbp4	0.00698235
Biofilm	Biofilm formation	spa	0.0072301
Biofilm	Biofilm formation	isaA	0.00730642
Biofilm	Biofilm formation	trxA	0.00766881
Biofilm	Biofilm formation	lexA	0.00870371
Biofilm	Biofilm formation	sodA2	0.00975432
Biofilm	Biofilm formation	sodA1	0.00975432
Biofilm	Biofilm formation	hemB	0.01110742
Biofilm	Biofilm formation	sdrD	0.01183919
Biofilm	Biofilm formation	sdrE	0.01256617
Biofilm	Biofilm formation	groEL	0.01290852
Biofilm	Biofilm formation	isdB	0.01371895
Biofilm	Biofilm formation	sei	0.01774429
Biofilm	Biofilm formation	isdA	0.03047125
Biofilm	Growth	icaA	1.14E-04
Biofilm	Growth	icaR	2.23E-04
Biofilm	Growth	sarA	2.67E-04
Biofilm	Growth	rsbU	7.55E-04
Biofilm	Growth	icaD	7.82E-04
Biofilm	Growth	geh	9.39E-04

Category	Processes	Gene/Protein	Significance
Biofilm	Pathogenesis	isdA	0.03516218
Biofilm	Pathogenesis	seb	0.04003479
Biofilm	Pathogenesis	sspA	0.04422304
Biofilm	Pathogenesis	kataA	0.04550966
Biofilm	Pathogenesis	groEL	0.04960678
Biofilm	Pathogenesis	tagD	0.05245852
Biofilm	Pathogenesis	rpoB	0.05432113
Biofilm	Pathogenesis	sei	0.07379956
Biofilm	Pathogenesis	guaA	0.07860734
Biofilm	Pathogenesis	ptsI	0.2629535
Biofilm	Quorum sensing	icaA	1.58E-04
Biofilm	Quorum sensing	icaR	1.73E-04
Biofilm	Quorum sensing	icaD	2.19E-04
Biofilm	Quorum sensing	sarA	2.24E-04
Biofilm	Quorum sensing	luxS	2.36E-04
Biofilm	Quorum sensing	geh	5.52E-04
Biofilm	Quorum sensing	agrB	7.21E-04
Biofilm	Quorum sensing	dltC	8.17E-04
Biofilm	Quorum sensing	mecA	0.00109239
Biofilm	Quorum sensing	clpP	0.00112885

Figure 2 Rundown of qualities downloaded for four procedures.

Network building and co-occurrence of genes

The nodes in the network represent the genes/protein while the source node represents the particular process. The edges are on the basis of significance attribute and highly significant edges are represented in large size. The color of nodes is on the basis of their degree. Lower the degree darker and smaller will be the node. Cystoscope uses statistical measures to make the network more significant. Since there was no overrepresentation of genes in the merged network as shown in Figure 3 therefore there were total 87 unique genes in the dataset. The overrepresentation of genes is due to their involvement in more than one process.

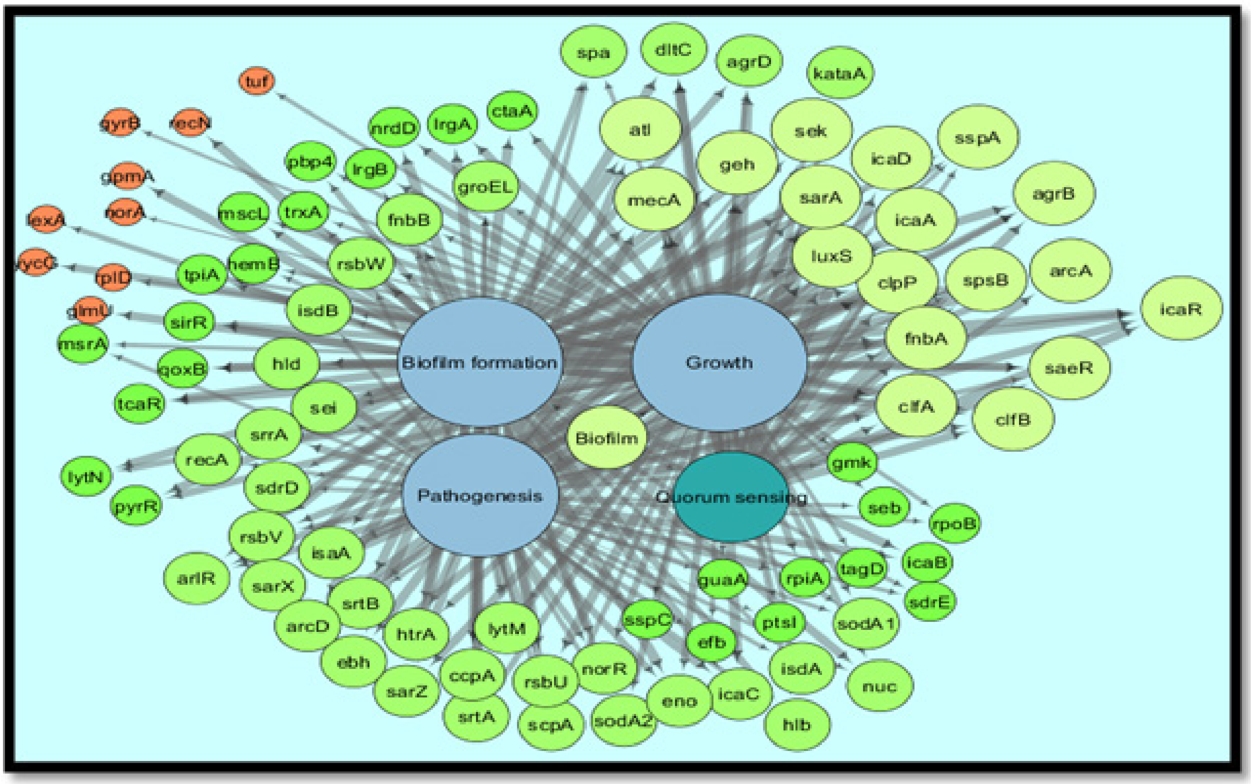


Figure 3 Merged network showing the four biofilm processes (source nodes) and corresponding genes (target nodes) connected on the basis of significance score (edges).

As this dataset was obtained using text mining so these overrepresented genes were actually the genes that were co-occurring in the titles/abstracts of the articles for more than one process. In this way the co-occurring genes can be selected from this network with degree layout which showed that the frequency of nodes sharing two neighbors is higher than those sharing four neighbors. It means that there were nodes representing genes that were involved in all four biological processes. Those genes involved in all of the four processes were selected using network analysis and degree layout. There were total 18 genes, obtained from the above merged network, that appear in all of the four processes of biofilm Figure 4.

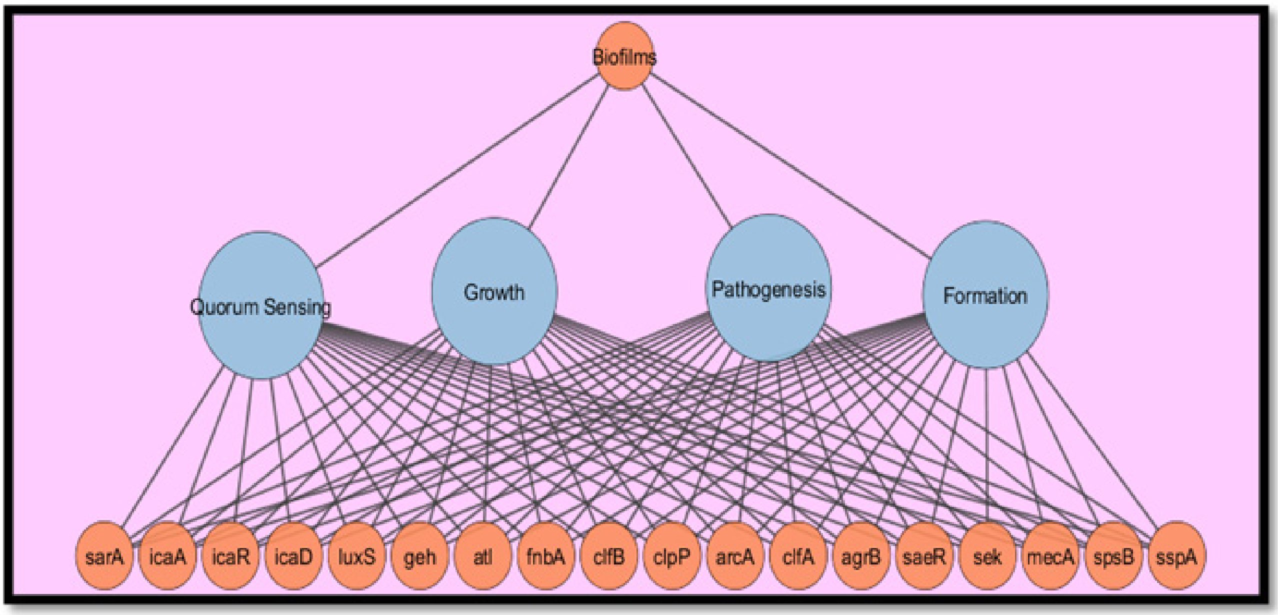


Figure 4 Network showing the genes that are co-occurring for all processes.

GO enrichment analysis

The selected genes were validated by using GO enrichment analysis. List of 18 genes were pasted under Enrichment analysis input box. Biological process was selected as it reduces the analysis time by selecting one process at a time. The pie chart shown in Figure 5 is prepared using the data obtained from GO enrichment analysis.

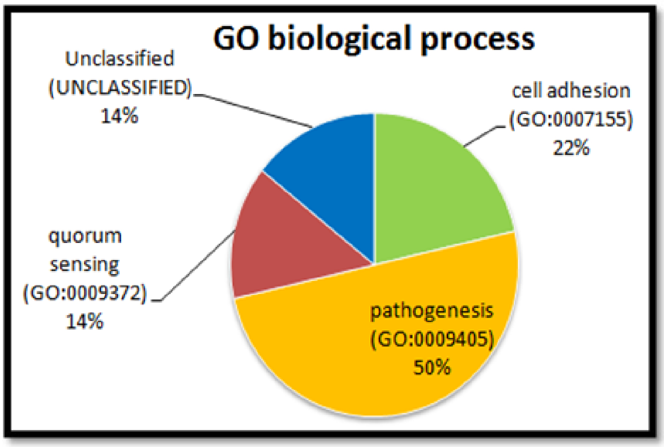


Figure 5 Pie chart showing the distribution of biological processes.

Out of 18 genes 16 genes were taken by GO consortium. Out of these 16 genes 2 were unclassified and rest classification of genes are shown in the form of pie chart. This enrichment validates that out of 18 genes, as obtained from text mining, 7 genes were there in Pathogenesis process while 2 genes were validated to be involved in quorum sensing process. Out of these 9 genes we were interested to find further common genes that could target both the processes simultaneously.

From this network Figure 6 we found *AgrB* gene that was involved in both processes.[27] This could be an interesting gene because by targeting this gene's activity both processes can be inhibited simultaneously.

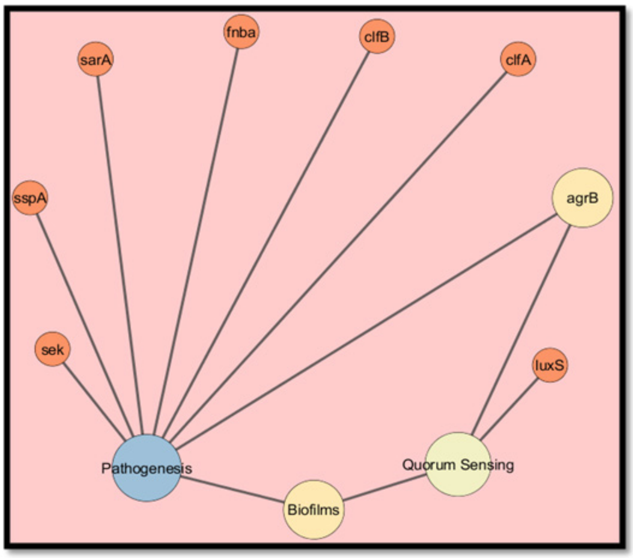


Figure 6 System demonstrating the accepted qualities.

Structure prediction and molecular docking

After finding the potential gene the next step was to predict the structure of gene encoded protein and perform molecular docking with common ligands obtained from both text mining and curated database searches. The structure of *AgrB* gene encoded protein i.e. AgrB (Accessory gene regulator protein B) was predicted using the online tool FALCON@ home. The Fasta sequence of protein with Uniprot id P0C1P7 was pasted in the FALCON@ home. It performs the template based modeling. The structure of AgrB protein was predicted using template with PDBid 1R8S. The alignment file of 1R8S and AgrB was provided by FALCON and the alignment score is 23.9459. The ligands to be docked was again extracted using both approaches i.e. text mining and database search. These searches showed 3-oxo-C12-HSL was highly significant using database search and Homoserine Lactone was highly cited drug as shown by text mining Figure 7.

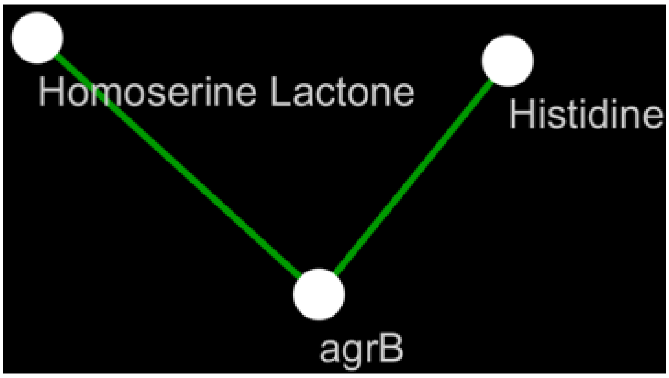
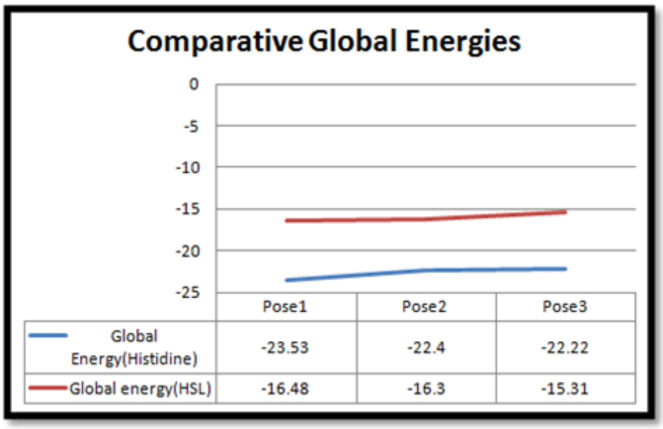


Figure 7 Merged network with intersecting nodes.

The structures of both the ligands were obtained from Drugbank. Also on comparing the properties of both ligands it was found that L-Histidine had more hydrophobic character than Homoserine Lactone (HSL) as logP value of former is highly negative. Moreover, number of Hydrogen bond donor and acceptor were also more in case of L-Histidine which means it binds with receptor more efficiently than HSL. So docking was performed using Patch Dock and further refined the results of docking using Fire Dock. Here three docking poses for both HSL and L-Histidine bound with AgrB protein were compared. The results were compared on the basis of global energy which predicts the stability of the complex. In both results pose1 had the lowest global energy. We also plotted a graph of comparative global energies for both docking poses. It was found that the complex of AgrB and L-Histidine was more stable than with HSL which was otherwise highly cited ligand for AgrB (Figure 8).



	Pose1	Pose2	Pose3
Global Energy(Histidine)	-23.53	-22.4	-22.22
Global energy(HSL)	-16.48	-16.3	-15.31

Figure 8 Examination of Global energies for different poses obtained for Histidine and HSL.

Conclusion

The present work is an application to text mining and network analysis approach. In the present work we started with 238 genes obtained using text mining which were then sorted down to 9 on the basis of their enrichment and validation processes. The co-occurring and validated genes found more interested than the rest as it can provide the way to novel receptor prediction that could target multiple processes simultaneously. Only 9 genes were validated for two important biological processes i.e. Quorum Sensing and Pathogenesis. These two processes play an important role in biofilm development and survival. Thus genes responsible for these processes are also of great importance. After further analysis, *AgrB* gene was found to be involved in both processes and blocking of this gene encoded protein might suggest the inhibition of both processes simultaneously. Again we found drugs for AgrB protein using text mining and found a less cited ligand L-Histidine that bound more efficiently to AgrB protein than HSL, a highly cited ligand for AgrB. Molecular docking and other curated databases were also used for this purpose. Thus, through text mining we could discover those relations that were less or never been discovered before.[28]

Acknowledgements

Authors are thankful to Vice Chancellor of Panjab University, Chandigarh and Coordinator of Centre for Systems Biology and Bioinformatics for providing all the facilities to carry out this work. We would also like to acknowledge the support provided by online tools, servers, databases and softwares for successful accomplishment of our work that is

PubMed Database (http://www.ncbi.nlm.nih.gov/pubmed)

Gene Ontology Consortium (http://geneontology.org/)

Coremine (https://www.coremine.com/medical/)

FireDock (http://bioinfo3d.cs.tau.ac.il/Fire Dock/)

Cytoscape (http://www.cytoscape.org/release_notes_3_3_0.html)

PatchDock (http:// bioinfo3d.cs.tau.ac.il/PatchDock/)

DrugBank (http://www.drugbank.ca/) and FALCON@home (http://protein.ict.ac.cn/FALCON/).

Conflict of interest

The author declares no conflict of interest.

References

1. Bravo A, Cases M, Queralt–Rosinach N, et al. A knowledge–driven approach to extract disease–related biomarkers from the literature. *BioMed research international*. 2014;2014:253128.
2. Jensen K, Panagiotou G, Kouskoumvekaki I. Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level. *PLoS Comput Biol*. 2014;10(1):e1003432.
3. Krallinger M, Valencia A. Text–mining and information–retrieval services for molecular biology. *Genome Biol*. 2005;6(7):224.
4. Fleuren WW, Alkema W. Application of text mining in the biomedical domain. *Methods*. 2015;74:97–106.
5. Abul Seoud RA, Mabrouk MS. TMT–HCC: a tool for text mining the biomedical literature for hepatocellular carcinoma (HCC) biomarkers identification. *Comput Methods Programs Biomed*. 2013;112(3):640–648.

6. Donlan RM. Biofilms: microbial life on surfaces. *Emerg Infect Dis*. 2002;8(9):881–890.
7. Doulgeraki AI, Di Ciccio P, Ianieri A, et al. Methicillin–resistant food–related Staphylococcus aureus: a review of current knowledge and biofilm formation for future studies and applications. *Res Microbiol*. 2016;S0923–2508(16)30083–3.
8. Gupta P, Sarkar S, Das B, et al. Biofilm, pathogenesis and prevention–a journey to break the wall:a review. *Arch Microbiol*. 2016;198(1):1–15.
9. Margarit y Ros I. Streptococcus pyogenes Pili. In: Ferretti JJ, Stevens DL, et al. editors. *Streptococcus pyogenes: Basic Biology to Clinical Manifestations*. Oklahoma City, USA; 2016.
10. Vuong C, Otto M. Staphylococcus epidermidis infections. *Microbes Infect*. 2002;4(4):481–489.
11. Vuong C, Voyich JM, Fischer ER, et al. Polysaccharide intercellular adhesin (PIA) protects Staphylococcus epidermidis against major components of the human innate immune system. *Cell Microbiol*. 2004;6(3):269–275.
12. Otto M. Staphylococcal biofilms. *Curr Top Microbiol Immunol*. 2008;322:207–28.
13. Goh KI, Cusick ME, Valle D, et al. The human disease network. *Proc Natl Acad Sci U S A*. 2007;104(21):8685–8690.
14. Yildirim MA, Goh KI, Cusick ME, et al. Drug–target network. *Nat Biotechnol*. 2007;25(10):1119–1126.
15. Durmus S, Cakir T, Ozgur A, et al. A review on computational systems biology of pathogen–host interactions. *Front Microbiol*. 2015;6:235.
16. Spasic I, Ananiadou S, McNaught J, et al. Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform*. 2005;6(3):239–251.
17. Roberts RJ. PubMed Central: The GenBank of the published literature. *Proc Natl Acad Sci U S A*. 2001;98(2):381–382.
18. Rogers FB. Medical subject headings. *Bull Med Libr Assoc*. 1963;51:114–116.
19. Jenssen TK, Laegreid A, Komorowski J, et al. A literature network of human genes for high–throughput analysis of gene expression. *Nat Genet*. 2001;28(1):21–28.
20. Chavalarias D, Wallach JD, Li AH, et al. Evolution of Reporting P Values in the Biomedical Literature, 1990–2015. *JAMA*. 2016;315(11):1141–1148.
21. Rebholz–Schuhmann D, Grabmuller C, Kavaliauskas S, et al. A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources. *Drug Discov Today*. 2014;19(7):882–889.
22. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498–2504.
23. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25–29.
24. Wang C, Zhang H, Zheng WM, et al. FALCON@home: a high–throughput protein structure prediction server based on remote homologue recognition. *Bioinformatics*. 2016;32(3):462–464.
25. Mashiach E, Schneidman–Duhovny D, Andrusier N, et al. FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res*. 2008;36(Web Server issue):W229–W232.
26. Schneidman–Duhovny D, Inbar Y, Nussinov R, et al. atchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res*. 2005;33(Web Server issue):W363–W367.
27. Vuong C, Saenz HL, Gotz F, et al. Impact of the agr quorum–sensing system on adherence to polystyrene in Staphylococcus aureus. *J Infect Dis*. 2000;182(6):1688–1693.
28. Li H, Liu C. Biomarker identification using text mining. *Computational and mathematical methods in medicine*. 2012;2012:135780.