CRRAO Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS)

Research Report

Author (s): S.B.Rao, P.L.Priyanka and P.Manimaran

Title of the Report: Candidate gene identification for SLE disease using network centrality measures and Gene Ontology

Research Report No.: **RR2014-11**

Date: **May 9, 2014**

Prof. C R Rao Road, University of Hyderabad Campus,
Gachibowli, Hyderabad-500046, INDIA.
www.crraoaimscs.org

This is the draft of the manuscript submitted to the Journal.
The revised version of this manuscript has been published in PLoS One Journal.
(http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0081766)

Candidate gene identification for SLE disease using network centrality measures and Gene Ontology

S. B. Rao,[a] P. L. Priyanka,[a,b] and P. Manimaran*[a]

[a]*C R Rao Advanced Institute of Mathematics, Statistics, and Computer Science, GachiBowli, Hyderabad - 500046, INDIA.*

[b]*Department of Biotechnology, University of Hyderabad, Hyderabad - 500046, INDIA.*

Abstract

Systemic lupus erythematosus (SLE) commonly accredited as “the great imitator” is a highly complex disease involving multiple gene susceptibility with non-specific symptoms. Many experimental and computational approaches have been used to investigate the disease related candidate genes. But the limited knowledge of gene function and disease correlation and also lack of complete functional details about the majority of genes in susceptible locus, encumbrances the identification of SLE related candidate genes. In this paper, we have studied the human immunome network using various graph theoretical centrality measures integrated with the gene ontology terms to predict the new candidate genes. As a result, we have identified 8 candidate genes which may act as potential targets for SLE disease.

Introduction

Systemic lupus erythematosus (SLE) is a polygenic and multi-factorial disease, which often manifests different clinical phenotypes \citep{Castro08}. The conventional therapies for SLE include antimalarial drugs, immunosuppressive medications, nonsteroidal anti-inflammatory drugs (NSAIDs), which are non-specific and immunosuppressive. So the research is concentrated on developing the targeted therapies. Investigation of genetic predisposition through gene expression profiling and linkage analysis in multiple populations generates large sets of potential candidate genes. This approach successfully predicts genes for the diseases that cause a single risk, but fails to identify the genes causing complex disease \citep{Glacier02}. This necessitated the development of in silico approaches such as ontology based, computation-based, and text based for the analysis of complex diseases\citep{Zhu07}. In silico methods utilize the information of protein interactions, GO terms, gene expression data, sequence features, protein domains, protein function, orthologous connections, chromosomal regions, pathways, mutations (SNPs), chemical components, disease probabilities etc for predicting the candidate gene.

Recently, several online tools have been developed for prioritizing candidate genes, which usually combine the different in silico approaches\citep{Tran10}. For example, SUSPECTS \citep{Adie06} ranks genes by matching sequence features, GO terms, interpro domains, and gene expression data. ToppGene \citep{Chen09} uses functional annotations, protein interaction networks to prioritize disease specific genes. Different tools like Polysearch \citep{Cheng08}, MimMiner \citep{Driel06}, and BITOLA \citep{Hris05} relies on biological data mining. Posmed, a computational based approach prioritizes candidate genes using an inferential process similar to artificial neural network comprising documentrons \citep{Yoshi09}. Some tools like Phenopred use disease phenotype information which employs/associates data from gene-disease relations \citep{Radi08}, protein-protein interaction data, protein functional annotation at a molecular level and protein sequence data to detect novel gene-disease associations in humans. All these online tools have been successfully used for the prediction of candidate gene in diseases like epilepsy \citep{Sha10}, and osteoporosis \citep{Huang08}, type II diabetes \citep{Tiffi06} and gene prioritization, depending on information of chromosomal location or genes differentially expressed in a tissue. But the above approaches have failed in case of SLE as it involves genes of differential expression patterns in tissues, influenced by various environmental factors. The limited information about the markers of SLE also contributed to their failure\citep{Ref crispin et al 2010}.

In such a situation, the network centrality measures coupled with the ontological terms favoured the identification of candidate genes for SLE. Analysis of Protein -protein interaction networks aids for inferring the function of uncharacterized proteins. In the recent past many network based analyses have been developed for protein function prediction, identification of functional modules and disease candidate gene prediction etc. With the advent of sophisticated technologies for the functional annotation of genes, the candidate genes prioritization have become increasingly facile. GO terms are used for the systematic annotation of genes.

In the present work, we study the human immunome network in combination with the graph theoretic centrality measures and GO terms in order to identify candidate genes for SLE disease. For this purpose we have adopted the procedure developed by CsabaOrtutay {\it et~al}. \citep{Ortu09}. In our work, we have applied more number of centrality measures such as eccentricity, centroid values, SP-betweenness, CF-closeness, CF-betweenness, Katz Status, Eigen vector and Page Rank apart from closeness and degree centrality used by Csaba Ortutay {\it et~al}. This study effectively analyzes candidate genes because of the consideration of more graph centrality measures. This method can be applied to investigate other immune diseases for which sufficient information is present in databases.

Results

We have constructed the human immunome network from the data collected from the HPRD \citep{Pra09} and the immunome database \citep{Ortu06}. The core immunome network contains 548 proteins with 1372 interactions (Figure: 1). In this work, we have adopted the method developed by CsabaOrtutay to predict the SLE related candidate genes \citep{Ortu09}. CsabaOrtutay and co-works in their work used three centrality measures such as degree, closeness and efficiency for the prediction of primary immunodeficiency (PID) related genes. In our study, we have applied ten centrality measures \citep{Jun06} to find the potential of individual proteins from the immunome network. From the Fig 1 which shows the Spearman correlation coefficients for all pairs centrality measures, we conclude that the centrality measures of the immunome network are not coincidental. The other topological properties like the degree distribution that follows power law behavior (γ = 1.86), the average degree is 4.995 and the diameter is 12 were also calculated \citep{Dor03}.

Gene enrichment analysis of top 50 high score genes

In order to verify the relevance of chosen centrality measures in identifying important immunome related GO terms, the gene enrichment analysis has been carried out for top 50 high score genes of each centrality measures. Enzyme linked receptor protein signalling pathway (GO: 0007167) and regulation of cytokine-mediated signalling pathway(GO:0001959) were dominant in majority of the centrality measures for biological process which indicates that signalling related ontology terms dominated the Biological process. The molecular function GO terms were dominated by the kinase activity among which Protein kinase activity (GO: 0004672), protein tyrosine kinase activity (GO: 0004713) and non-membrane spanning protein tyrosine kinase activity (GO: 0004715) were more significant. The cellular components were dominated by the membrane proteins (GO:0045121). From the above results, we can say that the centrality measures help in identification of important immunome related GO terms.

Gene enrichment analysis of known SLE genes

The gene enrichment analysis was carried out for the known SLE genes. In case of cellular component, only 3 terms (GO:0009897,GO:0044421,GO:0005615) were found to be enriched and are found to be associated with extracellular component. Growth factor activity (GO: 0008083) and cytokine receptor binding (GO:0005126) only were observed for molecular function. The regulatory processes were found to be dominant in case of GO biological processes of which most of them highly relates to immune processes with a high p value of 8.21×10^{-4}.

Prediction of new SLE candidate genes

We have combined the gene lists for the high network scores, knownSLE genes and genes with significant disease ontologies in order topredict the novel SLE genes. From the Venn diagram, the genes thatdon't fall under the known SLE genes but common in both lists ofgenes with high network scores and genes with significant diseaseontologies were predicted to be the novel genes. Since, thepredicted genes were obtained from the list of genes withsignificant ontologies, which infact was obtained through geneenrichment analysis. This implies that the predicted novel geneshave similar disease ontologies as that of know SLE genes i.e., theyare related to known SLE genes in their biological process,molecular function and cellular components. Thus, we can say thatthe predicted novel genes are significant to SLE disease.

By combining, the lists of the genes with top 50 high scores for allthe centrality measures, we obtained 113 unique genes out of which72 are known SLE genes. From the GO enrichment analysis we obtain131 genes with significant disease ontologies out of which 103 wereknown SLE genes. Altogether 145 of the 329 the known SLE genes wereselected either by GO terms or high scores of centrality measures.We found 38 genes had both high scores and significant GO terms inthat 30 genes are known SLE genes. Thus we were able to identify 8genes as a novel SLE candidate genes. This is clearly represented asa Venn diagram ~\ref{fig:02} and the details of the predicted genesare given in the table ~\ref{fig:03}.

Discussion

Human Protein-interaction networks have been used so far for the identification of genes responsible for causing disease and prediction of new disease candidate genes like PID, type II diabetes, osteoporosis, and neuro-degenerative diseases for drug discovery \citep{Ideker08,Oti06,Goh07,Chau09}. A recent review focuses on prediction of the protein networks and identification of the target genes that can be used for drug discovery \citep{Flor10}. Several Insilco methods are being used to reduce the cost of disease gene prioritization. For this, protein interaction networks were preferably used rather than other biological networks.

One can understand the functional importance of proteins from the network through different centrality measures. Centrality measures ranks network elements according to their importance within the network structure. In our study, we have used ten centrality measures to identify bottleneck proteins in the immunome network. The advantage of using 10 different centrality measures is that different centrality measures focuses on different important concepts and rank vertices differently based on their types such as Neighbourhood, Distance, Shortest-Path, Current-Flow, Feedback and Vitality. The biological relevance of the proteins can be extracted through GO terms, which provide information for protein functions, processes and localization. Our approach combines the ten centrality measures and the GO terms to find the potential genes that are responsible for SLE disease. It is worth mentioning that these 10 centrality measures give complementary information in ranking the nodes and this is evident from the correlation analysis.

In the present study we have predicted 8 novel SLE-related candidate genes, which are found to have numerous functions and several interaction partners of the immune system. These genes were further analysed and from the literature survey we have observed that among the 8 predicted genes, 3 genes SOCS3, TIRAP and PDGFRA were found to act as specific markers for the SLE disease [Ref]. Also, we have tried to identify the available drugs for the predicted SLE-related candidate genes using the online database DrugMapCentral, which is an unique informatics resource for translational studies on drugs (http://r2d2drug.org/DMC.aspx). But we found that no drugs have been discovered for these novel SLE-related candidate genes. The predicted novel SLE-related candidate genes can act as potential targets for the drug design and discovery because of their significant disease ontologies and high enrichment scores. The surface properties of target proteins can be studied by acquiring the structural information from PDB and one can identify the active site of the targets and design drugs for these targets.

The approach we have employed in our study uses the information about the interactions among the proteins of the human immunome and also their functional annotation. This method may be further applied to other immune related diseases which are completely dependent on malfunctioning of immune system such as autoimmune pancreatitis, Graves' disease, Multiple sclerosis, scleroderma.

Materials and methods

Collection of Immunome proteins and construction of Human immunome network

The immunome proteins were collected from the Immunome database \citep{Ortu06}. The database consists of 893 human proteins along with their gene annotations. Similarly, we have collected the human protein-protein interaction data from Human Protein Reference Database (HPRD) \citep{Pra09} which consists of 9617 proteins and 39,240 interactions. By mapping the immunome genes with the Human protein interaction data, we have extracted a sub-network named Human immunome network. After removal of all orphan nodes, the core human immunome network consists of 548 proteins with 1372 interactions (Figure~\ref{fig:01}) and which was used for further analysis.

Immunome related SLE disease genes

From the Auto-immune disease database \citep{Karo06}, 1402 SLE related proteins were collected, which contains not only the Human SLE genes but also microbial proteins as this disease is influenced by microbes. The microbial proteins were removed by comparing with human proteins taken from HPRD; as a result 1057 human related SLE genes were obtained. By comparing the human related SLE genes with immunome genes collected from the Immunome database, we found that only 329 genes were common and it is assumed to be known SLE genes.

Network properties and centrality measures

The scale-free nature of the human immunome network and other topological properties such as average degree, degree exponent, diameter, average clustering coefficient were calculated \citep{Dor03}. The importance of the nodes in a network can be quantified through the concept of centrality, which allows ranking of the nodes based on their network structure. Various centrality methods have been developed based on the concepts of connectivity, distance, current flow and feedback. In our study, we have used 10 graph centrality measures such as degree, eccentricity, closeness, centroid values, shortest-path betweenness, current-flow closeness, current-flow betweenness, Katz status index, Eigen vector, page rank. We have calculated the above centrality measures for Human Immunome Network using the tool CentiBIN \citep{Jun06}.

Degree centrality

The degree centrality measure ranks the potential of an individual node in the network based on its connectivity.

$$C_d(v) = deg(v).$$

Here 'deg' represents the degree of the node v.

Eccentricity

Eccentricity centrality ranks the potential of an individual node based on maximum shortest path length that a particular node takes to reach rest of the nodes 'w' in the network.

$$C_{ecc}(v) = \frac{1}{\max\{dist(v,w): v, w \in G\}}$$

Here dist (v, w) denotes the length of a shortest path between the nodes v and w.

Closeness centrality

According to closeness centrality the potential of an individual node based on how closely it is located to other nodes in the network. This centrality uses the sum of the minimal distances of a node to all other nodes for ranking.

$$C_{clo}(v) = \frac{1}{\sum_{w \in v} dist(v, w)}$$

Centroid values

The centroid value of a individual node 'v' is calculated by considering the number of nodes that have minimum shortest path that are closer to v than w.

$$C_{cen}\,(v) = \min\{\mathrm{f(v,w)}: \mathrm{V}\{\mathrm{v}\}\}$$

Where f (v, w):= $\gamma_v(w)$ - $\gamma_w(v)$ and $\gamma_v(w)$ denotes the number of node that are closer to v than w.

Shortest-path betweenness

The shortest path betweenness centrality ranks the potential of an individual node 'v' based on its ability to control signalling between any two nodes's' and 't' in the network.

$$C_{spb(v)=} \sum_{s \neq v \in G} \sum_{t \neq v \in G} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Here $\sigma_{st}(v)$ is the number of shortest paths between nodes 's' and 't' passing through the node 'v' and σ_{st} is the total number shortest paths between 's' and 't'.

Current Flow Closeness

The current flow closeness centrality ranks the potential of an individual node 'v' based on the concept of Kirchhoff's law in an electrical network. The centrality score for a node 'v' is calculated from the distance between two vertices 'v' and 'w' as the difference of their potentials p (v) − p (w).

$$C_{cfc}(v) = \frac{n-1}{\sum_{t \notin v} P_{vt}(v) - P_{vt}\,(t)}$$

Where P_{vt} (t) equals the potential difference in an electrical network and n-l is the normalization factor.

CF-Betweenness

The current-flow betweenness centrality ranks the potential of an individual node 'v' based on the amount of information flow through node 'v' averaged over the nodes's' and 't'.

$$C_{cfb}(v) = \frac{1}{(n-1)(n-2)} \sum_{s,t \in v} \tau_{st}(v)$$

Here τ_{st} (v) equals the fraction of information flow running over node ‘v’.

Katz status

Katz status centrality ranks the potential an individual node ‘v’ by considering its direct as well as indirect connections with the nodes in the network. Katz status centrality can be regarded as a generalization of the degree centrality.

$$C_{katz}(v) = \sum_{k=1}^{\infty} \alpha^k \, (A^T)^k \underset{\mathbf{1}}{\rightarrow}$$

Where α is a positive constant, A^T is the transpose of adjacency matrix and k is the degree of the node ‘v’.

Eigen vector

Eigenvector centrality ranks the potential of the individual nodes in the network through the Eigen vector elements of the largest Eigen value λ of the network.

$$\lambda C_{eiv} = AC_{eiv}$$

Page Rank

Page rank centrality measure ranks the potential of an individual node based on the concepts of the algorithm used internet search engine.

$$\boldsymbol{C_{pr}} = dpC_{pr} + (1 - d) \underset{\mathbf{1}}{\rightarrow}$$

Where P is the transition matrix and d is the damping factor.

GO terms enrichment and Predication of SLE candidate genes

The GO terms for all the 548 human immunome network genes were collected from Gene Ontology website ((http://www.geneontology.org)\citep{Ash00}. Using the Gene Ontology online tool, GOrilla \citep{Eden09} the gene enrichment was performed with the known SLE gene list as target and human immunome network genes as source using a p-value of 0.001. Significantly enriched GO terms were identified by using the threshold on the number of genes for each GO term i.e. the GO terms which were annotated for at least three but not for more than 50 immunome genes were considered. For each significantly enriched GO term of BiologicalProcess, Molecular Function and CellularComponent, the genes matching the corresponding GO terms of the Human immunome network were identified as the list of genes with specific disease ontologies.

The lists of the genes with the 50 highest scores for ten centrality measures were created for the Human immunome network. When combined these lists contained 113 unique genes. The significant SLE related GO terms were combined with the results for the protein-protein interaction network related scores in order to predict the new SLE related genes. A venn diagram was used in order to represent high network score genes, genes with significant disease ontologies and known SLE genes. The genes that were present in both the lists of high network scores and significant disease ontologies but were not in the list of known SLE genes were predicted as Novel SLE candidate genes.

Acknowledgement

The authors PLP, SBR and PM would like to thank Dept. of Science and Technology, Government of India, for their financial support (DST-CMS GoI Project No. SR/S4/MS: 516/07 Dated 21.04.2008).

References

1 E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous and B. S. Pickard, *Bioinformatics*, 2006, 22, 773-774.

2 M. Ashburne, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nature Genet*, 2000, 25,25-29.

3 J. Castro, E. Balada, J. Ordi-Ros, M. Vilardell-Tarrés, *Autoimmun Rev*, 2008, 7, 345-351.

4 J. Chen, E. E. Bardes, B. J. Aronow and A. G. Jegga,Nucleic Acids Res, 2006, 37, W305-W311.

5 D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju and D. S. Wishart, *Nucleic Acids Res*, 2008, 36 W399-W405.

6 J. C. Crispín, S. C. Liossis, K. Kis-Toth, L. A. Lieberman, V. C. Kyttaris, Y. Juang and G. C. Tsokos, *Trends Mol Med*, 2010, 16, 47-57.

7 S. N. Dorogovtsev and J. F. F. Mendes, Oxford University Press, Oxford, United Kingdom, 2003.

8 M. A. van-Driel, J. Bruggeman, G. Vriend, H. G. Brunner and J. A. M. Leunissen, Eur J Hum Genet, 2006, 14, 535-42.

9 E. Eden, R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini, BMC Bioinformatics, 2009, 10, 48.

10 A. F. Flórez, D. Park, J. Bhak, B. Kim, A. Kuchinsky, J. H. Morris, J. Espinosa and C. Muskus, BMC Bioinformatics, 2010, 11, 484.

11 A. M. Glazier, J. H. Nadeau and T. J. Aitman, Science, 2002, 298, 2345-2349.

12 K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A. L. Barabási, Proc Natl Acad Sci, 2007, 104, 8685-8690.

13 D. Hristovski, B. Peterlin, J. A. Mitchell, S. M. Humphrey, Int J Med Inform, 2005, 74, 289-98.

14 Q. Y. Huang1, G. H. Y. Li, W. M. W. Cheung1, Y. Q. Song and Annie W C Kung, J Hum Genet, 2008, 53, 644-655.

15 T. Ideker and R. Sharan, Genome Res, 2008, 18, 644-652.

16 B. H. Junker, D. Koschützki and F. Schreiber, BMC Bioinformatics, 2006, 7, 219.

17 T. Karopka, J. Fluck, H. T. Mevissen and Ä. Glass, BMC Bioinformatics,2006, 7, 325.

18 T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, Nucleic Acids Res, 2009, 37 D767-D772.

19 C. Ortutay and M. Vihinen, Cell Immunol, 2006, 244, 87-89.

20 C. Ortutay and M. Vihinen, Immunome Res, 2008, 4, 341-346.

21 C. Ortutay and M. Vihinen, Nucleic Acids Res, 2009, 37, 622-625.

22 M. Oti, B. Snel, M. A. Huynen, H. G. Brunner, J Med Genet, 2006, 43, 691-698.

23 P. Radivojac, K Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, S. D. Mooney, Proteins, 2008, 72, 1030-1037.

24 Y. Sha, Q. Liu, Y. Wang, C. Dong, L. Song , B J M G, 2010, 13, 4-10.

25 N. Tiffin, E. Adie, F. Turner, H. G. Brunner, M. A. van Driel, M. Oti, N. Lopez-Bigas, C. Ouzounis, C. Perez-Iratxeta, M. A. Andrade-Navarro, A. Adeyemo, M. E. Patti, C. A. M. Semple and W. Hide, Nucleic Acids Res, 2006, 34, 3067-3081.

26 L. C. Tranchevent, F. B. Capdevila, D. Nitsch, B. D. Moor, P. D. Causmaecker and Yves Moreau, Briefings in Bioinformatics, 2010, 12,22-32.

27 Y. Yoshida, Y. Makita, N. Heida, S. Asano, A. Matsushima, M. Ishii, Y. Mochizuki, H. Masuya, S. Wakana, N. Kobayashi and T. Toyoda, Nucleic Acids Res, 2009, 37, W147-W152.

28 M. Zhu and S. Zhao, Int J Biol Sci, 2007, 3, 420-427.