Journal of
Proteomics & Bioinformatics

Bragazzi et al. J Proteomics Bioinform 2011, 4:4
DOI: 10.4172/jpb.1000171

Research Article Open Access

LeaderGene: A Fast Data-mining Tool for Molecular Genomics

Nicola Luigi Bragazzi, Luca Giacomelli, Victor Sivozhelezov and Claudio Nicolini*

Chair of Biophysics University of Genova and Nanoworld Institute Fondazione EL.B.A.Nicolini , Italy

Abstract

DNA microarrays are one of the most promising methods for molecular genomics, but this technique is often associated with experimental complications and difficulties in the analysis. Moreover, the greatest part of genes displayed on an array is often not directly involved in the cellular process being studied. Recently, we proposed a data mining algorithm, based on the identification of genes involved in a given process, the calculation of interactions among them and their ranking according to number of interactions. Genes in the highest cluster are defined as "leader genes". These findings may lead to an ad hoc and therefore more significant experimentation. However, at present this complex process is performed manually. In this work, we present the general architecture of LeaderGene, an automated tool for ab-initio molecular genomics. Three different and independent parts: (1) Identification of gene list; (2) Calculation of weighted number of links; (3) Genes clustering. Initial inputs are provided by user; then, output of part 1 and part 2, respectively, become inputs of parts 2 and 3. The development of an user-friendly software capable to automatically compute leader genes in a given cellular system will allow further progresses in this field of molecular genomics.

Keywords: DNA microarrays; Molecular genomics; Data mining; Leader gene

Introduction

DNA microarrays have emerged as one of the most promising methods for the analysis of gene expression [1-4]. This technique allows the study of an immense amount of genes (over 10,000) with only one experiment and therefore can draw a picture of a whole genome. Anyway, the huge number of data coming out from microarray experiments may often raise experimental complications and difficulties in the analysis [2]. Moreover, the greatest part of genes displayed on an array is often not directly involved in the cellular process being studied. Commercial arrays with a lower number of gene usually of 150–200 — are currently available, but the genes displayed are usually once again chosen without a precise consideration of the particular target of the study.

Recently, we proposed a bioinformatics algorithm, based on the identification of genes involved in a given process and their scoring of importance via a cluster analysis, which allowed us to determine the most important genes, that we call "leader genes" [3-6]. The basis of the scoring system relies upon the calculation of interactions among genes, performed with software available in the web, such as STRING [7]. The number of links for each gene is then weighted and the final weighted numbers of links are clustered, in order to make a hierarchical classification of genes [5]. In this way, it becomes possible to draw and to update maps of the major biological control systems, and to integrate them in a concise manner to discern common patterns of interactions between gene expression and their correlated coding of proteins. These findings may lead to an ad hoc and therefore more significant experimentation [2,4,6]. However, at present this complex process is performed manually. Therefore, it is error-prone and time-consuming. In this work, we present the general architecture of LeaderGene, an automated tool for ab-initio molecular genomics.

The leader gene approach: general algorithm

The leader gene approach has been already described in detail [5]. The bioinformatics/data mining algorithm followed is summarized in Figure 1.

At first, the key genes involved in a given process are identified by iterative search of gene databases. In particular, several search strategies were implemented and iteratively repeated until convergence. First, research is performed in several databases such as PubMed, GenBank, Geneatlas, Genecards, using pertinent keywords and Mesh terms, as well as all their possible Boolean logics based combinations. In this way, it was possible to identify a preliminary set of genes representing every gene with an established role in the give process.

The preliminary set of genes was then expanded using the web-available software STRING (version 6.3), considering only direct interactions (i.e. physical contact between encoded proteins or

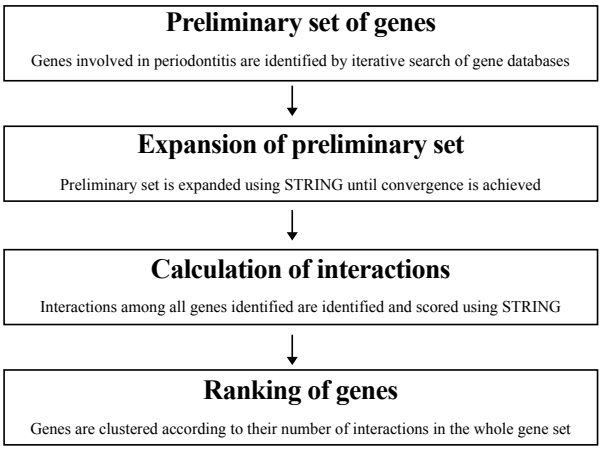

Figure 1: Flow chart of the leader gene approach.

***Corresponding authors:** Prof Claudio Nicolini, University of Genova, corso Europa 30, 16132 Genoa, Italy, Fax +39 010 353 38215; E-mail: manuscript@ibf.unige.it

Received April 02, 2011; **Accepted** April 23, 2011; **Published** April 25, 2011

Citation: Bragazzi NL, Giacomelli L, Sivozhelezov V, Nicolini C (2011) Leader Gene: A Fast Data-mining Tool for Molecular Genomics. J Proteomics Bioinform 4: 083-086. doi:10.4172/jpb.1000171

involvement in the same metabolic pathway), with a high degree of confidence. In this way, it is possible to identify new genes directly linked to those with an already established role in the process, and therefore potentially involved. In order to discard false positives, results are then filtered using a further search in PubMed records linked to every Gene record in the Entrez system. The process was repeated until no new gene potentially involved in the given process was identified.

Then, interaction map among identified genes are calculated using STRING. This software can give a combined association score to the given interaction, representing degree of confidence for each interaction. For every gene identified, we summed combined association scores, therefore obtaining a single score defined as weighted number of links.

Genes were then clustered, using hierachical or K-means algorithms, according to their weighted number of links. The genes belonging to the highest rank are termed leader genes; these genes have an highest weighted number of links if compared with the other. Therefore, they may be supposed to have an important role in the given keyword-specified cellular process such as human T lymphocytes cell cycle, human periodontitis, or human kidney transplant outcome.

The other ranks are termed class B, class C, class D genes and so on, according to their importance. Genes with no identified interactions (i.e. weighted number of links = 0) are defined as orphan genes.

Differences among various classes were statistically evaluated using an ANOVA test, with a Tukey- Kramer post hoc test. Statistical significance is usually set at a p value < 0.001, in order to ensure a high level of data reliability.

Main findings obtained with the LeaderGene approach

The leader gene theoretical approach has already been applied and validated on some human biological or pathological processes, such as human T lymphocytes cell cycle, human periodontitis and human kidney transplant tolerance and rejection [3-6, 8]. In this paragraph, we briefly present main findings obtained with this algorithm, with a special focus on human T lymphocytes cell cycle. In fact, theoretical results obtained in this system were extensively confirmed with targeted microarray analysis [3,6].

Leader gene identification was tried for the first time on human T lymphocytes cell cycle [5]. This particular system was chosen because it is much known and was quantitatively characterized time ago [9-12]. In total 238 genes involved in human were identified as involved in this process. The calculation of interactions among these genes and the subsequent clustering according to each weighted number of links identified 27 orphan genes (i.e. genes with no known interactions in the set being analyzed) and, invariably, six leader genes (MYC, CDK4, CDK2, CDC2, CDKN1A, CDKN1B). Interestingly, all leader genes are actually involved in the cell cycle control at important progression points, namely the most important four at the transition from G0 to G1 phase (MYC) [13], at the progression in G1 phase (CDK4) [14], and at the transitions from G1 to S (CDK2) [15], and from G2 to M phases (CDC2) [16,17]. The two remaining "leader genes" (CDKN1A and CDKN1B) are inhibitors of cyclin-CDK2 or –CDK4 complexes and thereby contribute to the control of G1/S transition and of G1 progression [18,19].

This theoretical prediction was validated with a DNA microarray gene expression analysis on human T lymphocytes stimulated with PHA to enter cell cycle, 24, 48 and 72 hours after the stimulation [2,6]. We chose a commercial array, displaying only 161 genes (among whom, 3 leader genes were displayed: MYC, CDK4 and CDC2), in order to limit experimental complications that may arise using a mass scale or a pangenomic microarray. Leader genes expression was monitored in order to identify changes over time: experimental analysis confirmed our bioinformatics predictions, therefore validating our approach. Moreover, our analysis was also focused on genes of lower importance than leader genes, but still with a high number of interactions calculated (namely, class B genes, class C genes and so on, with decreasing importance). This genes formed a close interaction network with leader genes [6]. Once again, the microarray analysis showed good agreement with theoretical prediction of their interactions with leader genes.

For what concerns periodontitis, 61 genes were identified as involved in the process [8]. Seven genes were listed as orphans; cluster analysis of the weighted number of links identified invariably the same 5 genes belonging to the highest cluster, i.e. to be leader genes: NFKB1, CBL, GRB2, PIK3R1 and RELA. We also identified 12 class B genes. A literature search confirmed that every gene we identified as a leader gene or class B gene could play an important role in periodontitis at a molecular level [8]. At present, we are carrying on a targeted RT-PCR experimentation on bioptical samples to further elucidate molecular mechanisms underlying their role in periodontitis.

When applied to human kidney transplant tolerance and rejection, the ledader gene approach was applied in combination with Statistical Analysis of Microarray (SAM) and with a new-developed non-statistical analysis in order to provide the most possible detailed picture of this process [5]. Overall, results suggested that leader gene and those largely changing expression form a unique network and that the mere changing in expression of a particular gene is not significant by itself, but only if it is put in a proper framework. This change can be often considered as a consequence of a more complex network of events, which starts from bioinformatics-identified leader genes. Noteworthy, leader genes do not always vary their expression so much to be identified as significant using pangenomic arrays. However, microarray technology is a necessary confirmation of every prediction made by the theoretical network analysis.

LeaderGene software: general architecture

LeaderGene software should be capable to perform all previously described procedures automatically. It is particularly important that the software must be user-friendly, to allow largest possible diffusion by researchers involved in different fields and possibly without a specific experience in data mining and advanced database querying. We are planning to develop codes in JAVA, for Windows platforms.

This software should be composed by three different and independent parts: (1) Identification of gene list; (2) Calculation of weighted number of links; (3) Genes clustering. Initial inputs are provided by user; then, output of part 1 and part 2, respectively, become inputs of parts 2 and 3. Final output is genes clustered according to WNL. Single outputs of part 1 and 2 are anyway saved and available to the user (Figure 2).

Part 1. Identification of LeaderGenes

The goal of part 1 is to identify all genes involved in a given process, on the base of a key-word based query in web-available databases. This part is supposed to be the most difficult to implement. Constant information exchange between software and the web is required, since most databases queried have only server-copy and do not allow a local copy to be downloaded.

Initial inputs are organism(s) in which the process being studied is occurring, general keywords to define search strategy and database

to be queried (GenBank, Geneatlas, Genecards). LeaderGene software will be capable to automatically query selected databases with chosen keywords and retrieve a preliminary list of genes involved in the process under study. Preliminary gene list is then submitted to STRING database in order to calculate the interaction map. This map is then expanded automatically by STRING. In this way genes directly linked with the ones in preliminary list, and therefore potentially involved in the process are identified. The user should provide a threshold of confidence for expansion. Newly identified genes will be then cross-checked against PubMed, using another set of more general keywords provided by user, in order to discard false positives. Genes with no citation according to keywords will be considered as unreliable and therefore discarded.

Expansion procedure will be performed recursively until no new gene is identified. Final output is a list of genes involved or potentially involved in the process under study. Flow chart of part 1 is represented in Figure 3.

However, some nomenclature problems may arise, since different databases uses several nomenclature annotations. This issue may be solved either with the development of an internal database of genes,

	Input	Output
Identification of **gene list**	Organism Keywords Databases Keywords for cross-check Confidence threshold	**List of relevant genes**
Calculation of **WNL**	**List of relevant genes** Confidence threshold Interaction map (yes/no) Map is just list of pairs in Medusa format	**Relevant genes + WNL** **Interaction map** (optional)
Genes **clustering**	List of relevant genes + WNL	**Genes clustered**

WNL: Weighted Number of Links. Input in red are provided by user

Figure 2: General architecture of the LeaderGene software.

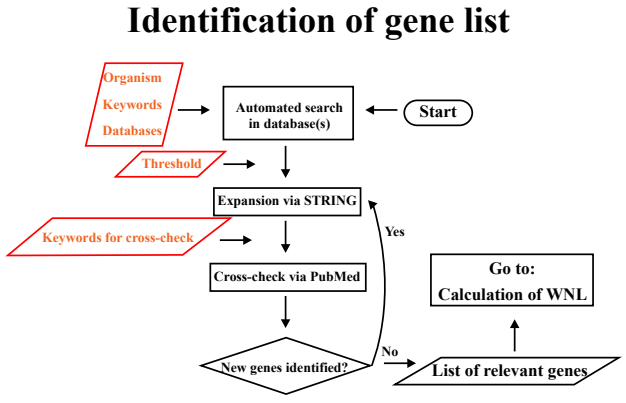

Figure 3: Flow chart of the LeaderGene part 1. Input in red are provided by user.

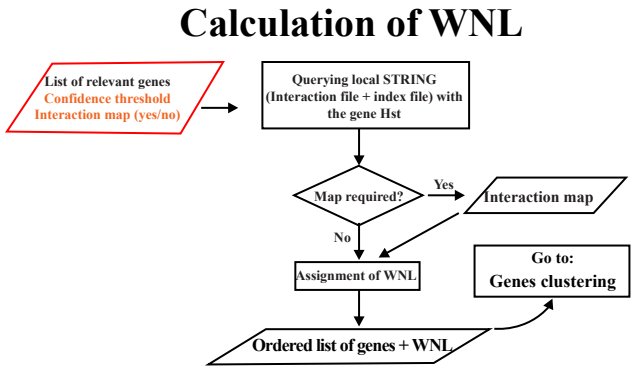

Figure 4: Flow chart of the LeaderGene part 2. Input in red are provided by user.

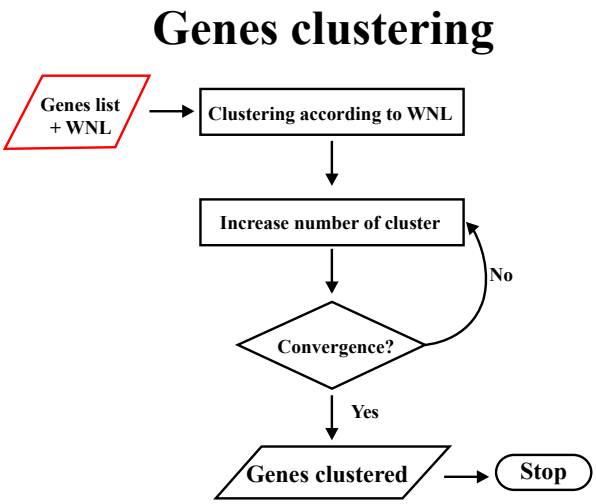

Figure 5: Flow chart of the LeaderGene part 3.

reporting all different annotation linked to a given object or via a BLASTP search on the identified genes, in order to retrieve encoded protein sequences, which will then submitted to STRING.

Part 2. Calculation of Weigthed Number of Links

Final list of genes involved in the process being studied is then submitted to STRING database, in order to calculate the final map of interaction and combined association scores for each interaction. User should provide a confidence threshold. From combined association scores, it is possible to calculate weighted number of links for each gene.

STRING database gives as output a map of interactions, in MEDUSA format, and a tab-delimited text file containing single gene pairs and their combined association score. If required by user, gene interaction map will be dumped and saved. Part 2 will perform a summation on text file provided by STRING database. Final list of genes with weighted number of links will be given as an output. Flow chart of the part 2 is represented in Figure 4.

Part 3. Gene clustering

In part 3, a K-means clustering algorithm is applied to genes, to rank genes according to the weighted number of links. Number of clusters will be increased by one until convergence is achieved (i.e. number of genes in highest cluster constant after 3 successive trials). It is noteworthy that, as a further validity proof, a significance test will be applied to ensure statistical difference between highest cluster and other ones in term of weighted number of links. Flow chart of the part 3 is represented in Figure 5.

Conclusions

Bioinformatics can give an added value to molecular biology. In particular, the detailed analysis of gene interaction maps and the ranking of genes according to their relevance may have great importance in the identification of new targets for a focused experimental analysis, which may suggest potential risk factor and therapy targets. At the same time, orphan genes should be characterized properly. Therefore, these findings can suggest targeted experimental analysis. In particular, it would be possible to create ad hoc arrays, both DNA- and protein-based, which can guarantee the best results in analyzing a particular cellular system and would allow to describe complex biomolecular pathways through the activity of a few, but highly important genes, which represent the real centre of interactions maps or are characterized by a significant change in expression. With respect to protein microarrays, the leader gene approach could simplify their analysis by reducing the protein displayed to the most important ones to be subsequently tested by mass spectrometry or by ad hoc NMR or X-ray crystallography experimentations.

The development of an user-friendly software capable to automatically compute leader genes in a given cellular system will allow further progresses in the entire fields of molecular genomics and proteomics.

Acknowledgments

This project was supported by grants to Fondazione Elba by MIUR for "Funzionamento" and by a FIRB RBIN04RXHS and FIRB Nanoitalnet from MIUR (Ministero dell'Istruzione, Università e Ricerca; Italian Ministry of Research and University) to Claudio Nicolini of the University of Genova.

References

1. Butte A (2002) The use and analysis of microarray data. Nat Rev Drug Discov 1: 951-960.
2. Nicolini C, Spera R, Stura E, Fiordoro S, Giacomelli L (2006) Gene expression of human T lymphocytes: experimental determination by DNASER technology. J Cell Biochem 97: 1151-1159.
3. Nicolini C (2006) Nanogenomics for medicine. Nanomedicine 1: 147-151.
4. Sivozhelezov V, Braud C, Giacomelli L, Pechkova E, Giral M, et al. (2008) Immunosuppressive drug-free operational immune tolerance in human kidney transplants recipients. Part II. Non-statistical gene microarray analysis. J Cell Biochem 103: 1693-1706.
5. Sivozhelezov V, Giacomelli L, Tripathi S, Nicolini C (2006) Gene expression of human T lymphocytes: predicted gene and protein networks. J Cell Biochem 97: 1137-1150.
6. Giacomelli L, Nicolini C (2006) Gene expression in the cell cycle of human T lymphocytes: experimental and bioinformatics analysis. J Cell Biochem 99: 1326-1333.
7. Von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 33: D433-435.
8. Covani U, Marconcini S, Giacomelli L, Sivozhelezov V, Barone A, et al. (2008) Bioinformatic prediction of leader genes in human periodontitis. J Period 79: 1974-1983.
9. Cantrell D (2002) Protein kinase B (Akt) regulation and function in T lymphocytes. Semin Immunol 14: 19-26.
10. Isakov N, Altman A (2002) Protein kinase C(theta) in T cell activation. Annu Rev Immunol 20: 761-794.
11. Oosterwegel MA, Greenwald RJ, Mandelbrot DA, Lorsbach RB, Sharpe AH (1999) CTLA-4 and T cell activation. Curr Opin Immunol 11: 294-300.
12. Abraham S, Vonderheid E, Zietz S, Kendall FM, Nicolini C (1980) Reversible (G0) and non-readily reversible (Q) noncycling cells in human peripheral blood. Immunological, structural, and biological characterization. Cell Biophysics 2: 353-371.
13. Oster SK, Ho CS, Soucie EL, Penn LZ (2002) The myc oncogene: Marvelousl Y Complex. Adv Cancer Res 84: 81-154.
14. Modiano JF, Mayor J, Ball C, Fuentes MK (2000) CDK4 expression and activity are required for cytokine responsiveness in T cells. J Immunol 165: 669-702.
15. Kawabe T, Suganuma M, Ando T, Rimura M, Hori H, et al. (2002) Cdc25C interacts with PCNA at G2/M transition. Oncogene 21: 1717-1726.
16. Baluchamy S, Rajabi HN, Thimmapaya R, Navaraj A, Thimmapaya B (2003) Repression of c-Myc and inhibition of G1 exit in cells conditionally over-expressing p300 that is not dependent on its histone acetyltransferase activity. Proc Natl Acad Sci USA 100: 9524-9529.
17. Torgler R, Jakob S, Ontsouka E, Nachbur U, Mueller C, et al. (2004) Regulation of activation-induced Fas (CD95/Apo-1) ligand expression in T cells by the cyclin B1/Cdk1 complex. J Biol Chem 279: 37334-37342.
18. Jerry DJ, Dickinson ES, Roberts AL, Said TK (2002) Regulation of apoptosis during mammary involution by the p53 tumor suppressor gene. J Dairy Sci 85: 1103-1110.
19. Chang BL, Zheng SL, Isaacs SD, Wiley KE, Turner A, et al. (2004) A polymorphism in the CDKN1B gene is associated with increased risk of hereditary prostate cancer. Cancer Res 64: 1997-1999.