Gene Expression in the Cell Cycle of Human T Lymphocytes: I. Predicted Gene and Protein Networks

V. Sivozhelezov,[2] L. Giacomelli,[1] S. Tripathi,[1] and C. Nicolini[1,2]*

[1]Nanoworld Institute and Biophysics Division, University of Genova, Corso Europa, 30, 16132 Genoa, Italy
[2]Fondazione Elba, Via delle Testuggini snc, 00100 Rome, Italy

Abstract The key genes involved in the cell cycle of human T lymphocytes were identified by iterative searches of gene-related databases, as derived also from DNA microarray experimentation, revealing and predicting interactions between those genes, assigning scores to each of the genes according to numbers of interaction for each gene weighted by significance of each interaction, and finally applying several types of clustering algorithms to genes basing on the assigned scores. All clustering algorithms applied, both hierarchical and K-means, invariably selected the same six ''leader'' genes involved in controlling the cell cycle of human T lymphocytes. Relations of the six genes to experimental data describing switching between stages of cell cycle of human T lymphocytes are discussed. J. Cell. Biochem. 97: 1137–1150, 2006.

Key words: cell cycle; lymphocytes; bioinformatics; gene interactions

Resting T lymphocytes stimulated to proliferate by phytohemagglutinin (PHA) represents an excellent model system for the cell cycle, studied at the chromatin level since long time in normal and transformed human cells by laser flow microfluorimetry [Abraham et al., 1980; Vonderheid et al., 1981]. Structural changes in chromatin independently suggested also from different sensitivity of quiescent versus cycling human T lymphocytes chromatin to acid denaturation [Darzynkiewicz et al., 1979] has been recently confirmed by the exhibition of the different spatial organization of genes in the nucleus [Brown et al., 1999]. The difference between cycling versus resting T cells has been recently also shown by effect of the beta-interferon, present in cycling T cells but not in resting T cells [Cooper et al., 2004]. Moreover expression of particular genes was shown to be cell cycle dependent, e.g., the link between ATR and p53 genes was present in cycling normal lymphocytes but absent in resting and malignant lymphocytes [Jones et al., 2004a,b] and similar effect was observed for DNA topoisomerase I [Bruno et al., 1992], ADP-ribosyl transferase [Scovassi et al., 1987], and nuclear antigen p105 [Clevenger et al., 1987].

The key question in cell biology is the effect on changes in the overall state of the cell such as the phase of the cell cycle on gene expression and its regulation. In this study, we perform such a search to identify interacting genes involved in human T cell lymphocyte activation. Human T lymphocytes constitute an ad hoc model due to the fact that their progression through the cell cycle is easily initiated by activation [Oosterwegel et al., 1999; Cantrell, 2002; Isakov and Altman, 2002] and, in particular, was quantitatively characterized time ago [Abraham et al., 1980]. Experimental investigations in this area used genome-wide measurements of gene expression levels with DNA microarrays from which it is possible to

Dr. Tripathi was on summer study at Nanoworld Institute. His current affiliation is: Indian Institute of Technology, Guwahati, India.

Grant sponsor: FISR-MIUR grant on DNA Nanotechnology to Nanoworld Institute-CIRNNOB of the University of Genova; Grant sponsor: FIRB-MIUR grant on Organic Nanoscience and Nanotechnology to the Nanoworld Institute, namely to the Fondazione EL.B.A and to the Centro Interuniversitario di Ricerca sulle Nanotecnologie e Nanoscienze Organiche e Biologiche of Genova University.

*Correspondence to: C. Nicolini, Nanoworld Institute and Biophysics Division, University of Genova, Corso Europa 30, 16132 Genova, Italy. E-mail: manuscript@ibf.unige.it

Received 28 July 2005; Accepted 20 September 2005

DOI 10.1002/jcb.20693

infer data on interactions between the genes and the resulting proteins [Butte, 2002]. To map such interactions directly from microarray experiments, researchers involved in the area compose sophisticated software [Jones et al., 2004a,b; Troyanskaya, 2005]. This resulted in accumulation of immense amount of data on gene and protein interaction (about 800,000 interacting gene and protein pairs are currently known). While database build-up continues, examples of focused searches of the already existing databases with statistics-based predictions of interactions aiming to reveal genes specific and important to the particular physiologically significant process are rare.

Cluster analysis is a common tool in interpreting the experimental microarray data [Reimers, 2005] but also was successfully used in mass scale text and data mining, for example to select disease gene candidates [Tiffin et al., 2005]. In the latter study, clustering was applied to pairs of interacting genes while we assign interaction-based scores to individual genes, and cluster genes according to those scores.

With this manuscript we use cluster analysis to determine the most important genes that we term "leader genes" and we approach the task to compile and update maps of the major biological control systems, in order to integrate them in a concise manner, to discern common patterns of interactions between gene expression and their correlated coding of proteins during cell cycle progression.

METHODS

Several existing and representative experimented databases (see Table I) are accessed via the search engine Entrez [www.ncbi.nlm.nih.gov/gquery/gquery/fcgi] in order to:

1. identify the genes involved in human T cells cell cycle,
2. predict possible interactions between the genes of Item 1,
3. identify the possible leader genes i.e. those having the maximum number of interactions.

Genes Involved in Human T Cells Cell Cycle

Several search strategies were implemented and iteratively repeated until the newly identified genes were already found in the previous search. These strategies included:

a) direct Genebank search with pertinent keywords (resulted in 43 genes),
b) branching of the results (a) with respect to all Internet-available genome databases with immediate cross-checking via the database PubMed (47),
c) scanning of the gene lists dedicated to lymphocyte activation and general cell cycle from commercially available DNA microarrays [www.superarray.com] with respect to pertinence to cell cycle for the "lymphocyte activation" SuperArray^{TM} microarray (84 more genes) and to lymphocyte activation for the "cell cycle" microarray (18),
d) branching of the search results from items (a–c) via the Gene Ontology database (46, totaling 254).

Particular attention should be paid to Item (c) from which it follows that not all the genes present in the dedicated microarrays are in fact related to the function specified by their manufacturers. For example, those microarrays often contain ribosomal proteins, which definitely belong to the catalytic machinery of gene expression, and thereby are involved in any biological process so their inclusion in the "cell cycle" microarray as "lymphocyte activation" microarray seems questionable.

Item (c) emerged because the experimental observations pertain to lymphocyte activation [Abraham et al., 1980; Oosterwegel et al., 1999; Cantrell, 2002; Isakov and Altman, 2002]. The adopted procedure as quoted in (c) implies that,

TABLE I. Databases Used for the Identification of Genes Involved in Human T Lymphocytes Cell Cycle and Their Leaders

Gene	Genes from GeneBank and associated information for a number of organisms including human.
HomoloGene	Contains homologs among the annotated genes of several completely sequenced eukaryotic genomes
MeSH	National Library of Medicine vocabulary of terms used for indexing articles in PubMed.
Nucleotide sequence database	A collection of nucleotide sequences from several sources, including GenBank, RefSeq, and PDB.
Protein sequence database	A collection of protein sequence entries compiled from a variety of sources including Swiss-Prot, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq.
PubMed	Access to all citations from MEDLINE database and additional life sciences journals.

although it is well established that the process of lymphocyte activation is closely linked to the lymphocyte cell cycle, not all genes involved in lymphocyte activation are directly involved in the lymphocyte cell cycle. Obvious examples are genes responsible for early stages of immune response, which have no direct relation to the lymphocyte cell cycle but are directly involved in lymphocyte activation. Note that the nomenclature used follows that currently adopted in GeneBank. For example the known gene FASLG is presented as TNFSF6, but the gene FAS listed in this report has a synonym of TNFRSF6.

Prediction of Possible Interactions

The weighted number of links is calculated for each gene using the program STRING [von Mering et al., 2005]. This value derives from the weighed sum of three types of interactions.

(1) co-occurrence of the names of gene or respective proteins in abstracts of papers available via Internet. The scores assigned are derived from benchmarked scoring system based on the frequencies and distributions of gene names in abstracts. The benchmarks themselves are set from manual evaluation of predictions of gene and protein interactions by experts [Donaldson et al., 2003], and are typically below 0.5.

(2) Scores derived from databases of gene networks, e.g., KEGG [Kanehisa et al., 2004] containing data on induction of a particular genes by other genes derived from microarray experiments or other high throughput techniques. The score of 1 is assigned if the link is already present in the database while putative links have lower values (typically 0.6–0.8).

(3) The same range of scores is assigned to gene interactions via physically observed interactions between proteins. The software used does not discriminate between in vivo or in vitro experiment derived data. Generally the scores are close to those of interaction type 2, but links of this type occur much rarely than of type 2 (see later).

Finally, the combined interaction scores S_{ij} for the gene pair (i,j) are integrated according to the formula:

$$S_{ij} = 1 - \Pi_k(1 - S_{ijk})$$

where S_{ijk} is the interaction score of the pair (i,j) of the type k. In this case $k = 1 \ldots 3$ according to the above list of score definitions.

Identification of Leader Genes

The combined association scores S_{ij} were summed for each gene i over its neighbors (i,j) giving the final weighted number of links for the gene i. Further, we applied clustering methods to the weighted number of links in order to identify the group of leader genes.

Cluster analysis [Datta and Datta, 2003; Tsai et al., 2005], also called segmentation analysis or taxonomy analysis, is a way to partition a set of objects into homogeneous and separated groups, or *clusters*, in such a way that the profiles of objects in the same cluster are very similar and the profiles of objects in different clusters are quite distinct. In particular, cluster analysis can be defined as follows:

Cluster analysis: Given a set S of n objects $\{x_1, x_2, \ldots, x_n\}$, where each object is described by m attributes $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, determine a classification that is most likely to have generated the observed objects.

Many different fields of study, such as engineering, zoology, medicine, linguistics, anthropology, psychology, and marketing, have contributed to the development of clustering techniques and their applications. For example, cluster analysis can be used to find two similar groups for the experiment and control groups in a study. In this way, if statistical differences are found in the groups, they can be attributed to the experiment and not to any initial difference between the groups.

The identifications of similar profiles are achieved through the comparison of the row or column vectors by means of a distance function d. The distance function is a particular form of a metric function.

A metric d is a function satisfying:

1. non-negativity: $d(a; b) \neq 0$;
2. symmetry: $d(a; b) = d(b; a)$;
3. $d(a; a) = 0$;
4. definiteness: $d(a; b) = 0$ if and only if $a = b$;
5. triangle inequality: $d(a; b) \geq d(b; c)\_d(a; c)$.

A function only satisfying 1–3 is called a distance.

Among the most used distance functions, we can find Euclidean distance and the Pearson correlation distance. If we define two vectors x and y, as:

$$x = (x_1, \ldots, x_n); \quad y = (y_1, \ldots, y_n)$$

then Euclidean distance can be defined as follows:

$$d(x, y) = (\Sigma(x_i - y_i)^2)^{1/2}$$

and the Pearson's correlation coefficient is defined as:

$$d_\rho(x, y) = (1 - \rho(x, y))/2$$

where:

$$\rho(x, y) = (x - x_{mean}) \cdot (y - y_{mean})/\sigma_x \cdot \sigma_y$$

Note that:

$$d_\rho(x, y) \in [0, 1] \text{ where } d_\rho(x, y) = 0$$

implies perfect similarity and that: $d_\rho(x, y) = 1$ implies maximal dissimilarity.

The Pearson measure in comparison to the Euclidean measure reflects more the "shape" rather than the "distance" of the vectors, since it is invariant under multiplication with a constant.

In cluster analysis, distance measure d, calculated between every single pair of object, must be extended to a measure of distance between clusters. There are several ways to do this (see Table II).

Once the distance function and the linkage methods have been set, it is necessary to define a strategy to build clusters. There are several methods to perform clustering analysis. We now will only consider two of the most important ones: hierarchical clustering and k-means clustering.

The hierarchical clustering algorithm either iteratively joins the two closest clusters starting from single clusters (bottom-up approach) or iteratively partitions clusters starting from the complete set (top-down approach) [Shannon et al., 2003]. After each step a new distance matrix between the newly formed clusters and the other clusters is recalculated.

Given a set of objects D, the algorithm can be summarized as follows:

1. Find a minimal entry d(i, j) in D and merge clusters i and j.
2. Compute D' from D by deleting row i and column j and adding a new row and column $i \cup j$ with the new entries being $d(k, i \cup j)$.
3. Repeat steps 1 and 2 until D' consists only of one entry.

The $d(k, i \cup j)$ are of course dependent on the choice of the method. For:

- Single Linkage: $d(k, i \cup j) = \min(d(k, i), d(j, k))$
- Complete Linkage: $d(k, i \cup j) = \max(d(k, i), d(j, k))$
- Average Linkage: $d(k, i \cup j) = (ni \cdot d(k, i) + nj \cdot d(j, k)/(ni + nj)$
- Centroid Linkage: $d(k, i \cup j) = ni \cdot d(k, i)/(n_i + n_j) + nj \cdot d(k, j)/(ni + nj) - ni \cdot nj \cdot d(i, j)/(ni + nj)^2$.

The resultant hierarchical clustering can be easily visualized using dendrograms, which represent all objects as leaves of a large, branching tree. The appropriate number of clusters can be obtained by cutting the dendrogram at a certain level to obtain the desired number of clusters.

There is no standard criterion or algorithm for choosing a cut-off point for a dendrograms, it is often set by the user considering the conditions and the goals of every single experiments.

K-means clustering uses a different approach [Tsai et al., 2005]. In k-means clustering, initial cluster centroids are selected, and the proximities (similarity or distance) from each object to all k centroids are calculated. Each object is then assigned to the cluster to which it is the closest. The k new centroids are formed with new cluster members and the objects are reallocated to one of the new k clusters. This iterative process stops if there is no reallocation of objects

TABLE II. Different Ways of Linkage in Cluster Analysis

Single linkage	The distance between two clusters is the minimal distance between two objects, one from each cluster
Average linkage	The distance between two clusters is the average of the pairwise distance between members of the two clusters
Complete linkage	The distance between two clusters is the maximum of the distances between two objects, one from each cluster
Centroid linkage	The distance between two clusters is the distance between their *centroids*. A centroid is the average point in the multidimensional space defined by the dimensions. Some time it is also stated as the *center of gravity* for the respective cluster.

or if the reassignment satisfies the criteria set by the user.

The aim of k-means clustering is to partition a set of clusters and to find the means μi of each cluster. The first step is to choose a dissimilarity measure d.

Algorithm:

1. Randomly pick $\mu_1, \ldots, \mu_k$, where μ are the means of the cluster C_i..
2. For each $x \in E$ compute $\min_i d(\mu_i, x)$.
3. Recompute new mean of clustering Ci: $\mu_i = 1/[C_i]\ \Sigma x$ where $x \in C_i$.
4. Repeat steps 2 and 3 until "convergence".

K-means clustering is a useful method if the user has an a priori idea about the number of clusters in which genes have to be divided.

Interactions between genes involved in human T lymphocyte cell cycle were evaluated

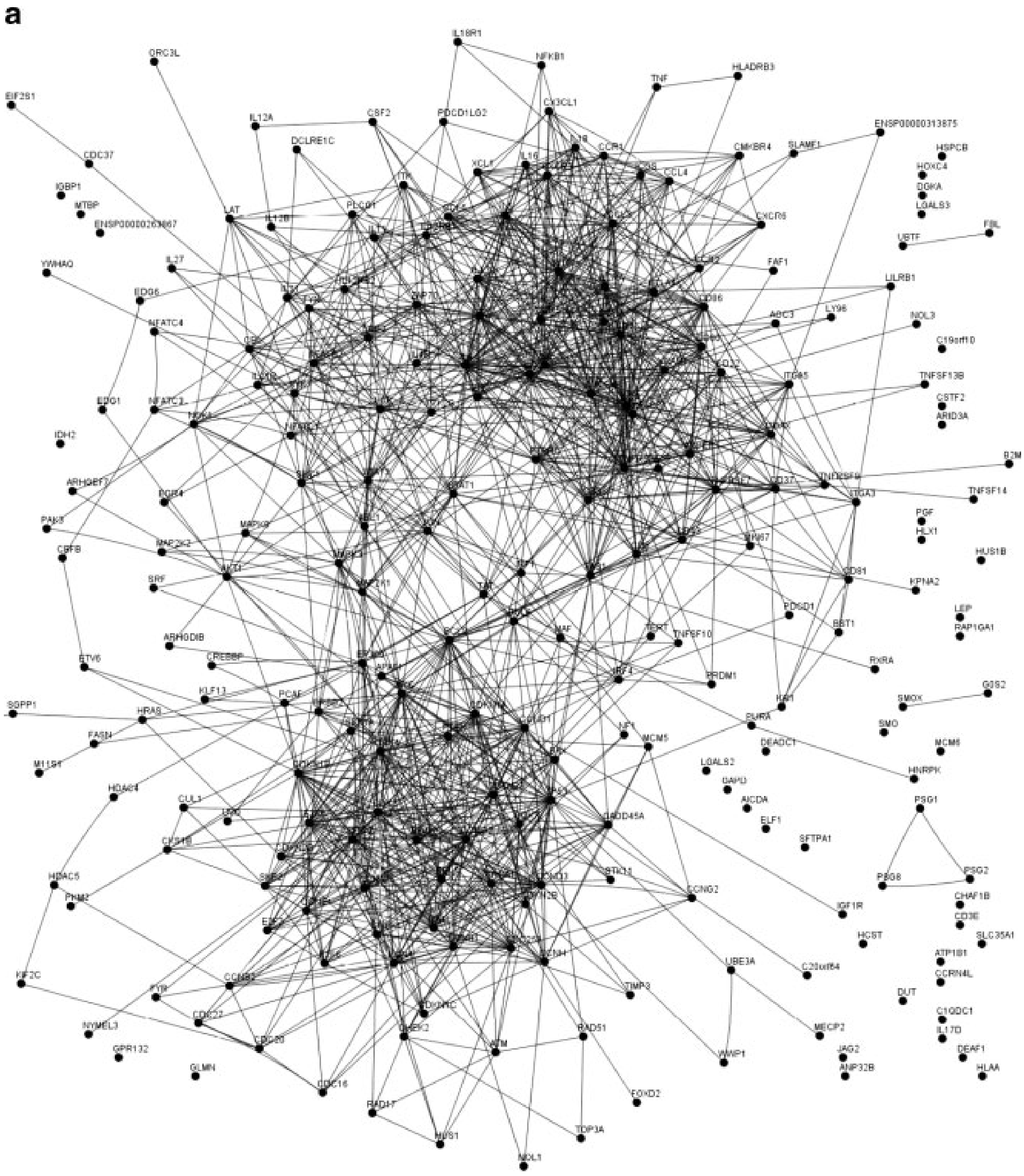

Fig. 1. **a**: Relations between the genes involved in human T lymphocyte cell cycle according to co-mentioning in paper abstracts. **b**: Relations between the genes involved in human T lymphocyte cell cycle according to gene–gene interaction (induction or suppression). **c**: Relations between the genes involved in human T lymphocyte cell cycle according to protein–protein interactions for proteins encoded by the corresponding gene.

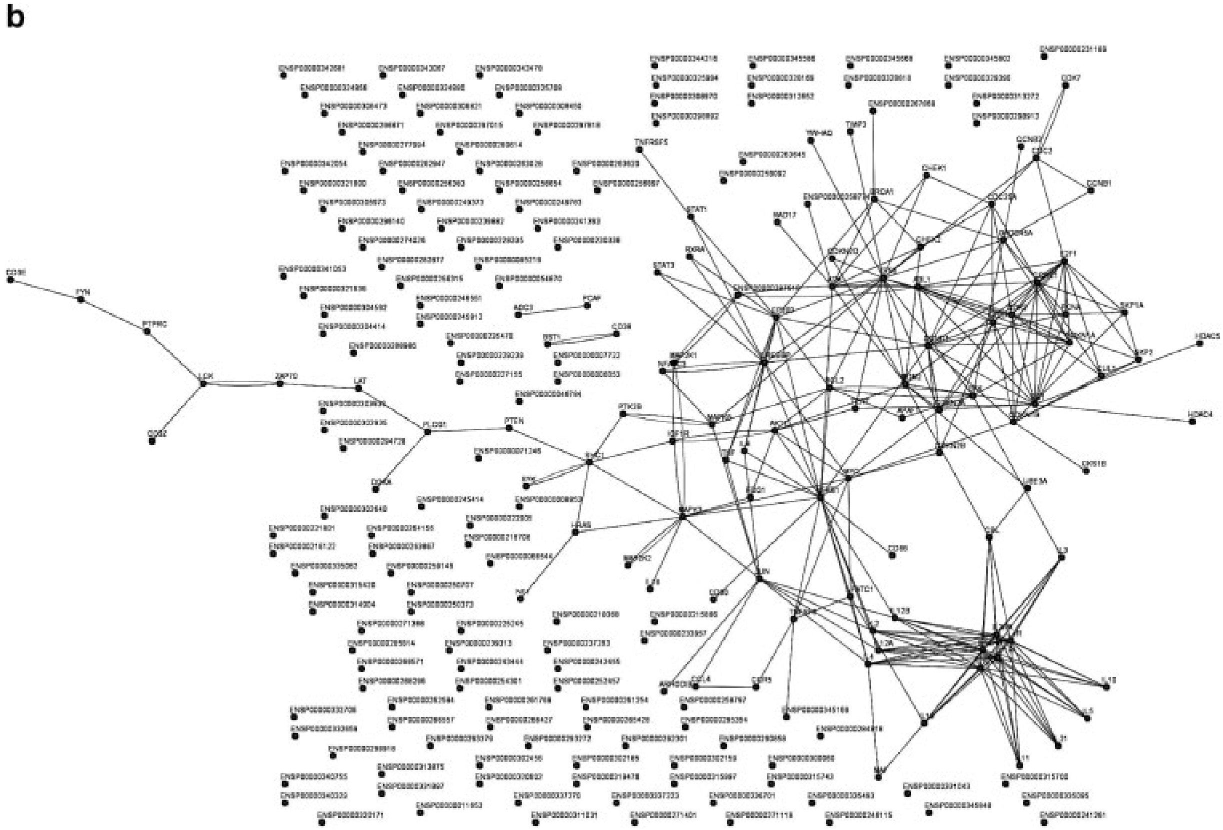

Fig. 1. (*Continued*)

using the program STRING (Figs. 1 and 2) (Search Tool for the Retrieval of Interacting Genes, Heidelberg, Germany) [von Mering et al., 2005].

The interactions among genes are summarized for each gene according to its successive number in the list we created (Fig. 3).

RESULTS

A total of 256 genes appear involved in the cell cycle of human T cells as shown in the methods.

Interactions between genes involved in human T lymphocyte cell cycle were evaluated using the program STRING (see Figs. 1 and 2) [von Mering et al., 2005]. The interactions among genes are summarized for each gene according to its successive number in the list we created (Fig. 3).

In order to identify the most important genes among the 256 we found, we decided to apply cluster algorithms on the weighted numbers. In particular, we wanted to determine a subset of "leader" genes, the ones with the highest withed number of links, which can be considered the most significant genes in human T lymphocytes cell cycle (Fig. 3). Other genes are thus dependent on these leader ones.

We used two different algorithms, hierarchical and k-means clustering (see Methods). In this way, we gave as an input the list of genes with their own weighted number of links. Genes are then grouped, according to the algorithm, in different subsets, on the base of the weighted number of links. First of all, we used K-means clustering. We started with two clusters and then increase their number, with a maximum of 500 cycles for every experiment. The cluster with the highest rank of weighted number of links is defined the leader cluster. By increasing the number of clusters, the number of genes belonging to the leader cluster is supposed to become lower and lower, until it becomes stable. Genes belonging to the leader cluster at this point are the leader genes (Fig. 4 and Table IV). It follows that there are only six genes in the leader cluster. In order to validate our data, we performed the Kruskal–Wallis test on the results of different experiments, to see if there is statistical significant difference among different clusters. The Kruskal–Wallis test is a more

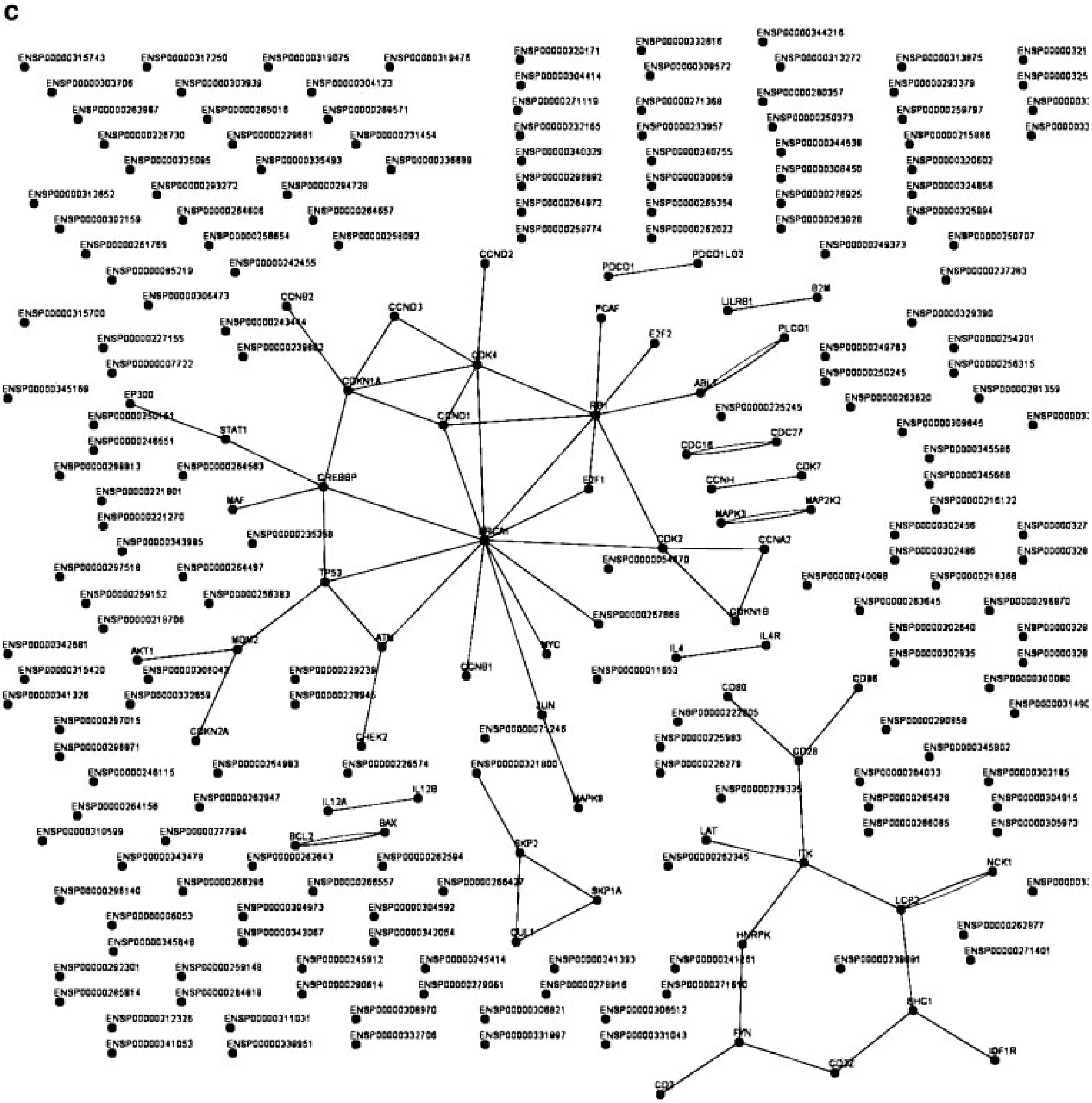

Fig. 1. (*Continued*)

general form of the ANOVA test, which does not require a Gaussian distribution of data. This test proved that different clusters medians always varied significantly ($P < 0.05$, data not shown).

Then, we applied to the whole set of data another algorithm, hierarchical clustering. We used different methods to compute linkage; we set an increasing cutoff to the maximum number of cluster to be calculated. The leader cluster contained the same six genes (Table IV) as the K-means calculated one, using the smallest distance linkage (2 clusters) and both the centroid and the complete linkage (3, 4, and 5 clusters each) (see Table III). Even for these experiments the Kruskal–Wallis test gave statistical significance ($P < 0.05$, data not shown).

Thus, it's possible to affirm that the six genes with the highest number of links are the most important gene in controlling human T lymphocytes cell cycle (see Table IV). A wider analysis of the interactions among leader genes, performed with STRING, revealed also genes important in their interactions, or neighboring genes (Fig. 5).

Moreover, nearly the entire list of the genes we created (about 90%) is involved in some interactions, leaving only 27 unlinked genes, or orphan genes (see Table V).

DISCUSSION

Cell cycle progression depends on a complex network of interacting proteins and genes, which regulate crucial activities such as DNA synthesis, gene expression, metabolism, and information processing. Disruptions in the intricate balance between the components of this network may lead to cancer, terminal differentiation, and/or aging; however, interfering with signals transmitted by bioregulatory networks is an important tool for the control of cell growth and of cancer therapy. In recent years,

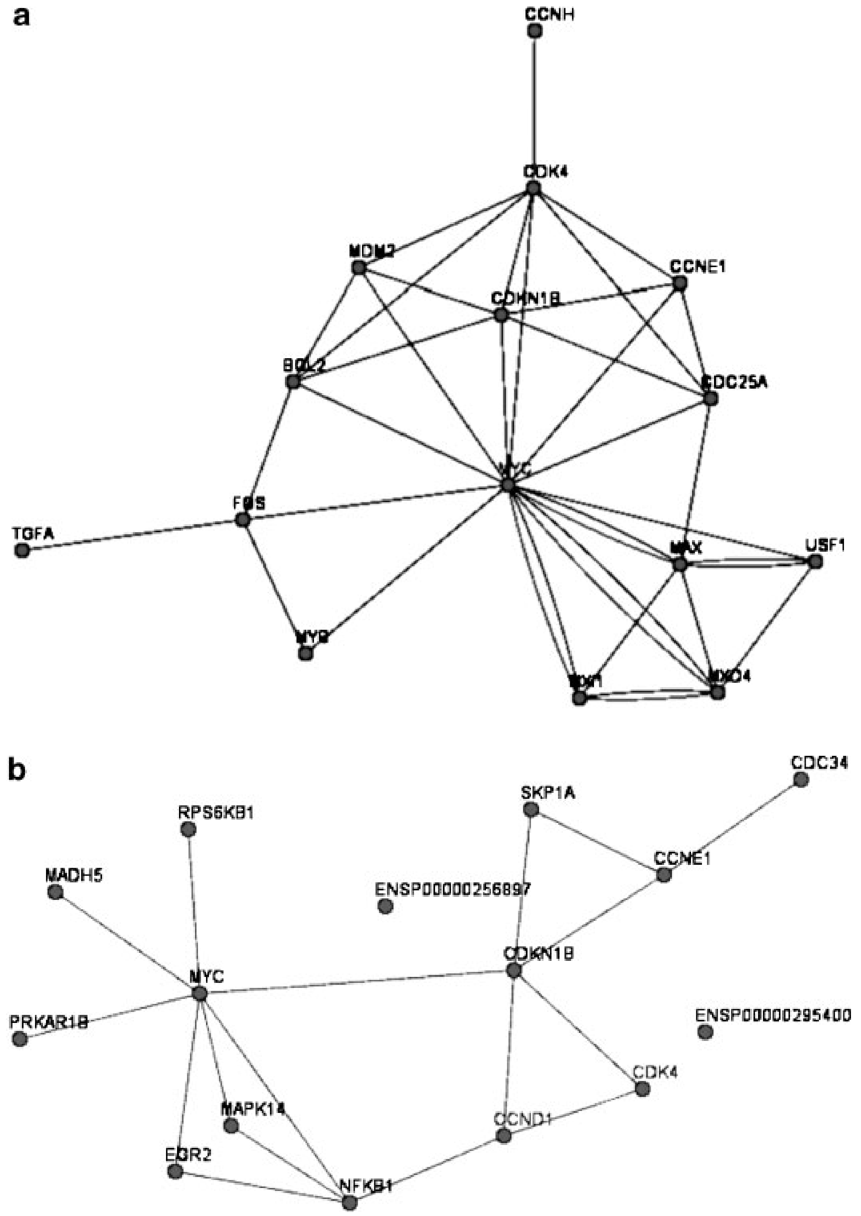

Fig. 2. **a**: Relations between the genes induced during lymphocyte activation according to co-mentioning in paper abstracts. **b**: Relations between the genes induced during lymphocyte activation according to gene–gene interaction (induction or suppression). **c**: Relations between the genes involved in human T lymphocyte cell cycle according to protein–protein interactions for proteins encoded by the corresponding gene.

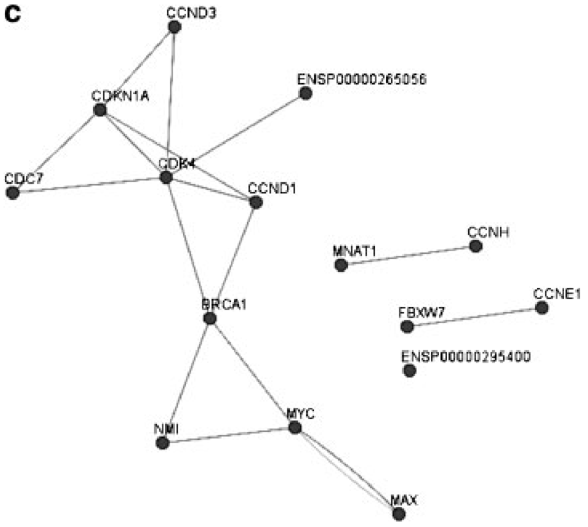

Fig. 2. (*Continued*)

knowledge about interacting molecules that regulate cell growth has increased exponentially, but our ability to make sense of this detailed information has not. Researchers interested in using modern biology need tools to organize a large collection of facts, including descriptions of bioregulatory molecules, their enzymatic modifications, and the complexes they form. Recently a common language was introduced that allows scientists to integrate data in a clear standardized, and computer-readable format, as Molecular Interaction Maps (MIMs) [Aladjem et al., 2004].

In the present work, we did not follow the MIMs method [Aladjem et al., 2004], since the latter is more suitable for a small set of genes/proteins than for a large map, like the one we constructed. Besides, the Molecular Interaction Maps method is dedicated to protein–protein interactions and analyzes different types of protein interactions in terms of a sophisticated set of descriptors. This study, instead, uses protein–protein interactions between proteins encoded by the studied genes together with other interactions between the genes, with scores assigned according to the ideology of STRING software and database, and utilizes fact that this scoring was assigned to each interaction type according to well-validated benchmarking systems (von Mering et al., 2005 and refs therein). It could be of interest to evaluate the contribution of each component of our interaction scores and see the emerging clustering patterns in view of relationships between the three interaction types. However, this would involve reconsidering the basics of the method and is therefore outside the scope of this paper but merits a separate study.

With the method described here, we succeed in the precise identification of “leader genes”

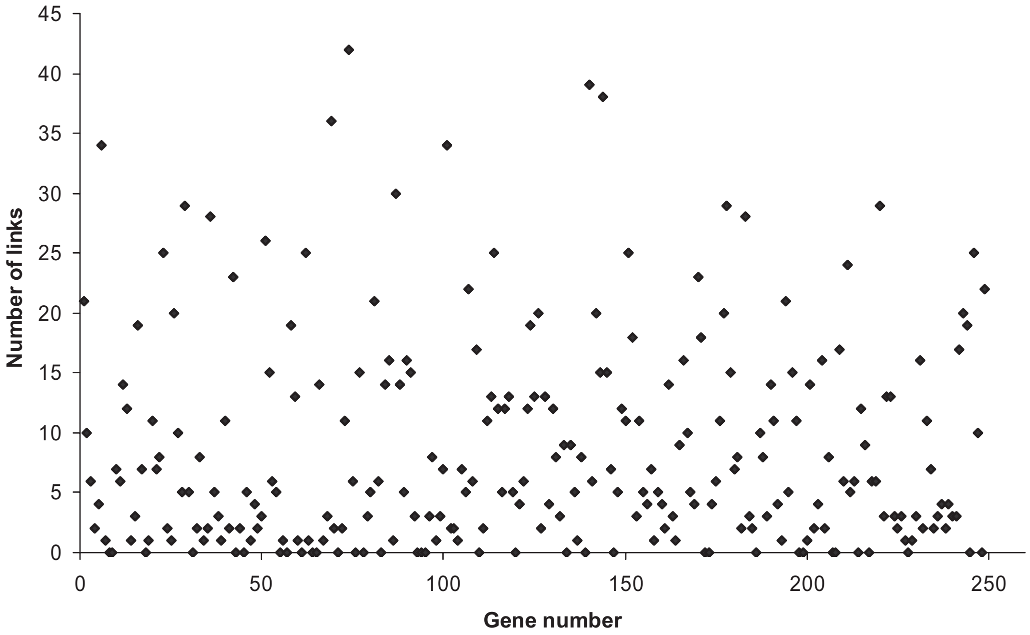

Fig. 3. Weighted number of links for the genes involved in cell cycle control of human T lymphocytes.

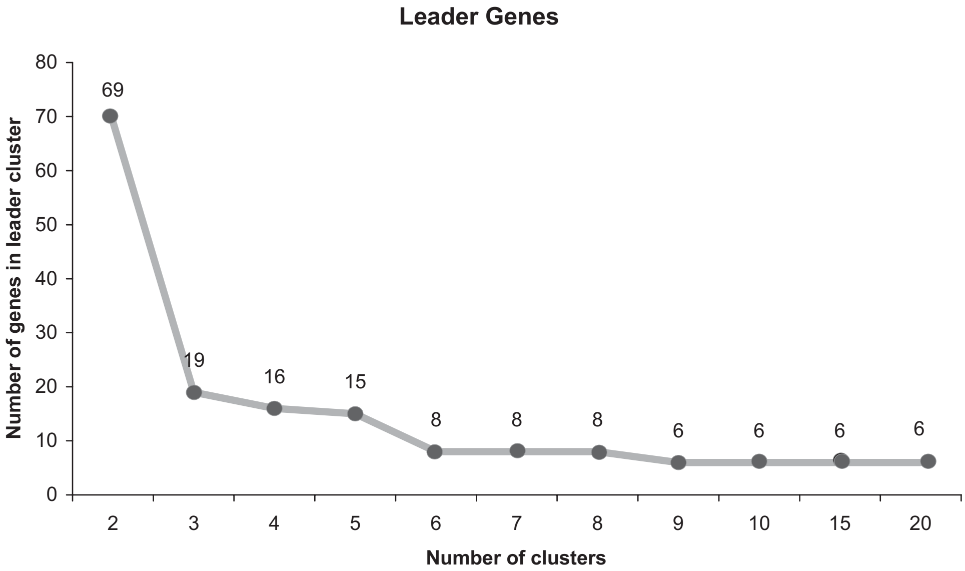

Fig. 4. Gene belonging to the leader cluster in different K-means clustering experiments, with an increasing number of cluster.

(Fig. 4 and Table III), which appear to control cell cycle progression (Table IV).

Such a small set of genes is composed of only six members, which have the highest number of interactions with other genes in the human T lymphocytes cell cycle process. Since their very high number of interactions, they are supposed to play a central role in the control of the cell cycle in human T lympocytes.

Moreover, our objective approach, based on weighed number of links and their clustering in order to identify leader genes, can give an added value to the methods described in Aladjem et al. [2004], giving a qualitative but also quantitative description of the relationships between genes.

A bibliographic research on cell cycle confirmed our results. Every gene we identified as a leader gene is known as a fundamental gene in the cell cycle control at important points (Table IV), namely the most important four at the transition from G0 to G1 phase (MYC) [Oster et al., 2002], at the progression in G1 phase (CDK4) [Modiano et al., 2000], and at the transitions from G1 to S (CDK2) [Baluchamy et al., 2003], and from G2 to M phases (CDC2) [Kawabe et al., 2002; Torgler et al., 2004].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 [Jerry et al., 2002; Chang et al., 2004].

TABLE III. Gene Belonging to the Leader Cluster in Different Hierarchical Clustering Experiments

	Linkage	Number of genes in the leader cluster
2 clusters	Centroid	35
	Average	35
	Smallest	6
3 clusters	Centroid	6
4 clusters	Centroid	6
5 clusters	Centroid	6
	Average	6

The six genes appearing in this table always compose the same subset (Table IVa).

It appears also clearly that leader genes (Fig. 4) are strictly connected, directly or through other interactions (Fig. 5). The identification of leader genes (Fig. 4 and Table IV) is indeed accompanied by the identification of neighboring genes, which can give a deeper insight on the leader genes role and their interactions, providing useful information. For example, the identification of orphan genes can suggest new targeted experimental researches in order to identify their partners (Table V).

Prediction of leader genes was found compatible with both our original experimental data obtained using commercially available DNA microarrays such as Human Starter, as well as with existing experimental data [Nicolini et al., 2005].

We found 27 unlinked genes (see Table V). Checking their Gene Ontology links revealed that six of them are involved in transcription

TABLE IV. Leader Genes According to the Clustering Experiments

Gene name	Id according to Figures 1 and 2	Weighted number of links	Gene description	Protein description	Apparent function in cell cycle
MYC	ENSP00000259523	27.81	v-myc myelocytomatosis viral oncogene homolog (avian); this region determines c-myc mRNA stability; v-myc avian myelocytomatosis viral oncogene homolog; v-myc myelocytomatosis viral oncogene homolog	Myc proto-oncogene protein (c-myc)	Entrance in G1 phase
CDK2	ENSP00000266970	26.65	Cyclin-dependent kinase 2; cdc2-related protein kinase; cell devision kinase 2; p33 protein kinase	Cell division protein kinase 2 (EC 2.7.1.-) (p33 protein kinase)	G1/S phase transition
CDC2	ENSP00000306043	26.47	Cell division cycle 2, G1 to S and G2 to M; cell cycle controller CDC2; cell division control protein 2 homolog; cell division cycle 2 protein; cyclin-dependent kinase 1; p34 protein kinase	Cell division control protein 2 homolog (EC 2.7.1.-) (p34 protein kinase) (Cyclin-dependent kinase 1) (CDK1)	G2/M phase transition
CDK4	ENSP00000316889	25.255	Cyclin-dependent kinase 4; cell division kinase 4; melanoma cutaneous malignant, 3	Cell division protein kinase 4 (EC 2.7.1.37) (Cyclin-dependent kinase 4) (PSK-J3)	Progression in G1 phase
CDKN1A	ENSP00000244741	25.08	Cyclin-dependent kinase inhibitor 1A (p21, Cip1); CDK-interaction protein 1; DNA synthesis inhibitor; cyclin-dependent kinase inhibitor 1A; melanoma differentiation associated protein 6; wild-type p53-activated fragment 1	Cyclin-dependent kinase inhibitor 1 (p21) (CDK-interacting protein 1) (Melanoma differentiation associated protein 6) (MDA-6)	Inhibitor of cyclin-CDK2 or −CDK4 complexes
CDKN1B	ENSP00000228872	23.90	Cyclin-dependent kinase inhibitor 1B (p27, Kip1); cyclin-dependent kinase inhibitor 1B	Cyclin-dependent kinase inhibitor 1B (Cyclin-dependent kinase inhibitor p27) (p27Kip1)	Inhibitor of cyclin-CDK2 or -CDK4 complexes

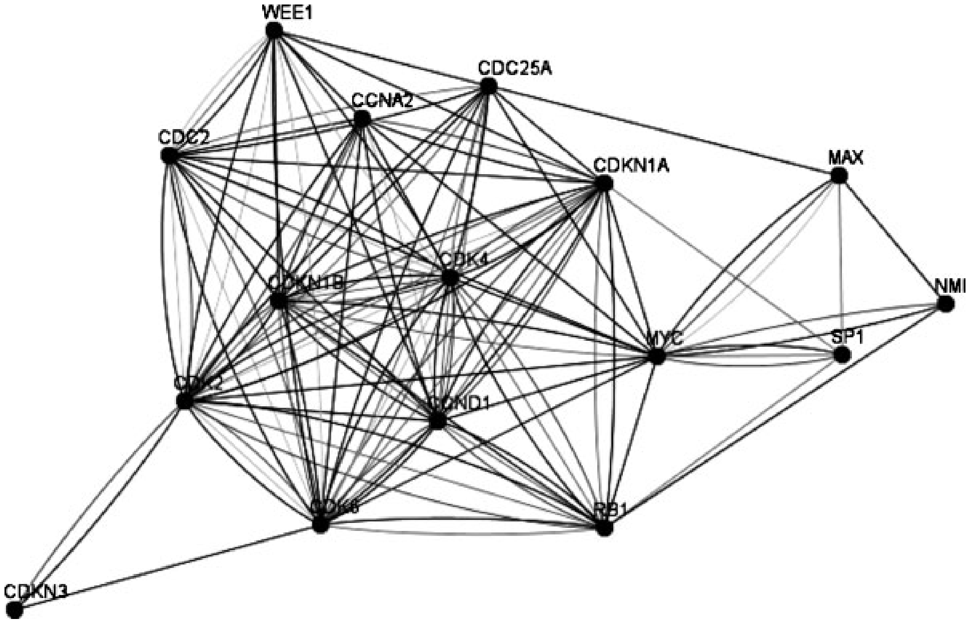

Fig. 5. Interaction map among leader genes and their neighborhood.

activation thereby possessing DNA binding properties while four are the key genes in cell cycle, while one (CHAF1B), the gene encoding subunit B of chromatin assembly factor 1, also known as p60, is both a DNA-binding and a cell cycle driving protein. One other gene known to be a cell-cycle switcher (G0S2), is also unlinked, suggesting that particular attention should be paid to experimentally derived correlations involving those genes, as well as the remaining genes.

The method we developed is based on a mathematical evaluation of the importance of genes in cell cycle. The role of the leader genes we identified in the cell cycle control was already well known, but the method we developed can give a more objective understanding of their importance among the several other genes involved in the same process. Leader genes approach can be useful in many cases. First of all, since their central role in cell cycle regulation, they can become promising pharmaceutical targets. It is possible to design appropriate drugs able to interact with leader genes, or to their neighbor, thus controlling cell cycle progression in a cascade process. This can be very useful in the therapy of several kinds of tumors. Moreover, the objective identification of leader genes and of gene interaction maps can suggest a more rational approach to experimental techniques and methods, as DNA microarrays have emerged to be a very powerful tool for gene expression analysis not only for the study of cell cycle [Butte, 2002; Nicolini et al., 2002; Nicolini et al., 2005]. Anyway, microarrays often display a very large number of genes, usually several thousand. This approach allows the study of a whole genome with a few experiments, but it also raise complication in experimental analysis, since the researcher has to confront with an enormous size of data. The application of bioinformatics studies and the identification of leader genes can predict which are the most important genes in a particular cellular process. In this way, it becomes possible to design smaller microarrays, which display only the interesting genes and thus are much easier to interpret. Protein microarrays are also used for the study of protein–protein and protein–gene interactions [Ramachandran et al., 2004]. Like the DNA microarrays, the leader gene approach can simplify their analysis, by reducing the protein displayed to the most important ones to be subsequently tested by mass-spectrometry or by ad hoc experimentation.

In conclusion, an approach based on bioinformatic and statistical analysis of already existing databases can really give an added value to the identification and the design of new pharmaceutical targets or to experimentation planning. Biology is becoming more and more an exact science, which cannot ignore the contribution of informatics and statistics in

TABLE V. Orphan Genes for Which the Weighted Number of Links Equals Zero

Gene name	Id according to Figures 1 and 2	Gene and gene product description
ANP32B	ENSP00000345848	Acidic leucine-rich nuclear phosphoprotein 32 family member B (PHAPI2 protein) (silver-stainable protein SSP29) (acidic protein rich in leucines)
C1QDC1	ENSP00000298892	C1q domain containing 1 isoform L
CCRN4L	ENSP00000280614	Nocturnin (CCR4 protein homolog)
CHAF1B	ENSP00000315700	Chromatin assembly factor 1 subunit B (CAF-1 subunit B) (Chromatin assembly factor I p60 subunit) (CAF-I 60 kDa subunit) (CAF-Ip60) (M-phase phosphoprotein 7)
CSTF2	ENSP00000263028	Cleavage stimulation factor, 64 kDa subunit (CSTF 64 kDa subunit) (CF-1 64 kDa subunit)
DEADC1	ENSP00000237283	deaminase domain containing 1
G0S2	ENSP00000243444	Putative lymphocyte G0/G1 switch protein 2
GLMN	ENSP00000311031	Glomulin (FKBP-associated protein) (FK506-binding protein-associated protein) (FAP)
GPR132	ENSP00000328818	Lysophosphatidylcholine receptor G2A (G2 accumulation protein)
HCST	ENSP00000246551	DNAX-activation protein 10; phosphoinositide-3-kinase adaptor protein
HLA-A	ENSP00000345802	HLA class I histocompatibility antigen, A-3 alpha chain precursor (MHC class I antigen A*3)
HLX1	ENSP00000259148	Homeobox protein HLX1 (Homeobox protein HB24)
HOXC4	ENSP00000305973	Homeobox protein Hox-C4 (Hox-3E) (CP19)
HUS1B	ENSP00000344216	HUS1 checkpoint protein B
IGBP1	ENSP00000337270	Immunoglobulin-binding protein 1 (CD79a-binding protein 1) (B cell signal transduction molecule alpha 4) (Alpha 4 protein)
IL31RA	ENSP00000297015	gp130-like monocyte receptor; soluble type I cytokine receptor CRL3; GP130 like receptor
KLF13	ENSP00000302456	Krueppel-like factor 13 (Transcription factor BTEB3) (Basic transcription element binding protein 3) (BTE-binding protein 3) (RANTES factor of late activated T lymphocytes-1) (RFLAT-1) (Transcription factor NSLP1) (Novel Sp1-like zinc finger transcription factor 1)
LY96	ENSP00000284818	Lymphocyte antigen 96 precursor (MD-2 protein) (ESOP-1)
MKI67	ENSP00000285814	MKI67 (FHA domain) interacting nucleolar phosphoprotein; nucleolar phosphoprotein Nopp34; nucleolar protein interacting with the FHA domain of pKi-67
PGF	ENSP00000256315	Placenta growth factor precursor (PlGF)
PSG1	ENSP00000308970	Pregnancy-specific beta-1-glycoprotein 1 precursor (PSBG-1) (Pregnancy-specific beta-1 glycoprotein C/D) (PS-beta-C/D) (Fetal liver non-specific cross-reactive antigen-1/2) (FL-NCA-1/2) (PSG95)
RAP1GA1	ENSP00000315420	Rap1 GTPase-activating protein 1 (Rap1GAP)
SFTPA1	ENSP00000242455	Pulmonary surfactant-associated protein A precursor (SP-A) (PSP-A) (PSAP) (Alveolar proteinosis protein) (35 kDa pulmonary surfactant- associated protein)
SHCBP1	ENSP00000306473	likely ortholog of mouse Shc SH2-domain binding protein 1
SMO	ENSP00000249373	Smoothened homolog precursor (SMO) (Gx protein)
TOP3A	ENSP00000321636	DNA topoisomerase III alpha (EC 5.99.1.2)
ENSP00000302185	TP53RK	p53-related protein kinase (EC 2.7.1.-) (Nori-2)

the preliminary phases of experimentation and in proper analysis of results.

REFERENCES

Abraham S, Vonderheid E, Zietz S, Kendall FM, Nicolini C. 1980. Reversible (G0) and nonreadily reversible (Q) noncycling cells in human peripheral blood. Immunological, structural, and biological characterization. Cell Biophys 2:353–371.

Aladjem MI, Pasa S, Parodi S, Weinstein JN, Pommier Y, Kohn KW. 2004. Molecular interaction maps—a diagrammatic graphical language for bioregulatory networks. Sci STKE Mar 2(222):pe8.

Baluchamy S, Rajabi HN, Thimmapaya R, Navaraj A, Thimmapaya B. 2003. Repression of c-Myc and inhibition of G1 exit in cells conditionally overexpressing p300 that is not dependent on its histone acetyltransferase activity. Proc Natl Acad Sci USA 100:9524–9529.

Brown KE, Baxter J, Graf D, Merkenschlager M, Fisher AG. 1999. Dynamic repositioning of genes in the nucleus of lymphocytes preparing for cell division. Mol Cell 3: 207–217.

Bruno S, Giaretti W, Darzynkiewicz Z. 1992. Effect of camptothecin on mitogenic stimulation of human lymphocytes: Involvement of DNA topoisomerase I in cell transition from G0 to G1 phase of the cell cycle and in DNA replication. J Cell Physiol 151:478–486.

Butte A. 2002. The use and analysis of microarray data. Nat Rev Drug Discov 1:951–960.

Cantrell D. 2002. Protein kinase B (Akt) regulation and function in T lymphocytes. Semin Immunol 14: 19–26.

Chang BL, Zheng SL, Isaacs SD, Wiley KE, Turner A, Li G, Walsh PC, Meyers DA, Isaacs WB, Xu J. 2004. A polymorphism in the *CDKN1B* gene is associated with increased risk of hereditary prostate cancer. Cancer Res 64:1997–1999.

Clevenger CV, Epstein AL, Bauer KD. 1987. Modulation of the nuclear antigen p105 as a function of cell-cycle progression. J Cell Physiol 130:336–343.

Cooper LJ, Topp MS, Pinzon C, Plavec I, Jensen MC, Riddell SR, Greenberg PD. 2004. Enhanced transgene expression in quiescent and activated human CD8+ T cells. Hum Gene Ther 15:648–658.

Darzynkiewicz Z, Traganos F, Andreeff M, Sharpless T, Melamed MR. 1979. Different sensitivity of chromatin to acid denaturation in quiescent and cycling cells as revealed by flow cytometry. J Histochem Cytochem 27: 478–485.

Datta S, Datta S. 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19:459–466.

Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K, Pawson T, Hogue CW. 2003. PreBIND and Textomy—mining the biomedical literature for protein–protein interactions using a support vector machine. BMC Bioinformatics 4:11.

Isakov N, Altman A. 2002. Protein kinase C(theta) in T cell activation. Annu Rev Immunol 20:761–794.

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.

Jones A, Hunt E, Wastling JM, Pizarro A, Stoeckert CJ Jr. 2004a. An object model and database for functional genomics. Bioinformatics 20:1583–1590.

Jones GG, Reaper PM, Pettitt AR, Sherrington PD. 2004b. The ATR-p53 pathway is suppressed in noncycling normal and malignant lymphocytes. Oncogene 23: 1911–1921.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–D280.

Kawabe T, Suganuma M, Ando T, Rimura M, Hori H, Okamoto T. 2002. Cdc25C interacts with PCNA at G2/M transition. Oncogene 21:1717–1726.

Modiano JF, Mayor J, Ball C, Fuentes MK, Linthicum DS. 2000. CDK4 expression and activity are required for cytokine responsiveness in T cells. J Immunol 165:6693–6702.

Nicolini C, Malvezzi AM, Tomaselli A, Sposito D, Tropiano G, Borgogno E. 2002. DNASER I: Layout and data analysis. IEEE Trans Nanobiosci 1:67–72.

Nicolini C, Spera R, Stura E, Fiordoro S, Giacomelli L. 2005. Gene expression in the cell cycle of human T lymphocytes: II. Experimental determination by DNAser technology. J Cell Biochem (in press).

Oosterwegel MA, Greenwald RJ, Mandelbrot DA, Lorsbach RB, Sharpe AH. 1999. CTLA-4 and T cell activation. Curr Opin Immunol 11:294–300.

Oster SK, Ho CS, Soucie EL, Penn LZ. 2002, The myc oncogene: Marvelousl Y Complex. Adv Cancer Res 84: 81–154.

Ramachandran N, Hainsworth E, Bhullar B, Eisenstein S, Rosen B, Lau AY, Walter JC, LaBaer J. 2004. Self-assembling protein microarrays. Science 305:86–90.

Reimers M. 2005. Statistical analysis of microarray data. Addict Biol 10:23–35.

Scovassi AI, Stefanini M, Lagomarsini P, Izzo R, Bertazzoni U. 1987. Response of mammalian ADP-ribosyl transferase to lymphocyte stimulation, mutagen treatment and cell cycling. Carcinogenesis 8:1295–1300.

Shannon W, Culverhouse R, Duncan J. 2003. Analyzing microarray data using cluster analysis. Pharmacogenomics 4:41–52.

Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA. 2005. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 33:1544–1552.

Torgler R, Jakob S, Ontsouka E, Nachbur U, Mueller C, Green DR, Brunner T. 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.

Troyanskaya OG. 2005. Putting microarrays in a context: Integrated analysis of diverse biological data. Brief Bioinform 6:34–43.

Tsai CA, Lee TC, Ho IC, Yang UC, Chen CH, Chen JJ. 2005. Multi-class clustering and prediction in the analysis of microarray data. Math Biosci 193:79–100.

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. 2005. STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res 33:D433–D437.

Vonderheid EC, Fang SM, Helfrich MK, Abraham SR, Nicolini C. 1981. Biophysical characterization of T-lymphocytes and Sezary cells. J Invest Dermatol 76: 28–37.