IL NUOVO CIMENTO Vol. 35 C, N. 5 Suppl. 1 Settembre-Ottobre 2012
DOI 10.1393/ncc/i2012-11340-4

Colloquia: PR PS BB 2011

Validation of a context analysis method for microRNA data

S. Rovetta(1)(*), F. Masulli(1)(2) and G. Russo(2)

(1) *Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova Genova, Italy*

(2) *Sbarro Institute for Cancer Research and Molecular Medicine, Temple University Philadelphia, PA, USA*

ricevuto il 30 Settembre 2011; approvato l' 1 Dicembre 2011

Summary. — We have previously presented a data-oriented approach to the study of microRNA-gene interactions, a hot topic of research in molecular biology, building heavily on methods from the document analysis field. This paper evaluates the performances of the method by means of a cross-validation approach. Latent information can be effectively exploited to suggest directions for laboratory experiments, an important topic in microRNA research, since these experiments are costly in both resources and time.

PACS 87.10.Vg – Biological information.
PACS 87.14.gn – RNA.
PACS 87.18.Vf – Systems biology.

1. – Introduction

A current topic of very active research [1, 2] is the study of microRNA, which acts on expressed genes by regulating their transcription in various ways. While modulation of gene expression at the transcriptional level is traditionally deemed responsible for the cell's operation, *post*-transcriptional modulation is increasingly acknowledged to play a number of important roles. It is now known that microRNAs are differentially found in many instances of major normal (*e.g.*, neurogenesis) and pathological (*e.g.*, many forms of cancer) events, which explains the great interest of this research area. In this contribution, we first present an overview of the method presented in [3], and then test it with an approach inspired from leave-one-out cross-validation.

(*) E-mail: rovetta@disi.unige.it

2. – MicroRNA target prediction and validation

2˙1. *MicroRNA.* – MicroRNAs (miRNAs) are short, non-coding RNAs, typically 22 nt long, that regulate gene expression by suppressing translation and destabilizing messenger RNAs (mRNAs) with specific target sequences. The action mechanism of each individual miRNAs is not always understood; however, regulation by miRNAs is peculiar in that it guarantees rapid and reversible changes in protein synthesis without altering transcription, and it is clear that collectively miRNAs play a central role in controlling both physiological and pathological processes [4].

As an example, several miRNAs have been found to have a significant involvement in the development of some types of cancer, for instance prostate cancer [5]. Decreased expression of some specific miRNAs has been consistently observed. As another example, hundreds of miRNAs are expressed in mammalian brain, and many of them are specifically expressed or enriched in brain tissues, such as let-7a, miR-218, and miR-125. Recent studies [6] suggest a possible role of miRNAs in neurogenesis. It has also been shown that synaptic plasticity is critically dependent upon regulation of specific protein synthesis near or within the pre- and/or post-synaptic sites, and numerous components of the microRNA machinery, including dicer, are expressed within dendrites and mature miRNAs and their precursors are detected in nerve terminals. A single miRNA may target hundreds of gene transcripts, including global regulators of translation, and the hypothesis of a combinatorial action of sets of miRNAs has been proposed, for instance in the case of so-called "microRNA clusters" [2]; so it is possible that miRNAs play an even more profound role than what is currently known.

2˙2. *Computational target prediction.* – Although the mechanisms for miRNA expression and maturation are quite well known, the post-transcriptional action mechanisms are not completely understood [7]. A miRNA does not require perfect complementary match to its target mRNA to perform its control action. In fact, imperfect complementarity and the resulting secondary conformation of the miRNA-target complex may explain some of the action mechanisms, which are probably varied. This situation makes it difficult to analyze miRNA-mRNA interactions in a purely computational way (target prediction), and this difficulty accounts for the relatively high number of computational prediction methods currently available, many of them disagreeing with others. Agreement is typically good only among methods based on similar hypotheses.

Most currently available prediction tools [8-10] are based on a given set of predefined hypotheses. These hypotheses often stem from either empirical observation, or statistical analysis of nucleotide sequence patterns (frequency of observation, enrichment analysis, phylogenetic conservation analysis). They are tested on a miRNA-mRNA site pair; partial scores are given to specific features measured on the pair, and if the overall score passes a threshold, then the mRNA site is predicted to be a target for the given miRNA. For instance, TargetScan [8] analyzes seed complementarity and phylogenetic conservation; RNAhybrid [9] finds sites with minimum free energy. PicTar [11] and miRanda [12] are somewhat more comprehensive tools. For instance, miRanda takes into account several effects, such as seed complementarity, non-complementarity of the bases following the seed region, overall complementarity, free energy, and conservation analysis. However, even miRanda does not provide perfectly consistent results with respect to available knowledge (targets which have been validated in lab experiments).

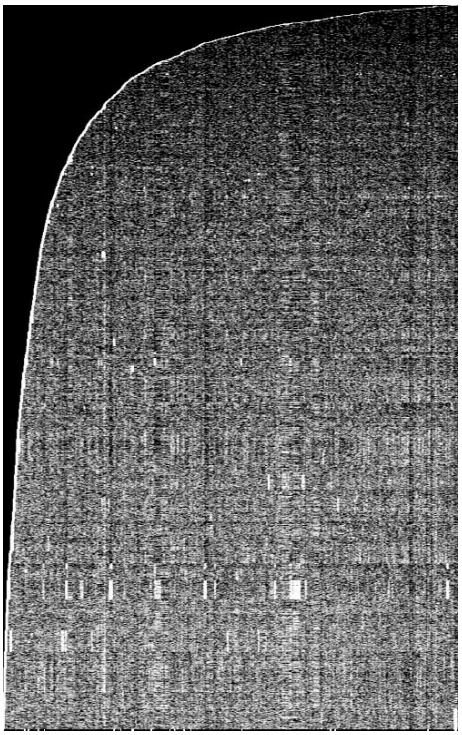

Fig. 1. – A visual representation of the data matrix.

3. – Context analysis from a Rosetta Stone

3˙1. *The data.* – A large quantity of data about miRNAs and their targets has been mined from several repositories, at the moment mainly from miRBase([1]), Argonaute([2]), and TarBase([3]), accessed using the Biomart([4]) R [13, 14] interface and *ad hoc* software. The result is a binary (0/1) data matrix indicating the very basic information of the simple presence/absence of one or more target sites on a given gene transcript for a given microRNA, which make up the rows and columns, respectively, of the matrix, as visually represented in fig. 1.

The accessed data sets contain different information. MiRBase and Argonaute report a large number of predicted gene-miRNA interaction targets, while TarBase is much less substantial numerically, but contains miRNA-gene pairs which have actually been validated. The number of miRNA sequences collected in our predicted database is 677. All sequences have actually been found to correspond to existing miRNAs. For these collected sequences, 23683 transcript sequences have been listed, with match (presence of at least one target site) for at least one, but typically more, of the collected miRNAs. This makes up a total of 16033391 matrix entries, of which 487409 (about 3%) are of

([1]) `http://www.mirbase.org/`
([2]) `http://www.ma.uni-heidelberg.de/apps/zmf/mirwalk/`
([3]) `http://diana.cslab.ece.ntua.gr/tarbase/`
([4]) `http://www.biomart.org/`

value 1. These indicate that, for the given combinations of gene transcript and a miRNA, a software prediction with sufficient confidence has been obtained.

The number of sequence in the validated database is 70, with 907 transcript sequences. This makes up a set of 63490 matrix entries, of which 1060 (about 1.7%) are of value 1. These indicate that, for the given combination, an experiment has been performed and the presence of at least one match has been actually validated, not simply predicted by software. In this work only the predicted data will be used, to ensure a sufficient number of cases.

3˙2. *A Rosetta stone.* – The data set contains values which refer to the interaction of mRNA and miRNA, formally similar to a term-document matrix in document analysis. This suggests the application of techniques from that field to extract knowledge from data, like the use of cosine distance and a vector space model [15].

In particular, the main goal of the overall activity is to improve detection of target sites on genes, mainly with a more accurate prediction program. To provide additional hints for the research, the same data can be used to suggest possible target genes for a given miRNA that have not yet been validated, so they appear as 0 in the data matrix. These suggestions can then be then submitted to the prediction program. Combining information from context (the data set) and local analysis (the prediction program), lab experiments for validation can be designed with increased confidence, which will lower their cost in both time and money.

As already stated, by construction the dataset contains asymmetric information. Validated targets are highly reliable, since they have been obtained by thorough laboratory experimentation. On the other hand, cells of the matrix with value 0 may represent either miRNAs for which a given mRNA does not actually have any target site; or untested miRNA-mRNA combinations.

This suggests an approach to target identification which is based on exploring the context to decide whether a given 0 cell might actually represent a true target, although not yet validated. The procedure is similar to the use of parallel texts to infer the meaning of an unknown word in a dead language. This can be described as a Rosetta stone approach. Some methods based on this approach will be outlined, although many more can be devised.

4. – The method and its experimental evaluation

4˙1. *Goals and procedure.* – In this study, we tested the method by examining some experimental hypotheses, and checking whether the method correctly answered them. The approach is of the leave-one-out type: we incorrectly delete one recorded interaction, and check whether the proposed criterion can infer it from context. Please refer to [3] for an example application of the method.

Two experiments are reported. The first experiment consists in checking whether a gene, once deleted from the list of matches with a given miRNA, can be recovered by the "Rosetta Stone" approach.

The second experiment consists in analyzing pairs of miRNAs and to check whether deletion of a gene from the list of matches for the first (query) miRNA could be compensated for by analyzing the second (reference).

The starting point for the procedure is a given miRNA, a "query" miRNA, for which we want to find new, possible interactions.

One or more reference miRNAs are then selected for comparison. These should be known to be involved in a given biological process of interest, *e.g.*, the development or the suppression of a tumor.

A set of genes which are targeted by this reference set of miRNAs is then selected. By "gene" we are really indicating the sequence of one specific transcript, among the several possible, for the terminal –3′ UTR– region of a gene of average length 2 kbases.

The resulting set is then split into two subsets: one contains all genes which are known to be targets to the query miRNA. This subset, which we call the MATCH set, can be expected not to be empty if the miRNA is involved in similar biological processes as the reference set, which is the case by hypothesis.

The second subset (the NON_MATCH set) contains all selected genes which *are not known* to be targets to the query miRNA. Among these, we are searching candidates for lab validation.

Each subset is composed of vectors whose length equals the number of available miRNAs (rows of the original data matrix). A rectangular affinity matrix is then built by selecting a similarity measure between vectors, and using it to compare all genes in the NON-MATCH set with all genes in the MATCH set.

Taking inspiration by document analysis, we can interpret this description of a miRNA as the "vector space model" of a text. The terms composing this text are all genes which interact with that miRNA, according to some rules (the "grammar" of an underlying language). As in the vector space model, the representation does not take grammatical structure into account, and only retains the used terms, in some arbitrary dictionary order.

Affinity is then naturally expressed by cosine similarity:

$$d(u, v) = u \cdot v, \tag{1}$$

chosen over the normalized version $d(u, v) = \frac{u \cdot v}{|u||v|}$ which does not take into account the number of matches.

To perform the leave-one-out experiments, we need to rank a given gene by similarity with the remaining ones. Therefore, we define a cumulative similarity between gene u and the set of genes V as the root mean square cosine similarity between u and all other genes in V:

$$cs(u, V) = \sum_{v \in V} \sqrt{d(u, v)}. \tag{2}$$

The advantage of this similarity definition is that, like mean similarity, it rewards larger numbers of similar items; however, unlike the mean, it can distinguish between many moderately similar items and fewer strongly similar ones, giving a larger reward to the former case.

5. – Experimental results

Software to perform the above operations was written in R and C [13, 14] using the GNU Scientific Library [16] and the MPI parallel environment [17].

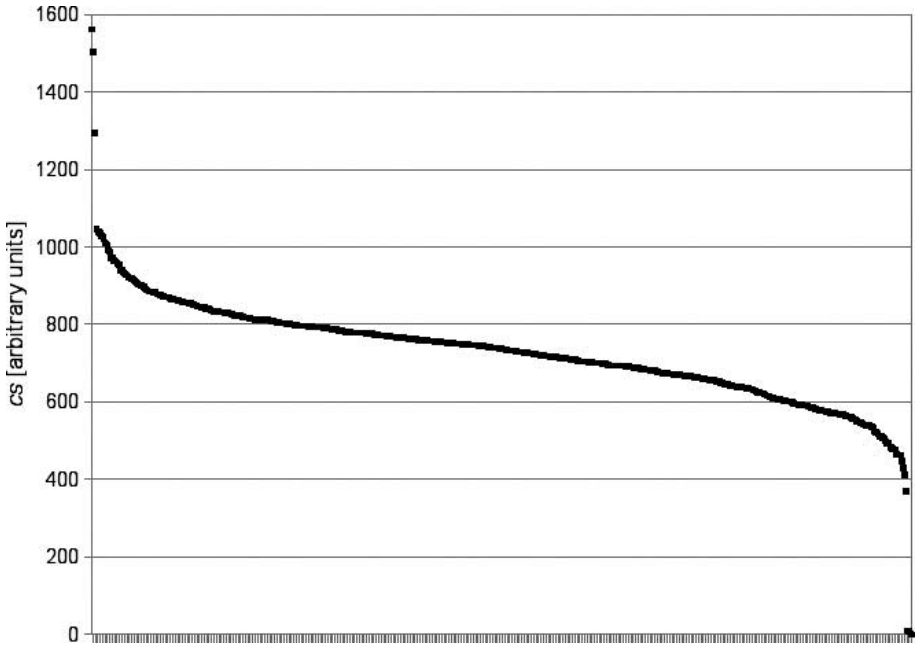

Fig. 2. – Experiment 1: Cumulative similarity *cs* of the worst gene for each miRNA w.r.t. all the remaining ones, as a measure of its likelihood to be recognized as match—sorted in decreasing order along the horizontal axis.

5˙1. *Experiment 1: Recovering matching genes from their own context.* – In this experimental setup, individual miRNAs are considered. The aim is to verify the hypothesis that the set of matching genes for a given miRNA can be used to infer a possible new matching gene.

To this purpose, one gene is deleted from the MATCH set of a miRNA. This is done for all matching genes in turn. For each gene, the similarity with respect to the remaining set is computed using the *cs* similarity.

The genes are then sorted by similarity. We consider a gene to be "recovered" if its value of *cs* is sufficiently high, according to some threshold.

The results, computed for all the available 677 miRNAs, indicate that all genes have $cs > 0$.

Not all genes feature a high degree of similarity. A worst-case statistic over the genes with the minimum *cs* for each miRNA (therefore 677 genes) reveals that few of them (5 of 677) are evaluated to a very low similarity ($cs \leq 10$, where the range of all others is $cs \in [369, 1561]$), see fig. 2.

Note that the source of many of the collected data is Miranda, which is known to have a higher false positive rate than other prediction tools [18]. Therefore, if used as exemplified in this experiment, our method can be used to point out possible false positives from the list of matches for a given miRNA, to be submitted to the lab or to other prediction methods for further inquiry.

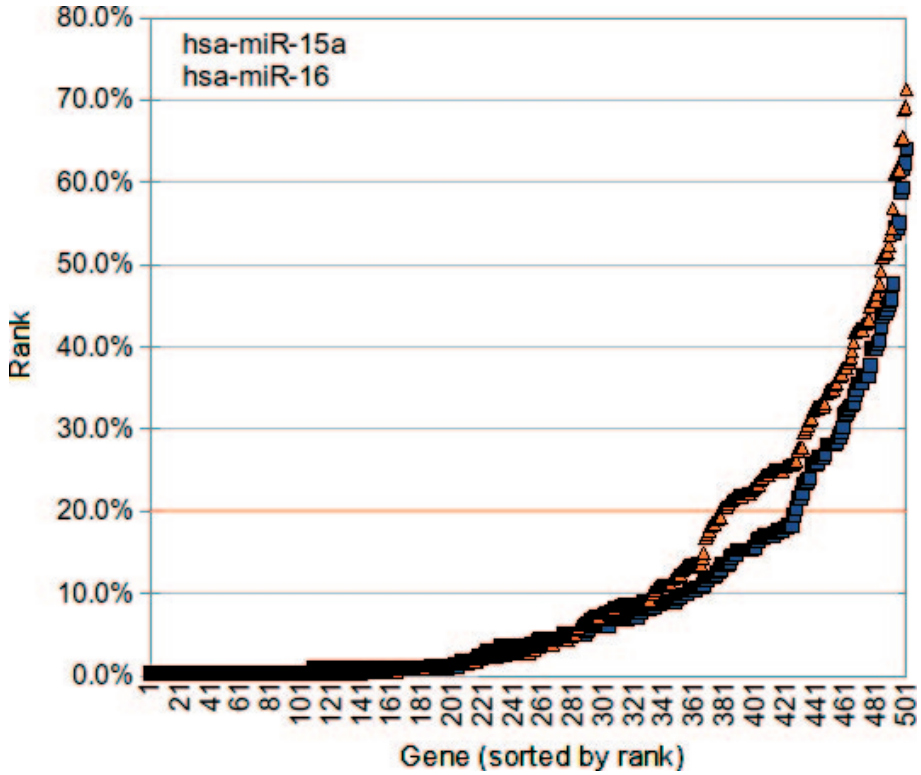

Fig. 3. – Experiment 2: hsa-miR-15a and hsa-miR-16 (see text for explanation).

5˙2. *Experiment 2: Inferring matching genes from external context*. – In the second experiment, miRNAs are evaluated in pairs. The number of pairs is $677 \times 676 = 457652$, so a data-parallel version of the software has been prepared and run on a multicore machine running the MPI middleware.

The aim of this experiment is to check whether the genes matching a given reference miRNA can be used as a context to infer new targets for a query miRNA.

The procedure described in [3] has been followed, by using the matching genes of a single miRNA as the reference set. The set of matching genes for the reference was split into a matching subset (genes which also match the query miRNA) and a nonmatching subset (genes which do not match the query miRNA).

The method from [3] can now be used to evaluate the nonmatching subset, to check whether some gene appears to be indicated as a possible match even if it comes from the nonmatching subset. However, the procedure was applied by removing one gene from the matching subset and putting it in the nonmatching subset—and this was repeated for each gene in the matching subset (and then for all possible pairs of miRNAs).

The results indicate that in general most genes are indicated quite clearly as matches. To make this evaluation, for each removed gene we check its rank in the nonmatching subset sorted by likelihood to be a match (this likelihood is evaluated according to [3]). The removed genes often appear among the top 20% in the list.

Two examples, among the 457 652 results available, are shown in figs. 3 and 4. The

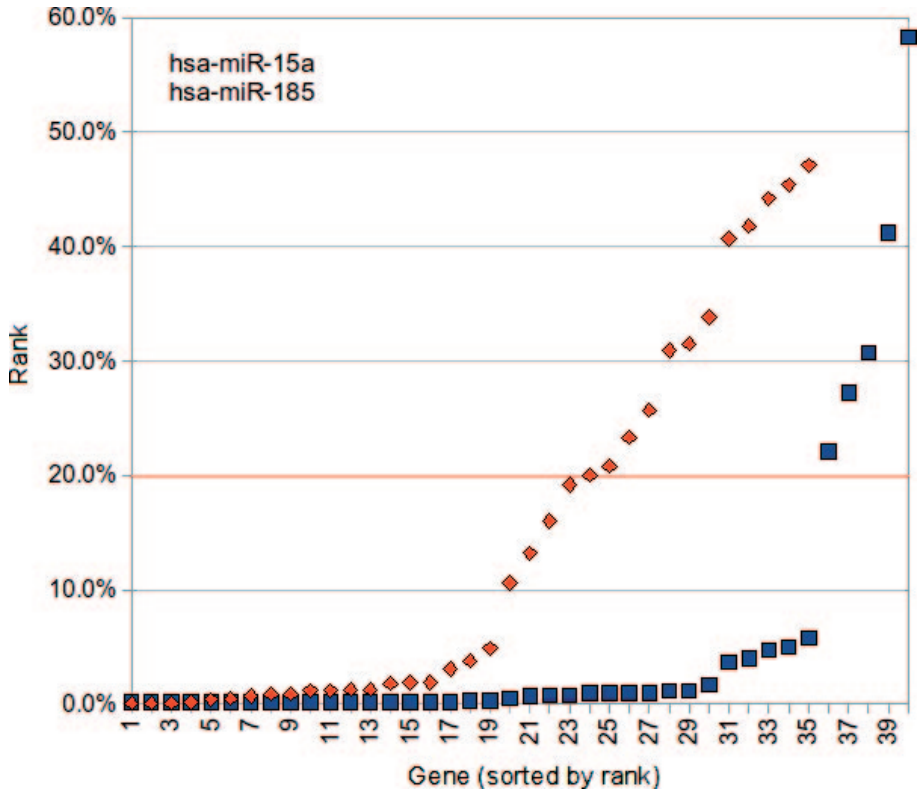

Fig. 4. – Experiment 2: hsa-miR-15a and hsa-miR-185 (see text for explanation).

first shows the ranks for hsa-miR-15a and hsa-miR-16. There are two data series because one has the first used as a query and the second as a reference, and the other has swapped roles.

It is possible to see that for both data series at least (roughly) 70% of genes come up among the 20% most likely to be matches (remember that they actually *are* matches). We also note that both series have a similar shape. This is probably due to the fact that these two miRNAs belong to the same miRNA cluster (they are transcribed from the same gene) and they are therefore related.

This is in contrast with fig. 4, which presents the result of the same procedure applied to hsa-miR-15a and hsa-miR-185. Here we can see that, firstly, the genes are a much lower number (40, *versus* 501 of the previous case), indicating that the sets of matching genes for the two are much less overlapping; in addition, while in one data series 90% of the genes are under the 20% threshold, in the other only 60% are under threshold. This asymmetry is another indicator of unrelatedness between the two selected miRNAs.

6. – Conclusion

The adopted testing methodology, inspired by "leave-one-out" cross-validation, seems to show how the proposed method is indeed sensitive to meaningful relationships in data. This implies that the context (defined as the set of interactions for a given miRNA or

gene) contains information that can be exploited to complement (where available) or supplement detailed information.

It would be interesting to exploit the Rosetta stone method to perform not only point-to-point similarity evaluations, but also group evaluations. Matrix-based clustering on a set of miRNAs can thus be performed, and the obtained clusters can be checked against known experimental relationships.

* * *

This research has been partially supported by the Human Health Foundation Onlus - Spoleto.

REFERENCES

[1] Hafner M., Landthaler M., Burger L., Khorshid M., Hausser J., Berninger P., Rothballer A., Ascano M., Jungkamp A.-C., Munschauer M., Ulrich A., Wardle G. S., Dewell S., Zavolan M. and Tuschl T., *Cell*, **141** (2010) 129.
[2] Bonauer A. and Dimmeler S., *Cell Cycle*, **8** (2009) 3866.
[3] Rovetta S., Masulli F., Monville M. E. and Russo G., *Context analysis in microrna data: A Rosetta stone approach*, in *Proceedings of the 2010 Conference on Neural Nets WIRN10: Proceedings of the 20th Italian Workshop on Neural Nets* (IOS Press, Amsterdam) 2011, pp. 135–143.
[4] Russo G., Puca A., Masulli F., Rovetta S., Cito L., Muresu D., Rizzolio F. and Giordano A., *Epigenetics, microrna and cancer: an update*, in *Cancer Epigenetics: Biomolecular Therapeutics in Human Cancer* (John Wiley) 2011, pp. 101–112.
[5] Bonci D., Coppola V., Musumeci M., Addario A., D'Urso L., Collura D., Peschle C., De Maria R. and Muto G., *European Urology Suppl.*, **7** (2008) 271.
[6] Krichevsky A. M., King K. S., Donahue C. P., Khrapko K. and Kosik K. S., *RNA*, **9** (2003) 1274.
[7] Jackson R. J. and Standart N., *Sci STKE*, **2007,** no. 367 (2007) .
[8] Lewis B. P., Shih I.-H., Jones-Rhoades M. W., Bartel D. P. and Burge C. B., *Cell*, **115** (2003) 787.
[9] Rehmsmeier M., Steffen P., Hochsmann M. and Giegerich R., *RNA*, **10** (2004) 1507.
[10] Miranda K. C., Huynh T., Tay Y., Ang Y.-S., Tam W.-L., Thomson A. M., Lim B. and Rigoutsos I., *Cell*, **126** (2006) 1203.
[11] Krek A., Grün D., Poy M. N., Wolf R., Rosenberg L., Epstein E. J., Macmenamin P., da Piedade I. d., Gunsalus K. C., Stoffel M. and Rajewsky N., *Nat. Genet.*, **37** (2005) 495.
[12] Enright A. J., John B., Gaul U., Tuschl T., Sander C. and Marks D. S., *Genome Biol.*, **5**, no. 1 (2003).
[13] Ihaka R. and Gentleman R., *J. Comput. Graphi. Stati.*, **5** (1996) 299.
[14] *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria) 2005.
[15] Salton G., Wong A. and Yang C. S., *Commun. ACM*, **18** (1975) 613.
[16] Galassi M. *et al.*, *GNU Scientific Library Reference Manual* (3rd Edition) 2009.
[17] Snir M., Otto S., Walker D., Dongarra J. and Huss-Lederman S., *MPI: The Complete Reference* (MIT press), 1995.
[18] Bentwich I., *FEBS Lett.*, **579** (2005) 5904.