W334–W338 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 20 May 2011*
doi:10.1093/nar/gkr289

PINTA: a web server for network-based gene prioritization from expression data

Daniela Nitsch[1], Léon-Charles Tranchevent[1], Joana P. Gonçalves[2,3], Josef Korbinian Vogt[4], Sara C. Madeira[2,3] and Yves Moreau[1,*]

[1]Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, 3001 Leuven, Belgium, [2]Knowledge Discovery and Bioinformatics group (KDBIO), INESC-ID, 1000-029 Lisbon, Portugal, [3]Instituto Superior Técnico (IST), Technical University of Lisbon, 1049-001 Lisbon, Portugal and [4]Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, 2800 Lyngby, Denmark

Received January 31, 2011; Revised March 15, 2011; Accepted April 12, 2011

ABSTRACT

PINTA (available at http://www.esat.kuleuven.be/pinta/; this web site is free and open to all users and there is no login requirement) is a web resource for the prioritization of candidate genes based on the differential expression of their neighborhood in a genome-wide protein–protein interaction network. Our strategy is meant for biological and medical researchers aiming at identifying novel disease genes using disease specific expression data. PINTA supports both candidate gene prioritization (starting from a user defined set of candidate genes) as well as genome-wide gene prioritization and is available for five species (human, mouse, rat, worm and yeast). As input data, PINTA only requires disease specific expression data, whereas various platforms (e.g. Affymetrix) are supported. As a result, PINTA computes a gene ranking and presents the results as a table that can easily be browsed and downloaded by the user.

BACKGROUND

A major challenge in human genetics is to identify novel disease genes to understand the mechanisms underlying genetic conditions and, in the long term, elaborate novel treatments for these disorders. Genetic studies, such as association studies and linkage analyses, identify chromosomal regions involved in a disease or phenotype of interest, but often result in large lists of candidate genes of which only one or a few are really associated to the disease or phenotype under study. Identifying, among such a list, the most promising candidate genes for a disease of interest has been defined as the gene prioritization problem. Candidate gene prioritization is key in genetics because it is generally too expensive and time-consuming to experimentally validate all candidate genes. Because of the huge amount of genomic data that is publicly available, computational approaches have been developed to avoid performing candidate gene prioritization manually.

In the past couple of years, several gene prioritization methods have been proposed by the bioinformatics community to address this problem. For a detailed review of web based gene prioritization tools and their information sources, the reader is referred to our recent review (1) and its associated web site (http://www.esat.kuleuven.be/gpp). Most of the available gene prioritization tools combine different data and information sources, among which the most commonly used data sources are literature, functional annotations, interactions, expression data and sequence information (2,3). Among these tools, ToppGene (4), SNPs3D (5), GeneDistiller (6) and Posmed (7) additionally include model organism data (mainly mouse data). Other tools, such as GeneWanderer (8), Prioritizer (9) and PhenoPred (10) make use exclusively of genome-wide protein–protein interaction networks. Prioritizer integrates several networks obtained from different databases (including expression data) and uses this huge network to investigate diseases for which several loci are known. Candidate genes from one locus that are connected to candidate genes in another locus are considered promising candidate genes. GeneWanderer uses a global network distance measure to define similarity in protein–protein interaction networks. PhenoPred is based on a human protein–protein interaction network and uses a supervised algorithm for detecting gene-disease associations, known gene-disease associations, protein sequence and protein functional information.

However, most of these tools are using a guilt-by-association concept (candidate genes that are similar

*To whom correspondence should be addressed. Tel: +32 (0)16 32 10 75; Fax: +32 (0)16 32 19 70; Email: yves.moreau@esat.kuleuven.be to the already confirmed disease genes are considered promising) and are therefore not applicable when little is known about the phenotype or when no confirmed disease genes are available beforehand. One common strategy to circumvent that problem is to rely on keywords to define the genetic condition under study. However, most of the existing tools that accept keywords also rely solely on text-mining of the literature and are therefore less suitable for novel discoveries (11–16). In a recent study (17,18), we have proposed a method that overcomes this limitation by representing prior knowledge about the biological process by experimental data on differential gene expression between affected and healthy individuals.

At the core of the method are a protein–protein interaction or association network and a disease specific expression data set. The method propagates the expression data over the network using an extended random walk approach. Candidate genes are then ranked based on the differential expression level (e.g. fold changes from a case–control study) of their neighborhood. Our method relies on the assumption that strong candidate genes tend to be surrounded by many differentially expressed genes in a genome-wide protein–protein interaction network. This allows the detection of a strong signal for a candidate even if its own differential expression value is too small to be detected by a standard analysis, as long as its interacting partners are highly differentially expressed. Our benchmark on 40 publicly available knockout experiments in mice showed that it outperforms a standard procedure in genetics that ranks candidate genes based solely on their own expression levels (18).

In this article, we describe a novel web server called PINTA (http://www.esat.kuleuven.be/pinta/) that implements the method we have developed previously (17,18). PINTA is a free, user-friendly and easy accessible web tool, which performs candidate gene prioritization (i.e. starting from a predefined set of candidate genes) as well as genome-wide gene prioritization using the method described above. To our knowledge, PINTA is the first web based tool that can prioritize candidate genes for diseases with only limited information about the phenotype or no confirmed disease genes by replacing this knowledge by expression data to model the disease under study. To be of use to a large range of biologists and geneticists, we have made PINTA available for multiple species besides human (mouse, rat, worm and yeast) that represent some of the most common model organisms for human. In addition to the approaches described in (17,18), PINTA can also propagate the expression over the network based on other extended random walk approaches [HITS with priors and k-step Markov (19)]. PINTA also supports the use of probe set names from various Affymetrix platforms.

PINTA WORKFLOW

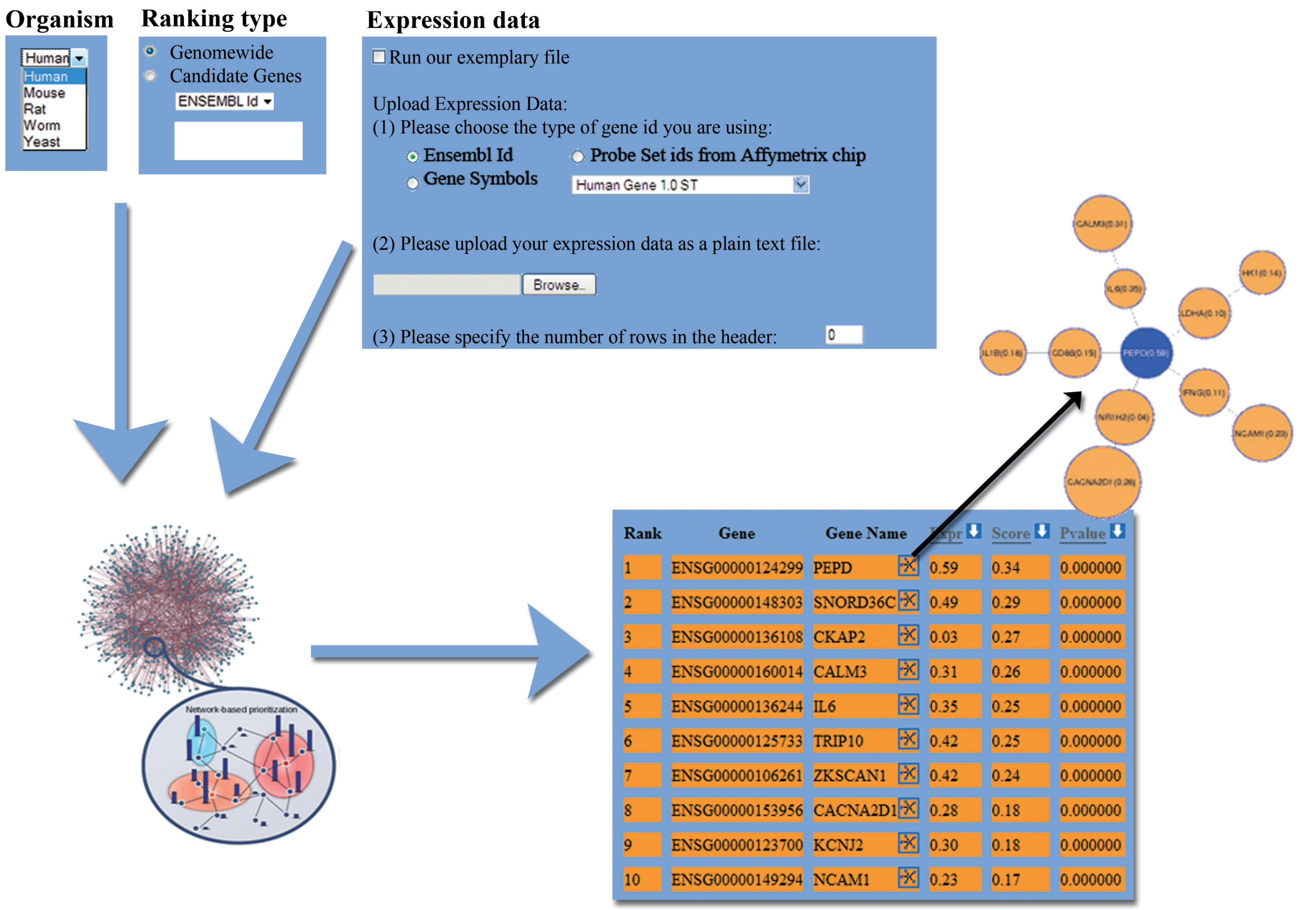
A four-step wizard guides the user through the gene prioritization procedure of PINTA (Figure 1). In the first step, the user selects the organism of interest among human, mouse, rat, worm or yeast. In the second step, the user defines whether PINTA needs to compute a genome-wide ranking or to rank a set of candidate genes. In the latter case, the user can choose to provide candidate genes using either official gene symbols or Ensembl identifiers. In the third step, the user can optionally choose whether he wants PINTA to perform a comparison analysis between its own ranking and the ranking obtained by simply ordering the differential expression values. The fourth step consists in uploading disease specific expression data, where probe set names of various Affymetrix platforms are also supported (see below). PINTA uses by default the best performing settings that were determined in our benchmark (18), which is convenient for non-expert users. Advanced users can fine tune the parameters to fit their needs and make the best of their data. Users can choose between five different prioritization algorithms [Heat Kernel Ranking (18), Arnoldi Diffusion Ranking (18), Random Walk on a graph (20), HITS with priors (19) or k-step Markov (19), for more details see below] and two different networks [STRING version 8.2 (21) and I2D version 1.8 (22)]. By default, the Heat Kernel Ranking and the STRING network are used.

RANKING WITH DIFFERENTIAL EXPRESSION DATA

As input data, PINTA only requires disease-specific expression data in the form of a text file in which each gene is assessed by a differential expression value. The user has to upload a tab delimited text file containing the expression data in the form of two columns: the first column contains the gene identifiers (Affymetrix probe identifiers, gene symbols or Ensembl identifiers) and the second column contains the gene's differential expression signal.

PINTA supports some of the most common Affymetrix chips to accommodate the user with an easier handling of the data and to avoid the need for manual mapping of probe id to gene: HG 1.0 ST, HEx 1.0 ST, HGU 133 Plus2, MouseGenome 430 2.0, MouseGene 1.0 ST, MouseExon 1.0 ST, RatExpressionArray 230 2.0, RatGene 1.0 ST, RatExon 1.0 ST, *Caenorhabditis elegans* Genome Array and Yeast 2.0 Array. If the user is not using one of these Affymetrix chips, he can still upload the expression data using Ensembl identifiers or gene symbols according to the drop down menu of PINTA.

PINTA requires one differential expression value per gene and is neither performing any kind of preprocessing nor normalization of the data; this step lies in the responsibility of the user. If the microarray experiment consists of several experiments or chips, the user has to provide one characteristic expression value for each gene/probe (e.g. the average expression value across all chips). The text file can be created in Microsoft® Excel and then converted into a tab delimited text file, or it can be directly created as a text file as an output of any preprocessing or normalization tool. Since the text file may contain headers



Organism

Human ▾
 Human
 Mouse
 Rat
 Worm
 Yeast

Ranking type

● Genomewide
○ Candidate Genes
ENSEMBL Id ▾

Expression data

□ Run our exemplary file

Upload Expression Data:
(1) Please choose the type of gene id you are using:
● Ensembl Id ○ Probe Set ids from Affymetrix chip
○ Gene Symbols Human Gene 1.0 ST

(2) Please upload your expression data as a plain text file:
Browse..

(3) Please specify the number of rows in the header: 0

Rank	Gene	Gene Name		Expr	Score	Pvalue
1	ENSG00000124299	PEPD		0.59	0.34	0.000000
2	ENSG00000148303	SNORD36C		0.49	0.29	0.000000
3	ENSG00000136108	CKAP2		0.03	0.27	0.000000
4	ENSG00000160014	CALM3		0.31	0.26	0.000000
5	ENSG00000136244	IL6		0.35	0.25	0.000000
6	ENSG00000125733	TRIP10		0.42	0.25	0.000000
7	ENSG00000106261	ZKSCAN1		0.42	0.24	0.000000
8	ENSG00000153956	CACNA2D1		0.28	0.18	0.000000
9	ENSG00000123700	KCNJ2		0.30	0.18	0.000000
10	ENSG00000149294	NCAM1		0.23	0.17	0.000000

Figure 1. The PINTA workflow. In the first step, the user selects the organism of interest. The second step consists in defining whether PINTA needs to compute a genome-wide ranking or to rank a defined set of candidate genes. In the last step, a disease-specific expression data is required in the form of a text file in which each gene is assessed by an expression value. PINTA computes a ranking of all candidate genes and presents the results in a table containing the gene ranks, internal scores, *P*-values and Bonferroni–Holm *P*-values (multiple testing correction). For the top ranked genes, PINTA provides (with the default settings) a graphical view of the strongest contributing interacting genes that lead to the candidate's strong scoring signal together with their expression signal in the network.

(e.g. internal describable notes), the user can specify how many rows of the file belong to the header and PINTA will ignore these rows in its internal computation. If the user is unsure how to choose the correct format for the text file, he can go back to the exemplary text file provided online by PINTA.

PINTA provides five different prioritization algorithms from which the user can choose in the advanced setting by activating the corresponding radio button. Performing a random walk on a network consists in taking successive random steps (20). In the Heat Kernel Ranking (18) (default algorithm in PINTA), the input signal given by the expression level of a candidate is adjusted according to a global network measure taking direct and indirect association into account, whereas the Arnoldi Diffusion Ranking performs network diffusion by applying the Arnoldi algorithm (18,23). The HITS with priors algorithm (19) is a random walk on the network based on mutual reinforcement relation between authorities and hubs using prior knowledge about the importance of nodes and the k-step Markov algorithm (19) is a random walk on the network defined by the probability transition of the graph.

PRIORITIZATION RESULTS AND DATA VISUALIZATION

PINTA computes a ranking of the candidate genes and presents the results in a table containing the gene ranks, internal scores, *P*-values and Bonferroni–Holm *P*-values (multiple testing correction). The *P*-values are calculated by random permutation on the expression data (random reassignment of the expression values to network nodes and computation of the corresponding randomized scores for all candidate genes). The user can sort this table by *P*-values, scores or expression values to check the differences in the corresponding rankings. PINTA provides for the top ranked genes (with the default settings) a graphical view of the strongest contributing interacting genes and their expression signal in the network. With this graphical view, the user can assess the importance of a top-ranked gene by investigating its neighborhood and the neighboring genes' expression levels. Highly influencing neighboring genes are not necessarily genes that are strongly connected with the candidate gene in the network, because our method based on a random walk approach considers both the strength of the interactions and the differential expression levels of the interactors. Therefore, a candidate gene might still be highly influenced by low interacting genes with high differential expression, leading to a high ranking score.

Furthermore, if chosen by the user, PINTA provides a comparison analysis between its own ranking and the ranking obtained by simply ordering the differential expression values decreasingly in a second table using different color schemes for an easy visualization. This comparison highlights that some highly ranked genes by PINTA have a strongly differentially expressed neighborhood, although their own differential expression level is low leading to a low ranking position in the differential expression ranking. These candidate genes would not be detected by this standard analysis as our benchmark could demonstrate (18).

All output tables and subnetworks containing the strongest contributing genes can be downloaded by the user for further use. The output tables are available as tab delimited text files that can be directly opened in Microsoft® Excel as well as in any text file reader. The graphical views of the subnetworks are available as pictures (.jpg).

SOFTWARE DOCUMENTATION

PINTA provides an online manual. In this manual, the workflow of PINTA is explained step-by-step using an exemplary expression data set. Additional pictures and screenshots can guide the user who wants to understand the details and to fine tune the prioritization parameters.

For testing purposes only, PINTA provides an exemplary expression data file that can easily be used by checking a dedicated checkbox during the fourth step of the wizard. By doing so, users find a quick way to examine PINTA's features. The data set is publicly available (GSE10849) and represents a mouse knockout experiment of the Cav1 gene. This data was RMA (24) preprocessed and as differential expression measures we computed the test statistic derived from cyberT (25). Performing a genome-wide ranking with default setting, PINTA ranks the knockout gene Cav1 on the first position, because this gene causes the most disrupted neighborhood in expression within the network.

IMPLEMENTATION

The gene prioritization method used by PINTA was implemented in MATLAB. To make it universally accessible, we have developed a PHP web-based interface that runs with the most common web browsers, for which Java does not have to be installed. PINTA is freely accessible and there is no login requirement.

CONCLUSION

We have developed the freely accessible web resource PINTA designed for the prioritization of candidate genes based on the differential expression of their neighborhood in a genome-wide protein–protein interaction network. PINTA is dedicated to the study of genetic disorders for which only little is known beforehand or when no background knowledge is assumed. PINTA relies on the presence of disease specific expression data, which makes it particularly attractive to study genetic conditions for which such expression data can easily be collected. PINTA propagates the expression data over the network using several random walk strategies. This allows the detection of a strong signal for a candidate gene even if its own differential expression value is small. PINTA is available for some prominent model organisms beside human and various array platforms are supported. PINTA ranks the candidate genes based on the assumption that strong candidate genes tend to be surrounded by many differentially expressed neighboring genes. A benchmarked on 40 mouse knockout experiments has shown that PINTA outperforms traditional approaches.

FUNDING

The Research of D.N., L.C.T. and Y.M. is supported by Research Council KUL (ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 *SymBioSys* en KUL PFV/10/016 *SymBioSys*, START 1, several PhD/postdoc & fellow grants); Flemish Government [FWO: PhD/ postdoc grants, projects, G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/ Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR), IWT: PhD grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3, FOD:Cancer plans, IBBT]; Belgian Federal Science Policy Office [IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: From Genomes to Networks, 2007–2011)]; EU-RTD (ERNSI: European Research Network on System Identification; FP7-HEALTH CHeartED); FCT PhD grant SFRH/BD/ 36586/2007 (to J.P.G.); FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and NEUROCLINOMICS (PTDC/EIA-EIA/111239/ 2009) (to S.C.M., in part).

Conflict of interest statement. None declared.

REFERENCES

1. Tranchevent,L.C., Capdevila,F.B., Nitsch,D., Moor,B.D., Causmaecker,P.D. and Moreau,Y. (2010) A guide to web tools to prioritize candidate genes. *Brief. Bioinform.*, doi:10.1093/bib/ bbq007.
2. Hutz,J.E., Kraja,A.T., Mcleod,H.L. and Province,M.A. (2008) CANDID: a flexible method for prioritizing candidate genes for complex human traits. *Genet. Epidemiol.*, **32**, 779–790.
3. Tranchevent,L.-C., Barriot,R., Yu,S., Van Vooren,S., Van Loo,P., Coessens,B., De Moor,B., Aerts,S. and Moreau,Y. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, **36**, W377–W384.

4. Chen,J., Xu,H., Aronow,B.J. and Jegga,A.G. (2007) Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, **8**, 392.
5. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 155.
6. Seelow,D., Schwarz,J.M. and Schuelke,M. (2008) GeneDistiller - distilling candidate genes from linkage intervals. *PLoS ONE*, **3**, e3874.
7. Yoshida,Y., Makita,Y., Heida,N., Asano,S., Matsushima,A., Ishii,M., Mochizuki,Y., Masuya,H., Wakana,S., Kobayashi,N. *et al.* (2009) PosMed (Positional Medline): prioritizing genes with an artificial neural network comprising medical documents to accelerate positional cloning. *Nucleic Acids Res.*, **37**, W147–W152.
8. Köhler,S., Bauer,S., Horn,D. and Robinson,P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
9. Franke,L., Bakel,H., Fokkens,L., de Jong,E., Egmont-Petersen,M. and Wijmenga,C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.
10. Radivojac,P., Peng,K., Clark,W.T., Peters,B.J., Mohan,A., Boyle,S.M. and Mooney,S.D. (2008) An integrated approach to inferring gene-disease associations in humans. *Proteins*, **72**, 1030–1037.
11. Hristovski,D., Peterlin,B., Mitchell,J.A. and Humphrey,S.M. (2005) Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inf.*, **74**, 289–298.
12. Yu,W., Wulf,A., Liu,T., Khoury,M.J. and Gwinn,M. (2008) Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics*, **9**, 528.
13. Cheng,D., Knox,C., Young,N., Stothard,P., Damaraju,S. and Wishart,D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36**, W399–W405.
14. Driel,M.A.van, Bruggeman,J., Vriend,G., Brunner,H.G. and Leunissen,J.A.M. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.: EJHG*, **14**, 535–542.
15. Van Vooren,S., Thienpont,B., Menten,B., Speleman,F., De Moor,B., Vermeesch,J. and Moreau,Y. (2007) Mapping biomedical concepts onto the human genome by mining literature on chromosomal aberrations. *Nucleic Acids Res.*, **35**, 2533–2543.
16. Xiong,Q., Qiu,Y. and Gu,W. (2008) PGMapper: a web-based tool linking phenotype to genes. *Bioinformatics*, **24**, 1011–1013.
17. Nitsch,D., Tranchevent,L.-C., Thienpont,B., Thorrez,L., Van Esch,H., Devriendt,K. and Moreau,Y. (2009) Network analysis of differential expression for the identification of disease-causing genes. *PLoS ONE*, **4**, doi:10.1371/journal.pone.0005526.
18. Nitsch,D., Gonçalves,J.P., Ojeda,F., Moor,B.D. and Moreau,Y. (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, **11**, 460.
19. White,S. and Smyth,P. (2003) Algorithms for estimating relative importance in networks. *Proceedings of SIGKDD*, pp. 266–275.
20. Lovász,S. (1996) Random walks on graphs: a survey. In Milos,D., Sos,V.T. and Szony,T. (eds), *Combinatorics, Paul Erdős is Eighty*. Budapest, Hungary: János Bolyai Mathematical Society, pp. 353–398.
21. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
22. Brown,K.R. and Jurisica,I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.
23. Saad,Y. (1992) Analysis of some Krylov subspace approximations to the matrix exponential operator. *SINUM*, **29**, 209–228.
24. Irizarry,R.A., Hobbs,B., Colin,F., Beazer-Barclay,Y.D., Antonellis,K., Scherf,U. and Speed,T.P. (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
25. Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.