Vincent Danos
Vincent Schachter (Eds.)

LNBI 3082

Computational Methods in Systems Biology

International Conference CMSB 2004
Paris, France, May 2004
Revised Selected Papers

Springer

Vincent Danos Vincent Schachter (Eds.)

Computational Methods in Systems Biology

International Conference CMSB 2004
Paris, France, May 26 - 28, 2004
Revised Selected Papers

Springer

Series Editors

Sorin Istrail, Celera Genomics, Applied Biosystems, Rockville, MD, USA
Pavel Pevzner, University of California, San Diego, CA, USA
Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

Vincent Danos
Université Paris 7
Equipe PPS Case 7014
2 place Jussieu, 75251 Paris Cedex 05, France
E-mail: Vincent.Danos@pps.jussieu.fr

Vincent Schachter
CNRG Genoscope
2 rue Gaston Cremieux, 91000 Evry, France
E-mail: vs@genoscope.cns.fr

Library of Congress Control Number: 2005922242

CR Subject Classification (1998): I.6, D.2.4, J.3, H.2.8, F.1.1

ISSN 0302-9743
ISBN 3-540-25375-0 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11409083 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 3082

Edited by S. Istrail, P. Pevzner, and M. Waterman

Editorial Board: A. Apostolico S. Brunak M. Gelfand
T. Lengauer S. Miyano G. Myers M.-F. Sagot D. Sankoff
R. Shamir T. Speed M. Vingron W. Wong

Subseries of Lecture Notes in Computer Science

Lecture Notes in Bioinformatics

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VII, 133 pages. 2005.

Vol. 3370: A. Konagaya, K. Satou (Eds.), Grid Computing in Life Science. X, 188 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VII, 115 pages. 2005.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004.

Vol. 3082: V. Danos, V. Schachter (Eds.), Computational Methods in Systems Biology. IX, 280 pages. 2005.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004.

Vol. 2812: G. Benson, R.D.M. Page (Eds.), Algorithms in Bioinformatics. X, 528 pages. 2003.

Vol. 2666: C. Guerra, S. Istrail (Eds.), Mathematical Methods for Protein Structure Analysis and Design. XI, 157 pages. 2003.

Preface

The Computational Methods in Systems Biology (CMSB) workshop series was established in 2003 by Corrado Priami. The purpose of the workshop series is to help catalyze the convergence between computer scientists interested in language design, concurrency theory, software engineering or program verification, and physicists, mathematicians and biologists interested in the systems-level understanding of cellular processes. Systems biology was perceived as being increasingly in search of sophisticated modeling frameworks whether for representing and processing system-level dynamics or for model analysis, comparison and refinement. One has here a clear-cut case of a must-explore field of application for the formal methods developed in computer science in the last decade.

This proceedings consists of papers from the CMSB 2003 workshop. A good third of the 24 papers published here have a distinct formal methods origin; we take this as a confirmation that a synergy is building that will help solidify CMSB as a forum for cross-community exchange, thereby opening new theoretical avenues and making the field less of a potential application and more of a real one. Publication in Springer's new Lecture Notes in Bioinformatics (LNBI) offers particular visibility and impact, which we gratefully acknowledge.

Our keynote speakers, Alfonso Valencia and Trey Ideker, gave challenging and somewhat humbling lectures: they made it clear that strong applications to systems biology are still some way ahead. We thank them all the more for accepting the invitation to speak and for the clarity and excitement they brought to the conference. We also wish to thank René Thomas for his keynote lecture on recent mathematical advances in the qualitative analysis of genetic regulation networks. As one can tell from the proceedings, his work has inspired many recent applications of formal methods to the engineering of biological models.

We are glad to take here the opportunity to express our gratitude to the members of the program committee and to the referees for their effort in the paper selection process and for their willingness to participate in the open-minded debate needed given the interdisciplinary nature of the area of computational systems biology. We would also like to thank the authors for their interest in the workshop and for their high-quality submissions and communications.

Finally, we wish to extend our warmest thanks to Monique Meugnier, Catherine Sarlande and Serge Smidtas for their invaluable help in organizing the workshop, and to the participating institutions, Genoscope, Genopole, CNRS, University of Paris 7, and the BioPathways Consortium, which provided financial support.

Conference web-site: `http://www.biopathways.org/CMSB04/`

Vincent Danos
Vincent Schachter

Table of Contents

Long Papers

Short Papers

Invited Contributions

An Explicit Upper Bound for the Approximation Ratio of the Maximum Gene Regulatory Network Problem

Sergio Pozzi[1], Gianluca Della Vedova[2], and Giancarlo Mauri[1]

[1] DISCo, Univ. Milano-Bicocca
[2] Dip. Statistica, Univ. Milano-Bicocca
sergio.pozzi@disco.unimib.it
{giancarlo.mauri, gianluca.dellavedova}@unimib.it

Abstract. One of the combinatorial models for the biological problem of inferring gene regulation networks is the MAXIMUM GENE REGULATORY NETWORK PROBLEM, shortly MGRN, proposed in [2]. The problem is **NP**-hard [2], consequently the attention has shifted towards approximation algorithms, leading to a polynomial-time 1/2-approximation algorithm [2], while no upper bound on the possible approximation ratio was previously known.

In this paper we make a first step towards closing the gap between the best known and the best possible approximation factors, by showing that no polynomial-time approximation algorithm can have a factor better than $1 - \frac{1/8}{1+e^2}$ unless **RP**=**NP**.

1 Introduction

The completion of the Human Genome project [9, 3] has only given more importance to the problem of determining the processes regulating the methabolism of living beings. The knowledge of all genetic sequences of an organism is just the first necessary step in understanding which of these sequences determine how those sequences are actually related to the phenotypes, as it is commonly believed that the dynamics of a living organism is determined throught some complicated and orchestrated interactions between thousands of genes and their products.

A Gene Network can be thought of as a set of molecular components such as genes, proteins and other molecules, interacting to collectively carry out some cellular functions. The advent of DNA microarray technology has led to easily obtaining huge amount of data regarding various aspects of cellular behavior, making possible to identify the interactions occuring among the various elements of a genetic system. Anyway the amount of data does not imply that the overall quality of data is sufficient to understand the various interaction, in fact these data are actually insufficient in granularity to uniquely determine the underlying network of interactions. Building the complex causal gene network of a genetic system on the basis of these sampled data is then a tipical inference and reverse engineering task.

V. Danos and V. Schachter (Eds.): CMSB 2004, LNBI 3082, pp. 1–8, 2005.

A number of different gene network models have been proposed in literature, each of them resorting to some simplifying assumptions either of biological or computational nature. In this paper we will study the boolean network models, where the state of each gene can be only dicotomic, that is active or not active. This model was already recognized to give a valid description of a genetic system in [7]. Boolean models are rich enough to represent interesting interactions among elements and, even if they are sometimes too simplistic [6], they allow to analyze briefly more complex systems. Actually this fact does not detract to the result of our paper, as we will show that a certain formulation of the gene network inference problem cannot be approximated efficiently, and this inapproximability result is very likely to be extended to more refined models.

In [1] modeling genes as boolean switches has allowed to study the problem of reverse engineering the gene networks by devising experiments in which, these switches are strategically manipulated (turned on and off) and then observing the behavior of the whole system. The main limit of this model is that the number of experiments that have to be performed in order to reconstruct a gene network of bounded in-degree D over n genes is $\Omega\left(n^D\right)$. In an other boolean GN model [5], the causal relations among network elements is derived on the basis of the mutual information among them. In our paper we will analyze a particular boolean model introduced in [2]. As will be explained in Sect. 2, this model is based on a simple combinatorial description with some biological evidence.

In [2], the problem of determining the causal relations among network elements has been proved to be **NP**-hard, consequently there has been much attention to designing approximation algorithms for the problem. The best known result in such direction is the 1/2-approximation algorithm of [2] (in this paper the approximation ratio of an algorithm is an upper bound of the ratio between the value of the approximate solution and the value of an optimal solution).

In this paper a first inapproximability result for the gene network inference problem based on this model is derived, by showing that it is unlikely that there exists an efficient approximation algorithm that can guarantee to obtain a $1 - \frac{(1/8)}{1+e^2}$ ratio.

Our paper is organized as follows: initially we will present formally the MGRN problem, together with some known approximability results.

Successively we will give a probabilistic reduction from instances of MaxE3Sat to MGRN ones. This reduction uses a previuolsy known reduction from instances of MaxE3Sat to instances of MaxE3Sat-B, originally proposed by Trevisan [8]. Our reduction extends the Trevisan reduction to the MGRN problem. We will conclude the paper by showing that a consequence of our reduction is that no polynomial-time approximation algorithm for the MGRN problem with approximation ratio $1 - \frac{1/8}{1+e^2}$ can exist, unless **RP**=**NP**.

2 The Maximum Gene Regulatory Problem

A Genetic Network in which an element can only activate or inhibit other elements, can be viewed as a directed graph in which the nodes represent the genes

and arcs represent the interactions between genes. Moreover each arc (v, w) is labeled by A or I, according to the fact the gene represented by v *activates* or *inhibits* the activity of the gene represented by w. Such graphs can be built using experiments data relative to gene expression dynamics [2]. In order to suggest the causal genetic network on the basis of the edge labeled directed graph, activating/inhibiting edges representing spurious interactions must be deleted. The task of deleting spurious interactions has to be done with the following constraints:

- A gene (a node on the graph) cannot be both of activating and inhibiting type.
- The number of genes that are *controlled* (that is vertices that have both A-labeled and I-labeled incoming arcs) must be maximized.

Both kinds of constraints find their justification in biological evidence and consistency with the parsimonious principle. As a final result of these two guiding assumptions a combinatorial optimization problem on graphs has been defined in [2]

Problem 1. MAXIMUM GENE REGULATION PROBLEM (shortly MGRN). The instance is a directed graph $G = (V, E)$, where each arc is labeled by either A or I. The goal is assigning to each vertex a label that is either A or I, so that, after deleting all arcs (v, w) with label different from that of v, the number of *controlled* vertices is maximized.

It is hopeless to devise efficient exact algorithm for the MGRN problem, since the problem is **NP**-hard, even for directed acyclic graphs of constant in/out-degree [2]. For this reason in the last few years the attention has been turned into finding efficient approximate solutions, showing that the solution having at least one half of the optimal number of controlled vertices, can be found in polynomial time [2], but it was not previously known if a polynomial-time approximation scheme (PTAS) was possible for such problem.

In our paper we will settle the question, by proving that it is not possible to describe a polynomial-time approximation algorithm with guaranteed ratio strictly better than $1 - \frac{1/8}{1+e^2}$, unless **RP**=**NP**.

3 A Better Reduction

In [2] it has been proved that MGRN is **NP**-hard. Here we will show a new reduction from MAXE3SAT-B to MGRN; our reduction is stronger, since it allows to prove a better inapproximation results.

The reduction associates to an instance of MAXE3SAT-B an instance of MGRN as follows: for each clause C_i we have a *clause gadget* consisting of two vertices C_i^1, C_i^2 and the A-labeled arc from C_i^1 to C_i^2. For each variable x_i we have the *variable gadget* consisting of two vertices x_i^T, x_i^F and no arc. If the total number of variables is n, then we have also $\lfloor n/2 \rfloor$ *assignment gadgets*, each gadget is made of $2(B+1)$ vertices, half of which are labeled red and half are labeled

blue. For $1 \leq i \leq \lfloor n/2 \rfloor$ all vertices x^T_{2i-1}, x^F_{2i-1}, x^T_{2i}, x^F_{2i} have an outgoing arc to each of the vertices of the i-th assignment gadgets. More precisely all red vertices have A-labeled arcs incoming from x^T_{2i-1} and x^F_{2i-1} and I-labeled arcs incoming from x^T_{2i} and x^F_{2i}, while all blue vertices have I-labeled arcs incoming from x^T_{2i-1} and x^F_{2i-1} and A-labeled arcs incoming from x^T_{2i} and x^F_{2i}. An assignment gadget and the two corresponding vertex gadgets are represented in Fig. 1.

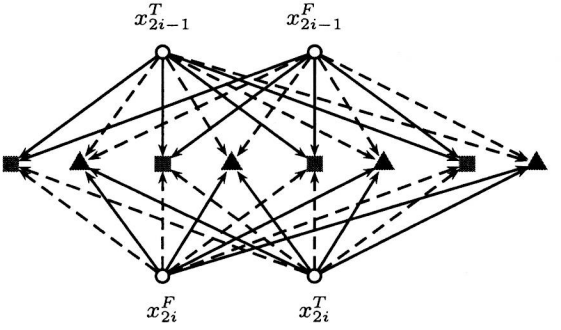

Fig. 1. Example of vertex and assignment gadget, red vertices are represented by squares and blue vertices by triangles. A-labeled edges in solid lines, I-labeled edges in dashed lines

Actually there is a minor problem if the number of variables is odd. In this case the last assignment gadget is different, as three variables are connected to it, as shown in Fig. 2.

The reduction can now be completed with the encoding of each clause $C_i = x^{\alpha_{i_1}}_{i_1} \vee x^{\alpha_{i_2}}_{i_2} \vee x^{\alpha_{i_3}}_{i_3}$, where each exponent α_{i_j} is equal to T or F, according to the fact that the corresponding variable is or is not negated in the clause. For each clause C_i there are three I-labeled arcs incoming in C^2_i and outgoing from vertex gadgets associated to the vertices appearing in the clause, more precisely the arcs are outgoing from the actual vertices encoding the variable and the fact that the variable is or is not negated in the formula. Formally for each clause $C_i = x^{\alpha_{i_1}}_{i_1} \vee x^{\alpha_{i_2}}_{i_2} \vee x^{\alpha_{i_3}}_{i_3}$ there are the three arcs $(x^{\alpha_{i_1}}_{i_1}, C^2_i)$, $(x^{\alpha_{i_2}}_{i_2}, C^2_i)$, $(x^{\alpha_{i_3}}_{i_3}, C^2_i)$. In Fig.3 is represented an example of encoding.

In the following of the paper we will denote with F an instance of MAXE3SAT-B and with G the instance of MGRN that is associated to F with the reduction we have just described. Moreover we will denote with $opt(F)$ and $opt(G)$ respectively the maximum number of clauses of F that are satisfiable by a single assignment and the optimum of the instance G.

The following lemma is the foundation of our inapproximability result.

Lemma 3.1. *Let F be an instance of the* MAXE3SAT-B *problem with n_B boolean variables and m_B clauses, and let G be the instance of* MGRN *that is associated to F. Then it is possible to associate to any solution of F with value x a solution of G of value $n_B(B+1)+x$. Vice versa it is possible to associate to any solution of G of value $n_B(B+1)+x$ a solution of F of value at least x.*

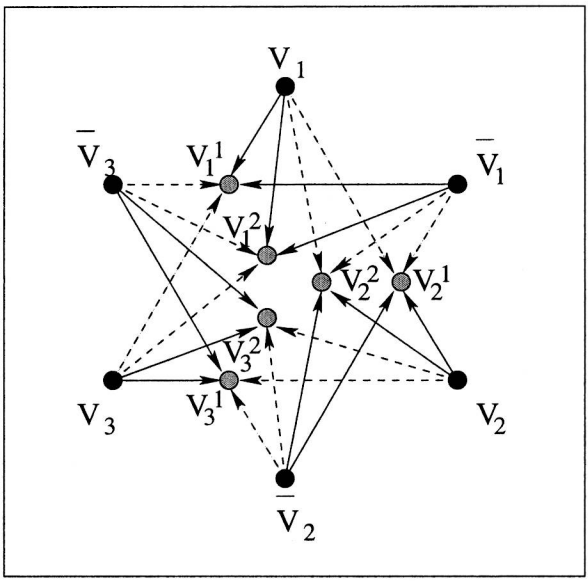

Fig. 2. Three variables connected to an assignment gadget with $B = 2$. A-labeled edges in solid lines, I-labeled edges in dashed lines

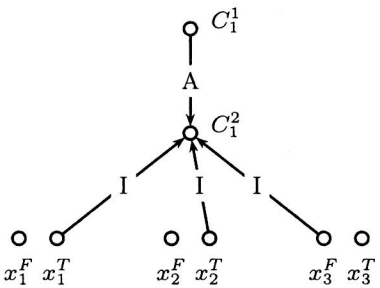

Fig. 3. Encoding of the clause $C_1 = x_1 \vee x_2 \vee \neg x_3$

Proof. Initially assume we have a solution of F. Please notice that the vertices that are relevant in computing the value of the solution are the vertices which have both A-labeled and I-labeled arcs and, by construction, the only such vertices are the assignment vertices and the vertices C_j^2.

Without loss of generality we can assume that all assignment vertices must be controlled. In fact if all assignment vertices of two certain variable x_i are not controlled and i is odd, then we obtain a better solution by A-labeling x_i^T, x_{i+1}^T and I-labeling x_i^F, x_{i+1}^F (if i is even, then we have to A-label x_i^T, x_{i-1}^T and I-label x_i^F, x_{i-1}^F). Now all $2B + 2$ assignment vertices of x_i are now controlled, but we do not know how many of the $2B$ clause gadgets to which x_i and x_{i+1} are connected are controlled. In the worst case they all were controlled before the

modification, and now none of them is controlled. Anyway after the modification the total number of controlled is increased by at least two, so the new solution is better than the previous one.

Now we can assume that all assignment vertices are controlled. This condition is equivalent to say that, for each odd i, exactly one of x_i^T and x_i^F is A-labeled and exactly one of x_{i+1}^T and x_{i+1}^F is A-labeled, which can be shown by trying all possibilities.

Since exactly one of x_i^T and x_i^F is A-labeled, we can assume that the labeling encodes a truth assignment, that is x_i is true if and only if x_i^T is A-labeled. By construction each vertex C_j^2 is controlled if and only if at least one of the variable vertices to which it is connected is A-labeled, which in turn means that the assignment of corresponding variables makes the clause C_j true. Consequently the number of vertices C_j^2 that are controlled is exactly x, where x is equal to the number of clauses that are satisfiable.

Now we are able to prove the second part of the lemma. Assume that we have a solution of G with value $n_B(B+1)+x$. Then, just as for the first part of the proof, it is immediate to obtain a solution of F with value at least x.

An immediate corollary of Lemma 3.1 is that if the instance F of MAXE3SAT-B is satisfiable, then the instance G of MGRN has optimum $n_B(B+1)+m_B$.

4 An Explicit Upper Bound

The starting point of our reduction is the MAXIMUM EXACT 3-SATISFIABILITY (MAXE3SAT) problem, where the instance is a boolean formula where each clause contains exactly 3 literals. For such problem some strong inapproximability results are known; in fact Håstad [4] has proved that for every $\delta > 0$, it is **NP**-hard to distinguish a satisfiable instance of MAXE3SAT from an instance where at most $7/8+\delta$ of the clauses can be simultaneously satisfied; we will call such problem a *gapped* version of MAXE3SAT.

Bulding upon the last result by Håstad, Trevisan [8] has devised a stochastic reduction from an instance I of MAXE3SAT to an instance F of MAXE3SAT-B, that is in F each literal appears in at most B clasuses. In the following we will denote by I, F and G respectively an instance of MAXE3SAT, MAXE3SAT-B and MGRN. Morever we will denote by n, m respectively the number of variables and of clauses of I, by n_B, m_B respectively the number of variables and of clauses of F. Consequently the maximum number of vertices of G that might be controlled is $(B+1)n_B+m_B$. The probabilistic reduction of [8] has the following properties:

1. if I is satisfiable then F is satisfiable;
2. for any sufficiently large B, then with probability at least $3/4-o(1)$ over the random choices made in the construction of F, if there is an assignment that satisfies at least a fraction $7/8+5/\sqrt{B}$ of the clauses of F, then there is an assignment that satisfies at least a fraction $7/8+1/\sqrt{B}$ of the clauses of I; furthermore $m_B > \left(\frac{B}{e^2}-4\right)n_B$.

Following the same ideas of [8], we will give a probabilistic reduction from instances of MAXE3SAT to instances of MGRN. Our reduction is actually a composition of the reduction in [8] (i.e. a reduction from MAXE3SAT to MAXE3SAT-B) and the reduction proposed in Sect. 4 (i.e. a reduction from MAXE3SAT-B to MGRN). Now we are ready to prove the fundamental feature of the our probabilistic reduction from MAXE3SAT to MGRN.

Lemma 4.1. *Let I be an instance of* MAXE3SAT *with n variables and m clauses, and let G be instance of* MGRN *associated to I by our reduction. Then for sufficiently large B and with probability at least $3/4-o(1)$, if there is a label assignment to the vertices of G such that at least* $\left(1-\frac{\frac{1}{8}-\frac{5}{\sqrt{B}}}{1+\frac{(B+1)e^2}{(B-4e^2)}}\right)((B+1)n_B+m_B)$ *vertices are actually controlled, then at least* $\left(7/8+1/\sqrt{B}\right)m$ *clauses of I can be satisfied.*

Proof. Let I be an instance of MAXE3SAT and let us suppose there exists a label assignment a to vertices of G such that its measure $m(a)$ is at least $\left(1-\frac{\frac{1}{8}-\frac{5}{\sqrt{B}}}{1+\frac{(B+1)e^2}{(B-4e^2)}}\right)((B+1)n_B+m_B)$, then $m_B > \left(\frac{B}{e^2}-4\right)n_B$ with probability at least $3/4-o(1)$. Consequently

$$
\begin{aligned}
m(a) &\geq \left(1-\frac{\frac{1}{8}-\frac{5}{\sqrt{B}}}{1+\frac{e^2(B+1)}{B-4e^2}}\right)((B+1)n_B+m_B) > \\
&> \left(1-\frac{\frac{1}{8}-\frac{5}{\sqrt{B}}}{1+(B+1)\frac{n_B}{m_B}}\right)((B+1)n_B+m_B) = \\
&= \left(1-\frac{\left(\frac{1}{8}-\frac{5}{\sqrt{B}}\right)m_B}{m_B+(B+1)n_B}\right)((B+1)n_B+m_B) = \\
&= m_b+(B+1)n_B-\frac{1}{8}m_B+\frac{5}{\sqrt{B}}m_B = (B+1)n_B+\left(\frac{7}{8}+\frac{5}{\sqrt{B}}\right)m_B
\end{aligned}
$$

By Lemma 3.1 there exists a solution of F satisfying at least $(7/8+5/\sqrt{B})m_B$ clauses. Applying the second property of the reduction in [8], there exists (with probability at least $3/4-o(1)$) a solution of I satisfying at least $(7/8+1/\sqrt{B})m$ clauses.

The following corollary is our main contribution.

Corollary 4.2. *For any $\delta > 0$, it is not possibile to approximate the* MGRN *problem within a factor $1-\frac{1/8}{1+e^2}+\delta$, unless* ***NP=RP****.*

Proof. First notice that it is not possible to approximate the MGRN problem within a factor $1-\frac{\frac{1}{8}-\frac{5}{\sqrt{B}}}{1+\frac{(B+1)e^2}{(B-4e^2)}}$ unless **NP=RP**. Otherwise we could solve

the gapped version MAXE3SAT with a polynomial-time probabilistic algorithm. In fact let I be an instance of such gapped version, and let G be the instance of MGRN associated to I. If the solution returned by the approximation algorithm has value more than $\left(1 - \frac{\frac{1}{8} - \frac{5}{\sqrt{B}}}{1 + \frac{(B+1)e^2}{(B-4e^2)}}\right)((B+1)n_B + m_B)$ then, by Lemma 4.1, with probability $3/4 - o(1)$, I is a satisfiable instance of gapped MAXE3SAT. Otherwise the solution returned by the algorithm has value at most $\left(1 - \frac{\frac{1}{8} - \frac{5}{\sqrt{B}}}{1 + \frac{(B+1)e^2}{(B-4e^2)}}\right)((B+1)n_B + m_B)$ consequently, with probability 1, at most a fraction $7/8 + \delta$ of the clauses are satisfied, hence we would have an algorithm in **RP**for the gapped version of MAXE3SAT. This would imply that **NP=RP**.

Without loss of generality we can restrict our interest only to large values of B. Since $\lim_{B\to\infty} 1 - \frac{\frac{1}{8} - \frac{5}{\sqrt{B}}}{1 + \frac{(B+1)e^2}{(B-4e^2)}} = 1 - \frac{1/8}{1+e^2}$, for any $\delta > 0$ taking a sufficiently large B completes the proof.

Acknowledgments

This work has been partially supported by FIRB project "Bioinformatica per la Genomica e la Proteomica".

References

1. T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano. Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. *Proc. 9th Symp. on Discrete Algorithms (SODA)*, pages 695–702, 1998.
2. T. Chen, V. Filkov, and S. S. Skiena. Identifying gene regulatory networks from experimental data. *Parallel Computing*, 27:317–330, 1999.
3. I. H. G. S. Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, February 2001.
4. J. Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48:798–859, 2001.
5. S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Proc. 5th Pacific Symposium on Biocomputing (PSB)*, pages 18–29, 1998.
6. C. Soulé. Graphic requirements for multistationarity. *ComPlexUs*, 1:123–133, 2003.
7. R. Thomas, A. Gathoye, and L. A. Lambert. A complex control circuit. regulation of immunity in temperate bacteriophage. *European Journal of Biochemistry*, 71:211–227, 1976.
8. L. Trevisan. Non-approximability results for optimization problems on bounded degree instances. *Proc. 33rd Symp. Theory of Computing (STOC)*, pages 453–461, 2001.
9. J. C. Venter, M. D. Adams, E. W. Myers, and et. al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.