SIGBio ACM

Bioinformatics, Computational Biology and Biomedical Informatics

# SIGBio Record

Newsletter of the SIGBio
ACM Special Interest Group

Volume 3, Issue 2, May 2013 ISSN 2159-1210

## Notice to Contributing Authors to SIG Newsletters

# Notes from Chair

## SIG name change:

Following the SIG Community Meeting held on Oct. 8, 2012 in Orlando, majority of members favor to change our SIG name to reflect more closely our full scope of Bioinformatics, Computational Biology and Biomedical Informatics. We requested a name change and the ACM SGB-EC has endorsed a name change for SIGBioinformatics to SIGBio. The scope in our bylaw is amended as follows:

New Article 1 of the SIGBio Bylaws is to read as follows:

**Article 1. Name and Scope**

a. This organization will be called the Special Interest Group on Bioinformatics, Computational Biology and Biomedical Informatics (SIGBio) of the Association for Computing Machinery, Inc. (the ACM).

b. The scope of the SIGBio specialty is bioinformatics, computational biology and biomedical informatics.

Aidong Zhang

# Editor in Chief's Notes:

This issue will present some novelties. First of all, we are pleased to announce the change of the name of the Newsletter in SIGBio record. The change reflects the change of the name of the SIG.Consequently newsletter will have a broader scope. We are also pleased to announce the collaboration of Prof. Pierangelo Veltri as novel associate editor.

This issue presents two contributed articles:

Information Entropy Based Methods for Genome Comparison that explores methods for comparison of Genomes

Computational regulatory network construction from microRNA and transcription factor perspectives, that discusses a novel research area on computational biology.

We thank contributors for this issue and hope that readers will find interesting references to their work in Bioinformatics and Health Informatics area.

Pietro Hiram Guzzi

Young-Rae Cho
Pierangelo Veltri

# Computational regulatory network construction from microRNA and transcription factor perspectives

Sungmin Rhee
School of Computer Science
and Engineering,
Bioinformatics Institute
Seoul National University
Seoul, Korea
lars@snu.ac.kr

Jinwoo Park
School of Computer Science
and Engineering,
Bioinformatics Institute
Seoul National University
Seoul, Korea
jw.park.bioinfo@gmail.com

Sun Kim
School of Computer Science
and Engineering,
Bioinformatics Institute
Seoul National University
Seoul, Korea
sunkim.bioinfo@snu.ac.kr

## ABSTRACT

As more genomic and epigenomic data becomes available, it has become possible to construct biological networks from the omics data. Among the biological networks, understand- ing gene regulatory mechanisms is a very important research problem that can reveal condition-specific, e.g., disease-specific, gene regulatory mechanisms. In this paper, we review the current development in the study of constructing gene reg- ulatory networks from microRNA and transcription factor (TF) perspectives. TFs and microRNAs play crucial role in gene regulatory networks since they regulate tens to hun- dreds of genes, which can be seen naturally as hubs in the network. This review consists of three parts. The first part summarizes recent works on TF regulatory network reconstruction in two sections, one on TF network reconstruction using time series gene expression data and the other on TF network construction by incorporating prior knowledge. The second part is about microRNA network construction in two sections, one on methods based on seed sequence matching and the other on the integrated analysis of gene and mi- croRNA expression data sets. The last part summarizes recent works on the integration of both TF and microRNA with target genes, which is a much more challenging research problem.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

## General Terms

Theory

## Keywords

MicroRNA, Transcription factor, Target prediction, Regulatory network

## 1. INTRODUCTION

Genome-wide data such as ChIP-based assay, ChIP-chip and more recently ChIP-seq, data for transcriptional networks or whole genome transcriptome data have been accumulat- ing rapidly. Such genome wide data can be efectively used to construct or infer gene regulatory networks. Gene reg- ulation is a dynamic process in cells, responding to stimuli by various agents such as drugs, bacteria, virus, and harsh weather conditions. Thus understanding the dynamics of the networks is very helpful studying biological mechanisms including diseases [7]. When analyzing the network, one of the most important part of the analysis is to identify the core or hub of the network. TFs and microRNAs (miRNAs) regulate up to several hundreds of genes, so they can be seen naturally as the core of the gene regulatory networks. Thus, in this survey, we classified gene regulatory networks into three categories: TF involved gene regulatory networks, miRNA involved gene regulatory networks and gene regula- tory network involving both TF and miRNA simultaneously. There have been many studies for the first and second cate- gories, but study on the last category is just in the beginning stage, probably because of its complexity of performing in- tegrated analysis.

## 2. TF INVOLVED GENE REGULATORY NETWORKS

There are many computational methods to infer transcriptional regulatory networks. Survey based on the network model architecture is the most popular way [4][7]. However, our survey is based on the types of data used to infer net- works so that the survey can be more practical for users. Network construction methods in this survey are classified into two categories: time-series gene expression data based approach and prior knowledge based approach.

### 2.1 Time-series gene expression data based approach

Time-series gene expression data is necessary to understand biological processes since the biological processes are dy- namic and often time dependent [1]. Thus many genome- wide gene expression data are time series data after pre- designed stimuli given, for example, drug treatment.

Ernst et al. introduced the DREM algorithm to model dy- namic gene regulatory events and it used input-output hidden Markov model to integrate time series expression data with static ChIP-chip or motif data [2]. It takes a binary matrix of predictions of TF-gene regulatory interactions and time-series log-ratio gene expression data against the unstressed control as inputs. The algorithm models expression patterns as series of bifurcation and assigns genes to paths according to each gene's expression pattern. Then, to each bifurcation points, DREM assigns TFs that regulate the genes. This method was used to construct networks using yeast response data and recovered many of the known gene interactions. It also predicted unknown interactions that were validated experimentally. DREM 2.0 [14] is the most recent version.

Li et al. introduced DELDBN [12] that integrated ordinary diferential equation (ODE) models with the dynamic Bayesian network analysis. Steady-state equation (1) is applied to short sampling time interval data and dynamic state equation (2) is applied to long sampling time interval data.

$$x_i(t + 1) = \beta_{ij}X_j(t) \tag{1}$$

$$\frac{x_i(t + 1) - x_i(t)}{\Delta t} = \beta_{ij}X_j(t) \tag{2}$$

Where $x_i(t)$, $x_j(t)$ are the expression level of gene $i$, gene $j$ at time $t$ respectively and $\beta_{ij}$ is the efect of gene $j$ on gene $i$.
Then DELDBN uses local causality based dynamic Bayesian network analysis to learn through the above two equations. Unlike other Bayesian network analysis, DELDBN uses a low time complexity algorithm and it is scalable to infer large networks. An *in vivo* benchmark data set from yeast was used to demonstrate the performance of the algorithm, showing the highest sensitivity and accuracy in comparison with other approaches. To show the scalability, DELDBN inferred the BRCA1 network using the human Hela cell time series gene expression data that is larger than the typically used yeast data set.

Song et al. introduced KELLER [16], a kernel-reweighted logistic regression method, to infer the latent time-evolving network of gene interactions. With the assumption that time-evolving networks change smoothly, similarity between networks measured in a close time interval is higher than networks measured in a far time interval. Therefore, the problem of estimating dynamic networks can be reduced to inferring a series of static networks by aggregating temporally adjacent networks by reweighting them. To evaluate the algorithm, a microarray gene expression data of *Drosophila melanogaster* with 66-step time series in a full life cycle of 4028 genes was used. The analysis focused 588 genes that were known to be related in the developmental process. KELLER successfully inferred time-evolving network of *Drosophila melanogaster* and showed that many genes had diverse functionalities that were diferent at each stages.

## 2.2 Prior knowledge based approach

Although the cost for generating genome-wide data is decreasing rapidly, it is still difcult to obtain enough omics data in one experiment due to the limited budget and time. More importantly, knowledge obtained from many previous studies are valuable. Thus developing computational methods that can incorporate prior knowledge is very important.

Li et al. [11] introduced an network-constrained regularization method that integrated prior knowledge with the form of networks like pathways. Predictors in the model are expression levels of genes with underlying network structures that can be obtained from prior knowledge. It presents a network-constrained penalty which is aggregated form of the lasso penalty and the penalty of Laplace matrix. Network-constrained regularization criterion is defined as follows,

$$L(\lambda_1, \lambda_2, \beta)=(\mathbf{y}-\mathbf{X}\beta)^{\mathbf{T}}(\mathbf{y}-\mathbf{X}\beta)+ \lambda_1|\beta_1|+ \lambda_2\beta^{\mathbf{T}}\mathbf{L}\beta \tag{3}$$

Where $\mathbf{y}$ is response vector, $\mathbf{X}$ is design matrix, $\lambda_1 > 0$ and $\lambda_2 > 0$ are user defined constants, $\beta_1 = {}_{j=1}^{p} |\beta_j|$ is the $L_1$-norm which leads sparseness of the result and $\beta^T L\beta$ leads smoothness. With the criterion, estimator $\beta = \hat{}$ $argmin_\beta L(\lambda_1, \lambda_2, \beta)$ is obtained. The algorithm was applied to the microarray gene expression data for glioblastoma. With two independent groups of clinical data, one was used for training samples and the other was used for the test set. As a prior knowledge based data, 33 KEGG regulatory pathways were used and the goal was to find disease related subnetworks. The analysis successfully discovered subnetworks that were known to be related with glioblastoma.

Greenfield et al. [3] developed two methods that incorporated prior knowledge for the analysis of time series or static gene expression data to infer dynamic gene regulatory networks. Both methods used the same ordinary diferential equation model below.

$$\frac{dx_i}{dt} = -\alpha x_{ii} + {}_{p} \beta_{i,p}x_p, i=1,...,N \tag{4}$$

Where $x_i$ is gene, $\alpha > 0$ is the first order degradation rate, $\beta$ is a set of parameters to be estimated and $P_i$ is the set of potential regulators for $x_i$. It is from the assumption that a gene is regulated in proportion to the amount of regulators and the gene itself. Based on this equation, two methods, Modified Elastic Net (MEN) and Bayesian Best Subset Regression (BBSR), are proposed. MEN is a modified form of existing regression application called *Elast-Net* and BBSR is a Bayesian regression based approach with Zellner's $g$ prior. It was shown that the proposed methods by utilizing prior knowledge were tolerant to the errors in the expression data.

## 3. MIRNA NETWORK INFERENCE

The miRNA network inference problem is to infer a network of miRNA and mRNA of protein coding genes. Genes targeted by miRNAs are down regulated since miRNA interferes with coding genes at the transcription and translation levels. The miRNA network inference problem can be largely divided into two sub-topics. The first one considers only sequence paring information between miRNA and mRNA since miRNAs interfere with mRNAs by hybridization, i.e., sequence pairing. This sequence only prediction method can be a good way for finding putative targets of miRNAs but they usually have very high false positive rates. The second method incorporates expression profiles of miR- NAs and mRNAs for the miRNA network construction. The main idea of the second method is to utilize negative rela- tionship between miRNA expression level and mRNA ex- pression level when a target relationship holds.

## 3.1 Sequence pairing algorithms for target finding

It is well known that miRNA binds to a reverse complement sequence in the 3 UTR region of mRNA and the corresponding mRNA is degraded. In eukaryotes, a seed re- gion of miRNA at the five prime end site that matches with mRNA is a predominant factor of mRNA repression. This information can be used for algorithms of finding uncovered miRNA target mRNAs.

TargetScan[10] utilizes that many miRNAs and their tar- get sites are conserved across the multiple species (human, mouse, puferfish). Short sequences of 2-7 nucleotides of miRNA are defined as the seed region and they are matched perfectly to 3 UTR. After the perfect matching, a thermo- dynamics based binding score between miRNAs and puta- tive targeted regions of mRNAs is calculated and it is used to rank the targeted genes. The final selection of targets are determined by a pre-selected rank threshold and a binding score threshold.

PicTar[9], like TargetScan, uses the conserved target site information across species and the thermodynamics based binding score for target inferencing. First, PicTar locates all possible target sites using nuclMap. Putative target sites are filtered out if the free energy between the miRNA and the targeted mRNA is higher than a preset cutof value. When there are multiple binding sites in 3 UTR region of mRNA, PicTar uses the maximum likelihood score to sort out true target sites. The score is based on the posterior probabilities of the binding sites that are targeted by the miRNAs compared to background of 3 UTR regions that are not putative miRNA binding sites. This score can sort out competition between diferent miRNAs on the same re- gions and can reduce the false positive rate by considering background probability of 3 UTR region.

Kertesz et al.[8] proposed another thermodynamics-model based approach, PITA, that consideed secondary structure opening energy for finding miRNA target recognition. Like TargetScan, PITA looks for perfect matches in the seed re- gion and calculates binding score between the miRNA and it's putative target mRNA. In addition, PITA considers the site accessibility of target sites. miRNA and their target mRNA can have a secondary structure and should be un- paired so that miRNA can attach to mRNA and then repress the transcription of mRNA. This structure based condition is enforced by calculating the free energy that is required to unpair the secondary structure of target sites and miRNA's secondary structure. RNAFold is used for this calculation. The final miRNA target interaction score is computed as the diference between the binding score and the secondary structure opening score.

## 3.2 Analysis with expression profiles

Although binding sites of miRNAs and their target regions of mRNAs have similar sequences, the sequence analysis alone cannot solve the high false-positive rate problem since core sequences are very short. miRNAs repress mRNA at both transcription and translational level. For the transcriptional repressing, mRNA transcript expression decreases as expres- sion levels of miRNA that targets the transcript increase. Thus a miRNA:mRNA pair that shows a strong negative correlation in their expression level have a high probability of being a genuine target pair. This negative correlation information is embedded in several computational methods.

MMIA[18] is a method for the integrated analysis of miRNA and RNA expression data. The integrated analysis is per- formed in two steps. The first step is to identify "difer- entially" expressed miRNAs by clustering analysis. In the second step, only genes that are targeted by diferentially expressed miRNAs are considered. The gene set is fur- ther reduced by using sequence based target finding algo- rithms such as TargetScan, PITA and PicTar, and also by using negative correlation information between miRNA and mRNA expression levels. MMIA divides the miRNA and mRNA's expression data to three clusters: a down-regulated group, an up-regulated group and an unchanged group. It predicts miRNA:mRNA target pairs when the miRNA be- longs to a down(up)regulated group and mRNA belongs to a up(down)regulated group. This approach can assure that finding genuine and actual working miRNA:mRNA pairs but also misses many actual miRNA:mRNA pairs whose expres- sion is not significantly up(down) in the cell.

Muniategui et al.[13] proposed a linear model for indicat- ing the degree of miRNA's repressing mRNA transcription. When sequence based algorithms report that $K$ miRNA are predicted to target mRNA $j$ and $c_{jk}$ is an indicator that miRNA $k$ putatively target $j$-th mRNA, a model as below is used:

$$x_j = \sum_{k=1}^{K} \beta_{jk} c_{jk} z_k + x^0 +_j \qquad j$$

where $_j$ is an error term and $x^0$ is an logarithm of the ex- $_j$ pression values when no miRNA targets the mRNA. With this model, Lasso regression is used to finding $\beta$ values min- imizing below equation.

$$\min_{\beta_j x_j^{0}} \{ x_j - \sum_{k=1}^{K} \beta_{jk} c_{jk} z_{jk} - x^0{}_j \quad {}_2 + \lambda_j * \sum_{k=1}^{K} |\beta_{jk} c_{jk}| \}$$

Lasso regression with a constraint that $\beta$ should be non-positive for$_K$ dicating only down-regulation of miRNA efect in and $\lambda_j *_{k=1} |\beta_{jk} c_{jk}|$ is the penalty term for enforcing the sparsity of solution.

GenMir++[5] uses a linear model for expected mRNA ex- pression values based on the equation below:

$$E[x_{gt} | \{s_{gk}\}, \{z_{kt}\}, \Lambda, \mu_t, \gamma_t] = \mu_t - \gamma_t \sum_{k} \lambda_k s_{gk} z_{kt}, \ \lambda_k > 0$$

when $x$ is mRNA expression values, $s_{gk}$ is an indicator that $k$ miRNA targets $g$ mRNA, $\mu$ is the background expression of mRNA, $\gamma$ is the tissue scaling factor. Using this lin- ear model, a Bayesian network model is proposed. In the

Bayesian network model, target transcript expression level $x$ is dependent on a tissue scaling parameter, the miRNA expression level, a regulatory weight, an indicator variable for whether miRNA $k$ truly targets transcript $g$ and this indicator is dependent on an indicator variable for miRNA $k$ putatively targets transcript $g$. $P(S| X, Z, C, \Theta)$ is esti- mated using the Bayesian inference and the expectation- maximization technique.

Joung et. al. [6] proposed a module based target finding algorithms using the co-evolutionary machine learning ap- proach and the estimation-of-distribution algorithm (EDA). The detection of modules, $(M, T)$ between miRNA set and mRNA set that best fit in terms of the fitness function (see below) needs to consider all subsets of $M$ and $T$ which is computationally infeasible. Thus an evolutionary algorithm was used to find the optimal solution. The fitness function
is:

$$F(M, T) = \alpha BS_{M} \quad {}^{T} + \beta EC_M + \gamma EC_T + VOL$$

when $BS_{M\,T}$ is a mean binding score between all pairs of $M$ and $T$, $EC_M$ and $EC_T$ are the expression coherence scores of $M$ and $T$ each, and $V\,OL$ is the volume term to prevent finding a solution with one or two miRNA and mRNAs. Given the module fitness function, a co-evolutionary approach is used to find an optimal solution. For miRNA and mRNA, two populations are managed and learned in the
context of each other. Individuals are selected based on the fitness function from the two populations and the probability vector is updated. The probability vector denotes the prob- ability of choosing a miRNA or mRNA to a miRNA:mRNA target module. The updated probability vector is used to generate new population.

## 4. TF-MIRNA INTEGRATED ANALYSES

A gene can be regulated by both miRNA and TF, thus inferencing target relationship should consider TF and miRNA simultaneously.

Shalgi et al. [15] integrated widely used algorithms for the miRNA target detection and the TF target detection to construct regulatory networks. They analyzed these con- structed networks to find local (network motif, hub genes) and global (connectivity distributions) architectures in the network. For the network construction, TargetScan and Pic- Tar were used for miRNA target detection and TRANSFAC was used for TF target finding. They used miRNAs and their target genes that were conserved between 4 species (hu- man, mouse, rat and dog). To reduce TF-gene interaction candidates, only TF promoter regions conserved in ortholo- gous genes from mouse and rat were used. Then they calculated hypergeometric p-value to finds significant miRNA-TF co-occuring pairs. They compared the constructed network with random models, detecting target hubs genes. In con- structing randomized network, they preserved the number of genes per miR but shuffled the assigned genes randomly to each miR. The analysis result showed that when miRNA-TF
works cooperatively, TF tends to be regulated by miRNA or

TF regulates the miRNA forming feed-forward loops.

Sun et al. [17] identified TF-miRNA regulatory networks consisting of 3-node FFL(feed-forward network) and 4-node FFLs in glioblastoma (GBM). First, GBM related genes and GBM related miRNAs from previous studies were collected. Human TFs were extracted from TRANSFAC. Then Tar- getScan was used for finding miRNA-TF/gene repression, and MATCH$^{TM}$ was used for TF-gene/miRNA interaction. Co-regulated relationship among genes were predicted using the ARACNE software. The process collected TF-miRNA pairs that cooperatively regulate the same target genes us- ing a cumulative hypergeometric test in a similar fashion to the method used in R. shalgi et. al. [15]. Based on the false discover rate, co-regulating pairs were further filtered out. The proposed method was able to find GBM specific network components.

The techniques that we surveyed so far built computational frameworks by utilizing existing tools as components. Zacher et al. [19] developed a joint Bayesian inference approach. In- dicator variables are used for MiRNA functional activities ($S$) and TF functional activities ($T$). MiRNA functional activities influence miRNA expression and the mRNA expression. TF functional activities also influence mRNA expression in the constructed model. Gene expression levels are approximated by a linear combination of miRNA and TF activities as an equation below:

$$o_{jlc}|S, T, b_j, \omega, v^2\ N(b_j + {}_j \underset{k\in miRNA(j)}{} s_{kc}\omega_{kj} + \underset{k\in TF(j)}{} t_{kc}\omega_{kj}, v^2)_j$$

where $o_{jlc}$ is expression for gene $j$ in $l$-th replicate of experimental condition $c$, $\omega_{kj}$ are relative influences of miRNA and TF regulators, $b_j$ are reference expression level of mRNA $j$,
$S$ and $T$ are functional activities of miRNA and TF. The limma algorithm and MCMC sampling were used to esti- mate the parameters $w, S, T$. This is a comprehensive mod- eling technique that considered expression profiles of both TF and miRNA. However, the proposed model did not consider the fact that miRNAs can repress TFs or TFs can influence miRNAs.

## 5. DISCUSSION

We reviewed the recent development in constructing regulatory networks of miRNA and TF. By nature, miRNA and TF are hubs that regulate up to hundreds of genes, thus they are very important for the correct inference of biological net- works. We categorized computational techniques in three groups: TF-involved networks, miRNA-involved networks, and TF-miRNA integrated networks. Although there have been many successful studies for constructing networks us- ing omics data, techniques for inferring regulatory networks needs much more eforts. First more accurate methods for component tools need to be developed. Examples include methods for more accurate miRNA target or methods for TF target prediction. Second these component tools or tech- niques needs to be incorporated into coherent computational models, e.g. a joint Bayesian inference approach by Zacher et al. [19].

## 6. ACKNOWLEDGMENTS

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No.NRF-2012M3C4A7033341).

## 7. REFERENCES

[1] Z. Bar-Joseph, A. Gitter, and I. Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13(8):552-564, August 2012.

[2] J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph. Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, 3(74), January 2007.

[3] A. Greenfield, C. Hafemeister, and R. Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060-1067, March 2013.

[4] M. Hecker, S. Lambecka, S. Toepferb, E. van Somerenc, and R. Guthkea. Gene regulatory network inference: Data integration in dynamic modelsˆA˘a a ˘T review. *Biosystems*, 96(1):86-103, April 2009.

[5] J. C. Huang, Q. D. Morris, and B. J. Frey. Bayesian inference of microrna targets from sequence and expression data. *Journal of Computational Biology*, 14(5):550-563, 10 2007.

[6] J.-G. Joung, K.-B. Hwang, J.-W. Nam, S.-J. Kim, and B.-T. Zhang. Discovery of microrna-rna modules via population-based probabilistic learning. *Bioinformatics*, 23(9):1141-1147, March 2007.

[7] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770-780, October 2008.

[8] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. The role of site accessibility in microrna target recognition. *Nature Genetics*, 37(5):1278-1284, 2007.

[9] A. Krek, D. Gr¨n, M. N. Poy, R. Wolf, L. Rosenberg, u E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stofel, and N. Rajewsky. Combinatorial icrorna target predictions. *Nature Genetics*, 37(5):495-500, 05 2005.

[10] B. P. Lewis, I. hung Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787 - 798, 2003.

[11] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175-1182, May 2008.

[12] Z. Li, P. Li, A. Krishnan, and J. Liu. Large-scale dynamic gene regulatory network inference combining diferential equation models with local dynamic bayesian network analysis. *Bioinformatics*, 27(19):2686-2691, August 2011.

[13] A. Muniategui, R. Nogales-Cadenas, M. V´zquez, a X. L.Araguren, X. Agirre, A. Luttun, F. Prosper, A. Pascual-Montano, and A. Rubio. Quantification of mirna-mrna interactions. *PLoS ONE*, 7(2):e30766, Feburary 2012.

[14] M. H. Schulz, W. E. Devanny, A. Gitter, S. Zhong, J. Ernst, and Z. Bar-Joseph. Drem 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC systems biology*, 6(104), August 2012.

[15] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel. Global and local architecture of the mammalian micrornaˆA¸transcription factor regulatory network. a˘S *PLoS Comput Biol*, 3(7):e131, 07 2007.

[16] L. Song, M. Kolar, and E. P. Xing. Keller: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128-i136, January 2009.

[17] J. Sun, X. Gong, B. Purow, and Z. Zhao. Uncovering microrna and transcription factor mediated regulatory networks in glioblastoma. *PLoS Comput Biol*, 8(7):e1002488, 07 2012.

[18] F. Xin, M. Li, C. Balch, M. Thomson, and M. Fan. Computational analysis of microrna pro¨ ˘Ales and ın their target genes suggests signi¨ ˘Acant involvement ın in breast cancer antiestrogen resistance. *Bioinformatics*, 25(4):430-434, December 2008.

[19] B. Zacher, K. Abnaof, S. Gade, E. Younesi, A. Tresch, and H. Fr¨hlich. Joint bayesian infrence of o condition-specific mirna and transcription factor activities from combined gene and microrna expression data. *Bioinformatics*, 28(13):1714-1720, May 2012.

# Information Entropy Based Methods for Genome Comparison

Mehul Jani
Department of Biological Sciences,
University of North Texas
Denton, Texas 76203
mehuljani@my.unt.edu

Rajeev K. Azad
Departments of Biological Sciences and
Department of Mathematics,
University of North Texas
Denton, Texas 76203
rajeev.azad@unt.edu

## ABSTRACT

A plethora of biologically useful information lies obscured in the genomes of organisms. Encoded within the genome of an organism is the information about its evolutionary history. Evolutionary signals are scattered throughout the genome. Bioinformatics approaches are frequently invoked to deconstruct the evolutionary patterns underlying genomes, which are difficult to decipher using traditional laboratory experiments. However, interpreting constantly evolving genomes is a non-trivial task for bioinformaticians. Processes such as mutations, recombinations, insertions and deletions make genomes not only heterogeneous and difficult to decipher but also renders direct sequence comparison less effective. Here we present a brief overview of the sequence comparison methods with a focus on recently proposed alignment-free sequence comparison methods based on Shannon information entropy. Many of these sequence comparison methods have been adapted to construct phylogenetic trees to infer relationships among organisms.

## General Terms

Algorithms, Measurement, Performance, Design, Reliability, Verification.

## Keywords

Genome comparison, Sequence alignment, Shannon entropy, Segmentation, Clustering

## 1. INTRODUCTION

Evolution has always intrigued humans. Early approaches of studying evolution were based on fossil study and other palaeontological methods. Developments in the field of molecular biology, especially DNA sequencing, led to methods based on comparing DNA and protein sequences for studying the relatedness between the DNA sequence and thus the organisms. Phylogenetic trees constructed using simple alignment methods were used to infer evolutionary relationships among organisms. Phylogenetic trees are used in diverse fields such as systems biology, molecular biology, Darwinian medicine and ecology. Often such studies have helped address disease problems, e.g., phylogenetic trees of bacteria have advanced our understanding of the spread of antibiotic resistance patterns and the emergence of virulence.

With the advent of high throughput and next generation sequencing technologies, we now have access to complete genomes of over 4000 prokaryotes and over 180 eukaryotes (http://www.genomesonline.org). Such vast amount of genomic data calls for development of robust bioinformatics approaches to studying evolution via inferring phylogeny relationships among organisms. There are broadly two approaches used for comparing DNA sequences- alignment based sequence comparison and alignment-free sequence comparison, which we discuss briefly in the next sections.

## 2. ALIGNMENT BASED SEQUENCE COMPARISON

Sequence alignment is the most common approach used for comparing sequences. In pairwise sequence alignment, an optimal alignment between two given sequences is searched for by maximizing a scoring function which is essentially the sum of the residue to residue alignment scores between the sequences. Dynamic programming methods were developed for the pairwise global alignment and for the local alignment between sequences. The former, the Needleman-Wunsch algorithm [10], aligns two sequences end to end, that is, the full span of the sequences are aligned against each other, while the latter, the Smith-Waterman algorithm [16], searches for best aligned motif (short conserved regions) in the alignment. Heuristic approaches for local sequence alignment were developed later for fast and efficient alignment of long biological sequences. Among these, FASTA [11] and BLAST [1] are the most frequently used heuristic algorithms. The visualization techniques such as Dot matrix were developed for visualizing the alignment between two sequences; this is among the oldest approaches for sequence comparison, first used by Gibbs and McIntyre in 1970. Biological sequences come in family; homologous sequences in a family share the common ancestry and often have similar functions. Pairwise alignment methods are limited in their ability to detect members of a sequence family. To circumvent these limitations and to detect remote homologs, multiple sequence alignment methods were developed. The progressive alignment methods first align the most closely related pair of sequences and then the next most similar sequence to this pair is aligned and the process is repeated iteratively to build a multiple sequence alignment (also sometimes referred to as 'profile'). CLUSTALW [17] is one of the most popular tools used for multiple sequence alignment. Multiple sequence alignment is a precursor to phylogenetic tree construction. Based on the alignment score between sequences in a profile, a distance matrix is created and a phylogeny tree is constructed using the distance matrix. Evolutionary relationships are thus inferred from relative positions of sequences in a phylogenetic tree.

Genome wide sequence alignment is a huge computational burden. Often the organismal relationships are inferred by constructing trees using highly conserved nucleotide sequences of RNA genes or the conserved sequences of proteins. The problem with using only conserved gene or protein sequences is that the evolutionary signals from rest of the genome are ignored. Since evolutionary signals are dispersed throughout the genome and not just restricted to a few genes, ignoring these signals may have confounding implications. In fact, trees made using conserved RNA or protein sequences have been shown to contradict each other. Further, most alignment methods do not account for long range interactions within genomes. Moreover, natural evolutionary processes like recombinations, mutations, deletions, insertions, rearrangements, etc., make direct alignment between sequences difficult, especially when such changes happen frequently leading to fast evolving genomes with little evolutionary signals for a reliable sequence alignment. In general, sequence alignment method works best when the sequences being compared share high homology. Therefore, there is a great need of methods that can adequately account for evolutionary signals underlying the genomes of organisms.

## 3. Alignment-free sequence comparison

Alignment-free approaches are especially useful if the sequences do not share high homology or are rapidly accumulating changes thus obfuscating the evolutionary
alignment-free methods
rearrangements, in particular, disrupt the sequence contiguity and thus render such sequences unalignable in order to assess their common ancestry. To circumvent the limitations of alignment based methods, several approaches that do not require alignment for sequence comparison have been proposed. These so called alignment-free methods are based on *k*-mer frequency for computing the similarity (or dissimilarity) score between the sequences. The goal of such methods is to assess the divergence between two sequences in terms of difference in the frequency distributions of *k*-mers in the sequences. The frequently used distance measures to assess the sequence divergence include Euclidean distance [2], *d2* distance [18], covariance or correlation function [12], Mahalanobis distance [21], Kullback-Leibler divergence [22] and Kolmogorov complexity metric [7]. In a different approach for alignment-free sequence comparison, methods based on substrings [5, 19] have been used. The average common substring (ACS) method by Ulitsky et al [19] calculates average length of maximum common substrings for every site of each sequence and then pairwise genome sequence distance is calculated [19]. B. Haubold et al [5] used the shortest unique substrings in a set of sequences being studied for sequence comparison. Recently, J. Cheng et al [6] have built a multi-methods web server for alignment-free genome phylogeny, which can implement 12 popular alignment-free methods in a user friendly web platform. We refer the readers to Vinga and Almeida [20] for a comprehensive review of some of the alignment free methods discussed above.

Sims et al [15] proposed a feature frequency profile (FFP) method, a method based on *k*-mer frequency approach, which was shown to outperform other methods including the average common substring and Gencompress [3] methods. In this method, the frequencies of all possible features (the *k*-mers) of size *k* are computed to make a feature frequency profile. The total number of possible features will be $4^k$ in DNA sequence comparison. The difference between two genomic sequences, quantified in terms of difference in their *k*-mer compositional biases, was computed using Shannon information entropy based measure (Eqn. 1 in Section 4). The most important contribution of this method, as noted by the authors, is obtaining the optimal *k*-mer size to be used for sequence comparison. The lower limit of the *k*-mer can be empirically obtained, whereas upper limit of *k*-mer is calculated based on cumulative relative entropy. In order to infer organismal relationships, the information-entropic measure, namely, the Jensen-Shannon divergence (Eqn. 1), was used to compute the distance between the genome sequences of organisms and then a phylogenetic tree was constructed using this distance matrix.

The performance of FFP method and other *k*-mer frequency based methods for alignment-free sequence comparison depends on the *k*-mer size [15]. While longer *k*-mers carry more information and therefore confer greater predictive power to the methods, it is, however, not practical to use longer *k*-mers if the sequences under comparison are not sufficiently long enough. In contrast, shorter *k*-mers provide reliable statistics, however, this may represent the inherent stochastic nature of genomes rather than having any biological or phylogenetic meaning.

Genomes are inherently heterogeneous. Bacterial genomes are chimeras of genes with different ancestry. Genomic mosaicism also arises when different segments of a genome are subject to
signals. Frequent different evolutionary pressures. All

including the FFP method represent a genome sequence as a *k*-mer frequency distribution, thus ignoring the inherent genomic mosaicism that requires multiple *k*-mer frequency distributions to represent uniquely distinct sequence classes within a mosaic genome. Methods that use a single oligomer distribution as the representation of a genome can yield confounding results when comparing two or more mosaic genomes. A single oligomer distribution averages out evolutionary signal from entire genome, disregarding the heterogeneity of the genome [13]. To overcome this problem, Azad and Li [13], first deconstructed the intragenomic heterogeneity using Shannon information entropy based recursive segmentation and clustering method, and then compared the compositionally homogenous regions from the genomes of interest.

## 4. Recursive segmentation and agglomerative clustering

An integrative framework of recursive segmentation and agglomerative clustering was developed recently to deconstruct the complex heterogeneities within genomic data [13]. Recursive segmentation for DNA sequence analysis has a history of over a decade [4,8]. The recursive segmentation and agglomerative clustering method interprets genomic data at the intrusive level of complexities using Shannon entropy [14]. This method uses Jensen-Shannon (JS) divergence measure for assessing divergence or dissimilarity between two sequences. The Jensen-Shannon divergence between two sequences $S_1$ and $S_2$ can be measured using the following formula [9],

$$D\,(S_1, S_2) = H(S) - \pi_1 H(S_1) - \pi_2 H(S_2). \qquad (1)$$

Here, Shannon entropy $H$ for a sequence is defined as $H = -\sum_x p(x)\log_2 p(x)$, where $p(x)$ is the probability of (oligo)nucleotide (residue for protein sequences) $x$ estimated from the count of $x$ in the sequence. $S$ is the concatenation $S_1$ and $S_2$, and $\pi_i$ is the weight factor proportional to the length of $S_i$, $\sum_i \pi_i = 1$. The entropy function measures the information stored in a sequence.

The genome complexity is decomposed successively by performing a binary segmentation recursively until none of the sequence segments or regions can be divided further [9] using following steps: (i) For a sequence $S$, the difference between sequence segments left and right to each sequence position in $S$ is calculated using Jensen-Shannon divergence measure. (ii) The position of maximum divergence between the left and right sequence segments is located. (iii) The sequence is segmented at this position to get two segments, $S_1$ and $S_2$ provided the segmentation is deemed statistically significant. (iv)The aforementioned procedure is repeated for segments $S_1$ and $S_2$ recursively until none of the resulting sequence segments can be divided further. (v) These compositionally homogeneous sequence segments are now considered as distinct clusters, each segment assigned to a distinct cluster. In this step, similar contiguous segment clusters are identified and grouped together (vi) These segment clusters are the seed clusters for the next step of the clustering procedure. The grouping of similar clusters is followed recursively until the difference between any two clusters becomes significantly large. This last step clusters even non-contiguous segments and thus account for long range interactions or relationships between different regions in a genome.

This recursive segmentation procedure can be accomplished within a hypothesis-testing framework [4] or a model-selection framework [8]. Azad and Li [13] allowed hyper-segmentation in the hypothesis testing framework. This helped to increase the sensitivity of the method in identifying the break points or segment boundaries. However, hyper-segmentation may cause fragmentation of biologically important domains. To reestablish the segmental structure, segmentation was followed by clustering (step v above) at a relaxed clustering stringency.

To assess the divergence between genomes, Genome Wide Distance (GWD) was calculated using following formula:

$$GWD = \frac{1}{2}\left[\frac{1}{M}\sum_{i=1}^{M}\min\{D(G_1^i, G_2^1), \ldots, D(G_1^i, G_2^N)\} + \frac{1}{N}\sum_{j=1}^{N}\min\{D(G_1^1, G_2^j), \ldots, D(G_1^M, G_2^j)\}\right] \qquad (2)$$

Here, $D(G^i{}_1, G^j{}_2)$ is the Jensen-Shannon divergence between clusters $i$ and $j$ of genomes $G_1$ and $G_2$. $M$ and $N$ are number of clusters for genome $G_1$ and $G_2$ respectively.
This method was reported to perform better than the FFP method for comparing genomes [13]. This validated the hypothesis that relationships among organisms could be better explained by first decomposing their genome complexities and then comparing compositionally distinct components of their genomes. In the recursive segmentation and agglomerative clustering approach, the global genomic heterogeneity is deciphered first; the earlier obtained split points thus guide the next rounds of segmentation to decipher the local heterogeneities and in this process, eventually, the distinct evolutionary signals encoded in biological domains within a genome are deciphered. This method thus captures the evolutionary patterns within genomes reflecting disparate evolutionary trajectories, thus helping in deducing the evolutionary relationships among organisms.

This method was used to address several other pressing issues in biology, such as, identification of alien genes in bacterial genomes and detection of copy number variations in cancer genomes [13]. In future, this method can be adapted to detect other biologically important features such as isochores or the origin and terminus of replication.

## 5. CONCLUSIONS

The vast number of methods developed for comparing genome sequences highlights the significance of deducing reliable phylogenetic relationships. Traditional sequence alignment methods though reliable for sequences which are highly related or share high homology often prove to be deficient when comparing rapidly evolving sequences.

Alignment-free approaches have made significant progress since it was first used by Blaisdell [2]. Many recent methods for sequence comparison have used alignment-free approach. The alignment-free approach allows computing distances between large genomes in relatively less time. Alignment-free methods are more robust for comparing highly evolved sequences, sequences which have undergone changes at multiple loci in a chromosome, and even shorter sequences.

The advantage of using recursive segmentation and agglomerative clustering method is that it first decomposes the complexities of heterogeneous genomes and then compares the homogeneous parts of the genomes, thus providing a better comparison tool for elucidating organismal relationships. This method can be used in concert with alignment based methods to construct robust phylogenetic trees.

## 6. ACKNOWLEDGEMENTS

This work was supported by a faculty start up fund from the University of North Texas to R.K.A.

## 7. REFERENCES

[1] Altschul SF, Gish W, Miller W., Myers EW and Lipman DJ 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

[2] Blaisdell BE 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl Acad. Sci. USA 83, 5155-5159.

[3] Chen X, Kwong S and Li M 2001. A compression algorithm for DNA sequences. IEEE Engineering in Medicine and Biology Magazine 20, 61-66.

[4] Grosse I, Bernaola-Galvan P, Carpena P, Roman-Roldan R, Oliver J, Stanley HE 2002. Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. Phys Rev E 65:041905.

[5] Haubold B, Pierstorff N, Möller F, Wiehe T 2005. Genome comparison without alignment using shortest unique substrings. BMC Bioinformatics 6:123.

[6] Cheng J, Cao F and Liu Z 2013. AGP: A multi-methods web server for alignment-free genome phylogeny. Mol. Biol. Evol. 30:1032-1037

[7] Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17:149-154.

[8] Li W 2001. New stopping criteria for segmenting DNA sequences. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 86:5815-5818.

[9] Lin J 1991. Divergence measures based on the Shannon entropy. IEEE Trans. Inform. Theory 37:145-151.

[10] Needleman SB and Wunsch CD 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443-453.

[11] Pearson WR 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 183, 63-98.

[12] Petrilli P 1993. Classification of protein sequences by their dipeptide composition. Comput. Appl. Biosci. 9:205-209.

[13] Azad RK and Li J 2013. Interpreting genomic data via entropic dissection. Nucleic Acids Res. 41: e23.

[14] Shannon CE 1948. A mathematical theory of communication. The Bell System Technical J. 27: 379-423.

[15] Sims GE, Jun SR, Wu GA, Kim SH 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl Acad. Sci. USA. 106, 2677-2682.

[16] Smith TF and Waterman MS 1981. Identification of common molecular subsequences. J Mol Biol. 147:195-197.

[17] Thompson JD, Higgins DG, Gibson TJ 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

[18] Torney DC, Burkes C, Davidson D, Sirkin KM 1990. In: Computation of d2: A measure of sequence dissimilarity, computers and DNA, SFI studies in the sciences of complexity. Bell G, Marr T, editors. VII. New York, NY: Addison-Wesley.

[19] Ulitsky I, Burnstein D, Tuller T, Chor B 2006. The average common substring approach to phylogenomic reconstruction. Journal of Computational Biology 13, 336-350.

[20] Vinga S, Almeida J 2003. Alignment free sequence comparison- a review, Bioinformatics 19: 513-523.

[21] Wu TJ, Burke JP, Davison DB 1997. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 53:1431-1439.

[22] Wu TJ, Hsieh YC, Li LA 2001. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. Biometrics. 57:441-448.

ACM BCB 2013
ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics
September 22-25, 2013 Washington DC, USA

# Call for Papers, Workshops and Tutorials

# ACM Conference On Bioinformatics, Computational Biology and Biomedical Informatics (ACM-BCB 2013)

September 22-25 2013

DoubleTree by Hilton Hotel Bethesda - Washington DC

ACM-BCB Website: http://www.cse.buffalo.edu/ACM-BCB2013

## Key Dates

| | Submission Deadline | Notification of Acceptance |
|---|---|---|
| Papers | May 15, 2013 | July 15, 2013 |
| Workshops | March 1, 2013 | March 15, 2013 |
| | March 30, 2013 | |

The ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM BCB) is the flagship conference of the ACM SIGBCB. This is the conference's fourth year, building upon the success of the first three meetings in Niagara Falls, Chicago, and Orlando.

The conference offers a forum for premier interdisciplinary research linking computer science, mathematics, statistics, biology, bioinformatics, and biomedical informatics. The past two decades have led to a tremendous growth in the size and dimensionality of biological and biomedical data. This conference serves to showcase leading-edge research in processing, modeling and analyzing these datasets for a variety of applications.

We invite contributed papers, workshops and tutorials in the topic areas listed below. Please refer to the

ACM BCB 2013 welcomes original submissions that have not been published and that are not under review by another conference or journal.

## Organization Committee

**Steering Committee**

Aidong Zhang, State University of New York at Buffalo, Chair
Vasant Honavar, Iowa State University, Conference Director

**General Chairs**

Sridhar Hannenhalli, University of Maryland
Cathy Wu, University of Delaware & Georgetown University

**Program Chairs**

Srinivas Aluru, Iowa State University
Donna Slonim, Tufts University

**Local Arrangement Chairs**

Amarda Shehu, George Mason University
Liliana Florea, Johns Hopkins University

**Workshop Chair**

Umit Catalyurek, Ohio State University

**Tutorial Chairs**

Clare Bates Congdon, University of Southern Maine
Vasant Honavar, Iowa State University

**Poster Chairs**

Dongxiao Zhu, Wayne State University
Yu-Ping Wang, Tulane University

**Industry Chair**

Anastasia Christianson, AstraZeneca Pharmaceutical

**Panel Chair**

Iosif Vaisman, George Mason University

**Publicity Chair**

Jianlin Cheng, University of Missouri, Columbia

**Registration Chair**

## Topics

- ○ **Genomics and Evolution**
- ○ **Protein and RNA Structure, Protein Function, and Proteomics**
- ○ **Computational Systems Biology**
- ○ **Next Generation Sequencing Data**
- ○ **Medical Informatics and Translational Bioinformatics**
- ○ **Cross-Cutting Computational Methods**
- ○ **Bioinformatics Infrastructure**
- ○ **Immunoinformatics and Computational Immunology**
- ○ **Computational Epidemiology**
- ○ **Biomedical Image Analysis**

acm Association for Computing Machinery
Advancing Computing as a Science & Profession

ACM SIGBio
Bioinformatics, Computational Biology and Biomedical Informatics

Preetam Ghosh, Virginia Commonwealth University

**Proceedings Chair**

Jing Gao, State University of New York at Buffalo

**Exhibit/System Demo Chair**

Nathan Edwards, Georgetown University

- **Knowledge Representation and Inference**
  - **Integration of Biomedical Data**
- **Databases, Knowledgebases & Ontologies**
  - **Text Mining and Natural Language Processing**

## *Workshops*

# ACM BCB Workshops

Five Workshops are planned to be held in conjunction with ACM-BCB 2013.

- Fourth Immunoinformatics and Computational Immunology Workshop (ICIW 2013)
- Computational Structural Bioinformatics Workshop (CSBW 2013)
- 6th International Workshop on Biomolecular Network Analysis (IWBNA 2013)
- Parallel and Cloud-based Bioinformatics and Biomedicine (ParBio 2013)
- Workshop on Epigenomics and Cell Function (ECF 2013)

## Health Informatics Symposium (HIS)

September 25, 2013, Washington, DC,
http://www.cse.buffalo.edu/ACM-BCB2013/HIS.html

This symposium will be held in conjunction with the annual ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM BCB), which is the main flagship conference of the ACM SIGBio (ACM Special Interest Group on Bioinformatics, Computational Biology and Biomedical Informatics). The goal of this symposium is to bring together bring computational scientists together with researchers and professionals to discuss the problems in healthcare, public health, and everyday wellness. The symposium highlights the most novel technical contributions in computing oriented toward health informatics and the related social and ethical
implications.

Specific topics of interest cover various facets of health informatics research,
including but not limited to the following:

- Information technologies for clinical and healthcare delivery and management
- Health data acquisition, management, and visualization
- Healthcare knowledge management and decision support
- Healthcare modeling and simulation
- Data analytics, data mining, and machine learning
- Health information systems
- Healthcare communication networks and environments
- Interactions with health information technologies

Submitted manuscripts should not exceed 10 pages in ACM template on 8.5 x 11 inch paper (see ACM templates). All accepted papers of registered authors will be included in the proceedings published by ACM digital libraries. The authors of selected papers will be invited to adapt their papers for being
published in a special issue of a journal (to be determined).

## SIGBIO Record - Submission Guidelines

### Submission categories

Submissions to the newsletter can be either on a special issue topic or on topics of general interest to the SIGBIO community.

These can be in any one of the following categories:

- Survey/tutorial articles (short) on important topics.
- Topical articles on problems and challenges •
Well-articulated position papers.
- Review articles of technical books, products and .
- Reviews/summaries from conferences, panels and special meetings within 1 to 4 pages [1500-2500 words]
- Book reviews and reports on relevant published technical books
- PhD dissertation abstracts not exceeding 10 pages
- Calls and announcements for conferences and journals not exceeding 1 page
- News items on the order of 1-3 paragraphs

Brief announcements Announcements not exceeding 5 lines in length can simply be sent as ASCII text to the
editors by e-mail. SIGBIO Record publishes announcements that are submitted as is without review.

Announcements cannot be advertisements and should be of general interest to the wider community. The Editor reserves the right to reject any requests for announcements at his discretion.

Authors are invited to submit original research papers or review papers in all areas of bioinformatics and computational biology. The papers published in SIGBioinformatics Record will be archived in ACM Digital Library. Papers should follow the ACM format, and there is no page limitation.

http://www.acm.org/sigs/publications/proceedings-templates

Submissions should be made via email to the editors Pierangelo Veltri *(University Magna Graecia of Catanzaro, Italy),*
Young-Rae Cho *(Baylor University )*