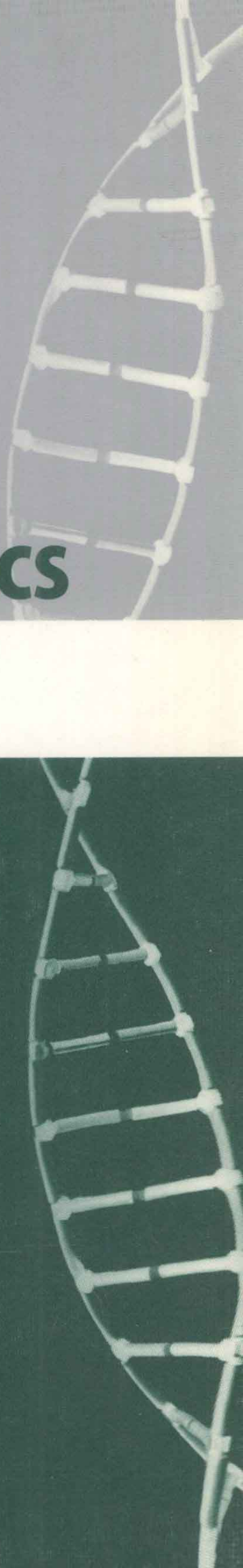
Rita Casadio
Gene Myers (Eds.)

LNBI 3692

Algorithms in Bioinformatics

5th International Workshop, WABI 2005
Mallorca, Spain, October 2005
Proceedings



Springer

Rita Casadio Gene Myers (Eds.)

Algorithms in Bioinformatics

5th International Workshop, WABI 2005
Mallorca, Spain, October 3-6, 2005
Proceedings

Springer

Volume Editors

Rita Casadio
University of Bologna, Department of Biology/CIRB
Via Irnerio 42, 40126 Bologna, Italy
E-mail: casadio@alma.unibo.it

Gene Myers
Howard Hughes Medical Institute
4000 Jones Bridge Road, Chavy Chase, MD 20815-6789, USA
E-mail: gene@eecs.berkeley.edu

Library of Congress Control Number: 2005932938

CR Subject Classification (1998): F.1, F.2.2, E.1, G.1-3, J.3

ISSN	0302-9743
ISBN-10	3-540-29008-7 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-29008-7 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: 11557067 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 3692

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

Subseries of Lecture Notes in Computer Science

Preface

We are pleased to present the proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI 2005) which took place in Mallorca, Spain, October 3–6, 2005. The WABI 2005 workshop was part of the five ALGO 2005 conference meetings, which, in addition to WABI, included ESA, WAOA, IWPEC, and ATMOS. WABI 2005 was sponsored by EATCS (the European Association for Theoretical Computer Science), the ISCB (the International Society for Computational Biology), the Universitat Politècnica de Catalunya, the Universitat de les Illes Balears, and the Ministerio de Educación y Ciencia. See http://www.lsi.upc.edu/~wabi05/ for more details.

The Workshop on Algorithms in Bioinformatics highlights research work specifically developed to address algorithmic problems in biosequence analysis. The emphasis is therefore on statistical and probabilistic algorithms that address important problems in the field of molecular and structural biology. At present, given the enormous scientific and technical efforts in functional and structural genomics, the relevance of the problem is therefore constrained by the need for sound, efficient and specialized algorithms, capable of achieving solutions that can be tested by the biological community. Indeed the ultimate goal is to implement algorithms capable of extracting real features from real biological data sets. Therefore the workshop aims to present recent research results, including significant work in progress, and to identify and explore directions of future research.

Original research papers (including significant work in progress) or state-of-the-art surveys were solicited on all aspects of algorithms in bioinformatics, including, but not limited to: exact and approximate algorithms for genomics, genetics, sequence analysis, gene and signal recognition, alignment, molecular evolution, phylogenetics, structure determination or prediction, gene expression and gene networks, proteomics, functional genomics, and drug design. We received 94 submissions in response to our call for papers, and were able to accept 35 of these. In addition, WABI 2005 hosted a distinguished lecture by Dr. Marino Zerial of the Max Planck Institute for Molecular Cell Biology and Genetics in Dresden, given to the entire ALGO 2005 conference.

We would like to sincerely thank all the authors of submitted papers, and the participants of the workshop. We also thank the Program Committee and their sub-referees for their hard work in reviewing and selecting the papers for the workshop. The Program Committee consisted of the following 40 distinguished researchers:

Pankaj Kumar Agarwal (Duke University)
Tatsuya Akutsu (Kyoto University)
Amir Amihood (Bar-Ilan University)

Alberto Apostolico (Purdue University)
Craig Benham (University of California, Davis)
Gary Benson (MSSN, New York)
Mathieu Blanchette (McGill University)
Nadia El-Mabrouk (University of Montreal)
Olivier Gascuel (LIRMM, Montpelier)
Raffaele Giancarlo (University of Palermo)
Roderic Guigo (IMIM, Barcelona)
Michael Hallet (McGill University)
Daniel Huson (University of Tuebingen)
Gregory Kucherov (INRIA Nancy)
Michelle Lacey (Tulane University)
Jens Lagergren (KTH Stockholm)
Giuseppe Lancia (Univeristy of Udine)
Gad M. Landau (University of Haifa)
Thierry Lecroq (Université de Rouen)
Bernard Moret (University of New Mexico)
Shinichi Morishita (University of Tokyo)
Elchanan Mossel (Univeristy of California, Berkeley)
Vincent Moulton (University of Uppsala)
Lior Pachter (University of California, Berkeley)
Knut Reinert (Free University of Berlin)
Isidore Rigoutsos (IBM Watson)
Marie-France Sagot (INRIA Rhône-Alpes)
David Sankoff (University of Ottawa)
Sophie Schbath (INRIA Jouv-en-Josas)
Eran Segal (Rockefeller University)
Charles Semple (University of Canterbury)
Joao Carlos Setubal (Virginia Polytechnic Institute)
Roded Sharan (Tel Aviv Univeristy)
Steven Skiena (University of New York, Stony Brook)
Jens Stoye (University of Bielefeld)
Esko Ukkonen (University of Helsinki)
Lisa Vawter (Aventis Inc., USA)
Alfonso Valencia (CNB-CSIC, Spain)
Tandy Warnow (University of Texas)
Lusheng Wang (City Univeristy of Hong Kong)

Finally we would like to especially thank Bernard Moret, the de facto steering committee, for answering questions on history and precedence, for his advice on difficult protocol issues, and for setting up and hosting the EasyChair refereeing system used by the Program Committee.

July 2005

Rita Casadio and Gene Myers
WABI 2005 Program Co-chairs

Vol. 3695: M.R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005.

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005.

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005.

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

Vol. 3380: C. Priami (Ed.), Transactions on Computational Systems Biology I. IX, 111 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.

Table of Contents

Expression

1. Hybrid Methods

2. Time Patterns

Phylogeny

1. Quartets

2. Tree Reconciliation

3. Clades and Haplotypes

Networks

Genome Rearrangements

1. Trasposition Model

2. Other Models

Sequences

1. Strings

2. Multi-alignment and Clustering

3. Clustering and Representation

Structure

1. Threading

2. Folding

Spectral Clustering Gene Ontology Terms to Group Genes by Function

Nora Speer, Christian Spieth, and Andreas Zell

University of Tübingen, Centre for Bioinformatics Tübingen (ZBIT),
Sand 1, D-72076 Tübingen, Germany
nspeer@informatik.uni-tuebingen.de

Abstract. With the invention of biotechnological high throughput methods like DNA microarrays, biologists are capable of producing huge amounts of data. During the analysis of such data the need for a grouping of the genes according to their biological function arises. In this paper, we propose a method that provides such a grouping. As functional information, we use Gene Ontology terms. Our method clusters all GO terms present in a data set using a Spectral Clustering method. Then, mapping the genes back to their annotation, genes can be associated to one or more clusters of defined biological processes. We show that our Spectral Clustering method is capable of finding clusters with high inner cluster similarity.

1 Introduction

In the past few years, high-throughput techniques like microarrays have become major tools in the field of genomics. In contrast to traditional methods, these technologies enable researchers to collect tremendous amounts of data, whose analysis itself constitutes a challenge. Since these techniques provide a global view on the cellular processes as well as on their underlying regulatory mechanisms, they are quite popular among biologists. After the analysis of such data, using filtering methods, clustering techniques or statistical approaches, researchers often end up with long lists of interesting candidate genes that need further examination. Then, in a second step, they categorize these genes by known biological functions.

In this paper, we address the problem of finding functional clusters of genes by clustering Gene Ontology terms. Based on methods originally developed for semantic similarity, we are able to compute a functional similarity between GO terms [13]. This information is fed into a spectral clustering algorithm [15]. This has the advantage, that after mapping the genes back to the GO terms, a gene with more than one associated term (function) can be present in more than one cluster which seems biologically plausible.

The organization of this paper is as follows: a brief introduction to the Gene Ontology is given in section 2. Related Work is discussed in section 3. Section 4 explains our method in detail. The experimental setup and the results on real world data sets are shown in section 5. Finally, in section 6, we conclude.

R. Casadio and G. Myers (Eds.): WABI 2005, LNBI 3692, pp. 1–12, 2005.

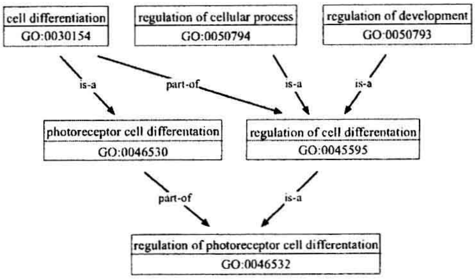

cell differentiation
GO:0030154
regulation of cellular process
GO:0050794
regulation of development
GO:0050793
is-a
part-of
is-a
is-a
photoreceptor cell differentiation
GO:0046530
regulation of cell differentiation
GO:0045595
part-of
is-a
regulation of photoreceptor cell differentiation
GO:0046532

Fig. 1. Relations in the Gene Ontology. Each node is annotated with a unique accession number.

2 The Gene Ontology

The Gene Ontology (GO) is one of the most important ontologies within the bioinformatics community and is developed by the Gene Ontology Consortium [21]. It is specifically intended for annotating gene products with a consistent, controlled and structured vocabulary. Gene products are for instance sequences in databases as well as measured expression profiles. The GO is independent from any biological species. It represents terms in a Directed Acyclic Graph (DAG), covering three orthogonal taxonomies or "aspects": *molecular function*, *biological process* and *cellular component*. The GO-graph consists of over 18.000 terms, represented as nodes within the DAG, connected by relationships, represented as edges. Terms are allowed to have multiple parents as well as multiple children. Two different kinds of relationship exist: the "is-a" relationship (*photoreceptor cell differentiation* is, for example, a child of *cell differentiation*) and the "part-of" relationship that describes, for instance, that *regulation of cell differentiation* is part of *cell differentiation*.

Providing a standard vocabulary across any biological resources, the GO enables researchers to use this information for automatic data analysis done by computers and not by humans.

3 Related Work

While GO analysis is an increasingly important field, existing techniques suffer from some weaknesses: Many methods consider the GO simply as a list of terms, ignoring any structural relationships [2,7,17,23]. Others regard the GO primarily as a tree and convert the GO graph into a tree structure for determining distances between nodes [11]. Again others use a pseudo-distance that does not fulfill all metric conditions and relies on counting path lengths [3]. This is a delicate approach in unbalanced graphs like the GO those subgraphs have different degrees of detail.

Besides, the aim of some methods is primary either to use the GO as preprocessing [1] or as visualization tool [6]. Only few approaches utilize its structure for computation. Many methods are scoring techniques describing a list of genes annotated with GO terms [2,6,7,11,17,23]. But to our knowledge and apart from our earlier publications [20,19], there exists no automatic functional GO-based clustering method. One method is related to clustering and can be used to indicate which clusters are present in the data [3]. However, it suffers from the weaknesses that come along with using pseudo-distances as mentioned earlier.

4 Methodology

Our method consists of different steps that will be explained separately in this section: the mapping of the genes to the Gene Ontology, the calculation of functional similarities on GO terms, the spectral clustering algorithm and finally how the appropriate number of clusters is determined.

4.1 Mapping the Genes to the Gene Ontology

The functional similarity measure operates on pairs of GO nodes in a DAG, whereas in general, researchers are dealing with database ids of genes or probes. Therefore, a mapping M that relates the genes of a microarray experiment to nodes in the GO graph is required. Many databases (e.g. TrEMBL (GOA-project)) provide GO annotation for their entries and companies like Affymetrix provide GO mappings to their probe set ids as well. We used GeneLynx [8] to map the genes of dataset I. Hvidsten *et al.* [9] provide a mapping for dataset II.

4.2 Similarities Within the Gene Ontology

To calculate functional similarities between GO nodes, we rely on a technique that was originally developed for other taxonomies like WordNet to measure semantic similarities between words [12].

Following the notation in information theory, the information content (IC) of a term t can be quantified as follows [13]:

$$IC(t) = -\ln P(t) \tag{1}$$

where $P(t)$ is the probability of encountering an instance of term t in the data.

In the case of a hierarchical structure, such as the GO, where a term in the hierarchy subsumes those lower in the hierarchy, this implies that $P(t)$ is monotonic as one moves towards the root node. As the node's probability increases, its information content or its informativeness decreases. The root node has a probability of 1, hence its information content is 0. As the three aspects of the GO are disconnected subgraphs, this is still true if we ignore the root node "Gene Ontology" and take, for example, "biological process" as our root node instead.

To compute a similarity between two terms, one can use the IC of their common ancestor. As the GO allows multiple parents for each term, two terms can share ancestors by multiple paths. We take the minimum $P(t)$, if there is more than one ancestor. This is called P_{ms}, for *probability of the minimum subsumer* [13]. Thereby, it is guaranteed, that the most specific parent term is selected:

$$P_{\mathrm{ms}}(t_i, t_j) = \min_{t \in S(t_i, t_j)} P(t) \tag{2}$$

where $S(t_i, t_j)$ is the set of parental terms shared by both t_i and t_j. Based on Eq. 1 and 2, Lin extended the similarity measure, so that the IC of each single node was also taken into account [12,13]:

$$s(t_i, t_j) = \frac{2 \ln P_{ms}(t_i, t_j)}{\ln P(t_i) + \ln P(t_j)} \ . \tag{3}$$

Since $P_{ms}(t_i, t_j) \geq P(t_i)$ and $P_{ms}(t_i, t_j) \geq P(t_j)$, its value varies between 1 (for similar terms) and 0.

One should note, that the probability of a term as well as the resulting similarity between two terms differs from data set to data set, depending on the distribution of terms. Therefore, our clustering differs from a general clustering of the GO and a subsequent mapping of the genes to such a general clustering. Due to our approach, we are able to arrange the resulting cluster boundaries depending on the distribution of the GO terms either more specific (if the terms concentrate on a specific part of the GO) or more general (if the terms are widely spread).

4.3 Spectral Clustering

We decided to cluster GO terms, not genes, because of two reasons: first, we do not face the problem of combining different similarities per gene like in earlier publications [19,20] and second, after mapping the genes back to the GO, they can be present in more than one functional cluster which is biologically plausible, since they can also fulfill more than one biological function.

Recently, Spectral Clustering methods haven been growing in popularity. Several new algorithms have been published [22,18,14,15]. A set of objects (in our case GO terms) to be clustered will be denoted by T, with $|T| = n$. Given an affinity measure $A_{ij} = A_{ji} \geq 0$ for two objects i, j, the affinities A_{ij} can be seen as weights on the undirected edges ij of a graph G over T. Then, the matrix $A = [A_{ij}]$ is the real-valued adjacency matrix for G. Let $d_i = \sum_{j \in T} A_{ij}$ be called the degree of node i, and D be the diagonal matrix with d_i as its diagonal. A clustering $C = \{C_1, C_2, \ldots, C_K\}$ is a partitioning of T into the nonempty mutually disjoint subsets $C_1, C_2, \ldots, C_K$. In the graph theoretical paradigm a clustering represents a multiway cut in the graph G.

In general, all Spectral Clustering algorithms use Eigenvectors of a matrix (derived from the affinity matrix A) to map the original data to the K-dimensional vectors $\{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ of the spectral domain $\Re^K$. Then, in a second step, these vectors are clustered with standard clustering algorithms. Here, we use K-means. We chose the newest Spectral Clustering algorithm by Ng *et al.* [15] and we will now review it briefly:

1. From the affinity matrix A and its derived diagonal matrix D, compute the Laplacian matrix $L = D^{-1/2}AD^{-1/2}$.
2. Find $v^1, v^2, \dots, v^K$, the Eigenvectors of L, corresponding to the K largest Eigenvalues.
3. Form the matrix $V_{n\times k} = \left[v^1, v^2, \dots, v^K\right]$ with these Eigenvectors as columns.
4. Form the matrix Y from V by renormalizing each of X's rows to have unit norm.
5. Cluster the rows of $Y = [\gamma_1, \gamma_2, \dots, \gamma_n]$ as points in a K-dimensional space.
6. Finally assign the original object i to cluster j if and only if row γ_i of the matrix Y was assigned to j.

Since Spectral Clustering relies on the affinity matrix A, it is easy to apply it to any kind of data, where affinities can be computed. For numerical data, affinities are usually computed with a kernel function, e.g. $A_{ij} = \exp(\frac{-d(i,j)^2}{2\sigma^2})$, with $d(i,j)$ denoting the Euclidean distance between point i and j and σ denoting the kernel width. For non-numerical data, like GO terms, affinity can either be defined in the same way, given a distance measure d. This approach has the advantage of non-linearity, controlled by the kernel width σ, which allows for sharper separation between clusters. But it has also disadvantages: the question of how to deduce σ in a meaningful way arises and additionally, for many data types, especially the GO, similarity is much easier to define since it does not need to fulfill any metric conditions. As noted in [16], there is nothing magical about the definition of affinity. Therefore, we directly apply our similarity matrix as affinity matrix.

4.4 Cluster Validity

We selected the number of clusters K in our data according to the Davies-Bouldin index [5]. Given a clustering $C = \{C_1, C_2, \dots, C_K\}$, it is defined as:

$$DB(C) = \frac{1}{K}\sum_{i=1}^{K} \max\left\{\frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)}\right\} \tag{4}$$

where $\Delta(C_i)$ represents the inner cluster distance of cluster C_i and $\delta(C_i, C_j)$ denotes the inter cluster distance between cluster C_i and C_j. K is the number of clusters. Small values of $DB(C)$ indicate a good clustering.

$\Delta(C_i)$ and $\delta(C_i, C_j)$ are calculated as the sum of distances to the respective cluster mean and the distance between the centers of two clusters, respectively. Since we use similarities, not distances, and cannot compute means in the GO, we apply the DB-Index in the spectral domain $\Re^K$ (after the Eigenvector decomposition) where we are dealing with simple numerical data.

5 Computational Experiments

5.1 Data Sets

One possible scenario where researchers would like to group a list of genes according to their function is when they received lists of up- or down-regulated genes from the analysis of an DNA microarray experiment. Thus, we chose two publicly available microarray data sets, annotated the genes with the GO and used them for functional clustering. We only use the taxonomy *biological process*, because we are mainly interested in gene function in a more general sense. However, our method can be applied in the same way for the other two taxonomies.

The authors of the first data set examined the response of human fibroblasts to serum on cDNA microarrays in order to study growth control and cell cycle progression. They found 517 genes whose expression levels varied significantly, for details see [10]. We used these 517 genes for which the authors provide NCBI accession numbers. The GO mapping was done using GeneLynx [8]. After mapping to the GO, 238 genes showed one or more mappings to *biological process* or a child term of *biological process*. These 238 genes were used for the clustering.

In order to study gene regulation during eukaryotic mitosis, the authors of the second data set examined the transcriptional profiling of human fibroblasts during cell cycle using microarrays [4]. Duplicate experiments were carried out at 13 different time points ranging from 0 to 24 hours. Cho *et al.* [4] found 388 genes whose expression levels varied significantly. Hvidsten *et al.* [9] provide a mapping of the data set to GO. 233 of the 388 genes showed at least one mapping to the GO *biological process* taxonomy and were thus used for clustering.

5.2 Experimental Design

In the experiments, we had the problem of how to compare our method to other known clustering algorithms, because to our best knowledge, there is no clustering method that does a clustering only due to a similarity matrix. Instead, most algorithms need distances. Beside that, most clustering techniques were originally developed for numerical data and therefore utilize means during the clustering process which we cannot compute in the GO. Only linkage methods work on a proximity matrix, although this is also usually a distance matrix. Average Linkage clustering is known to be its most robust, non-means based representative. Therefore, we compare our approach to a modified version of an Average Linkage algorithm that joins the most similar clusters, instead of joining those with the smallest distance. Inner cluster similarity of cluster C_i is computed as follows:

$$s(C_i) = \frac{1}{|C_i|(|C_i - 1|)} \sum_{t_i,t_j \in C_i, t_i \neq t_j} s(t_i, t_j) \tag{5}$$

with $s(t_i, t_j)$ denoting the similarity between term t_i and t_j and $|C_i|$ denoting the number of terms in cluster C_i.

For Spectral Clustering, K-means was carried out 25 times and the solution with the minimum distortion was taken as proposed in [15]. For both algorithms, we performed runs for different values of K, ranging from $K = 5, 6, \ldots, 25$.