Michael R. Berthold
Robert Glen
Ingrid Fischer (Eds.)

LNBI 4216

Computational Life Sciences II

Second International Symposium, CompLife 2006
Cambridge, UK, September 2006
Proceedings

Springer

Michael R. Berthold Robert Glen
Ingrid Fischer (Eds.)

Computational Life Sciences II

Second International Symposium, CompLife 2006
Cambridge, UK, September 27-29, 2006
Proceedings

Springer

Volume Editors

Michael R. Berthold
Ingrid Fischer
University of Konstanz
Department of Computer and Information Science
Box M712, 78457 Konstanz, Germany
E-mail: berthold@ieee.org, Ingrid.Fischer@inf.uni-konstanz.de

Robert Glen
Unilever Center
University of Cambridge
Lensfield Road, Cambridge CB2 1EW, UK
E-mail: rcg28@cam.ac.uk

Library of Congress Control Number: 2006932818

CR Subject Classification (1998): H.2, H.3, H.4, J.3

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN	0302-9743
ISBN-10	3-540-45767-4 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-45767-1 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11875741 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 4216

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

Subseries of Lecture Notes in Computer Science

Preface

Since our first CompLife symposium last year, we have seen the predicted trends in the life and computer science areas continue with ever-increasing production of high-quality data mated to novel analysis methods. The integration of the most advanced computational methods into experimental design and in particular the validation of these methods will remain a challenge. However, there is increasing appreciation between the different scientific communities in computer science and biology that each has substantial goals in common and much to gain by collaboration on complex problems. Providing a forum for an open and lively exchange between computer scientists, biologists, and chemists remains our goal. To encourage precisely this type of exchange, crossing the borders of the sciences, we organized the First Symposium on Computational Life Science in Konstanz, Germany in September 2005 (the proceedings were published in this series as LNBI 3695). Due to the success of the symposium, especially in bringing together scientists with diverse backgrounds, a second symposium was held in Cambridge (September 27-29, 2006).

The conference program shows that the scientific mix worked out very well again. We received higher quality submissions (56 this time) and selected 23 for oral presentation. As a supplement to the normal conference program we arranged for a "Free Software Session," where a dozen open source tools and toolkits were presented. Due to the nature of such software projects it seemed inappropriate to cover them in printed form but the conference Web site will continue to link to the respective pages (www.complife.org). Adding this session to the symposium also educated attendees on how to use some of the methods presented and shed some light on the wealth of free tools available already.

Selecting the papers included in this volume would not have been possible without the help of our Area Chairs and an international Program Committee that put in countless hours to create a minimum of three detailed reviews for each paper! And, of course, a successful conference relies on many individuals working hard behind the scenes. We would like to thank first and foremost Susan Begg and Heather Fyson for conference and local organization and keeping everybody on track. Peter Burger worked on the Web pages promoting the conference and Thorsten Meinl was the man behind the free software session and, together with Andreas Bender, also took care of publicity. Last, but certainly not least, thanks go to Ingrid Fischer and Richard van de Stadt for putting together this volume!

July 2006

Michael R. Berthold
Robert Glen

Organization

General Chair	Michael R. Berthold University of Konstanz, Germany Michael.Berthold@uni-konstanz.de
Program Chair	Robert Glen Unilever Center, University of Cambridge, UK rcg28@cam.ac.uk
Publication Chair	Ingrid Fischer University of Konstanz, Germany Ingrid.Fischer@uni-konstanz.de
Publicity Chairs	Andreas Bender Novartis Institutes for BioMedical Research, USA andreas.bender@complife.org
	Thorsten Meinl University of Konstanz, Germany Thorsten.Meinl@uni-konstanz.de
Conference Chair	Heather Fyson University of Konstanz, Germany heather.fyson@uni-konstanz.de
Local Chair	Susan Begg Unilever Center, University of Cambridge, UK smb28@cam.ac.uk
Webmaster	Peter Burger University of Konstanz, Germany Peter.Burger@uni-konstanz.de

Program Committee

Area Chairs

Giuseppe Di Fatta, University of Konstanz, Germany and CNR Palermo, Italy
Aldo Faisal, University of Cambridge, UK
Hans-Christian Hege, Zuse Institute Berlin, Germany
Janette Jones, Unilever Research, UK
Oliver Kohlbacher, Tübingen University, Germany
Peter Murry-Rust, University of Cambridge, UK
Jeremy Nicholson, Imperial College, UK
Gisbert Schneider, Frankfurt University, Germany
Brian Shoichet, UC San Francisco, USA
Arno Siebes, Utrecht University, NL
Hong Yan, City University, Hong Kong
Daniel Zaharevitz, NIH, USA

Program Committee

Alexander Bockmayr, FU Berlin, Germany
Sebastian Böcker, Friedrich Schiller University Jena, Germany
Tim Clark, Friedrich Alexander University Erlangen-Nuremberg, Germany
Thomas Exner, University of Konstanz, Germany
Peter Haebel, ALTANA Pharma Konstanz, Germany
Lawrence Hall, University of South Florida Tampa, USA
Joost Kok, Leiden University, The Netherlands
Hans-Peter Lenhof, Saarland University Saarbrücken, Germany
Xiaohui Liu, Brunel University, UK
Vladimir Marik, Czech Technical University Prague, Czech Republic
Srinivasan Parthasarathy, The Ohio State University, USA
David Patterson, Vistamont Consultancy, USA
Matthias Rarey, University of Hamburg, Germany
Knut Reinert, FU Berlin, Germany
Klaus Schäfer, ALTANA Pharma Konstanz, Germany
Hannu Toivonen, University of Helsinki, Finland
Allan Tucker, Brunel University, UK
Alexandre Urzhumtsev, University H. Poincare Nancy, France
Peter Willet, The University of Sheffield, UK
Mohammed Zaki, Rensselaer Polytechnic Institute, USA
Ralf Zimmer, LMU Munich, Germany
Albert Y. Zomaya, The University of Sydney, Australia

Additional Reviewers

Fatih Altiparmak, Sitaram Asur, Robert Banfield, Daniel Baum, Fabian Birzele, Andreas Döring, Caroline C. Friedel, Jan Gewehr, Clemens Gröpl, Volkhard Helms, Andreas Hildebrandt, Wilhelm Huisinga, Hans-Michael Kaltenbach, Andreas Keller, Robert Kueffner, Stefan Kurtz, Jan Küntzer, Hans Lamecker, Abdelhalim Larhlimi, Zsuzsanna Liptak, Andreas Moll, Ozgur Ozturk, Heike Pospisil, Sven Rahmann, Alexander Rurainski, Johannes Schmidt-Ehrenberg, Larry Shoemaker, Selina Sommer, Jens Stoye, Wiebke Timm, Duygu Ucar, Chao Wang, Hui Yang

Sponsoring Institutions

We thank our sponsors for their support in making the second International Symposium on Computational Life Science a successful event: Molecular Graphics Modelling Society (MGMS), ALTANA Pharma AG, AstraZeneca R&D, Boehringer Ingelheim, Inpharmatica, Unilever R&D Port Sunlight and Unilever Corporate Research.

Lecture Notes in Bioinformatics

Vol. 4216: M.R. Berthold, R. Glen, I. Fischer (Eds.), Computational Life Sciences. XIII, 269 pages. 2006.

Vol. 4205: G. Bourque, N. El-Mabrouk (Eds.), Comparative Genomics. X, 231 pages. 2006.

Vol. 4175: P. Bücher, B.M.E. Moret (Eds.), Algorithms in Bioinformatics. XII, 402 pages. 2006.

Vol. 4146: J.C. Rajapakse, L. Wong, R. Acharya (Eds.), Pattern Recognition in Bioinformatics. XIV, 186 pages. 2006.

Vol. 4115: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence and Bioinformatics, Part III. XXI, 803 pages. 2006.

Vol. 4075: U. Leser, F. Naumann, B. Eckman (Eds.), Data Integration in the Life Sciences. XI, 298 pages. 2006.

Vol. 4070: C. Priami, X. Hu, Y. Pan, T.Y. Lin (Eds.), Transactions on Computational Systems Biology V. IX, 129 pages. 2006.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006.

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006.

Vol. 3886: E.G. Bremer, J. Hakenberg, E.-H.(S.) Han, D. Berrar, W. Dubitzky (Eds.), Knowledge Discovery in Life Science Literature. XIV, 147 pages. 2006.

Vol. 3745: J.L. Oliveira, V. Maojo, F. Martín-Sánchez, A.S. Pereira (Eds.), Biological and Medical Data Analysis. XII, 422 pages. 2005.

Vol. 3737: C. Priami, E. Merelli, P. Gonzalez, A. Omicini (Eds.), Transactions on Computational Systems Biology III. VII, 169 pages. 2005.

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. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 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

Genomics

Data Mining

Molecular Simulation

Molecular Informatics

Systems Biology

Biological Networks / Metabolism

Computational Neuroscience

Improved Robustness in Time Series Analysis of Gene Expression Data by Polynomial Model Based Clustering

Michael Hirsch[1,*], Allan Tucker[1], Stephen Swift[1], Nigel Martin[2], Christine Orengo[3], Paul Kellam[4], and Xiaohui Liu[1]

[1] School of Information Systems Computing and Mathematics, Brunel University, Uxbridge UB8 3PH, UK

[2] School of Computer Science and Information Systems Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK

[3] Department of Biochemistry and Molecular Biology, University College London, Gower Street, London, WC1E 6BT, UK

[4] Department of Infection, University College London, Gower Street, London, WC1E 6BT, UK

Abstract. Microarray experiments produce large data sets that often contain noise and considerable missing data. Typical clustering methods such as hierarchical clustering or partitional algorithms can often be adversely affected by such data. This paper introduces a method to overcome such problems associated with noise and missing data by modelling the time series data with polynomials and using these models to cluster the data. Similarity measures for polynomials are given that comply with commonly used standard measures. The polynomial model based clustering is compared with standard clustering methods under different conditions and applied to a real gene expression data set. It shows significantly better results as noise and missing data are increased.

1 Introduction

Microarray experiments are widely used in medical and life science research [11]. This technology makes it possible to examine the behaviour of thousands of genes simultaneously. Moreover, microarray time series experiments provide an insight into the dynamics of gene activity as an essential part of cell processes.

Despite efforts to produce high quality microarray data, such data is often burdened with a considerable amount of noise. Attempts to reduce the noise are manifold, including intelligent experimental design, multiple repeats of the experiment and noise reduction techniques in the data preprocessing [13]. In addition to the noise problem, parts of the data often can not be retrieved properly so that the dataset contains missing values. For example, a dataset of several experiments with yeast (about 500,000 values) [10] has more than 11% missing values.

[*] This work is in part supported by the BBSRC in UK (BB/C506264/1).

M.R. Berthold, R. Glen, and I. Fischer (Eds.): CompLife 2006, LNBI 4216, pp. 1–10, 2006.

With decreasing quality the direct clustering (DC) of the data with standard methods [5] becomes less reliable. If the data has considerable missing data, the straightforward calculation of the score functions homogeneity and separation [4] for the cluster quality becomes impossible. To overcome these problems this paper suggests the modelling of the data with continuous functions. The model based clustering is done not on the original dataset directly, but on models learnt from it. The models reduce random noise and interpolate missing values, thereby increasing the robustness of clustering.

In this paper the polynomial model based clustering (PMC) is introduced. In contrast to the DC of the data, which calculates the similarity matrix directly from the data, PMC comprised of three steps: the modelling, the calculation of the similarity matrix from the models and the grouping.

2 Methods

The application of continuous functions in time series modelling is motivated by some specific assumptions. Time series result from measurements of a quantity at different time points (TP) over a certain time period. The quantity changes continuously if it could be measured at any time in the presumed time period. Measurement restrictions are due to extrinsic factors such as technical restrictions. Moreover, if a continuous quantity has the value x at TP a and the value y at TP b, then the quantity has any value between x and y at some TP between a and b. Often time series or functions have no sharp edges in the time response, i.e. they are differentiable or smooth.

Any smooth function can be approximated by the Taylor expansion, i.e. by a polynomial. Polynomials are easy to handle since basic operations can be done by simple algebraic manipulations on the parameters. Therefore polynomials are a natural choice in time series modelling. Nevertheless, other classes of functions might be used as well. Previously, polynomials have also been used in other applications of gene expression data modelling [8,12].

2.1 Modelling

Consider series of observations, $y_l(t_i)$ $(l = 1 \dots N,\ i \in I = \{1, \dots, T\})$, of N quantities at T TPs. The time elapsed between two measurements at t_i and t_{i+1} might be different through the series. A sub-series of $y_l(t_i)$ in which the missing values are omitted is denoted by $\tilde{y}_l(t_i)$ $i \in J$, where the index set J is the subset of I that contains these time-indices, where a value is available. If J is equal to I, then $\tilde{y}_l(t_i) = y_l(t_i)$.

Polynomials have the general form

$$P(t) = \sum_{i=0}^{n} \alpha_i t^i \ , \tag{1}$$

where n is the degree of the polynomial. To fit a polynomial to the data, the least squares method is used [9]. This method optimises the parameter, α_i, of a function $f(t, \alpha_0, \ldots, \alpha_n)$, $n+1 < |J|$, $|J|$ is the number of elements in J, such that the function $Q(\alpha_0, \ldots, \alpha_n) = \sum_{i \in J} (f(t_i, \alpha_0, \ldots, \alpha_n) - \tilde{y}_l(t_i))^2$ becomes minimal. Therefore the equations $\partial Q / \partial \alpha_k = 0$, $k = 0 \ldots n$ have to be solved. Applying this equation to polynomials yields

$$\sum_{k=0}^{n} \alpha_k \sum_{j \in J} t_j^{k+i} = \sum_{j \in J} t_j^i \tilde{y}_l(t_j) \quad i = 0, \ldots, n \ . \qquad (2)$$

These are $n+1$ linear equations for the $n+1$ parameters $\alpha_0, \ldots, \alpha_n$. To solve these equations an inverse matrix of the $(n+1) \times (n+1)$ matrix $\sum_{j \in J} t_j^{k+i}$ has to be calculated for each distinct subset J of I that occurs in the data set. To avoid large numbers in the calculation and hence a loss of precision, the time series are scaled to the time interval $[-1, 1]$.

The modelling is done using polynomials with degrees ranging from 2 to 12. Figure 1 shows examples for the degrees 4, 8 and 12. With increasing degree the models fits the data better, but also may over-fit the data.

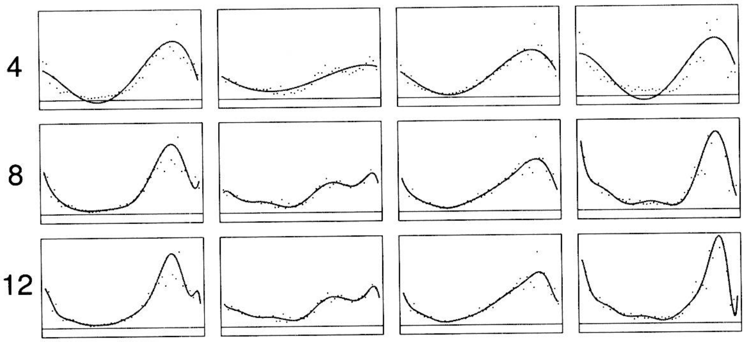

Fig. 1. Modelling of gene expression data with polynomials of different degrees

2.2 Similarity Measures

To calculate the similarity between polynomials, distance measures for functions have to be used. Usually these distance measures involve integration, which replace the sum in the equations for discrete measures. Polynomials are expandable into a Taylor-series, so that a large class of distance measures can be applied. For polynomials it is possible to calculate the anti-derivative, so that numerical integration can be avoided. Each polynomial is represented by the vector of its parameters, $(\alpha_0, \alpha_1, \ldots, \alpha_n)$. Therefore the sum of two polynomials, represented by (α_i) and (β_i), can be written as $(\alpha_0 + \beta_0, \alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n)$ and the anti-derivative of (α_i) is represented by $(0, \alpha_0, 1/2\alpha_1, \ldots, 1/(n+1)\alpha_n)$. The representations for the products and derivatives of polynomials can be found analogously. Therefore, the calculation of the integrals can be reduced to some simple algebraic operations on the $n+1$ parameters α_i, which keeps the computational complexity for the distance measures low. The calculation of the derivative, the anti-derivative, the sum and the function value of polynomials takes $O(n)$ operations, the calculation of the product of two polynomials takes $O(n^2)$ operations. For the DC the calculation effort depends on the number of TPs T. Because the number of parameters has to be considerably smaller than the number of TPs (otherwise the models would be over-fit), the calculation of the similarity matrix takes less operations for the PMC than for the DC. Two distance measures are considered, the L_p distance and the distance based on a continuous Pearson correlation coefficient.

$\mathbf{L}_p$ Distance. The L_p distance is a standard distance in the space of continuous functions and is equivalent to the p-distance (Minkowski distance) for finite dimensional spaces, such as Euclidean distance,

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum (x_i - y_i)^2} \ , \tag{3}$$

or the Manhattan distance. Let $\boldsymbol{x} = x(t)$ and $\boldsymbol{y} = y(t)$ be continuous functions over the closed interval [a,b], then the L_p distance is given by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt[p]{\int_a^b |x(t) - y(t)|^p \mathrm{d}t} \ . \tag{4}$$

Usual choices for the exponent are $p = 2$, which is analogous to the Euclidean distance or $p = 1$, which is analogous to the Manhattan distance.

Continuous Correlation Coefficient. Using the mean value theorem of calculus it is possible to formulate the Pearson correlation coefficient of series $\boldsymbol{x} = \{x_i\}$ and $\boldsymbol{y} = \{y_i\}$,

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum x_i y_i - \frac{1}{N} \sum x_i \sum y_i}{\sqrt{\left(\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right)\left(\sum y_i^2 - \frac{1}{N}(\sum y_i)^2\right)}} \ , \tag{5}$$

for integrable functions. Let $\boldsymbol{x} = x(t)$ and $\boldsymbol{y} = y(t)$ be continuous functions over the closed interval $[a, b]$ and $L = b - a$, then the correlation r can be calculated by

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\int_a^b xy \mathrm{d}t - \frac{1}{L} \int_a^b x \mathrm{d}t \int_a^b y \mathrm{d}t}{\sqrt{\left(\int_a^b x^2 \mathrm{d}t - \frac{1}{L}(\int_a^b x \mathrm{d}t)^2\right)\left(\int_a^b y^2 \mathrm{d}t - \frac{1}{L}(\int_a^b y \mathrm{d}t)^2\right)}} \ . \tag{6}$$

2.3 Grouping

The similarity matrices can be used with any standard clustering technique. To compare the method presented in this paper the Partitioning Around Medoids (PAM) [6] and two variations of hierarchical clustering algorithms were used, the average-linkage cluster analysis and the complete-linkage algorithm [3]. These methods are well-established and have been used for clustering microarray data with some success.

3 Data Set

The PMC is tested with a subset of the gene expression data of the malaria intraerythrocytic developmental cycle [2]. This subset was chosen, because a functional interpretation of the genes is known and can be used to assess the clusterings. It comprises 530 genes in 14 functional groups. The gene expression is measured in 48 TPs with 1 hour time differences. The data set contained 0.32% of missing data and had a low noise level, which has been verified through [2] and by visually plotting many of the functional groups.

4 Experiments

In every experiment the clustering is done with PAM, the average-linkage method and the complete-linkage method. For DC the methods were always applied to both the Euclidean and the correlation based similarity matrix. Polynomials of degrees from 2 to 12 were fitted to each variation of the data set and both the L_2 distance (4) and the correlation (6) were used for clustering. The following experiments were conducted.

1. The data set was clustered without any variations.
2. Normal distributed noise was added to the data. The standard deviation varied between 2% and 66% of the overall mean of the original gene expression values. The experiment was repeated 25 times.
3. The data set was changed by randomly deleting values. The number of missing values varied between 2% and 50%. The experiment was repeated 25 times.

To validate the clustering results, the weighted κ (WK) method [1,7] and quotient of homogeneity and separation (H/S) [4] were used. The WK is a similarity metric between clusters, with possible values between -1 and 1. The larger the WK value the better the agreement between the cluster results. For a clustering $\mathcal{C} = \{C_1, \ldots, C_K\}$ and a distance measure d H/S is given by

$$H(\mathcal{C}) = \sum_{k=1}^{K} H(C_k) = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in C_k} d(\boldsymbol{x}, \boldsymbol{r}_k)^2 \tag{7}$$

and

$$S(\mathcal{C}) = \sum_{1 \le l < k \le K} d(\boldsymbol{r}_j, \boldsymbol{r}_k)^2 \ , \tag{8}$$

where $\boldsymbol{r}_k = 1/n_k \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}$ are the cluster centres. A good clustering should have a low homogeneity value and a high separation value, hence a low H/S quotient. Because it is a quotient of sums of distances, H/S is always non-negative.