Raffaele Giancarlo
Sridhar Hannenhalli (Eds.)

LNBI 4645

Algorithms in Bioinformatics

7th International Workshop, WABI 2007
Philadelphia, PA, USA, September 2007
Proceedings

Springer

Raffaele Giancarlo Sridhar Hannenhalli (Eds.)

Algorithms in Bioinformatics

7th International Workshop, WABI 2007
Philadelphia, PA, USA, September 8-9, 2007
Proceedings

Springer

Volume Editors

Raffaele Giancarlo
Università degli Studi di Palermo
Department of Mathematics
via Archirafi 34, 90123 Palermo, Italy
E-mail: raffaele@math.unipa.it

Sridhar Hannenhalli
University of Pennsylvania
Penn Center for Bioinformatics and Department of Genetics
1409 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA
E-mail: sridharh@pcbi.upenn.edu

Library of Congress Control Number: 2007932232

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

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-74125-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-74125-1 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: 12103256 06/3180 5 4 3 2 1 0

Lecture Notes in Bioinformatics 4645

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

Subseries of Lecture Notes in Computer Science

Preface

We are very pleased to present the proceedings of the Seventh Workshop on Algorithms in Bioinformatics (WABI 2007), which took place in Philadelphia, September 8–9, 2007, under the auspices of the International Society for Computational Biology (ISCB), the European Association for Theoretical Computer Science (EATCS), the Penn Genomics Institute and the Penn Center for Bioinformatics.

The Workshop on Algorithms in Bioinformatics covers research in all aspects of algorithmic work in bioinformatics. The emphasis is on discrete algorithms that address important problems in molecular biology, that are founded on sound models, that are computationally efficient, and that have been implemented and tested in simulations and on real datasets. The goal is to present recent research results, including significant work-in-progress, and to identify and explore directions of future research. Specific topics of interest include, but are not limited to:

- Exact, approximate, and machine-learning algorithms for genomics, sequence analysis, gene and signal recognition, alignment, molecular evolution, polymorphisms and population genetics, protein and RNA structure determination or prediction, gene expression and gene networks, proteomics, functional genomics, and drug design.
- Methods, software and dataset repositories for development and testing of such algorithms and their underlying models.
- High-performance approaches to computationally hard problems in bioinformatics, particularly optimization problems.

A major goal of the workshop is to bring together researchers in areas spanning the range from abstract algorithm design to biological dataset analysis, so as to enable a dialogue between application specialists and algorithm designers, mediated by algorithm engineers and high-performance computing specialists. We believe that such a dialogue is necessary for the progress of computational biology, inasmuch as application specialists cannot analyze their datasets without fast and robust algorithms and, conversely, algorithm designers cannot produce useful algorithms without being conversant with the problems faced by biologists.

Part of this mix has been achieved for all seven WABI events. For six of them, WABI was collocated with the European Symposium on Algorithms (ESA), along with other occasional conferences or workshops, so as to form the interdisciplinary scientific meeting known as ALGO. As agreed by the WABI and ALGO Steering Committees, starting this year WABI will be part of ALGO only every two years, alternating between Europe and other continents.

We received 133 submissions in response to our call for WABI 2007 and were able to accept 37 of them, ranging from mathematical tools to experimental studies of approximation algorithms and reports on significant computational analyses. Numerous biological problems were dealt with, including genetic mapping, sequence alignment and sequence analysis, phylogeny, comparative genomics, and protein structure. Both machine-learning and combinatorial optimization approaches to algorithmic problems in bioinformatics were represented.

We want to thank all authors for submitting their work to the workshop and all presenters and attendees for their participation. We were particularly fortunate in enlisting the help of a very distinguished panel of researchers for our Program Committee, which undoubtedly accounts for the large number of submissions and the high quality of the presentations. Our sincere thanks go to all:

Piotr Berman, Penn. State U., USA
Mathieu Blanchette, McGill U., Canada
Paola Bonizzoni, U. Milano-Bicocca, Italy
Philipp Bücher, EPFL, Switzerland
Rita Casadio, U. Bologna, Italy
Maxime Crochemore, U. Marne-la-Vallée, France
Nadia El-Mabrouk, U. Montréal, Canada
Liliana Florea, George Washington U., USA
Olivier Gascuel, LIRMM-CNRS, France
David Gilbert, U. Glasgow, UK
Concettina Guerra, U. Padova, Italy & Georgia Tech, USA
Roderico Guigo, CRG, U. Barcelona, Spain
Daniel Huson, U. Tübingen, Germany
Shane Jensen, U. Penn., USA
Jens Lagergren, KTH Stokholm, Sweden
Arthur Lesk, Penn. State U., USA
Ming Li, U. Waterloo, Canada
Stefano Lonardi, UC Riverside, USA
Webb Miller, Penn. State U., USA
Satoru Miyano, Tokyo U., Japan
Bernard Moret, EPFL, Switzerland
Burkhard Morgenstern, U. Göttingen, Germany
Gene Myers, HHMI Janelia Farms, USA
Uwe Ohler, Duke U., USA
Laxmi Parida, IBM T.J. Watson Research Center, USA
Kunsoo Park, Seoul National U., S. Korea
Graziano Pesole, U. Bari, Italy
Ron Pinter, Technion, Israel
Cinzia Pizzi, INRIA, France
Knut Reinert, Freie U. Berlin, Germany
Mikhail Roytberg, Russian Academy of Sciences, Russia
Marie France Sagot, INRIA, France
David Sankoff, U. Ottawa, Canada
Roded Sharan, Tel-Aviv U., Israel

Adam Siepel, Cornell U., USA
Mona Singh, Princeton U., USA
Saurabh Sinha, UIUC, USA
Steven Skiena, SUNY Stony Brook, USA
Peter Stadler, U. Leipzig, Germany
Jens Stoye, U. Bielefeld, Germany
Granger Sutton, J. Craig Venter Institute, USA
Anna Tramontano, U. Roma "La Sapienza", Italy
Olga Troyanskaya, Princeton U., USA
Alfonso Valencia, U. Autonoma, Spain
Gabriel Valiente, Tech U. Catalonia, Spain
Li-San Wang, U. Penn., USA
Lusheng Wang, City U. Hong Kong, Hong Kong
Haim Wolfson, Tel-Aviv U., Israel

We would also like to thank Alessandra Gabriele, Giusué Lo Bosco and Cesare Valenti, all of University of Palermo, for providing assistance in assembling this volume. Last but not least, we thank Junhyong Kim and his colleagues Stephen Fisher and Li-San Wang, all at U. Penn, for doing a superb job of organizing the first edition of the conference in the USA and for the continuous technical support during all phases of the conference.

We hope that you will consider contributing to future WABI events, through a submission or by participating in the workshop.

September 2007

Raffaele Giancarlo
Sridhar Hannenhalli

Organization

The WABI 2007 Program Committee gratefully acknowledges the valuable input received from the following external Reviewers:

Edo Airoldi
J. A. Amgarten Quitzau
Lars Arvestad
Marie-Pierre Béal
Vincent Berry
Enrique Blanco
Guillaume Blin
Serdar Bozdag
Kajia Cao
Ildefonso Cases
Robert Castelo
Cedric Chauve
Giovanni Ciriello
Jordi Cortadella
Gianluca Della Vedova
Pietro Di Lena
Riccardo Dondi
Iakes Ezkurdia
Piero Fariselli
Alfredo Ferro
Oxana Galzitskaia
Claudio Garutti
Stoyan Georgiev
Robert Giegerich
Osvaldo Graña
Clemens Gröpl
Roderic Guigo i Serra
Bjarni Halldorsson
Michael Hallett
Sylvie Hamel
Elena Harris
Robert Harris
M. Helmer-Citterich
Matthew Hibbs
Curtis Huttenhower
Seiya Imoto
Yuval Inbar
Dmitry Ivankov
Katharina Jahn
Jieun Jeong
Tao Jiang
Raya Khanin
Jong Kim
Gunnar W. Klau
Tobias Kloepper
Arun Konagurthu
Mathieu Lajoie
Florian Leitner
Gonzalo Lopez
Antoni Lozano
Bill Majoros
Mohamed Manal
Florian Markowetz
Pier Luigi Martelli
David Martin
Efrat Mashiach
Jon McAuliffe
Julia Mixtacki
Chad Myers
Luay Nakhleh
Heiko Neuweger
Giulio Pavesi
Ernesto Picardi
M. Sohel Rahman
Sven Rahmann
Vincent Ranwez
Christian Rausch
Antonio Rausell
Daniel Richter
Romeo Rizzi
Jairo Rocha
Allen Rodrigo
Oleg Rokhlenko
Ivan Rossi
Bengt Sennblad
Maria Serna
Maxim Shatsky
Tomer Shlomi
Michael Shmoish
A. Shulman-Peleg
Jijun Tang
Ali Tofigh
Vladimir Vacic
Marco Vassura
Stphane Vialette
Jordi Villa i Freixa
Robert Warren
Tobias Wittkop
Stefan Wolfsheimer
Yonghui Wu
Joseph Wun-Tat Chan
Nir Yosef

Lecture Notes in Bioinformatics

Vol. 4645: R. Giancarlo, S. Hannenhalli (Eds.), Algorithms in Bioinformatics. XIII, 432 pages. 2007.

Vol. 4643: M.-F. Sagot, M.E.M.T. Walter (Eds.), Advances in Bioinformatics and Computational Biology. IX, 177 pages. 2007.

Vol. 4544: S. Cohen-Boulakia, V. Tannen (Eds.), Data Integration in the Life Sciences. XI, 282 pages. 2007.

Vol. 4532: T. Ideker, V. Bafna (Eds.), Systems Biology and Computational Proteomics. IX, 131 pages. 2007.

Vol. 4463: I. Măndoiu, A. Zelikovsky (Eds.), Bioinformatics Research and Applications. XV, 653 pages. 2007.

Vol. 4453: T. Speed, H. Huang (Eds.), Research in Computational Molecular Biology. XVI, 550 pages. 2007.

Vol. 4414: S. Hochreiter, R. Wagner (Eds.), Bioinformatics Research and Development. XVI, 482 pages. 2007.

Vol. 4366: K. Tuyls, R.L. Westra, Y. Saeys, A. Nowé (Eds.), Knowledge Discovery and Emergent Complexity in Bioinformatics. IX, 183 pages. 2007.

Vol. 4360: W. Dubitzky, A. Schuster, P.M.A. Sloot, M. Schroeder, M. Romberg (Eds.), Distributed, High-Performance and Grid Computing in Computational Biology. X, 192 pages. 2007.

Vol. 4345: N. Maglaveras, I. Chouvarda, V. Koutkias, R. Brause (Eds.), Biological and Medical Data Analysis. XIII, 496 pages. 2006.

Vol. 4316: M.M. Dalkilic, S. Kim, J. Yang (Eds.), Data Mining and Bioinformatics. VIII, 197 pages. 2006.

Vol. 4230: C. Priami, A. Ingólfsdóttir, B. Mishra, H.R. Nielson (Eds.), Transactions on Computational Systems Biology VII. VII, 185 pages. 2006.

Vol. 4220: C. Priami, G. Plotkin (Eds.), Transactions on Computational Systems Biology VI. VII, 247 pages. 2006.

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

Vol. 4210: C. Priami (Ed.), Computational Methods in Systems Biology. X, 323 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. 4023: E. Eskin, T. Ideker, B. Raphael, C. Workman (Eds.), Systems Biology and Regulatory Genomics. X, 259 pages. 2007.

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.C. 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

Shotgun Protein Sequencing

Pavel A. Pevzner

Ronald R. Taylor Professor of Computer Science, University of California, San Diego,
La Jolla, CA 92093

Abstract. Despite significant advances in the identification of known proteins, the analysis of unknown proteins by tandem mass spectrometry (MS/MS) still remains a challenging open problem. Although Klaus Biemann recognized the potential of mass spectrometry for sequencing of unknown proteins in the 1980s, low-throughput Edman degradation followed by cloning still remains the main method to sequence unknown proteins. The automated spectral interpretation has been limited by a focus on individual spectra and has not capitalized on the information contained in spectra of overlapping peptides. Indeed, the powerful Shotgun DNA Sequencing strategies have not been extended to protein sequencing yet. We demonstrate, for the first time, the feasibility of Shotgun Protein Sequencing of protein mixtures and validate this approach by generating highly accurate de novo reconstructions of various proteins in western diamondback rattlesnake venom. We further argue that Shotgun Protein Sequencing has the potential to overcome the limitations of current protein sequencing approaches and thus catalyze the otherwise impractical applications of proteomics methodologies in studies of unknown proteins. We further describe applications of this technique to analyzing proteins that are not directly inscribed in DNA sequences (like antibodies and fusion proteins in cancer).

This is a joint work with Nuno Bandeira (UCSD) and Karl Clauser (Broad).

R. Giancarlo and S. Hannenhalli (Eds.): WABI 2007, LNBI 4645, p. 1, 2007.

Locality Kernels for Protein Classification

Evgeni Tsivtsivadze, Jorma Boberg, and Tapio Salakoski

Turku Centre for Computer Science (TUCS)
Department of Information Technology, University of Turku
Joukahaisenkatu 3-5 B, FIN-20520 Turku, Finland
firstname.lastname@it.utu.fi

Abstract. We propose kernels that take advantage of local correlations in sequential data and present their application to the protein classification problem. Our locality kernels measure protein sequence similarities within a small window constructed around matching amino acids. The kernels incorporate positional information of the amino acids inside the window and allow a range of position dependent similarity evaluations. We use these kernels with regularized least-squares algorithm (RLS) for protein classification on the SCOP database. Our experiments demonstrate that the locality kernels perform significantly better than the spectrum and the mismatch kernels. When used together with RLS, performance of the locality kernels is comparable with some state-of-the-art methods of protein classification and remote homology detection.

1 Introduction

One important task in computational biology is inference of the structure and function of the protein encoded in the genome. The similarity of protein sequences may imply structural and functional similarity. The task of detecting these similarities can be formalized as a classification problem that treats proteins as a set of labeled examples which are in positive class if they belong to the same family and are in negative class otherwise.

Recently, applicability of this discriminative approach for detecting remote protein homologies has been demonstrated by several studies. For example, Jaakkola et al. [1] show that by combining discriminative learning algorithm and Fisher kernel for extraction of the relevant features it is possible to achieve a good performance in protein family recognition. Liao and Noble [2] further improve results presented in [1] by proposing combination of pairwise sequence similarity feature vectors with Support Vector Machines (SVM) algorithm. Their algorithm called SVM-pairwise is performing significantly better than several other baseline methods such as SVM-Fisher, PSI-BLAST and profile HMMs.

The methods described in [1] and [2] use an expensive step of generating vector valued features for protein discrimination problems, which increases computational time of the algorithm. The idea to use a simple kernel function that can be efficiently computed and does not depend on any generative model or separate preprocessing step is considered by Leslie et al. in [3]. They show that

R. Giancarlo and S. Hannenhalli (Eds.): WABI 2007, LNBI 4645, pp. 2–11, 2007.

simple sequence based kernel functions perform surprisingly well compared to other computationally expensive approaches.

In this study, we address the problem of protein sequence classification using the RLS algorithm with locality kernels similar to the one we proposed in [4]. The features used by the locality kernels represent sequences contained in a small window constructed around matching amino acids in the compared proteins. The kernels make use of the range of similarity evaluations within the windows, namely *position insensitive matching*: amino acids that match are taken into account irrespective of their position, *position sensitive matching*: amino acids that match but have different positions are penalized, *strict matching*: only amino acids that match and have the same positions are taken into account. By incorporating information about relevance of local correlations and positions of amino acids in the sequence into the kernel function, we demonstrate significantly better performance in protein classification on Structural Classification of Proteins (SCOP) database [5] than that of the spectrum and the mismatch kernels [3,6,7].

Previously, we have shown that the locality-convolution kernel [4] can be successfully applied to parse ranking task in natural language processing. The similarity of the data representation in cases of biological sequence and text, as well as results obtained in this study, suggest that locality kernels can be applied to tasks where local correlations and positional information within the sequence might be important.

The paper is organized as follows. In Section 2, we present overview of the RLS algorithm. In Section 3, we define notions of locality window, positional matching, and present locality kernels. In Section 5, we evaluate the applicability of the locality kernels for the task of protein classification and compare their performance with the spectrum and the mismatch kernels. We conclude this paper in Section 6.

2 Regularized Least-Squares Algorithm

Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_t, y_t)\}$, where $\mathbf{x}_i = (x_1, \ldots, x_n)^{\mathrm{T}}$, $\mathbf{x}_i \in S$ and $y_i \in \{0, 1\}$ be the set of training examples. The target output value y_i is a label value which is either 0, indicating that $\mathbf{x}_i$ does not belong to the class or 1 otherwise. The target output value is predicted by the regularized least-squares (RLS) algorithm [8,9]. We denote a matrix whose rows are $\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_t^{\mathrm{T}}$ as X and a vector of output labels as $\mathbf{y} = (y_1, \ldots y_t)^{\mathrm{T}}$. The RLS algorithm corresponds to solving following optimization problem:

$$\min_{\mathbf{w}} \sum_{i=1}^{t} (y_i - f(\mathbf{x}_i))^2 + \lambda \|\mathbf{w}\|^2, \qquad (1)$$

where $f : S \to \mathbb{R}$, $\mathbf{w} \in \mathbb{R}^n$ is a vector of parameters such that $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$, and $\lambda \in \mathbb{R}_+$ is a regularization parameter that controls the trade-off between fitting the training set accurately and finding the smallest norm for the function f.

Rewriting (1) in matrix form and taking derivative with respect to $\mathbf{w}$, we obtain

$$\mathbf{w} = (X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}\mathbf{y}, \tag{2}$$

where I denotes identity matrix of dimension $n \times n$. In (2) we must perform matrix inverse in dimension of feature space, that is $n \times n$. However, if the number of features is much larger than the number of training data points, a more efficient way is to perform inverse in the dimension of training examples. In that case, following [9], we present (2) as a linear combination of training data points:

$$\mathbf{w} = \sum_{i=1}^{t} a_i \mathbf{x}_i, \tag{3}$$

where

$$a = (K + \lambda I)^{-1}\mathbf{y} \tag{4}$$

and $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel matrix that contains the pairwise similarities of data points computed by a kernel function $k : S \times S \to \mathbb{R}$. Finally, we predict an output of new data point as follows:

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{y}^{\mathrm{T}}(K + \lambda I)^{-1}\mathbf{k}, \tag{5}$$

where $k_i = k(\mathbf{x}_i, \mathbf{x})$. Kernel functions are similarity measures of data points in the input space S, and they correspond to the inner product in a feature space H to which the input space data points are mapped. The kernel functions are defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle,$$

where $\Phi : S \to H$. Next we formulate the locality kernel functions that are used with the RLS algorithm for protein classification task.

3 Locality Kernels

There are three key properties of the locality kernels that make them applicable to the task of remote homology detection in the proteins. Firstly, the features used by these kernels contain amino acids that are extracted in the order of their appearance in the protein sequence. Secondly, local correlations within the protein sequence are taken into account by constructing a small window around the matching amino acids. Finally, positional information of the amino acids contained within window is used for similarity evaluation.

Let us consider proteins $\mathbf{p}, \mathbf{q}$ and let $\mathbf{p} = (p_1, \ldots, p_{|\mathbf{p}|})$ and $\mathbf{q} = (q_1, \ldots, q_{|\mathbf{q}|})$ be their amino acid sequences. The similarity of $\mathbf{p}$ and $\mathbf{q}$ is obtained with kernel

$$k(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{|\mathbf{p}|} \sum_{j=1}^{|\mathbf{q}|} \kappa(i, j). \tag{6}$$