

# Lecture Notes in Bioinformatics

2666

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

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

Subseries of Lecture Notes in Computer Science

**Springer**

*Berlin*

*Heidelberg*

*New York*

*Hong Kong*

*London*

*Milan*

*Paris*

*Tokyo*

Concettina Guerra Sorin Istrail (Eds.)

# Mathematical Methods for Protein Structure Analysis and Design

C.I.M.E. Summer School

Martina Franca, Italy, July 9-15, 2000

Advanced Lectures



Springer

## Series Editors

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

## Volume Editors

Concettina Guerra  
Università degli Studi di Padova  
Dipartimento di Ingegneria dell'Informazione  
via Gradenigo 6a, 35131 Padova, Italy  
E-mail: guerra@dei.unipd.it

Sorin Istrail  
Celera Genomics, Applied Biosystems  
45 West Gude Drive, Rockville, MD 20850, USA  
E-mail: sorin.istrail@celera.com

## Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek.  
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;  
detailed bibliographic data is available in the Internet at <<http://dnb.ddb.de>>.

CR Subject Classification (1998): J.3, F.2, H.2, G.2, I.3.5, I.4

ISSN 0302-9743

ISBN 3-540-40104-0 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York  
a member of BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2003  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign  
Printed on acid-free paper      SPIN: 10927403      06/3142      5 4 3 2 1 0

---

## Preface

The papers collected in this volume reproduce contributions by leading scholars to an international school and workshop which was organized and held with the goal of taking a snapshot of a discipline under tumultuous growth. Indeed, the area of protein folding, docking and alignment is developing in response to needs for a mix of heterogeneous expertise spanning biology, chemistry, mathematics, computer science, and statistics, among others.

Some of the problems encountered in this area are not only important for the scientific challenges they pose, but also for the opportunities they disclose in terms of medical and industrial exploitation. A typical example is offered by protein-drug interaction (docking), a problem posing daunting computational problems at the crossroads of geometry, physics and chemistry, and, at the same time, a problem with unimaginable implications for the pharmacopoeia of the future.

The school focused on problems posed by the study of the mechanisms behind protein folding, and explored different ways of attacking these problems under objective evaluations of the methods. Together with a relatively small core of consolidated knowledge and tools, important reflections were brought to this effort by studies in a multitude of directions and approaches. It is obviously impossible to predict which, if any, among these techniques will prove completely successful, but it is precisely the implicit dialectic among them that best conveys the current flavor of the field. Such unique diversity and richness inspired the format of the meeting, and also explains the slight departure of the present volume from the typical format in this series: the exposition of the current sediment is complemented here by a selection of qualified specialized contributions.

The topics covered in this volume pinpoint major issues arising in the development and analysis of models, algorithms and software tools centered on the structure of proteins, all of which play crucial roles in structural genomics and proteomics. The study of 3D conformations and relationships among proteins is motivated by the belief that the spatial structure, more than the primary sequence, dictates the function of a protein. The largest repository of

3D protein structures is the Protein Data Bank (PDB), currently containing about 17,000 proteins. The PDB has experienced a sustained growth and is expected to continue to grow at an increasing pace in the near future. The available structures are classified into a relatively small number of families and folds, according to their three-dimensional conformation. While the number of proteins will continue to grow, it is widely believed that the number of new folds will remain relatively stable. Structural comparisons involving these structures are at the core of docking and the classification of proteins and sub-aggregates, and motif searches in sequence and protein databases, and ultimately they contribute to understanding the mechanics of folding in living organisms.

The first three chapters of this volume contain material that was presented at the school. The chapter entitled “Protein Structure Comparison: Algorithms and Applications,” by G. Lancia and S. Istrail, focuses on the algorithmic aspects and applications of the problem of structure comparison. Structure similarity scoring schemes used in pairwise structure comparison are discussed with respect to the ability to capture the biological relevance of the chemical and physical constraints involved in molecular recognition. Particular attention is paid to the measures based on *contact map* similarity.

The chapter “Spatial Pattern Detection in Structural Bioinformatics,” by H.J. Wolfson, discusses the task of protein structural comparison as well as the prediction of protein-protein, protein-DNA or protein-drug interaction (docking). Different protein shape representations are used in biological pattern discovery. The paper discusses the shape representations best suited to each computational task, then outlines some rigid and flexible protein structural alignment algorithms, and discusses the issues of rigid bound versus unbound and flexible docking.

The chapter “Geometric Methods for Protein Structure Comparison,” by C. Ferrari and C. Guerra, discusses, from a theoretical point of view, geometric solutions to the problem of finding correspondences between sets of geometric features, such as points or segments. After reviewing existing methods for the estimation of rigid transformations under different metrics, the paper focuses on the use of the secondary structures of proteins for fast retrieval of similarity. It also deals with the integration of strategies using different levels of protein representations, from atomic to secondary structure level.

The chapter “Identifying Flat Regions and Slabs in Protein Structures,” by M.E. Bock and C. Guerra, presents geometric approaches to the extraction of planar surfaces, which is motivated by the problem of identifying packing regions in proteins.

The two contributions, “OPTIMA: a New Score Function for the Detection of Remote Homologs,” by M. Kann and R.A. Goldstein, and “A Comparison of Methods for Assessing the Structural Similarity of Proteins,” by D.C. Adams and G.J.P. Naylor, deal with the problem of protein comparison, focusing on different similarity functions for sequence and structure comparison.

The next three papers, “Prediction of Protein Secondary Structure at High Accuracy Using a Combination of Many Neural Networks,” by C. Lundegaard, T.N. Petersen, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G. Gippert and O. Lund, “Self-consistent Knowledge-Based Approach to Protein Design,” by A. Rossi, C. Micheletti, F. Seno, A. Maritan, and “Learning Effective Amino-Acid Interactions,” by F. Seno, C. Micheletti, A. Maritan and J.R. Banavar, discuss techniques and criteria for protein folding and design.

The paper “Protein structure from solid-state NMR,” by J.R. Quine and T.A. Cross, presents a mathematical analysis for solid-state nuclear magnetic resonance (NMR). Finally, the contribution “Protein-like Properties of Simple Models,” by Y.-H. Sanejouand and G. Trinquier, focuses on properties relevant to the sequence-structure relationships.

The school was attended by 56 participants from 10 countries. Lectures were given by Prof. Ken Dill, University of California (USA), Prof. Arthur Lesk, University of Cambridge Clinical School (UK), Prof. Michael Levitt, Stanford University School of Medicine (USA), Prof. John Moult, University of Maryland (USA), and Prof. Haim Wolfson, Tel Aviv University (Israel). Invited talks at the workshop were given by Prof. Mary Ellen Bock, Purdue University (USA) and Dr. Andrea Califano, IBM Yorktown (USA).

Concettina Guerra  
Sorin Istrail

---

# Contents

## **Protein Structure Comparison: Algorithms and Applications**

<i>Giuseppe Lancia, Sorin Istrail</i> .....	1
1 Introduction .....	1
2 Preliminaries .....	4
3 Applications of Structure Comparisons .....	6
4 Software and Algorithms for Structure Comparison .....	11
5 Problems Based on Contact Map Representations .....	20
6 Acknowledgements .....	30
References .....	30

## **Spatial Pattern Detection in Structural Bioinformatics**

<i>Haim J. Wolfson</i> .....	35
1 Introduction .....	35
2 Protein Shape Representation .....	37
3 Protein Structural Alignment .....	38
4 Protein-Protein Docking .....	46
5 Summary .....	52
References .....	53

## **Geometric Methods for Protein Structure Comparison**

<i>Carlo Ferrari, Concettina Guerra</i> .....	57
1 Introduction .....	57
2 Protein Description .....	59
3 Structural Comparison: Problem Formulation .....	62
4 Representation of Rigid Transformations .....	63
5 Determination of 3D Rigid Transformations .....	68
6 Geometric Pattern Matching .....	71
7 Indexing Techniques .....	73
8 Graph-Theoretic Approaches .....	76
9 Integration of Methods for Protein Comparison Using Different Representations .....	77



10 Conclusions .....	78
11 Acknowledgements .....	79
References .....	79

## **Identifying Flat Regions and Slabs in Protein Structures**

<i>Mary Ellen Bock, Concettina Guerra</i> .....	83
1 Introduction .....	83
2 A Geometric Algorithm .....	85
3 An Improved Geometric Algorithm .....	87
4 Hough Transform .....	88
5 Performances of the Two Algorithms .....	90
6 Plane Detection in Proteins .....	92
7 Acknowledgements .....	95
References .....	96

## **OPTIMA: A New Score Function for the Detection of Remote Homologs**

<i>Maricel Kann, Richard A. Goldstein</i> .....	99
1 Abstract .....	99
2 Introduction .....	99
3 Methods .....	100
References .....	107

## **A Comparison of Methods for Assessing the Structural Similarity of Proteins**

<i>Dean C. Adams, Gavin J.P. Naylor</i> .....	109
1 Introduction .....	109
2 The DALI Algorithm .....	109
3 The Root Mean Square Algorithm .....	110
4 Geometric Morphometrics .....	111
5 Comparison of Methods .....	111
6 Discussion .....	113
References .....	114

## **Prediction of Protein Secondary Structure at High Accuracy Using a Combination of Many Neural Networks**

<i>Claus Lundegaard, Thomas Nordahl Petersen, Morten Nielsen, Henrik Bohr, Jacob Bohr, Søren Brunak, Garry Gippert, Ole Lund</i> .....	117
1 Summary .....	117
2 Introduction .....	117
3 Methods .....	118
4 Results .....	119
References .....	121

**Self-consistent Knowledge-Based Approach to Protein Design**

<i>Andrea Rossi, Cristian Micheletti, Flavio Seno, Amos Maritan</i> .....	123
1 Introduction .....	123
2 The Design Strategy .....	124
3 Results and Discussion .....	125
4 Summary .....	127
References .....	128

**Protein Structure from Solid-State NMR**

<i>John R. Quine, Timothy A. Cross</i> .....	131
1 Discrete Curves .....	131
2 Tensors and NMR .....	133
3 Structure from Orientational Constraints .....	134
4 Acknowledgment .....	136
References .....	136

**Learning Effective Amino-Acid Interactions**

<i>Flavio Seno, Cristian Micheletti, Amos Maritan, Jayanth R. Banavar</i> ..	139
1 Introduction .....	139
2 Models and Techniques .....	141
3 Results .....	142
4 Conclusions .....	144
References .....	144

**Proteinlike Properties of Simple Models**

<i>Yves-Henri Sanejouand, Georges Trinquier</i> .....	147
1 The 3x3x3 Cubic Lattice Model .....	147
2 N-Soft-Spheres Models .....	151
References .....	152

<b>List of Participants</b> .....	155
-----------------------------------	-----