

# Contents

Preface	page	v
List of protocols		xvii
Abbreviations		xix

## 1 Threading methods for protein structure prediction 1

*David Jones and Caroline Hadley*

1	Introduction	1
2	Threading methods	1
	1-D-3-D profiles: Bowie <i>et al.</i> (1991)	5
	Threading: Jones <i>et al.</i> (1992)	5
	Protein fold recognition using secondary structure predictions: Rost (1997)	7
	Combining sequence similarity and threading: Jones (1999)	7
3	Assessing the reliability of threading methods	8
	Alignment accuracy	9
	Post-processing threading results	10
	Why does threading work?	10
4	Limitations: strong and weak fold-recognition	11
	The domain problem in threading	11
5	The future	12
	References	12

## 2 Comparison of protein three-dimensional structures 15

*Mark S. Johnson and Jukka V. Lehtonen*

1	Introduction	15
2	The comparison of protein structures	16
	General considerations	16
	What atoms/features of protein structure to compare?	17
	Standard methods for finding the translation vector and rotation matrix	20
	Standard methods to determine equivalent matched atoms between structures	25
	Quality and extent of structural matches	29
3	The comparison of identical proteins	31
	Why compare identical proteins?	31
	Comparisons	31

## CONTENTS

4	The comparison of homologous structures: example methods	32
	Background	32
	Methods that require the assignment of seed residues	34
	Automatic comparison of 3-D structures	35
	Multiple structural comparisons	41
5	The comparison of unrelated structures	42
	Background	42
6	Large-scale comparisons of protein structures	46
	References	48
<b>3</b>	<b>Multiple alignments for structural, functional, or phylogenetic analyses of homologous sequences</b>	<b>51</b>
	<i>L. Duret and S. Abdeddaim</i>	
1	Introduction	51
2	Basic concepts for multiple sequence alignment	53
	Homology: definition and demonstration	53
	Global or local alignments	54
	Substitution matrices, weighting of gaps	54
3	Searching for homologous sequences	56
4	Multiple alignment methods	57
	Optimal methods for global multiple alignments	59
	Progressive global alignment	61
	Block-based global alignment	63
	Motif-based local multiple alignments	65
	Comparison of different methods	65
	Particular case: aligning protein-coding DNA sequences	68
5	Visualizing and editing multiple alignments	69
	Manual expertise to check or refine alignments	71
	Annotating alignments, extracting sub-alignments	71
	Comparison of alignment editors	72
	Alignment shading software, pretty printing, logos, etc.	72
6	Databases of multiple alignments	72
7	Summary	73
	References	74
<b>4</b>	<b>Hidden Markov models for database similarity searches</b>	<b>77</b>
	<i>Ewan Birney</i>	
1	Introduction	77
2	Overview	78
3	Using profile and profile-HMM databases	79
	Pfam	80
	Prosite profiles	80
	SMART	81
	Other resources and future directions	81
	Limitations of profile-HMM databases	81
4	Using PSI-BLAST	81

5	Using HMMER2	82
	Overview of using HMMER	83
	Making the first alignment	83
	Making a profile-HMM from an alignment	84
	Finding homologues and extending the alignment	84
6	False positives	85
7	Validating a profile-HMM match	85
8	Practical issues of the theories behind profile-HMMs	86
	Overview of profile-HMMs	86
	Statistics for profile-HMM	87
	Profile-HMM construction	89
	Priors and evolutionary information	89
	Technical issues	90
	References	91

## **5 Protein family-based methods for homology detection and analysis**

*Steven Henikoff and Jorja Henikoff*

1	Introduction	93
	Expanding protein families	93
	Terms used to describe relationships among proteins	93
	Alternative approaches to inferring function from sequence alignment	94
2	Displaying protein relationships	95
	From pairwise to multiple-sequence alignments	95
	Patterns	96
	Logos	97
	Trees	97
3	Block-based methods for multiple-sequence alignment	98
	Pairwise alignment-initiated methods	98
	Pattern-initiated methods	99
	Iterative methods	99
	Implementations	100
4	Position-specific scoring matrices (PSSMs)	101
	Sequence weights	102
	PSSM column scores	102
5	Searching family databases with sequence queries	103
	Curated family databases: Prosite, Prints, and Pfam	105
	Clustering databases: ProDom, DOMO, Protomap, and Prof_pat	105
	Derived family databases: Blocks and Proclass	106
	Other tools for searching family databases	107
6	Searching with family-based queries	108
	Searching with embedded queries	108
	Searching with PSSMs	108
	Iterated PSSM searching	109
	Multiple alignment-based searching of protein family databases	110
	References	110

**6 Predicting secondary structure from protein sequences 113**

*Jaap Heringa*

- 1 Introduction 113
  - What is secondary structure? 113
  - Where could knowledge about secondary structure help? 114
  - What signals are there to be recognized? 114
- 2 Assessing prediction accuracy 118
- 3 Prediction methods for globular proteins 120
  - The early methods 120
  - Accuracy of early methods 122
  - Other computational approaches 122
  - Prediction from multiply-aligned sequences 123
  - A consensus approach: JPRED 129
  - Multiple-alignment quality and secondary-structure prediction 131
  - Iterated multiple-alignment and secondary structure prediction 132
- 4 Prediction of transmembrane segments 133
  - Prediction of  $\alpha$ -helical TM segments 134
  - Orientation of transmembrane helices 136
  - Prediction of  $\beta$ -strand transmembrane regions 136
- 5 Coiled-coil structures 137
- 6 Threading 138
- 7 Recommendations and conclusions 138
  - References 139

**7 Methods for discovering conserved patterns in protein sequences and structures 143**

*Inge Jonassen*

- 1 Introduction 143
- 2 Pattern descriptions 144
  - Exact or approximate matching 144
  - PROSITE patterns 145
  - Alignments, profiles, and hidden Markov models 146
  - Pattern significance 148
  - Pattern databases 150
  - Using existing pattern collections 153
- 3 Finding new patterns 154
  - A general approach 154
  - Discovery algorithms 155
- 4 The Pratt programs 156
  - Using Pratt 157
  - Pratt: Internal search methods 159
  - Scoring patterns 161
- 5 Structure motifs 162
  - The SPratt program 162
- 6 Examples 164
- 7 Conclusions 164
  - References 165

## **8 Comparison of protein sequences and practical database searching 167**

*Golan Yona and Steven E. Brenner*

- 1 Introduction 167
- 2 Alignment of sequences 168
  - Rigorous alignment algorithms 169
  - Heuristic algorithms for sequence comparison 171
- 3 Probability and statistics of sequence alignments 173
  - Statistics of global alignment 174
  - Statistics of local alignment without gaps 175
  - Statistics of local alignment with gaps 177
- 4 Practical database searching 178
  - Types of comparison 178
  - Databases 179
  - Algorithms 181
  - Filtering 181
  - Scoring matrices and gap penalties 182
  - Command line parameters 185
- 5 Interpretation of results 187
- 6 Conclusion 188
  - References 188

## **9 Networking for the biologist 191**

*R. A. Harper*

- 1 Introduction 191
- 2 The changing face of networking 192
  - Networking in Europe 194
  - The way we were . . . e-mail servers for sequence retrieval 195
  - Similarity searches via e-mail 199
  - Speed solutions for similarity searches 201
- 3 Sequence retrieval via the WWW 203
  - Entrez from the NCBI 205
  - SRS from the EBI 205
- 4 Submitting sequences 208
  - Bankit at NCBI 209
  - Sequin from NCBI 209
  - Webin from EBI 210
  - Sakura from DDJB 212
- 5 Conclusions 212
  - References 213

## **10 SRS—Access to molecular biological databanks and integrated data analysis tools 215**

*D. P. Kreil and T. Etzold*

- 1 Introduction 215
  - SRS fills a critical need 215
  - History, philosophy, and future of SRS 216

2	A user's primer	217
	A simple query	219
	Exploiting links between databases	220
	Using Views to explore query results	221
	Launching analysis tools	223
	Overview	225
3	Advanced tools and concepts	225
	Refining queries	225
	Creating custom Views	230
	SRS world wide: using DATABANKS	232
	Interfacing with SRS over the network	233
4	SRS server side	236
	User's point of view	236
	Administrator's point of view	238
5	Where to turn to for help	240
	Acknowledgements	241
	References	241

**List of suppliers** 243

**Index** 247