

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Why is Structure Comparison Important? . . . . .	2
1.2	Outline of this Work . . . . .	4
<b>2</b>	<b>Protein Structures</b>	<b>6</b>
2.1	Amino Acid Properties . . . . .	7
2.2	Peptide Bond . . . . .	9
2.3	Protein Folding . . . . .	11
2.4	Secondary Structure Elements . . . . .	13
2.4.1	Helices . . . . .	13
2.4.2	Strands and Sheets . . . . .	15
2.4.3	Identifying Secondary Structure Elements . . . . .	16
2.4.4	Amino Acid Distributions in SSEs . . . . .	18
2.5	Protein Domains and Structural Motifs . . . . .	19
2.6	Protein Structure Evolution . . . . .	20
2.6.1	Divergent Evolution . . . . .	21
2.6.2	Convergent Evolution . . . . .	21
2.7	Protein Structure Representations . . . . .	22
2.7.1	Contact Maps . . . . .	22
2.7.2	Protein Graphs . . . . .	23
2.7.3	Contact Definitions . . . . .	24
<b>3</b>	<b>Protein Topologies</b>	<b>30</b>
3.1	Introduction . . . . .	30
3.2	Protein Topologies as Protein Graphs . . . . .	32
3.3	Linear Notations and Graphical Representation . . . . .	34
3.3.1	The KEY Notation . . . . .	36
3.3.2	Adjacent and Reduced Notation . . . . .	37
3.3.3	The Sequence Notation . . . . .	39
3.3.4	Bifurcated and Non-bifurcated Graphs . . . . .	39
3.4	The PTGL Database . . . . .	41
3.4.1	Database Design . . . . .	41
3.4.2	Statistics . . . . .	41
3.5	Supersecondary Structure Motifs . . . . .	42
3.5.1	$\alpha$ -helical Motifs . . . . .	43
3.5.2	$\beta$ -Sheet Motifs . . . . .	44
3.5.3	$\alpha/\beta$ Motifs . . . . .	47
3.5.4	$\alpha+\beta$ Motifs . . . . .	48

3.5.5	Comparison PTGL versus TOPS . . . . .	48
3.6	Discussion . . . . .	49
<b>4</b>	<b>Structure Comparison and Classification</b>	<b>51</b>
4.1	The Structure Alignment Problem . . . . .	51
4.2	Similarity Measures . . . . .	55
4.2.1	Geometry-based Measures . . . . .	56
4.2.2	Contact Map Overlap . . . . .	57
4.2.3	Other Measures . . . . .	58
4.3	Statistical Significance . . . . .	58
4.3.1	Geometry-based Random Models . . . . .	59
4.3.2	Use of Databases . . . . .	59
4.4	Protein Structure Classification . . . . .	59
4.4.1	SCOP . . . . .	60
4.4.2	CATH . . . . .	61
4.4.3	FSSP/DALI/DDD . . . . .	61
4.5	Protein structures with Non-trivial Relationships . . . . .	62
<b>5</b>	<b>Non-sequential Structure Alignment</b>	<b>64</b>
5.1	Introduction . . . . .	64
5.2	The GANGSTA Method . . . . .	66
5.2.1	Protein Graph Representation . . . . .	66
5.2.2	Structure Alignment on SSE Level . . . . .	68
5.2.3	The Genetic Algorithm . . . . .	69
5.2.4	Structure Alignment on Residue Level . . . . .	75
5.2.5	The GANGSTA Score . . . . .	76
5.2.6	Statistical Significance . . . . .	77
5.2.7	Database Search . . . . .	78
5.3	Results . . . . .	79
5.3.1	Implementation . . . . .	79
5.3.2	Example for a Non-sequential Alignment . . . . .	80
5.3.3	Statistical Significance . . . . .	82
5.3.4	Comparison with other Methods . . . . .	84
5.3.5	Non-sequential Structure Alignments . . . . .	86
5.3.6	Different Contact Definitions . . . . .	92
5.4	Discussion . . . . .	93
<b>6</b>	<b>Exact Protein Graph Alignment</b>	<b>96</b>
6.1	Introduction . . . . .	96
6.2	Graph-theoretical Methods . . . . .	98
6.2.1	The Maximal Common Subgraph Problem . . . . .	98
6.2.2	Transformation of the MCS Problem . . . . .	101
6.2.3	The Clique Problem . . . . .	103
6.3	The ExactGANGSTA Method . . . . .	108
6.4	Results . . . . .	110
6.4.1	Protein Graph Properties . . . . .	111
6.4.2	ExactGANGSTA versus GA . . . . .	112
6.5	Discussion . . . . .	117
<b>7</b>	<b>Conclusion and Future Work</b>	<b>118</b>

<b>Appendix</b>	<b>123</b>
<b>A Naming Convention</b>	<b>123</b>
<b>B Statistical Potentials</b>	<b>124</b>
<b>C Graph-theoretical Definitions</b>	<b>125</b>
<b>D Datasets</b>	<b>126</b>
D.1 ASTRAL SCOP40 Dataset . . . . .	126
D.2 Four-Helix-Bundle Dataset . . . . .	126
D.3 TRAF Dataset . . . . .	127
D.4 C2 Dataset . . . . .	127
D.5 Rossmann-Fold Dataset . . . . .	127
D.6 CP Dataset . . . . .	128
D.7 DIFFAL Dataset . . . . .	128
D.8 Novotny Dataset . . . . .	128
D.9 Fischer Dataset . . . . .	129
<b>E List of Structural Alignment Methods</b>	<b>132</b>
<b>F Additional Figures</b>	<b>136</b>
F.1 Protein Structure . . . . .	136
<b>G Additional Tables</b>	<b>143</b>
<b>References</b>	<b>146</b>