

https://dspace.org/

A conformational analysis of signal peptides

Chantson, Tracy, Elizabeth

1998

http://hdl.handle.net/10539/22401

Chantson, Tracy, Elizabeth (1998) A conformational analysis of signal peptides, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/22402>

Downloaded from a DSpace repository. https://dspace.org/

A CONFORMATIONAL ANALYSIS OF SIGNAL PEPTIDES

Tracy Elizabeth Chantson

A thesis submitted to the

Faculty of Science

University of the Witwatersrand

in fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry

University of the Witwatersrand

Johannesburg

South Africa

March, 1998

DECLARATION

I hereby declare that this thesis represents my own work. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. This thesis has not been submitted before for any other degree or examination at any other university.

J. E. Chantson

T. E. Chantson

13th day of March, 1998

ACKNOWLEDGEMENTS

The following contributions to this thesis are gratefully acknowledged:

Professor L. Glasser who supervised this project. I am most grateful for his invaluable support, advice and guidance throughout this work.

Mr P. Stephens for assistance regarding the use of the genetic algorithm and for guidance during optimisation of the genetic algorithm.

Dr D. Ripoll of The Theory Center, Cornell University, Ithaca, NY, for supplying the source code for ECEPPAK and for extensive help with information thereof.

Dr R. Veale of the Zoology Department, University of the Witwatersrand, for generous advice regarding signal peptide function.

Mr N. De Villiers and The Vector Group (Vector Network Computers (PTY) Ltd), for the generous donation of processor time and disk space on their Cray computer, and for assistance with the use of the system.

The *Foundation for Research and Development*, the *University of the Witwatersrand*, the *German Academic Exchange*, and *Professor L. Glasser* for financial support during the realisation of this study.

ABSTRACT

Conformational analysis of portions of functionally-active and functionally-inactive signal peptides (incorporating the wild-type and mutants thereof) has been performed using a variety of computational prediction techniques based on both statistics and molecular mechanics. Molecular mechanics conformational studies are generally plagued by the problem of combinatorial explosion; this problem was addressed with a systematic searching procedure as well as a recently developed genetic algorithm, both utilising the ECEPP/3 force field. The genetic algorithm, in combination with a gradient minimiser, proved to be successful in finding low-energy conformations for each peptide sequence studied. Analysis was performed in both simulated hydrophobic and hydrophilic environments, under distance-constraints.

The molecular mechanics results and statistical predictions generated from the study were compared with existing experimental observations. The reliability of statistical predictions proved to be dependent on prediction method; the more consistent predictions were produced by methods based on membrane proteins, as opposed to those based on globular proteins. The physical property of hydrophobicity of signal peptide sequences, explored in these statistical predictions, was determined to be an important factor in relating sequence to functional activity. Molecular mechanics calculations produced either interrupted or non-interrupted α-helical secondary structures both for functionally-efficient and for functionally-inefficient signal peptides, indicating that α-helix formation alone cannot be correlated with protein export competence. It was concluded from our overall results that both α-helicity and hydrophobicity are required for the efficient functioning of signal peptides.

CONTENTS

	Page
Declaration	ii
Acknowledgements	iii
Abstract	iv
Contents	v
List of Abbreviations	ix
List of Tables	xii
List of Figures	xvi
CHAPTER 1 INTRODUCTION	1
1.1. Protein conformational prediction	1
1.1.1. The protein folding problem	1
1.1.2. Knowledge-based prediction	3
1.1.2.1. Prediction of tertiary structure	*3*
1.1.2.2. Prediction of secondary structure	*4*
1.1.3. Molecular modelling prediction	6
1.1.3.1. Ab initio and semi-empirical	*6*
1.1.3.2. Molecular mechanics	*7*
1.1.3.3. Conformational searching	*8*
1.1.3.4. Solvent modelling	*18*
1.1.3.5. Vibrational entropy	*19*
1.2. Signal peptides	**21**
1.2.1. Function and structure	21
1.2.2. Concepts in protein secretion	22
1.2.2.1 Prokaryotic and eukaryotic secretion	*22*
1.2.2.2. Cotranslational and posttranslational transfer	*23*
1.2.2.3. Molecular chaperones	*23*

1.2.3. Protein secretion mechanisms 24
1.2.3.1. The signal hypothesis 25
1.2.3.2. The membrane trigger hypothesis 28
1.2.3.3. Other hypotheses 29
1.2.4. Conformational analysis 30
1.2.4.1. Experimental 31
1.2.4.2. Knowledge-based modelling 33
1.2.4.3. Molecular modelling 34
1.3. Membrane proteins **36**
1.3.1. Conformational analysis 36
1.3.1.1. Knowledge-based modelling 36
1.3.1.2. Molecular modelling 38
1.4. Objectives of this study **39**

CHAPTER 2 PROGRAM METHODS **40**
2.1. Knowledge-based modelling **40**
2.1.1. Globular protein-based predictions 40
2.1.2. Membrane protein-based predictions 41
2.2. Molecular modelling **42**
2.2.1. The ECEPPGA program 42
2.2.1.1. ECEPP/3 42
2.2.1.2. The genetic algorithm 43
2.2.1.3. Interface with ECEPP/3 44
2.2.1.4. Optimisation of the genetic algorithm 45
2.2.2. Distance-constraints modelling 46

CHAPTER 3 EXPERIMENTAL METHODS **48**
3.1. Signal peptide systems studied **48**
3.1.1. LamB 48
3.1.2. CPY 49
3.1.3. gC 50

3.2. Knowledge-based modelling 51

3.3. Molecular modelling 52

3.3.1. Systematic conformational searches 52

3.3.2. ECEPPGA 54

3.3.2.1. Optimisation of the performance of ECEPPGA 57

3.3.2.2. ECEPPGA conformational searches 58

CHAPTER 4 RESULTS AND DISCUSSION 60

4.1 Knowledge-based modelling 60

4.1.1. Globular protein-based predictions 60

4.1.1.1. Chou-Fasman 60

4.1.1.2. Consensus 67

4.1.2. Membrane protein-based predictions 73

4.1.2.1. Hydrophobicity 73

4.1.2.2. PHD methods 80

4.1.2.3. TMpred 82

4.1.2.4. PSA 86

4.2. Molecular modelling 91

4.2.1. Systematic conformational searches 91

4.2.1.1. SYBYL 91

4.2.1.2. ECEPPAK 93

4.2.2. Optimisation of the performance of ECEPPGA 95

4.2.3. Comparison of search methods 99

4.2.4. ECEPPGA conformational searches 101

4.2.4.1. LamB 103

4.2.4.2. CPY 109

4.2.4.3. gC 113

4.2.5. ECEPPGA conformational searches with modified modelling conditions 118

4.2.5.1. Restricted conformational space 118

4.2.5.2. Varying sequence length 120

4.2.5.3. Varying dielectric permittivity 122

4.2.5.4. Varying proline configuration 129

4.2.6. Calculation of vibrational entropy 131

CHAPTER 5 CONCLUSIONS 133

5.1. The performance of knowledge-based modelling techniques 133

5.2. The performance of molecular modelling techniques 134

5.3. Signal peptide conformation during protein secretion 135

5.4. Suggestions for future work 136

REFERENCES 137

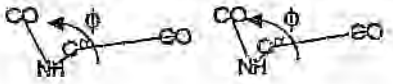
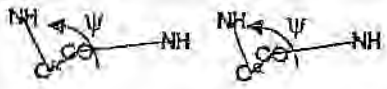
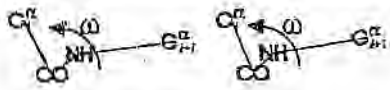
APPENDIX A EXAMPLES OF ECEPPGA INPUT FILES 146

A.1. Example of an ECEPPGA main input file 146

A.2. Example of an ECEPP/3 main input file 155

A.3. Example of a distance-constraints input file 156

LIST OF ABBREVIATIONS

ϕ	-	Phi torsional angle as depicted in the following backbone stereogram: 
ψ	-	Psi torsional angle as depicted in the following backbone stereogram: 
ω	-	Omega torsional angle as depicted in the following backbone stereogram: 
χ	-	Amino acid side chain chi torsional angle
α-helix	-	Characteristic helical peptide secondary structural element with 3.6 residues per turn of the helix and defining angles of $\phi = -57°$, $\psi = -47°$
β-sheet	-	Characteristic sheet-like peptide secondary structural element with defining angles of $\phi = -139°$, $\psi = 135°$ for the anti-parallel, and $\phi = -119°$, $\psi = 113°$ for the parallel
ε	-	Dielectric constant or relative permittivity (a measure of the polarity of some medium)
3-D	-	Three dimensional
AMI	-	Austin Model (a semi-empirical quantum mechanical method to model molecules)
AMBER	-	Assisted Model Building using Energy Refinement
C-F	-	Chou-Fasman (a statistical prediction tool)
C-terminus	-	Carboxy terminus
CD	-	Circular dichroism (a spectroscopic technique)
CPY	-	*Saccharomyces cerevisiae* carboxypeptidase Y protein
DEM	-	Diffusion Equation Method (a global optimisation methodology)
DPM	-	Double Prediction Method (a statistical prediction tool)
$E(\phi,\psi)$	-	Ramachandran map (a description of the potential energy surface for an amino acid)
ECEPP	-	Empirical Conformational Energy Program for Peptides

ECEPPAK	-	A molecular modelling suite of programs incorporating ECEPP/3
ECEPPGA	-	The modified ECEPP containing the genetic algorithm extensions
EDMC	-	Electrostatically Driven Monte Carlo (a global optimisation methodology)
EFF	-	Empirical force field
ER	-	Endoplasmic reticulum
GMEC	-	Global minimum energy conformation
GA	-	Genetic Algorithm (a global optimisation methodology)
gC	-	Swine herpesvirus glycoprotein C protein
GOR	-	Garnier-Osguthorpe-Robson (a statistical prediction tool)
GROMOS	-	Groningen Molecular Simulation
h-core	-	Hydrophobic core (h-region of signal peptide)
LamB	-	*Escherichia coli* λ phage receptor protein
MC	-	Monte Carlo (a global optimisation methodology)
MCM	-	Monte Carlo plus Minimisation (a global optimisation methodology)
MD	-	Molecular Dynamics (a global optimisation methodology)
MM	-	Molecular Mechanics (an empirical method to model molecules)
MNDO	-	Modified Neglect of Differential Overlap (a semi-empirical quantum mechanical method to model molecules)
MS	-	Microsoft
N-terminus	-	Amino terminus
NMR	-	Nuclear magnetic resonance (a spectroscopic technique)
NN	-	Neural Network (a statistical prediction tool)
PHD	-	Profile network from HeiDelberg (a statistical prediction tool)
PDB	-	Brookhaven Protein Data Bank
PM3	-	Parametrized Model (a semi-empirical quantum mechanical method to model molecules)
PRISM	-	Pattern-Recognition Importance Sampling Minimisation (a global optimisation methodology)
PSA	-	Protein Sequence Analysis (a statistical prediction tool)
SA	-	Simulated Annealing (a global optimisation methodology)
SOPMA	-	Self-Optimised Prediction Method from Alignment (a statistical prediction tool)
SCEF	-	Self-Consistent Electrostatic Field (a global optimisation methodology)
SP	-	Signal peptide

SYBYL - The Sybyl molecular modelling suite of programs by Tripos Associates

WT - Wild-type (the naturally occurring signal peptide)

ABBREVIATIONS OF AMINO ACIDS

met (M) - methionine

ile (I) - isoleucine

thr (T) - threonine

leu (L) - leucine

arg (R) - arginine

lys (K) - lysine

pro (P) - proline

ala (A) - alanine

val (V) - valine

gly (G) - glycine

ser (S) - serine

gln (Q) - glutamine

cys (C) - cysteine

asp (D) - aspartate

phe (F) - phenylalanine

tyr (Y) - tyrosine

LIST OF FIGURES

		Page
Figure 1.1:	Hierarchical levels of structural organisation in globular proteins	2
Figure 1.2:	A general Ramachandran plot depicting commonly observed (ϕ,ψ) regions for secondary structural conformational states	10
Figure 1.3:	A schematic representation of the crossover operator	16
Figure 1.4:	A schematic representation of a potential energy surface with two minima	19
Figure 1.5:	The basic design of signal peptides	21
Figure 1.6:	The aqueous channel is gated	26
Figure 1.7:	Hypothesised arrangement of proteins and lipids in the protein-conducting channel	27
Figure 1.8:	Model for initial interaction of a signal sequence with a membrane	32
Figure 2.1:	Flow diagram of the interface between the GA and ECEPP/3	45
Figure 3.1:	Aligned amino acid sequences of the LamB system signal peptides	49
Figure 3.2:	Aligned amino acid sequences of the CPY system signal peptides	50
Figure 3.3:	Aligned amino acid sequences of the gC system signal peptides	51
Figure 4.1:	Probable secondary structure of the sequences of the LamB signal peptide system determined with the Chou-Fasman prediction method	61
Figure 4.2:	Probable secondary structure of the sequences of the CPY signal peptide system determined with the Chou-Fasman prediction method	64
Figure 4.3:	Probable secondary structure of the sequences of the gC signal peptide system determined with the Chou-Fasman prediction method	66
Figure 4.4:	Hydrophobicity plots of the signal sequences of the LamB signal peptide system determined from the normalised hydropathy scales of (a) Hopp and Woods, (b) Rose *et al.* and (c) Sweet and Eisenberg	75
Figure 4.5:	Hydrophobicity plots of the signal sequences of the CPY signal peptide system determined from the normalised hydropathy scales of (a) Hopp and Woods, (b) Rose *et al.* and (c) Sweet and Eisenberg	77

Figure 4.6: Hydrophobicity plots of the signal sequences of the gC signal peptide system determined from the normalised hydropathy scales of (a) Hopp and Woods, (b) Rose *et al.* and (c) Sweet and Eisenberg 79

Figure 7: TMpred prediction of transmembrane helical regions of the sequences of the LamB signal peptide system 83

Figure 4.8: TMpred prediction of transmembrane helical regions of the sequences of the CPY signal peptide system 84

Figure 4.9: TMpred prediction of helical transmembrane regions of the sequences of the gC signal peptide system 85

Figure 4.10: Probable secondary structure of the sequences of the LamB signal peptide system determined with the PSA prediction method 87

Figure 4.11: Probable secondary structure of the sequences of the CPY signal peptide system determined with the PSA prediction method 88

Figure 4.12: Probable secondary structure of the sequences of the gC signal peptide system determined with the PSA prediction method 89

Figure 4.13: Comparison of local minima energy wells of the LamB Δ78 signal sequence constrained in the α-helical and β-sheet secondary structures during systematic searching with SYBYL 91

Figure 4.14. Ramachandran plots of the lowest energy-minimised structures for the (a) Δ78r2 and (b) Δ78 signal sequences of the LamB system determined with SYBYL systematic searches 92

Figure 4.15: Minimised energy *versus* distance-constraint curves for the (a) Δ78r2 and (b) Δ78 signal sequences of the LamB system determined with ECEPPAK systematic searches 94

Figure 4.16: Ramachandran plots of the lowest energy-minimised structures for the (a) Δ78r2 and (b) Δ78 signal sequences of the LamB system determined with ECEPPAK systematic searches 95

Figure 4.17: Effect of parameter variation on the minimisation achieved at various generations for the Δ78r2 signal sequence of LamB 97

Figure 4.18: Effect of parameter variation on the minimisation achieved at various generations for the Δ78r1 signal sequence of LamB 98

Figure 4.19: Effect of parameter variation on the minimisation achieved at various generations for the Δ78 signal sequence of LamB 98

Figure 4.20: A comparison of minimised energy *versus* distance-constraint curves for the (a) Δ78r2 and (b) Δ78 signal sequences of the LamB system determined with the ECEPPAK systematic search and ECEPPGA 100

Figure 4.21: Stereoviews of minimised energy structures for the WT signal sequence of the LamB system determined with ECEPPGA at various distance-constraint values 102

Figure 4.22: Minimised energy *versus* distance-constraint curves for the signal sequences of the LamB system determined with ECEPPGA with a dielectric constant of 2 104

Figure 4.23: Ramachandran plots of the α-helical regions of energy-minimised structures of the signal sequences of the LamB system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 11.00 Å 106

Figure 4.24: Helical wheel plots of energy-minimised structures of the signal sequences of the LamB system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 11.00 Å 108

Figure 4.25: Minimised energy *versus* distance-constraint curves for the signal sequences of the CPY system determined with ECEPPGA with a dielectric constant of 2 110

Figure 4.26: Ramachandran plots of the α-helical regions of energy-minimised structures of the signal sequences of the CPY system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 13.75 Å for CPYm2, CPYm8 and WT, and 12.25 Å for CPYm6 and CPYm12 111

Figure 4.27: Helical wheel plots of energy-minimised structures of the signal sequences of the CPY system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 13.75 Å for CPYm2, CPYm8 and WT, and 12.25 Å for CPYm6 and CPYm12 112

Figure 4.28: Minimised energy *versus* distance-constraint curves for the signal sequences of the gC system determined with ECEPPGA with a dielectric constant of 2 115

Figure 4.29: Ramachandran plots of the α-helical regions of energy-minimised structures of the signal sequences of the gC system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 14.00 Å for WT, A10P, L12P and L14P, and 12.50 Å for ΔA10 116

Figure 4.30: Helical wheel plots of energy-minimised structures of the signal sequences of the gC system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 14.00 Å for WT, A10P, L12P and L14P, and 12.50 Å for ΔA10 117

Figure 4.31:Minimised energy *versus* distance-constraint curves for the WT and A13D signal sequences of the LamB system determined with ECEPPGA, using varying selected sequence lengths 121

Figure 4.32:Minimised energy *versus* distance-constraint curves for the CPYm2 and WT signal sequences of the CPY system determined with ECEPPGA, using varying selected sequence lengths 122

Figure 4.33:Minimised energy *versus* distance-constraint curves for the signal sequences of the LamB system determined with ECEPPGA with varying dielectric constants 125

Figure 4.34:Minimised energy *versus* distance-constraint curves for the WT and L14P signal sequences of the gC system determined with ECEPPGA with dielectric constants of 2 and 80 127

Figure 4.35:Minimised energy *versus* distance-constraint curves for the signal sequences of the LamB system determined with ECEPPGA with a dielectric constant of 80 128

LIST OF TABLES

		Page
Table 1.1:	Examples of molecular chaperones implicated in protein transport	24
Table 3.1:	Sequences and distance-constraint end points selected for GA modelling	55
Table 4.1:	Summary of the activities and α-helical contents of the LamB signal peptides	63
Table 4.2:	Secondary structure of the sequences of the LamB signal peptide system determined by joint prediction	68
Table 4.3:	Secondary structure of the sequences of the CPY signal peptide system determined by joint prediction	70
Table 4.4:	Secondary structure of the sequences of the CPY signal peptide system determined with the PHDsec prediction method	71
Table 4.5:	Secondary structure of the sequences of the gC signal peptide system determined by joint prediction	72
Table 4.6:	Summary of the translocation efficiencies and hydrophobicities of the signal peptides in the LamB system	76
Table 4.7:	Summary of the mammalian translocation efficiencies and hydrophobicities of the signal peptides in the CPY system	78
Table 4.8:	Summary of the translocation efficiencies and hydrophobicities of the signal peptides in the gC system	80
Table 4.9:	Secondary structure prediction of the sequences of the LamB, CPY and gC systems determined with the PHDhtm prediction method	81
Table 4.10:	Solvent accessibility prediction of the sequences of the LamB, CPY and gC systems determined with the PHDacc prediction method	81
Table 4.11:	Translocation efficiencies and membrane-spanning probabilities of the sequences of the LamB, CPY and gC systems determined with the PSA prediction method	86
Table 4.12:	The variation of relevant ECEPPGA parameters during GA optimisation runs	96

Table 4.13: A comparison of the lowest-energy minimised structures for the LamB signal peptide system determined with ECEPPGA, with and without distance-constraints 103

Table 4.14: A comparison of the lowest-energy minimised structures for the CPY signal peptide system determined with ECEPPGA, with and without distance-constraints 109

Table 4.15: A comparison of the lowest-energy minimised structures for the gC signal peptide system determined with ECEPPGA, with and without distance-constraints 114

Table 4.16: The effect of α-helical restricted conformational space on unconstrained conformational energies of the deletion mutants of the LamB signal peptide system determined with ECEPPGA 119

Table 4.17: The effect of β-strand restricted conformational space on unconstrained conformational energies of the deletion mutants of the LamB signal peptide system determined with ECEPPGA 119

Table 4.18: The varying sequences and distance-constraint end points selected for GA modelling 120

Table 4.19: The effect of dielectric constant on the lowest-energy minimised structures for the LamB signal peptide system determined with ECEPPGA, with and without distance-constraints 124

Table 4.20: The effect of dielectric constant on the lowest-energy minimised structures for the CPYm2 and WT signal sequences of the CPY system determined with ECEPPGA, without distance-constraints 126

Table 4.21: The effect of dielectric constant on the lowest-energy minimised structures for the WT and L14P signal sequences of the gC system determined with ECEPPGA, with and without distance-constraints 126

Table 4.22: The effect of proline conformation on unconstrained conformational energies of the Δ78 signal sequence of the LamB system determined with ECEPPGA 130

Table 4.23: Values calculated for the lowest-energy conformations of the signal sequences of the LamB, CPY and gC systems determined with ECEPPGA 132

CHAPTER 1
INTRODUCTION

1.1. Protein conformational prediction

1.1.1. The protein folding problem

Anfinsen's[1] classic experiment with the enzyme ribonuclease revealed the reversible and spontaneous nature of protein folding. Since then, the myste[illegible]ust how a protein arranges itself into its native 3-D conformation[a] (natural conformation *in vivo*) ...o continued to puzzle scientists. The need to solve the protein folding problem is fuelled by the fact that protein activity is attributed to its folded form. Knowing how proteins fold can help in the determination and prediction of these native structures, which can in turn promote understanding of structure-function relationships in biological processes. The ultimate aim is to contribute fundamental information to the fields of protein engineering and rational drug design.

A protein can adopt more than one folded conformation while traversing its folding pathway, the interconversions between different conformations resulting from vibrations within the macromolecule. The choice of native structure from among these conformations is governed by thermodynamic and kinetic factors. Thermodynamically, the native structure is hypothesised to correspond to the protein's most stable conformation (the global free energy minimum).[3] However, this absolute minimum may not always be kinetically accessible and the native state may instead correspond to a local minimum.[4] Thus, when *predicting* folding patterns theoretically, cognisance must be taken of the fact that the native 3-D conformation is the global minimum only if it meets both thermodynamic and kinetic requirements.

The spontaneous nature of protein folding[1] *in vitro* yields the view that all the information required to direct the folding of a protein is encoded in the amino acid sequence of the protein. Other factors, such as molecular chaperones and catalysts, which come into play *in vivo*, are believed merely to facilitate

[a] The traditional definition of "molecular conformations" is a set of 3-D arrangements of the atoms of a molecule in space whose interconvertibility is due solely to rotation about single bonds.[2]

the folding process.[5] Using this concept, the hierarchical nature of proteins (see Figure 1.1) can be exploited to investigate conformations along the folding route. Protein secondary structure is predicted from primary structure as an intermediate step towards the final goal of specific tertiary structure prediction. This approach to protein folding is called the framework model[7] and has been validated by experimental data. In investigating the *in vitro* folding of the hen lysozyme protein, Dobson *et al.*[5] used a variety of complementary experimental techniques, including nuclear magnetic resonance spectroscopy (NMR), electrospray mass spectrometry (ESMS) and circular dichroism (CD), to obtain a detailed folding model. They discovered that the denatured protein initially collapses and then rapidly forms a stabilised secondary structure before folding proceeds to the native conformation.

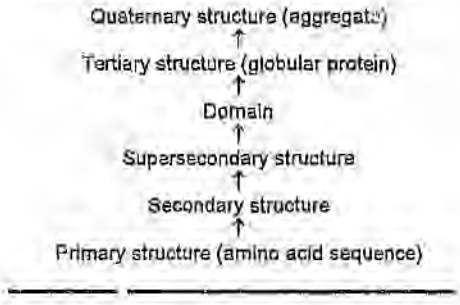


Figure 1.1: Hierarchical levels of structural organisation in globular proteins.[6]
Primary structure: covalent structure defined by the amino acid sequence; secondary structure: local, regular conformations of the polypeptide backbone; supersecondary structure: physically preferred aggregates of secondary structure; domains: parts of the protein which display distinct globular regions; tertiary structure: 3-D folded arrangement of the polypeptide backbone together with the spatial dispositions of its side chains; quaternary structure: structures of aggregates of globular proteins.

The major complication encountered in protein folding simulations is that of combinatorial explosion. This refers to the exponential increase in the number of conformations possibly adopted by a protein as the number of amino acid residues and rotatable bonds in its peptide chain increases. A recent model for protein folding, developed by Wolynes *et al.*,[8,9] deals with this problem in an interesting way. The number of conformations along the folding path is reduced by guiding the protein through a funnel-shaped energy landscape. The protein moves through many possible shapes present in the energy funnel until it reaches its most energetically stable conformation at the bottom of the funnel.

1.1.2. Knowledge-based prediction

The most widely used technique in the attempt to predict protein folding theoretically is molecular taxonomy[10] or knowledge-based modelling.[11] In essence, the technique evaluates structural and evolutionary relationships between proteins by comparing new protein sequences with proteins of known structure, using a combination of statistical theory and empirical rules.

1.1.2.1. Prediction of tertiary structure

The knowledge-based method of homology modelling[11-13] is currently the most successful way of predicting the unknown tertiary structure of a protein from its amino acid sequence alone. Homologous proteins are proteins that are believed to have evolved from a common ancestor and will therefore have substantial similarities in their primary structures and perhaps display similar folding conformations. In the first step of this modelling process, the structurally unknown peptide sequence is aligned with sequences of other proteins of known structure to search for patterns of homology. Several databases of known protein X-ray crystallographic structures, such as the Brookhaven Protein databank (PDB)[14] and SwissProt,[15] are presently available. Sequence alignment algorithms, such as those of Needleman and Wunsch[16] and of Smith and Waterman[17] use dynamic programming methods[b] to carry out a global comparison of two sequences. When more than one sequence of known structure exists within a protein family, multiple sequence alignments can be performed. Once the alignment is completed, the backbone and loops of the new protein are constructed, the conformations of the side chains are determined, and the structure is subjected to refinement using energy minimisation techniques.

An obvious limitation of the homology modelling method is the need for known structures that are related to the structurally unknown sequence; the success of the method depends on the degree of similarity between the sequences. For those that are distantly related, the limitation results in predictions of new protein structures which are, as yet, unreliable.

A protein conformation modelling method that has lately received much interest because of the hope that it may assist with the identification of distantly related sequences is knowledge-based potentials. Knowledge-based potentials are derived either from a statistical analysis of protein structural features and amino acid sequence data (potentials constructed within the framework of

[b] "Dynamic programming" is a mathematical technique for optimisation which avoids complete enumeration of all solutions by solving subproblems in a series of stages.

statistical mechanics),[18] or from an optimisation of potential functions with known structure and sequence data (such as interresidue contacts).[19,20] Reduced representations are often used to describe protein conformation in this context.

1.1.2.2. Prediction of secondary structure

As mentioned earlier, the levels of hierarchy which proteins display can also be employed to predict protein tertiary architecture. Secondary structure elements are first predicted, followed by the assembly of these elements into a compact structure. The elements are the regular conformations of α-helices, β-strands and reverse turns, which appear to be adopted by 90% of the amino acid residues in most proteins.[21]

Representative examples of secondary structure prediction methods are those of Chou and Fasman,[22] Garnier *et al.*[23] and Lim.[24] The first two schemes are based on the principle that different amino acid residues demonstrate different conformational preferences. The probabilistic Chou-Fasman[22] scheme is the most popular because of its simplicity and ease of application. For each of the 20 naturally occurring amino acids, it assigns conformational propensity values for the formation of α-helices, β-strands and coils (β-turns are included in the coil regions).[25] These values were originally obtained from a statistical analysis of a limited set of protein X-ray crystal structures. Prediction of secondary structure for a particular peptide sequence is then accomplished by a simple procedure which is dependent upon the successive occurrence of large propensity values in the sequence. The GOR (Garnier-Osguthorpe-Robson)[23] method is also derived from a statistical study of known tertiary structures. Directional and positional information from the study was used to compute not only conformational propensities for single residues, but also to explicitly include the effects of short and medium range interactions in the calculations.

Although the Chou-Fasman and GOR schemes, among others, rely on amino acid sequence (groups of amino acids) to conduct their predictions, secondary structure prediction can also be achieved by taking only amino acid composition into account.[26,27] The predictive algorithm of Eisenhaber and co-workers[27] combines vector decomposition techniques with the amino acid compositions of secondary structural types present in known protein tertiary structures to calculate secondary structural content. The algorithm is also able to assign secondary structural class[28] to new sequences. The concept of secondary structural class originated with an intuitive

classification of proteins into certain structural classes or folding types (all-α, all-β, α+β, α/β).[29]

Lim's[24] prediction method was developed from his *a priori* theory of secondary structure of globular proteins. Stereochemical criteria such as the compact packing of polypeptide chains and the presence of hydrophobic cores and polar shells in water-soluble globular proteins form the basis of the theory. Rules and formulae were devised to search for spans in primary sequences which display secondary structure characteristics. In his approach to the modelling problem, Lim took two important considerations into account: the hydrophobic nature of amino acids, and the dependence of secondary structure determination not only on local amino acid sequence, but also on tertiary structure.

The hydrophobic nature of amino acids plays a significant role in the formation of α-helices and β-sheets. It is thus a factor which has been well examined in the attempt to predict structure. Hydrophobicity scales,[30] as well as amino acid solvent accessibilities (determined from multiple sequence alignment information),[31,32] are used to locate regions of a peptide chain which are either buried in the interior or exposed on the surface of a folded protein. As a general rule, hydrophobic residues tend to occur in the interior, while hydrophilic residues tend to lie on the surface.[6]

One of the reasons for the generally low prediction accuracies (50 to 65%) achieved with secondary structure prediction methods is the fact that long-range interactions are not considered in the calculations. Kabsch and Sander[33] established that the local conformation of a protein is determined during the folding process and is thus dependent on tertiary structure. Although local interactions (interactions between the side groups of the amino acids and the backbone), in particular hydrogen bonding, may define secondary structure, it would appear that long-range interactions are necessary to ensure stability of the structure. Hydrogen bonding, van der Waals interactions, electrostatic interactions, disulphide bridge formation, packing density caused by the hydrophobic effect, and amino acid intrinsic propensities for secondary structure are factors that play a role in stabilising secondary structure elements in native globular proteins.[34,35]

A recent knowledge-based technique that appears to be particularly successful in terms of prediction accuracy (>70%) is that formulated by Rost and Sander.[36] Their method employs a

neural network (NN)[c] algorithm and protein evolutionary information in the form of multiple sequence alignments. Here, the input to the net is information derived from an alignment and the output is secondary structure type. This output then serves as the input for the next network, the process being repeated through the system of networks.[37]

1.1.3. Molecular modelling prediction

An alternate approach to knowledge-based modelling is molecular modelling. The latter uses first principles and energy calculations to analyse protein conformation directly, and is useful in helping to glean a physical understanding of how intra- and intermolecular interactions determine 3-D structure.

Molecular modelling calculations are based on the assumption that the most stable conformation of a molecule, the thermodynamic native state in the case of proteins, corresponds to the global minimum of the Gibbs function (free energy) of the system (molecule and solvent). The Gibbs function, G, comprises the potential energy of all intramolecular interactions, U, the free energy of all solvent interactions, V, including the free energy of solvation, and a vibrational entropy free energy term, S.[38]

$$G = U + V + S \qquad (1)$$

In this study, the term molecular modelling refers to a collective description of the various computational techniques and approximations used to determine the energy of a molecular system. These fall into one of three categories: *ab initio*, semi-empirical and empirical.

1.1.3.1 Ab initio and semi-empirical

In the quantum mechanical *ab initio*[39] and semi-empirical[40] approaches, molecular orbital theory is most commonly implemented to provide a description for the molecule. *Ab initio* employs non-parameterised treatment for small molecules and an approximation in the form of basis sets which allow for numerical solutions. Gaussian[41] appears to be the most widely used *ab initio* program at present. Semi-empirical methods are less time-consuming than *ab initio* and can be applied to medium-sized molecules. They introduce additional simplifying assumptions and approximations in their methodology. Examples of programs which incorporate this approach are MNDO,[42] AM1[43] and PM3.[44]

[c] Neural networks are designed to operate in a manner corresponding to that of neurons in the brain.

1.1.3.2. Molecular mechanics

For larger molecular structures such as proteins, computational limitations warrant the use of empirical methods, *i.e.*, methods based on both theoretical and experimental data, to calculate conformational energy. The molecular mechanics (MM) procedure[45] uses an empirical force field (EFF) to provide us with a description for the potential energy hypersurface of a molecule as a function of atomic positions. It considers molecules to be collections of atoms which interact with each other *via* classical forces. The potential energy of these interactions is described in terms of an analytical function corresponding to the summation of energy terms due to bond stretching, angle bending, torsional angle strain, nonbonded van der Waals interactions and coulomb electrostatic interactions (equation 2).

$$E_{total} = E_{bond} + E_{angle} + E_{torsion} + E_{nonbond} + E_{coulomb} \qquad (2)$$

Additional terms such as out-of-plane bending and hydrogen bonding potentials may be added. Each energy term has a preferred equilibrium position of a generalised co-ordinate (bond length, bond angle, dihedral angle, van der Waals interaction distance, *etc.*) and, coupled with it, a force constant which associates an energetic penalty with any deviation from this equilibrium value.[46] Examples of some extensively applied force fields include MM2[47] and MM3[48] for small organic molecules, ECEPP[49-51] for peptides, and AMBER[52] and CHARMM[53] for peptides and nucleic acids.

To obtain the geometry of a molecular structure at the energy minimum, the potential energy function for the molecule must be optimised. This is achieved with an iterative MM energy minimisation procedure, beginning with a elected initial molecular conformation. Optimisation algorithms can be classified into three groups:[54] (a) simple search procedures which do not involve the evaluation of derivatives of the potential energy function, *e.g.*, Rosenbrock,[55] (b) procedures which involve the evaluation of first derivatives, *e.g.*, steepest descent[56] and conjugate gradient,[57] and (c) procedures which involve the evaluation of both first and second derivatives, *e.g.*, Newton-Raphson[58] and variants thereof.

1.1.3.3. Conformational searching

One of the major problems confronting the explicit MM approach to macromolecular structure prediction is that of multiple minima. The potential energy hypersurface of a molecule comprises many local minima which may be similar in energy, and which complicate the search for the global minimum. Since conventional energy minimisation procedures merely place a molecule into the minimum corresponding to the energy well in which it is placed in the energy landscape, other strategies are necessary to surmount the local "barriers" and so reach the global minimum energy conformation (GMEC). These strategies involve searching the conformational space of the molecule to locate the energy well possibly containing the GMEC, followed by minimisation of the structures whose energies lie within that well.[59]

An exhaustive scanning of conformational space is possible for small molecules which possess relatively few rotatable bonds. For larger molecules, however, the large number of torsional degrees of freedom and subsequent exponential increase in the number of available conformational states (the combinatorial explosion problem) makes exhaustive searches computationally unmanageable. Ways to reduce the scope of a search, such as the imposition of constraints on the conformations generated, and the biasing of the search towards regions of low energy have therefore been incorporated into many scanning algorithms.[2] Another important aspect is the selection of the initial set of points on the energy surface to be explored. Information obtained from knowledge-based studies, *e.g.*, amino acid secondary structure propensities,[22,23,60] residue conformational pattern clustering,[61] and conformational filtering of small peptides,[62] is regularly used to assist the selection.

In view of the abundance of material available in the literature pertaining to conformational sampling techniques, only a brief discussion of some of the search methodologies will be given here. For recent reviews on the subject, the papers by Vásquez *et al.*,[59] Eisenhaber *et al.*,[13] Piela *et al.*[63] and Scheraga[64] are recommended.

Polypeptide conformational sampling techniques appear to be broadly classified as being either deterministic (successive conformations are determined by preceding conformations) or stochastic (conformations are generated in a random fashion).[65] They can be further organised into the following categories:[2]

- ***Systematic searching***

A complete enumeration of all possible conformations in a systematic manner is the most direct way of exploring conformational space. This method, also called grid searching, normally involves the variation in turn of the torsion angles of a protein sequence along a specified grid while the remaining internal co-ordinates (bond lengths, bond angles) are fixed. Since all possible combinations of allowed torsion values are examined, combinatorial explosion soon leads to an overwhelming number of generated structures for a sequence of medium length (> 10 residues). If we assume that there are n rotatable bonds in a sequence and that the angular increment on the grid is θ, the number of conformations that would be produced equals $(360/\theta)^n$. This is the major disadvantage of the technique.

The efficiency of systematic searches can be improved by explicitly decreasing the number of conformations to be searched. One way of achieving this is to perform separate optimisations and constructions of the peptide's backbone and side chain dihedral angles.[66,67] Optimisation of the one set of dihedrals is performed while the angles of the other set are preserved. In addition, instead of predicting the structure of the entire protein, prediction of only short segments (loops) of the protein can be performed. Another way of eliminating unnecessary conformations is use of the dead-end theorem,[68] an algorithm that discards any side chain rotamers that are, in principle, incompatible with the GMEC.

Reduction of the conformational space in a systematic search may also be achieved by considering the intrinsic propensity of amino acids to occur in favoured regions of the Ramachandran map, $E(\phi,\psi)$.[a] Energy minima and locations of the 20 naturally occurring amino acid residues have been characterised *via* MM calculations.[69-71] Results indicate the existence of large regions of (ϕ,ψ) space that can be regarded as energetically forbidden. Each amino acid can therefore be assigned an ensemble of conformational states, these states represent defined regions of the Ramachandran plot as shown in Figure 1.2.

[a] The conformation of a polypeptide chain can be described by specifying all the dihedral angles of each residue: ϕ, ψ and ω for the backbone, and all χ's for the side chain. $\omega = 180°$ is normally assumed, except for proline where $\omega = 0°$ is also permitted.

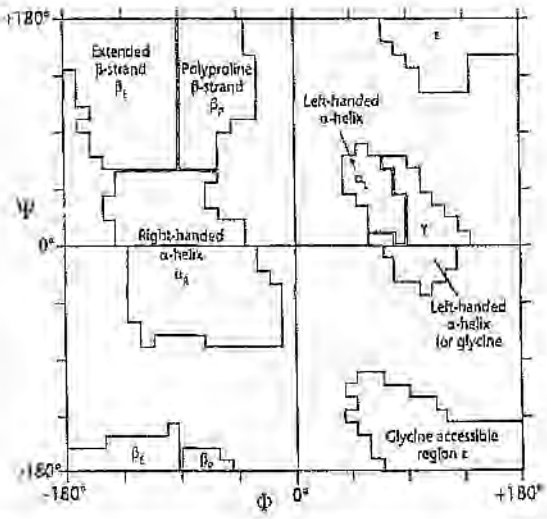


Figure 1.2: A general Ramachandran plot depicting commonly observed (ϕ,ψ) regions for secondary structural conformational states[37]

- ***Model building***

As mentioned above, individual amino acids have probabilities of occurring in more than one conformation.[69-71] Knowledge of their low-energy conformers led to the development of build-up[72] or fragment-based approaches to peptide conformational searching. These low-energy structures serve as the initial building blocks for the formation of larger low-energy polypeptide structures and protein fragments. The fragments are combined to ultimately form the entire molecule. During each stage of the construction process, the intermediate structures are energy minimised and an ensemble of their lower energy conformers retained for use in the following stage.

The enormous number of conformations that require minimisation and storage at each step is one dilemma encountered in the build-up approach. Introduction of an energy cut-off, the retention of only those minima with significantly different backbone conformations, and the utilisation of geometry constraints are among the proposed solutions to the problem.[2] Another restriction relates to the fundamental assumption that short-range interactions between neighbouring residues play a more dominant role than long-range forces in determining conformation. The application of build-up procedures is thus limited to relatively small unconstrained oligopeptides where long-range interactions are overshadowed by local ones.[59]

- *Symbolic representations*

Peptide conformations can be represented by symbolic descriptors in an effort to reduce conformational space. Symbolic structure representations, also referred to as reduced parameter representations, lead to simplified protein models[73] which are able to facilitate faster sampling of conformations due to a decrease in the number of degrees of freedom.

Several conformational searching techniques have been extended to incorporate simplified representations. Koca *et al.*[74] have developed a program which performs a restricted grid search on a set of elementary fragments of a peptide. The fragments are subsets of the dihedral angles of the peptide. The PRISM[75] and "representative"[76] methods are two examples of model-building procedures that incorporate reduced representations. They use variations of a finite-state model[a] which depicts conformational states as points in the (ϕ,ψ) plane. PRISM also uses statistical information derived from known protein structures in a pattern-recognition technique which predicts tripeptide probabilities.

Lattice models also represent a symbolic description of polypeptide chains.[13,59,73] They depict the chains as pseudoatom positions at discrete points on a multidimensional lattice. The confinement of a peptide in this manner drastically reduces conformational space and, for medium-sized chains, allows for complete enumeration of the space. Computational time is further decreased due to the use of simplified potential functions for interresidue contact. Interactions between such residues are often described with a two-state square model, corresponding to native-non-native contacts, or with a three-state cubic model, corresponding to hydrophobic-hydrophobic, hydrophilic-hydrophilic and hydrophobic-hydrophilic contacts.[73]

These simplified forms of interaction potentials, together with Monte Carlo and molecular dynamics simulations (neural networks is a recent addition[78]), have made lattice calculations a popular choice in the field of protein modelling. Many attempts have been made to delineate the qualitative features of protein conformation[79,80] and folding,[81-85] *e.g.*, folding cooperativity, compactness and the balance of local and nonlocal forces. Current studies with lattice models focus either on the prediction of secondary and tertiary structure from sequence,[86,87] or on the thermodynamics and kinetics of protein folding.[88,89]

[a] A "finite state model" is a representation of protein backbone conformations using a finite number of values for the backbone dihedral angles.[77]

- *Random searching*

In random or stochastic sampling techniques, structures are randomly changed to generate new structures, some of which are selected for the next iteration in the process. In the case of proteins, these changes are perturbations of the torsional angles of the rotatable bonds. The cycle of random change and selection continues until sufficient sampling has been performed (the predefined number of iterations is exceeded or no new structures are produced). The selection criteria for the starting structures of each iteration are usually energy-based. Stochastic approaches frequently involve faster search speeds than deterministic approaches, and thus provide a way of finding low-energy minima on the conformational surface without having to execute an exhaustive search.[2,90]

The Monte Carlo (MC) stochastic method explores protein conformational space by simulating the conformational motion of proteins. Selection of trial conformations is conducted with the Metropolis algorithm[91] which ensures that the appearance of new conformations is proportional to their Boltzmann factors $exp(-\Delta E/kT)$, where ΔE is the energy difference between the old and the new conformations being compared, k is the Boltzmann constant and T is the simulation temperature. To prevent MC methods from being trapped in local energy minima (a phenomenon caused by increasingly low Boltzmann acceptance ratios for new conformations), small steps along the conformational path are forced to be taken. Thus, although searching of the entire conformational space is possible, the process is extremely slow.[63]

The modified MCM (Monte Carlo plus minimisation) procedure is an attempt by Li and Scheraga[92] to circumvent the inefficiency of pure Metropolis MC methods. It enables the use of very large steps and samples only a discrete space of local energy minima. This is achieved by energy minimising each conformation after its random alteration and then subjecting it to Metropolis selection. The EDMC (electrostatically driven Monte Carlo)[93] method comprises the random generation of new conformations based on information obtained from the SCEF (self-consistent electrostatic field)[94] method, followed by minimisation and application of the Metropolis test. The SCEF procedure assumes that each amino acid residue has an optimal electrostatic energy, and uses the orientation of their dipole moments to bias dihedral angle jumps. An alternative way to bias dihedral jumps is included in the MC procedure of Abagyan and Totrov.[95] Knowledge derived from statistical analyses of residues in terms of their

conformational preferences is utilised to formulate probability distribution functions. These functions are then used in the torsion randomisation step of the search procedure.

Simulated annealing (SA), introduced by Kirkpatrick *et al.*,[96] is a stochastic procedure for conformational sampling that is based on the thermodynamic process of annealing. This is a technique by which a many-body system, initially at high temperature, is brought to a low-energy state by slowly lowering the temperature of the system; SA thus simulates the thermal motion of proteins in conformational space. It is usually implemented in conjunction with the Metropolis MC algorithm. At each temperature in the cooling process, step-wise random changes are made to the conformations. Selection of the lower-energy conformers is then made. The main drawback of this method is the fact that it depends heavily on the carefully chosen parameters of temperature interval size (determining the cooling schedule) and step size (controlling the size of each step).[59]

Genetic algorithms (GAs) form another class of probabilistic optimisation methods. They differ from the class of MC methods in that they produce an arbitrary collection of conformations.[97] Since the searching procedure employed in this study is a GA, a detailed discussion of the method is warranted and is given towards the end of this section.

- ***Distance geometry and related methods***

The maintenance of distance constraints between atoms in a molecule forms the basis of distance geometry methods. Examples of these constraints include the equilibrium bond length separation between covalently bonded atoms, the distance between nonbonded atoms (≥ sum of their van der Waals radii) and supplied upper and lower bounds on interatomic distances.[98] The distance geometry algorithm does not contain any energy criterion; distances instead of potential energies are the primary variables and conformations are represented by matrices of interatomic distances. Consequently, its "best" structure, *i.e.*, the one that best satisfies the distance constraints, may be relatively high in energy. Further optimisation of structures derived in this manner is thus inevitable.[2]

Crippen[99] was the first to apply fundamental distance geometry concepts to the problem of multiple minima. He addressed the issue of structure optimisation with an energy embedding method.[100] The theory behind the method is that atoms in a molecule have more freedom at higher dimensions, thereby decreasing the number of potential barriers and easing the relaxation

of the molecule into low-energy conformations. This results in fewer minima on the potential energy surface, *viz.*, a smoothing of the surface. The method entails initial energy minimisation of conformations in a higher than 3-D space, followed by contraction of the space and projection of a resultant single minimum onto the 3-D "plane". Purisima and Scheraga's[101] relaxation of dimensionality technique is an extension of the idea of energy embedding. The two methods differ in the way they use interatomic distances in their energy functions and in the procedures used for dimensionality relaxation.

Methods of conformational sampling which are related to the methods described above are target function minimisation (derives 3-D structures consistent with distance constraints),[102] the antlion method (deformation of energy surface by alteration of force field components)[103] and the diffusion equation method or DEM (smoothing of potential energy surface by mathematical transformation of the energy function).[104] DEM makes use of a diffusion (or heat conduction) equation to deform the complex potential hypersurface of a polypeptide. The deformation is performed in successive stages and ideally leads to the disappearance of higher-energy minima from the surface, leaving behind a single minimum. This minimum is hopefully related to the global one, which can be attained by gradual back-propagation of the smoothing process. The GMEC of the original potential function is thus, in principle, recovered.

- ***Molecular dynamics***

Dynamic search methods simulate nature by modelling the actual, instantaneous motion of a molecular system. The trajectory of conformational fluctuations experienced by a molecule is traced by integrating Newton's (or Lagrange's) equations of motion over time for all of its atoms. New atom positions and velocities are computed at every step in time. Molecular dynamics (MD) is a deterministic process in that it uses previous atom positions to compute the new positions. In theory, if the temperature of the system (simulations are coupled to a "temperature bath") is assigned a high enough value so that the kinetic energy present in the system allows it to surmount energy barriers, and sufficient time intervals are used, low-energy minima should be obtained.[2,105]

The necessity for integration of the equations of motion by numerical methods forces the time steps taken during simulation to be small. With larger systems, this causes an exponential increase in the volume of conformational space and impractically high demands on computer resources. Long simulation trajectories and efficient searching of the entire space become

impossible. Ways of improving the effectiveness of MD sampling are the enlarging of the time step size by freezing internal degrees of freedom,[106,107] the use of multiple steps for certain interactions,[108] and the reduction of conformational space size by using NMR distance restraints.[109]

To execute comprehensive searches of the conformational space, classic MD (which is more suited to searching local regions of space) is often used in conjunction with other methods. In these instances, MD either completes the searching process on previously procured local minima or refines final structures to relax conformational strain.[2,13] Apart from their application to conformational sampling studies, MD simulations have also been used extensively to analyse the dynamics and thermodynamics of *in vacuo* and solvated biomolecular systems.[110-112] Relative conformation stabilities and transition energetics are investigated by the estimation of the free energies involved. In this manner, peptide folding pathways can be emulated and folding/unfolding mechanisms derived. MD thus fulfils an important role in the elucidation of protein folding.

- *Genetic algorithms*

The GA programming technique was introduced by Holland[113] in 1975 and since then has been applied to a wide range of global optimisation problems. Its popularity is due both to its suitability for solving such problems and its ease of implementation. GA methodology is based on analogies to the current theory of biological evolution and hereditary. It mimics natural selection strategies from evolution, embracing the "survival of the fittest" principle.

Many variants of GAs have been developed for the purposes of accommodating different applications. It has only recently been employed in molecular modelling to search for low-energy conformations. The standard GA method which constitutes the basis of these variants can be described as follows:[114-116]

1. An initial population of parental individuals (feasible solutions) is randomly created. Each individual is represented by a chromosome, a string of characteristic genes.
2. All the individuals are evaluated and ranked with a fitness function appropriate to the problem in hand. The most fit of these passes directly to the following generation; a process referred to as "elitism".

3. A breeding population is formed by selecting top-ranking individuals from those that remain. This is the "natural selection" step.
4. These selected individuals undergo certain transformations *via* genetic operators to reproduce children for the next generation. Operators include recombination by crossover (two chromosomes mate by partly exchanging genes to form two new chromosomes) and mutation (one randomly chosen gene, or more, of a chromosome is altered). Mutation ensures a level of genetic diversity in the population.

The process (from 2 to 4) is repeated until a certain number of generations or some termination criterion is reached. As optimisation continues, subsequent generations will consist of increasingly fit or "superior" individuals. Relatively "strong" individuals survive and reproduce, while relatively "weak" individuals die. With reference to peptide conformational searching, the terms individual or solution, chromoseme, gene and fitness function can be interpreted respectively as conformation, set of physical variables encoded as a string of binary (or integer, or real number) digits, physical variable such as dihedral angle, and decreasing function of the potential energy. The crossover operation can therefore be depicted as in Figure 1.3.

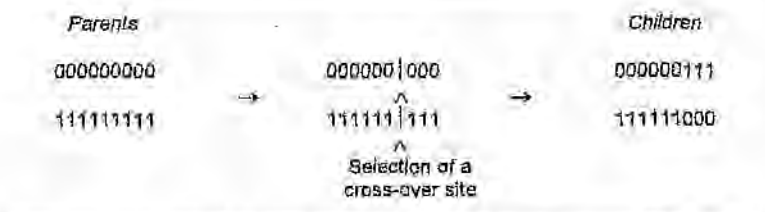


Figure 1.3: A schematic representation of the crossover operator[116]

The primary advantage of the GA technique is its ability to explore diverse regions of the conformational surface; it operates independently of initially selected surface points.[117] Genetic diversification of a population improves its probability of creating superior individuals and, in the case of conformational sampling, prevents the premature convergence of the search to one local minimum structure. The mutation genetic operator accounts for diversity to some extent, but to ensure a sufficiently high degree of diversity, tools such as *space sharing*[116] and *niche interaction*[118,119] are often introduced. *Space-sharing* is akin to forced mutation. New conformations are checked against one another for diversity and those displaying too many similarities are then mutated. *Niche interaction* is the sharing of genetic information between

niches which are subpopulations of the initial population of a search. Each niche undergoes independent searching.

Although the GA has proved to be robust and efficient in searching for local optima, it appears to be less efficient in locating the global optimum. The latter may be achieved by combining the GA with a minimisation method (for instance, gradient minimisation, SA, NN) to form a *hybrid* system.[118,120] Minimisation can either be executed after the generation of each new population, the fitness function being accordingly modified to become a function of minimised potential energy, or after generation of the final population. The GA therefore performs a global search of the energy surface, while the complementary method refines the structures resulting from the search.

Further variations to the GA method for searching conformational space, specifically that of proteins,[120a] include reduced representations of proteins[121,122-122b] and highly simplified EFFs.[121,122,123] These modifications help to speed up GA sampling time, a factor which becomes restrictive when large numbers of conformations must be examined and their energies computed or minimised.

A principal disadvantage of GA optimisation techniques is the configuration of the algorithm.[117] For every different application, optimal values for each parameter of the GA, *e.g.*, population size, number of generations, crossover and mutation rate, must be assigned. The attainment of optimal parameter values is often tedious and the choice of values can influence significantly the performance of the GA.[118]

- *Conclusions*

Much progress has been made in the effort to overcome the obstacles of multiple minima and combinatorial explosion. Several conformational searching methods appear to be effective for small to medium-sized peptides (5-20 residues). For larger globular proteins, these methods become prohibitively expensive with respect to computer time and have not met with complete success thus far. However, the ongoing improvement in experimental techniques, development of new theory, and rapid advances in computer hardware and software may in the future provide solutions to the problems encountered in these demanding applications.

1.1.3.4. Solvent modelling

Since proteins are surrounded by solvent molecules in physiological systems, the influence of solvent environment on peptide energy and structure must be addressed in molecular modelling calculations. Water, in particular, has a high dielectric constant and may therefore markedly affect peptide conformation. When modelling solvation, factors which impact on the energy of the system such as peptide-solvent interactions, solvent-solvent interactions and cavity formation need to be considered.[13]

Solvation, or hydration in many cases, has been treated in a number of ways within the MM framework. The most direct approach includes solvent molecules explicitly in the system.[59,124] Potential functions are also used to describe these molecules, leading to the development of several water models.[125,126] The polypeptide is placed in a box of solvent molecules and all interaction energies are then calculated *via* either energy minimisation, MC, or MD procedures. Periodic boundary conditions at constant volume or constant pressure are usually implemented. The long computational times required in these explicit solvent calculations have prompted the investigation of implicit solvent treatments. Examples of the latter are empirical hydration models[127] in which free energy of solvation for the system is approximated by assigning a solvation free energy to every functional group of the peptide (these values are obtained by averaging over all interactions of a specific group with a nearby hydration shell of molecules), and the integral equation model[128] in which potentials of mean force are calculated for all peptide-solvent interactions.

Another example of a model in which solvent is included implicitly is the computationally favourable modified dielectric.[59,124] In addition to local solvation effects, solvation also influences the polarity of the peptide. Since the electrostatic potential of the peptide incorporates the dielectric constant (ε), any change in polarity will also cause a change in the overall potential energy of the molecule. For *in vacuo* (gas-phase) calculations or calculations which explicitly include solvent, ε is normally set to unity. To simulate the charge screening effect of solvent in implicit treatments, the value of the effective dielectric constant is increased, resulting in a reduction of the Coulomb potential. The interior of a folded protein is generally assumed to be a low-dielectric medium (effective ε from 2 to 4), while the charges on the protein surface are associated with a high-dielectric medium ($\varepsilon \geq 80$). This transition from a low ε value in the interior to a high ε in the solvent has prompted the use of a distance-dependent

dielectric function in many investigations. Approximations adopted in this modified dielectric approach appear to result in the incorrect treatment of weak electrostatic forces.[129] This limitation has been addressed with the application of classical continuum electrostatics to the model, which allows for the management of both intramolecular electrostatic interactions and ionic strength effects.[130]

1.1.3.5. Vibrational entropy

Vibrational entropic contributions to the free energy of a peptide is a factor that is often omitted in peptide modelling studies. Conformational entropy arises when a molecule, or parts of it, undergoes small oscillations about each of its energy minima. In the case where the bond lengths and bond angles of a peptide have been fixed for MM purposes, the oscillations result from changes in the peptide's backbone and dihedral angles. The number of accessible vibrational states for a particular minimum is related to the width of its conformational energy well. The importance of this entropy factor is illustrated in Figure 1.4 which represents a potential energy surface with two minima. The broad well of the high-energy minimum indicates its greater flexibility when compared to the low-energy one, and could result in a lower free energy for the higher minimum. Thus both the depth of the conformational well, *i.e.*, the potential energy, and its shape and width at the bottom, *i.e.*, the vibrational entropy, are determinants of the free energy of a molecule.[38,59,131]

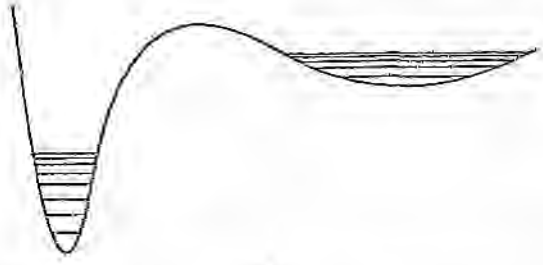


Figure 1.4: A schematic representation of a potential energy surface with two minima.[131]
Because of vibrational motion, the broad minimum of a higher-energy conformer may contain enough vibrational entropy to reduce its free energy below that of a lower-energy conformer with an energy well which is lower but narrower.

Classical statistical mechanics has been used by Gō and Scheraga[132,133] to determine conformational entropy. They defined librational entropy S_i, the entropy at the minimum point, i, in terms of the equation:

$$S_i = -1/2\ R\ ln\ |F_i|, \qquad (3)$$

where F is the matrix of second derivatives of the conformational energy with respect to the independent variables (the torsion angles) calculated at i, and R is the gas constant. The relative entropy ΔS of a conformation which displays small torsion angle fluctuations about its energy minimum can be calculated from the equation:

$$\Delta S = (1/T)(\Delta U - \Delta G), \qquad (4)$$

where T is the temperature, and ΔH *(enthalpy change)* $\approx \Delta U$ *(relative conformational energy)* under gas-phase conditions at constant pressure and volume. To evaluate the relative free energy ΔG of the conformation, Gō *et al.*[132,133] applied an harmonic approximation with statistical weighting. The statistical weights, which incorporate both a potential energy function and an entropy function, are indicative of the relative populations of the different energy minima. Free energies are thus computed by direct calculation of the probability distribution of the system.

1.2. Signal peptides

1.2.1. Function and structure

Following the synthesis of proteins by free ribosomes in the cytoplasm of the cell, those proteins that function elsewhere within or outside the cell move to their respective locations by crossing membranes, a phenomenon known as translocation or topogenesis.[134] Many of the secretory proteins, consisting of long, polar peptide chains, would be unable to traverse the nonpolar membranes were it not for an extension of between 15 to 26 amino acids at the amino-terminal end of the chains. Those proteins that are destined for translocation comprise two sections: the N-terminal signal peptide (SP) or leader peptide extension and the mature or nascent polypeptide section; they are often referred to as preproteins or precursors[135]. The temporary N-terminal extensions appear to direct and facilitate polypeptide translocation.

SPs are particular to each protein and, despite their unity of function, are not closely sequence-related. However, although SPs lack primary sequence homology, they do have certain properties in common.[134,136] Their primary structural features have been analysed in terms of physico-chemical amino acid properties, *e.g.*, hydrophobicity, size, polarity and ionic nature, by both statistical and artificial intelligence-based methods.[137,138] Thus, they typically have three distinct domains (Figure 1.5): an amino-terminal positively charged region (n-region, of 1-5 residues); a central hydrophobic part (h-region, of 7-15 residues); and a carboxy-terminal part (c-region, of 3-7 residues). It seems that the n- and h-regions are responsible for the function of selectively targeting and translocating newly synthesised proteins to their appropriate locations, with the c-region only being needed for proper cleavage of the SP from the mature chain.

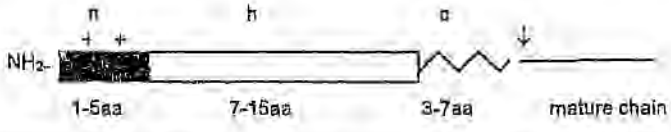


Figure 1.5: The basic design of signal peptides[136]

Despite much research in the field, an understanding of the exact role that SPs play in the secretory process remains elusive. Some experiments have demonstrated that signal sequences merely perform an indirect function, that of inhibiting protein folding and providing a tag identifying the protein to be transported. Contrary to this, other experiments have shown the peptides to fulfil a direct role as true targeting signals that are specifically recognised by the secretory apparatus (the different cellular and membranous components implicated in protein transfer).[139] It has been speculated that overall common characteristics such as conformation, *i.e.*, secondary structure, and hydrophobicity may account for this recognition.[140]

1.2.2. Concepts in protein secretion

Before describing some of the models that have been proposed[141] in the attempt to describe the molecular mechanism of protein secretion, certain relevant concepts require definition.

1.2.2.1. Prokaryotic and eukaryotic secretion

Prokaryotes are living organisms whose cells do not contain well-defined nuclei and organelles. Examples are bacteria and blue-green algae. The cytoplasm in prokaryotic cells is bounded by two membranes, inner and outer, which are themselves separated by a periplasmic space. Generally, the only secretion of proteins which occurs is that across the inner cytoplasmic membrane.

In contrast, eukaryotes are higher organisms, such as yeast and mammals, that possess definite cellular compartments. Protein secretion thus occurs across diverse membranes: endoplasmic reticulum (ER), mitochondrial, chloroplast, peroxisomal, nuclear.[134] In the literature, SPs refer to the N-terminal extensions of proteins destined for the ER, while N-terminal extensions of mitochondrial proteins are called presequences. Besides being even more diverse than SPs, they vary greatly in length and display basic and amphiphilic characteristics. Extensions to chloroplast proteins are similar in composition to presequences and are known as transit peptides. Although presequences and transit peptides are similar to SPs in their basic design, SPs are significantly more hydrophobic.[142] Only SPs will be considered in the present study.

1.2.2.2. Cotranslational and posttranslational transfer

Cotranslational transfer occurs when a polypeptide is transferred across a membrane prior to completion of its synthesis or translation. It embraces vectorial (residue by residue) transfer of nascent chains. Investigation of energy factors[143,144] which drive the export of proteins across membranes has led to the conclusion that no external energy source, other than that of protein synthesis, is required for cotranslational translocation. Ribosomes use the energy associated with protein elongation to push nascent chains through the membrane.

Posttranslational transfer occurs after the synthesis of a polypeptide has been completed. In such circumstances, mature proteins often cross membranes in domains (series of residues) rather than vectorially. It has been surmised that external energy sources are essential for posttranslational translocation.[145] Feasible thermodynamic sources of energy include transmembrane electrochemical gradients, *e.g.*, pH, ionic strength (which result in a membrane potential), adenosine triphosphate (ATP)[r] and post-translocational protein folding.[146]

In higher eukaryotes, the mode of translational transfer appears to be organelle-dependent: posttranslational for mitochondria and chloroplasts, and cotranslational for the ER. Although it has been observed that some proteins of the prokaryotic bacterium, *Escherichia coli*, are secreted both co- and posttranslationally, bacterial proteins are generally secreted posttranslationally.[141]

1.2.2.3. Molecular chaperones

Although *in vitro* studies of isolated proteins show folding to be a spontaneous process,[1] studies of proteins in their physiological environments indicate that folding *in vivo* is facilitated by other recently-identified cellular components called molecular chaperones.[139,147,148] These are proteins that have the ability to temporarily stabilise non-native conformations of other proteins by binding rapidly to conformations of their substrates which may be either unfolded or partially folded. In this way, the timing and location of folding can be controlled. This kinetic modulation of folding may be an important factor contributing to the efficient transport of proteins across membranes. Most membrane systems cannot transport proteins that are in their

[r] ATP is a reservoir of chemical potential energy; its hydrolysis to adenosine diphosphate (ADP) is a highly-exergonic reaction.

fully folded native states[140] and it has been experimentally determined[149,150] that certain molecular chaperones (see Table 1.1) do indeed interact with portions of secretory proteins, thereby impeding premature folding. Such interactions also prevent the aggregation of newly synthesised polypeptide chains. The nature of these interactions is unclear. GroEL, a bacterial molecular chaperone, is thought to recognise incompletely folded proteins through structural motifs, specifically amino-terminal α-helices on nascent chains.[151] However, another bacterial chaperone, SecB, appears to bind weakly to multiple sites on preproteins,[152] thus demonstrating a lack of specific recognition and non-involvement of SPs. In this case, SPs are believed to retard the folding of proteins and so facilitate binding between preproteins and SecB.[153]

Table 1.1: Examples of molecular chaperones implicated in protein transport[147,151]

Chaperone	Subcellular localisation	Organism	Function
Eukaryotic			
SRP	cytosol	mammals	Binds SPs prior to translocation
BiP/ Grp78	ER	mammals	Binds subunits of ER proteins
Kar2p	ER	*Saccharomyces cerevisiae*	Promotes protein translocation into ER
Hsp60	mitochondria	*Saccharomyces cerevisiae*	Promotes folding and assembly of imported proteins
Cpn60	chloroplasts	plants	Promotes folding and assembly of imported proteins
Prokaryotic			
SecA	plasma membrane	*Escherichia coli*	Targets precursors for translocation
SecB	cytosol	*Escherichia coli*	Stabilises precursors by binding with them
DnaK	cytosol	*Escherichia coli*	Stabilises newly made proteins; preserves folding competence of proteins
GroEL	cytosol	*Escherichia coli*	Binds proteins during translocation and folding

1.2.3. Protein secretion mechanisms

The mechanism of the protein secretory process at the molecular level can be divided into three stages: entry into the transport pathway (targeting of the secretory protein to the membrane and their initial association - translocation initiation); translocation across the membrane (the translocation mechanism); and release on the opposite side (cleaving of the SP from the mature protein and subsequent folding and assembly of the released protein).[140] Of the many hypotheses that exist, no single one can account for all the experimental data (from *in vivo* genetic studies and from *in vitro* biochemical studies) collected thus far. However, it must be noted that the existence of many different export pathways, which may operate in parallel to one other, and the apparent variety of functions

performed by signal sequences, may render the achievement of a single concordant mechanism impossible.[141]

Translocation is the most controversial stage of the export mechanism, there being much speculation concerning the composition of the translocation apparatus (the translocon). There are two opposing suppositions on the nature of the environment surrounding the secretory protein during translocation: the protein either passes through an aqueous channel which may be formed from proteins or from both proteins and lipids (the signal hypothesis),[154] or inserts spontaneously into the hydrophobic membrane bilayer (the membrane trigger hypothesis).[155] Both mechanisms presume direct involvement of the SP in the secretory process. The ideas embodied in these hypotheses of ..n form the basis of further model development.

1.2.3.1. The signal hypothesis

The signal hypothesis is generally accepted for eukaryotic secretion, specifically in mammalian systems. It is supported by extensive evidence from studies in eukaryotes. The initially proposed mechanism[154] entails (1) recognition of the signal sequence, attached to a polypeptide chain emerging from a ribosome, by a ribonucleoprotein present in the cytosol called the signal recognition particle (SRP), (2) interaction of the sequence with the SRP and subsequent cessation of peptide elongation, (3) diffusion of the so-formed complex to the ER membrane where it interacts with a "docking" integral membrane protein (the SRP receptor); translation of the preprotein then resumes, (4) formation of a proteinaceous pore in the membrane through which the nascent peptide is vectorially extruded into the aqueous lumen of the ER and, finally, (5) cleavage of the SP in the aqueous medium by a signal peptidase enzyme. Elongation of the nascent chain provides the energy required for translocation. An interesting observation recently made by Engelhard[156] is that cleaved signal sequences play a function in the construction of antigens in living organisms.

Some aspects of the signal hypothesis that have been probed and expanded upon by several researchers are discussed below.

- ***The translocation site***

The signal hypothesis proposes that proteins are transferred across the ER membrane *via* a channel or translocon. Electrophysiological techniques have been used by Simon and Blobel[157] to demonstrate the existence of an aqueous protein-conducting channel in the ER. Their results

indicate that this channel is of a fixed size and is not freely permeable to ions when occupied by a translocating peptide. The size of the channel and its insensitivity to membrane lipid changes suggested that it may be formed of proteins. The findings of Crowley and colleagues[158,159] have reinforced the proposal of a hydrophilic channel. They incorporated fluorescent probes into nascent chains during translation to identify the environment of the chains in the ER membrane. Besides discovering that the channel is indeed aqueous and that it spans the entire membrane, they also found that it is sealed off from the ER lumen, only opening after the nascent chain reaches a certain residue length (Figure 1.6). The preprotein is also sealed off from the cytoplasm by a tight binding of the ribosome to the translocon.[158,160] Translocation thus occurs directly from an aqueous tunnel in the ribosome into an aqueous pore in the membrane. It is assumed that this closed tunnel-pore system dictates movement of the nascent chain across the membrane.

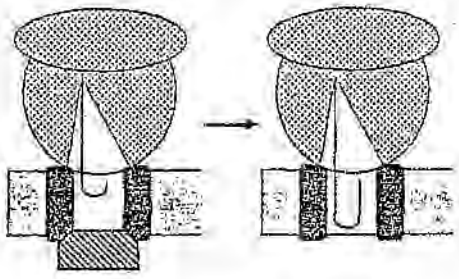


Figure 1.6: The aqueous channel is gated.[159]
Initially, the signal sequence and nascent chain (which enter the channel as a loop) are sealed off from both the cytoplasm (by the tight ribosome-translocon junction) and the ER lumen (by a lumenal protein). After translation of about 70 nascent chain residues, the aqueous channel is opened to the lumen.

- ***Translocation site components***

Identification of the components of the translocon has been the subject of numerous investigations. There has always been the question of whether signal peptides interact with lipids or with proteins in the membrane pore. Photo-crosslinking studies[161,163] have revealed the immediate proximity of a specific set of ER membrane proteins to the nascent chain during translocation in both yeast and mammalian systems. Reconstitution experiments[164] have suggested that, of these proteins, the Sec61p trimeric complex (consisting of α, β, and γ

subunits) is the major component of the protein-conducting channel. The multispanning α-subunit has been found to contact continuously with the nascent polypeptide from one end of the membrane to the other; Sec61α thus forms a lining in the channel.[160,165] Beckmann *et al.*[165a] have recently determined the cryo-electronic structure of the Sec61 complex bound to the ribosome. Their findings corroborate the notion that the translocon extends from the ribosomal tunnel and that the ribosome-translocon interface is both dynamic and regulated.

Contrary to the notion that SPs only interact with proteins while traversing the protein-conducting channel, Martoglio *et al.*[166] have suggested that interaction with lipids is also possible. They hypothesise that the channel, which is formed by proteins, is open laterally toward the lipid bilayer during early stages of protein insertion. In this model (depicted in Figure 1.7), the nascent polypeptide is arranged in a loop-like conformation, with the hydrophobic h-region of the signal sequence facing the lipid bilayer, while the hydrophilic, translocating portion of the nascent protein is in a proteinaceous environment. The suggestion of lipid interaction in the channel concurs with that of Rapoport[167] in his earlier amphiphilic tunnel hypothesis. He postulated that the membrane channel was amphiphilic and is so able to bind both hydrophobic and hydrophilic parts of the precursor.

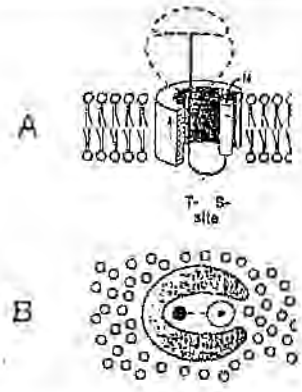


Figure 1.7: Hypothesised arrangement of proteins and lipids in the protein-conducting channel.[166]
(A) side view; (B) top view. The secretory protein inserts into the channel in a loop-like configuration with the nascent chain facing protein (T site) and the signal sequence contacting both protein and lipid (S site).

- ***Parallel prokaryotic mechanism***

The eukaryotic and prokaryotic protein secretion mechanisms display striking analogies due probably to the existence of an evolutionary relationship between protein transport across the ER membrane in eukaryotic cells and transport across bacterial membranes. Similarly structured signal sequences have been shown to be exchangeable between the two classes of organisms, and homologous complexes implicated in their export pathways have been discovered. For example, bacterial SRP and FtsY[168] are the *E. coli* homologues of eukaryotic SRP and the SRP receptor, and the *E. coli* integral membrane protein SecY[135] is analogous to the eukaryotic Sec61p. SecY is structurally similar to Sec61p, SecY also constituting the central translocation component, while SRP/FtsY operate in a parallel manner to their mammalian counterparts.

Since results consistent with both transfer through an aqueous pore and direct phospholipid transfer have been reported for bacterial systems, the nature of the immediate environment of the precursor as it crosses the membrane has been speculated to be both phospholipid and proteinaceous.[135]

1.2.3.2. The membrane trigger hypothesis

In the membrane trigger hypothesis,[155] specific transport apparatus such as receptor proteins and membrane channels are postulated to be unnecessary for protein translocation. Preproteins insert directly into the phospholipid bilayer in a spontaneous manner due to the thermodynamics of protein folding. The energy needed for secretion is acquired from the membrane potential. The role of the signal sequence is to modulate folding by interacting with the remainder of the precursor protein so that the latter folds into a conformation that is export competent, *i.e.*, able to partition into the nonpolar part of the membrane. Conformation is therefore an important determinant in this hypothesis and will be addressed in detail at a later stage.

The idea behind the membrane trigger model was formulated by considering primarily the ability of synthetic SPs to interact with lipids. Biophysical studies of isolated SPs in model membrane systems (mostly representative of prokaryotic *E. coli* systems) have demonstrated the effect of signal sequence interaction on membrane lipid fluidity,[144] lipid vesicle aggregation,[169,170] lipid monolayer surface pressure[171] and lipid packing and orientation.[172] The affinity of signal sequences for phospholipid monolayers has been shown to correlate with their *in vivo*

activity.[172-174] The thermodynamics of peptide incorporation into phospholipid bilayers has also been examined,[175,176] and recent results[176] show the hydrophobic effect to be the overall driving force for transmembrane insertion.

The hydrophobicity of the h-region and the positive residues of the n-region on each SP are proposed to assist spontaneous insertion: the hydrophobic core partitions into the membrane lipids and spans the bilayer, while the amino-terminal end binds electrostatically to anionic phospholipids at the membrane surface.[177] Although the direct interaction of anionic phospholipids with preproteins is speculative, it has been observed that their involvement is essential for efficient translocation.[178] Hydrophobicity appears to play a crucial role in signal sequence functioning. The composition and length of the h-core have been shown to influence markedly SP conformation and protein translocation ability.[173,179,180]

1.2.3.3. Other hypotheses

Descriptions of translocation mechanisms that are variations on the signal hypothesis and the membrane trigger hypothesis are given below.

- ***The loop model***

Inouye and Halegoua's[181] loop model for protein export proposes that, after the signal sequence is anchored to the membrane surface *via* electrostatic interactions, it inserts into the membrane bilayer as a loop. The reverse turn in the loop is induced by the presence of proline and glycine residues in the sequence. This loop structure projects further into the membrane bilayer as the preprotein lengthens. Shaw *et al.*[182] have provided supporting evidence for the loop model. They analysed signal sequence topology during membrane insertion and found that the N-terminus of a secr y protein remains in the cytoplasm ,with the growing C-terminus being continuously translocated across the membrane.

Extensions of the loop model are the helical hairpin hypothesis[143] and the direct transfer model.[183] Both propose that signal sequences partition into the hydrophobic part of the membrane in helical conformations and that this partitioning is thermodynamically-based. The initial driving force for protein export is the favourable free energy involved when hydrophobic SPs are transferred from the aqueous cytoplasm to the lipidic membrane. The helical hairpin hypothesis postulates that the loop which inserts into the membrane comprises two helical regions which form a side-by-side "hairpin" structure. One region consists of the signal

sequence, and the other is the first 15-25 residues of the mature protein. Both regions span the membrane.

Although the loop model was originally devised as a modification of the membrane trigger hypothesis, it has been extrapolated to the signal hypothesis.[137,165,166] In the latter case, the loop structure of the nascent chain enters a proteinaceous channel in the membrane.

- ***The domain model***

This export model[164] combines and modifies the targeting stage of the signal hypothesis and the translocation stage of the membrane trigger hypothesis. The nascent chain only enters the membrane once its synthesis is almost complete; it then crosses the membrane in domains Translocation is thus posttranslational. Membrane potential or a conformational change in the secretory protein provides the energy for transfer.

- ***Non-bilayer lipid structures***

A more active role for membrane lipids in the translocation process has been suggested.[172,185] The lipids arrange themselves into inverted hexagonal structures, *i.e.*, the membrane bilayer organisation is disrupted, and so forms a hydrophilic tunnel through which the preprotein is extruded. Extrusion is assisted by the elongation of the protein as well as by the motion of the lipids. Subsequent work by de Kruijff and co-workers[186,187] has endorsed this hypothesis and has further demonstrated that functional SPs may induce the local formation of these intermediate lipid structures.

1.2.4. Conformational analysis

Although there appears to be a lack of consensus in the literature about the nature of the environment surrounding nascent polypeptide chains as they translocate across membranes, one certainty is the important part played by secondary structure conformation in the secretory process. The conformational changes that signal sequences undergo when either interacting with the secretory apparatus or moving between different environments must influence strongly the ability of nascent chains to be translocated. Differences in translocation efficiency of wild-type and mutated SPs have usually been interpreted in terms of conformational changes experienced by the peptides on mutation. A related factor is the common characteristic of conformational homology which functional signal sequences appear to possess.

1.2.4.1. Experimental

Direct determinations of signal sequence conformation generally involve the application of the experimental techniques of CD, infrared, ultraviolet and NMR spectroscopy to synthetic SPs, SP fragments, and peptides resembling SPs. The determinations are performed in polar or nonpolar bulk solvents, or in membrane-mimetic substances such as phospholipid monolayers, lipid vesicles and micelles. Experimental conditions, *e.g.*, peptide concentration, lipid/peptide ratio and phospholipid type (anionic *vs.* zwitterionic), have been demonstrated by Keller *et al.*[188] to be critical in accurate conformational analysis.

The use of isolated signal sequences, *i.e.*, without [illegible] of the precursor protein, in investigations of conformation has been ratified by [illegible] are transferable from one protein to another without loss of export fun[illegible] periments conducted in lipid vesicles have shown that the SP and adjacent [illegible] *oli* LamB protein are conformationally independent of each other.[189] If tran[illegible] to occur between signal sequences and portions of the corresponding mature protein, a[illegible] been surmised by some,[140,155] these are probably less significant to function than sequence conformation.

Results from conformational studies where bulk solvents have been used indicate a predominance of either β-sheet[190] or random[169,191] conformation in polar solvents and α-helical[169,191-194] conformation in nonpolar solvents. Intramolecular hydrogen bonds are able to form in hydrophobic solvents, thus inducing α-helix formation. Further observations[191,194] have disclosed that hydrophobicity and helical propensity are related to *in vivo* signal sequence function; sequences which are sufficiently hydrophobic and which have a high tendency to form α-helices in nonpolar solvents will function efficiently. The stability of these helices has been identified as the main helical property implicated in SP activity,[193] with the helices being most stable in the hydrophobic regions of the peptides.

The conformational response of signal sequences to different environments has also been noted in studies where heterogeneous solutions are employed to simulate the membrane bilayer and its interface with water. The membrane trigger hypothesis[155] is the protein secretion mechanism assumed in these experiments. Several investigators[177,195,196] have suggested that changes in SP conformation are induced on insertion into the membrane lipid region. These conformational changes are illustrated in Figure 1.8. Unstructured signal sequences in an aqueous milieu are

proposed to adopt a β-structure when interacting with the membrane, *i.e.*, at the lipid interface with the aqueous medium, and an α-structure when inserted into the membrane (the lipid phase). The orientations of the β- and α-structure with regard to the lipid-water interface are coplanar[177] and perpendicular,[172] respectively. The transition between β-sheet and α-helical forms of the SP may be necessary for its operation.

The conformational flexibility of SPs has prompted the postulation of another secretion mechanism, the "unlooping model".[127] This model expands the loop model[181] by proposing that the SP unravels from its looped helix (helix-turn-helix) conformation which it assumed on entering the membrane, so moving the N-terminus of the mature protein across the bilayer. The SP thus assumes a stretched conformation sometime during translocation. In their conformational studies of signal sequences in lipophilic environments, Yamomoto *et al.*[192] have found that it is the helix at the C-terminus of the sequence that adopts this extended conformation, while others[199,200] have confirmed the existence of a kink in the helix between the h-core and the C-terminus.

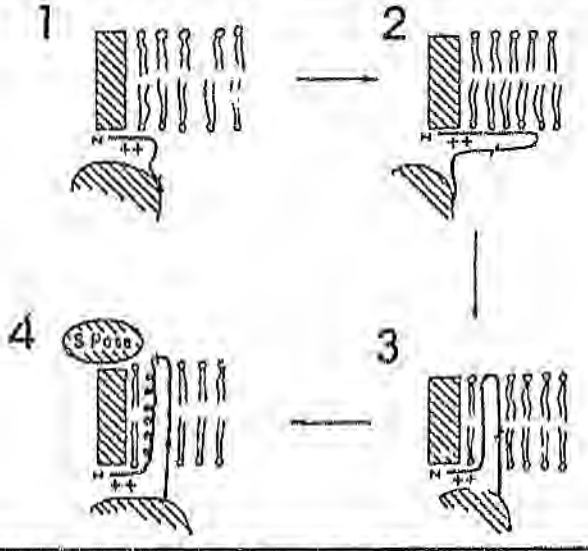


Figure 1.8: Model for initial interaction of a signal sequence with a membrane.[141]
(1) The signal sequence emerges from the ribosome and enters the aqueous milieu in a random conformation. (2) It interacts with the membrane surface and adopts a β-structure. (3 & 4) It inserts into the membrane bilayer, where it undergoes a conformational change from an extended β-structure to an α-helical structure. The cleavage site falls on the opposite face of the membrane (S Pase = signal peptidase).

1.2.4.2. Knowledge-based modelling

There is limited structural data available on signal sequences due to the difficulties involved in crystallising them for diffraction studies. Their inherent hydrophobicities cause them to be almost invariably insoluble in water. Fortunately, this data is not required for the prediction of signal sequence conformation by statistical procedures since tertiary structure can effectively be ignored for membrane-bound proteins.[201] Strong constraints imposed by the environment on the peptides appear to reduce their structural degrees of freedom to secondary structure status only.

The Chou-Fasman approach[22] to predicting protein conformation from primary structure has been applied to several SPs.[179,190,203,204] In general, the h-cores of the peptides are predicted to form either α-helical or β-sheet conformations, or both! However, the validity of these predictions is dubious if spontaneous insertion of the SP into the hydrophobic membrane is assumed; the Chou-Fasman[22] conformational parameters are based on water-soluble, globular proteins. This may explain the contradictory conformations detected for signal sequences, *viz.*, either α-helix or β-sheet, whereas it would be expected that a common conformation would dominate. Nevertheless, the predictions do corroborate the experimental data in suggesting the formation of highly regular secondary structures for SPs.

Similar conclusions have been reached in an investigation by Prabhakaran:[197] amino acids in the hydrophobic core of SPs show preferences for both α-helix and β-sheet conformations. Here, adjusted Chou-Fasman parameters (including information about interresidue interactions)[202] were used to calculate the statistical distribution of conformational preferences of amino acids along the length of signal sequences and along adjacent portions of their nascent peptides.

Emr and Silhavy[203] have used the wild-type signal sequence and derivative strains thereof to relate export of the *E. coli* LamB protein with Chou-Fasman[22] *predicted conformations*. Functional SPs display a tendency to adopt α-helices in their central, hydrophobic regions, while the same region in the non-functional SP displays a largely random conformation. These findings were later supported by CD analysis of the relevant conformations under apolar conditions.[191] Thus, the presence of α-helical conformation appears to be critical for efficient translocation.

1.2.4.3. Molecular modelling

Few MM-based predictions of signal sequence conformation have been performed. All favour an α-helical conformation throughout the hydrophobic region of a sequence in lipid-simulated environments.

Conformational energy calculations have been applied to the 3D structures of the wild-type signal sequences and several variants thereof for murine pre-κ-light chain,[205] *E. coli* LamB,[206,207] yeast invertase,[203] and human apolipoprotein B.[205] In a study of the murine pre-κ-light chain SP, Pincus and Klausner employed potentials from the ECEPP force field,[50] the Powell energy minimisation method (which uses conjugate gradients)[208] and a global conformational searching technique based on the build-up procedure.[72] Only nondegenerate[8] local energy minima of component di- and tripeptides of the signal sequence were combined in the construction of longer polypeptides. Solvent effects were excluded as it was assumed that SPs function inside highly nonpolar environments during the translocation process.

Perez and colleagues[206] applied the updated ECEPP/2 potentials[51] in their MM investigation of portions of the wild-type SP and mutated SPs of the *E. coli* LamB protein. The Powell minimising algorithm[208] was again used. Three strategies were used to explore the conformational space: a build-up procedure; a random search of points generated on the hypersurface; and a search based on the adoption of specific secondary structures. The majority of the resultant low-energy structures were obtained with the last strategy. The relative helicities displayed by the peptides agreed with previous results gleaned from CD experiments[191] and statistical predictions.[203] Hydrophobicity values were also calculated for each peptide. These values equated well with SP activity, emphasising the importance of hydrophobicity in SP function.

The wild-type signal sequence of the *E. coli* LamB protein and a 50% active mutant sequence have been the subject of another molecular modelling study.[207] The study attempts to model peptides, in random-coiled and helical-constrained structures, at a cytoplasm/membrane interface. THOR, a program that was developed specifically for the modelling of biomolecules

[8] "Nondegenerate" conformations not only have different side chain conformations, but also have distinct backbone conformations.

in water, nonpolar media, and at the interface between them, was used. It incorporates the GROMOS force field[209] as well as MD and energy minimisation (steepest descent)[57] procedures. The interface is represented by a discontinuity in the dielectric constant; $\varepsilon = 80$ in the polar medium, and $\varepsilon = 2$ in the nonpolar medium. In the case of helical-constrained structures, conformations are maintained by applying high force constants to the appropriate backbone torsions. The following observations were derived from this work: both random-coiled wild-type and mutant sequences evolve from a stretched conformation to an increasingly folded one during MD simulations in the polar medium, and have an affinity for the lipidic phase; the helical form of the mutant is unstable under the chosen modelling conditions; the wild-type peptide (in both random and helical conformations) is inclined to be more stable at the interface than is the mutant peptide (in the random conformation); and the helical-constrained wild-type experiences a conformational change at the interface (the helix is partially disrupted and a backbone turn occurs).

Brasseur and co-researchers[265-267] have performed extensive MM calculations both on SPs (yeast invertase and apolipoprotein B (apoB)) and on other lipid-associating α-helical proteins, under conditions mimicking the phospholipid matrix of biological membranes (using values of ε varying linearly from 3 to 30 across the interface). A three-step procedure was used: energy minimisation to model assumed α-helical conformations of isolated molecules and optimisation of torsional angles *via* a simplex procedure;[268] calculation of molecular hydrophobicity potentials[269] of the molecules; and energy minimisation to model orientations of the now-rigid molecules, inserted into organised lipid bilayers.[268] A semi-empirical theoretical model provides a molecular description of the lipid bilayers. The hydrophobicity of the SP affects the lipid bilayer by disrupting its structure, so easing transfer of material across the bilayer. A correlation between the angle of insertion of the SP into the membrane and its ability to direct secretion was identified: the oblique orientation of the SP to the membrane is important for its proper functioning.

1.3. Membrane proteins

Integral membrane proteins contain at least one polypeptide segment that is embedded in the membrane lipid bilayer.[134] When two or more of the segments are present, they are connected with aqueous-located loops. In the literature, it has generally been accepted that these membrane-embedded segments are long, hydrophobic α-helices. Exceptions to this are the highly polar β-barrel sheets found in porin membrane proteins.[210] Membrane proteins and signal sequences thus share common characteristics, namely, their hydrophobic natures (with the exclusion of porins) and their abilities to adopt similar conformations upon integration into membranes. In view of these similarities, techniques used to predict conformations of transmembrane sequences may assist in conformational predictions of signal sequences.

Apart from the fact that transmembrane segments of membrane proteins are more hydrophobic than SPs, a major difference between the two is the type of amphiphilicity that they exhibit. With SPs, amphiphilicity results from the segregation of hydrophobic and hydrophilic residues in the peptide sequences while, with membrane proteins, hydrophobic and hydrophilic faces form in the secondary structural regions of the segments.[201] Differences in primary structure between SPs and transmembrane sequences[211] may play a vital role in their differentiation by components of the protein secretory apparatus. The components must be able to identify those sequences that are destined for translocation across the membrane and those that are destined for integration into the membrane.

1.3.1. Conformational analysis

1.3.1.1. Knowledge-based modelling

As with signal peptides, knowledge-based analysis of membrane protein secondary structure using methods derived from water-soluble, globular proteins are likely to be unreliable. Therefore, several predictive algorithms have been developed which are particular to integral membrane proteins.

Typical prediction schemes focus on the delineation of lipid-buried segments by investigating the presence or absence of hydrophobic stretches in the proteins. Many assume that the transmembranal stretches are synonymous with α-helices. The most widely utilised scheme is that of Kyte and Doolittle.[212] It identifies potential membrane-spanning segments on the basis of residue hydropathy (accounting for both hydrophilicity and hydrophobicity). To improve the

accuracy of predictions, several modifications of standard hydrophobicity-based methods have been made. The modifications encompass measurements of helix amphiphilicity by constructing helical wheels[h][213] or hydrophobic moment plots,[i][214] consideration of flanking residues that disrupt transmembrane helices,[215] superimposition of the "positive-inside" rule,[j][216] and utilisation of multiple sequence alignments of related proteins.[217] TMPRED[218] is another prediction algorithm that attempts to improve accuracy. It embodies a "consensus procedure" which combines results of six different prediction techniques with transmembrane helical properties.

Hydropathy plots have been further applied to topology prediction of multi-spanning membrane proteins[219-222] where the locations of transmembrane helices with respect to the rest of the protein, and their orientations with respect to the cell, are deduced. The entire conformation of the protein can be derived once transmembrane organisation is established. Topology prediction techniques also rely on statistical studies of amino acid distribution in cytoplasmic and extracellular membrane protein regions since these regions exhibit differences in residue composition.

High accuracy predictions of putative transmembrane helical sequences have been produced with NN models. Lohmann *et al.*[223] use an evolutionary algorithm to develop and optimise NN architecture and weights, and physicochemical amino acid properties to depict sequences. Rost and colleagues[224] derive input data from multiple sequence alignments for their NN system, so achieving a prediction accuracy of 95%. NN-based methods have also been used to evaluate membrane protein topology. Fariselli and Casadio[225] describe one such method, HTP, which realises a success rate of 77%.

Few secondary structure prediction methods have been designed where helical transmembrane segments are not implicitly assumed. Furthermore, there appear to be no currently available prediction methods for the second type of membrane protein (represented by the porin structure). This is probably due to the scarcity of structurally known membranal β-strands.

---

h A "helical wheel" is a projection down the helical axis of the positions of the side chains.

i A "hydrophobic moment plot" displays the hydrophobic moment of a helix as a function of its hydrophobicity. The periodicity of the hydrophobicity is calculated.

j The "positive-inside" rule states that positively charged basic residues occur with much higher frequency in cytoplasmic protein loops that connect membrane-spanning segments than in extracellular loops.

1.3.1.2. Molecular modelling

In a parallel molecular modelling study to that of the murine pre-κ-light chain signal sequence, Pincus *et al.*[226] have calculated the 3-dimensional structure of the membrane-active protein, melittin. In this case, solvent effects were omitted since melittin is known to fold properly in nonpolar environments and become denatured in water. Two low-energy, α-helical conformations were found for the membrane-bound portion of the peptide, a result which agreed with experimental [227] The native form of melittin has also been successfully determined with MM by Head-Gordon and Stillinger.[228] Their calculations incorporated the antlion method[103] to sample the potential energy hypersurface for conformations, parameters from the CHARMM force field,[54] and gas phase conditions.

The identification of membrane protein structural features (α-helices, loops, terminal segments) has been explored [229] Simulations were performed in a dielectric continuum with a relative dielectric [illegible]ivity of 2, and results were verified with experimental NMR measurements. The simulations were able to detect both amphipathic and hydrophobic membrane-spanning helices. MM has also recently been employed to analyse properties of transmembrane α-helices such as their spatial hydrophobic nature (using MC simulations of nonpolar and polar solvents around the helical peptides),[230] and their relative stabilities in different environments (using MD simulations in explicit solvent environments).[231] Another interesting application has been the modelling of α-helix bundles in integral membrane proteins. Evidence suggests that helix bundles, composed of varying numbers of helices, constitute the transmembrane regions. Tuffery *et al.*[232] developed a MM-based strategy (conformational sampling with energy minimisations) to optimise the packing of helices in a bundle by predicting their relative positions and rotational orientations within the bundle.

1.4. Objectives of this study

The primary objective of the current study was to use computational procedures to analyse in detail the conformations assumed during the translocation process by both export-effective signal sequences and by those mutated sequences which do not facilitate export. The literature survey has revealed the necessity for a molecular modelling study which can provide a more quantitative basis for results obtained from previous predictive and MM calculations and which can resolve the various conflicting conclusions. It was hoped to gain a better understanding of the definitive role played by SPs in the protein secretory mechanism. It was also hoped to produce a general method for accurate secondary structure prediction of signal sequences.

Computational procedures which we have utilised to predict protein secondary structure or analyse peptide conformations include knowledge-based methods (available from the literature and the Internet) derived from both globular proteins and membrane proteins, and molecular modelling. The latter incorporates systematic searches (using both the SYBYL[233] and ECEPPAK[234] programs) and GA searches (using the ECEPPGA[120] program). MM calculations were performed with a selected sequence of distance-constraints, the use of which will be rationalised in the following chapter, applied to each peptide sequence. Both polar and nonpolar solvent environments were considered. Since the SP systems (consisting of wild-type and mutated peptides) investigated in this work had been previously subjected to experimental conformational analysis, current findings can be compared with experimental data. Thus, the validity of calculated results can be evaluated.

CHAPTER 2
PROGRAM METHODS

2.1. Knowledge-based modelling

A brief description of each knowledge-based method used to predict SP secondary structure is given in this section. All the prediction programs are available for use on the Internet.

2.1.1. Globular protein-based predictions

The Chou-Fasman (C-F) algorithm[22] was selected for prediction in the current study because of its previous application to two of the SP systems investigated in this project, namely, LamB and CPY (*cf.* Chapter 3). The comparative affinities of signal sequences for the α-helical and β-sheet conformations were evaluated by means of C-F profiles which were obtained from ProtScale, a primary sequence analysis tool accessible from the University of Geneva's ExPASy WWW server (http://www.expasy.ch). The ProtScale program calculates profiles of various kinds, *e.g.*, polarity, hydrophobicity, Chou-Fasman, for selected proteins using predefined amino acid scales from the literature. The C-F conformational parameter scales were taken from a compilation made by Chou and Fasman[25] from a statistical survey of 29 proteins.

The performance of the C-F prediction method was compared with that of more recent methods of predicting globular protein secondary structure. Five such methods are available from the IBCP (Lyon, France) server at the web page http://www.ibcp.fr/predict.html. The server makes available joint protein sequence prediction; five predictive methods are used to produce a consensus secondary structure. The methods selected are based on different approaches: Gibrat (information theory);[235] Levin (sequence similarity);[236] DPM (class prediction);[237] SOPMA (self-optimised prediction from alignment);[238,239] PHD (neural networks).[36,240,241] The program supplies a number of states to describe secondary structure, the number varying with each prediction method. The states are α-helix, β-sheet, coil, β-turn, bend and bridge. Joint analysis permits the cross-validation of methods, so enhancing overall prediction accuracy.

2.1.2. Membrane protein-based predictions

Patterns in amino acid hydrophobicities of the signal sequences were explored with hydrophobicity plots which distinguish between transmembranal and non-membranal proteins. The three hydropathy scales of Hopp and Woods,[242] Rose *et al.*[243] and Sweet and Eisenberg[244] were used by Bird and co-workers[204] in their study of CPY mutant signals, and so were also employed in this work. Plots were computed with the ProtScale program (ExPASy server). In addition, the average hydrophobicity of each SP h-region was calculated by adding component residue hydrophobicity values and dividing by the number of residues in the region.

A comprehensive, automatic service for protein structure prediction is provided by PredictProtein (http://www.embl-heidelberg.de/predictprotein/predictprotein.html). The server allows access to seven prediction-related methods: MaxHom (for multiple sequence alignment), PHDacc (for solvent accessibility prediction), EvalSec (for prediction accuracy evaluation) and the secondary structure prediction programs PHDsec (for water-soluble globular proteins), PHDhtm (for integral membrane proteins), PHDtopology (for helical transmembrane assembly) and PHDthreader (for fold recognition). The PHD methods are based on profile neural networks and use multiple sequence alignments as input. PHDsec[36,240,241] is one of the five methods used in the joint analysis procedure described above. It yields average three-state (α-helix, β-strand and loop or coil) accuracies of >72%. Because PHDacc and PHDhtm are associated with the prediction of hydrophobic proteins, they were applied to the signal sequences in the present study. PHDacc[245] computes ten relative accessibility states and projects them onto three states (buried, intermediate and exposed). PHDhtm[224] has an expected two-state (transmembrane and non-transmembrane) output accuracy of >95%.

Two programs that are also suitable for transmembrane protein prediction are TMpred and PSA. TMpred is a conformational prediction program from the Bioinformatics Group at ISREC in Lausanne, Switzerland, and is available from the ISREC server at the web address http://ulrec3.unil.ch/software/TMPRED.html. TMpred predicts membrane-spanning regions of proteins as well as orientations (inside → outside and outside → inside)[k] of these region from statistical data extracted from TMbase,[246] a database of naturally occurring transmembrane proteins.

[k] These orientations describe the position of the N-terminus of a protein. Depending on the type of organelle, "inside" normally refers to the cytoplasmic face of the membrane and "outside" refers to the lumenal face.

The Protein Sequence Analysis (PSA) server (http://bmerc-www.bu.edu/psa/) is maintained by the BioMolecular Engineering Research Center of Boston University, Boston, MA. PSA is based on probabilistic discrete state-space models (DSMs)[247,248] which recognise patterns of α-helices, β-strands, turns and loops (or coils) in specific structural classes. It analyses sequences *via* Type-1 or Type-2 DSMs. Type-1 models are applicable to monomeric, single-domain, water-soluble, globular proteins, while Type-2 models are suitable for multimeric, multidomain proteins and membrane-spanning proteins. In this investigation, signal sequences were subjected to the latter type of analysis.

2.2. Molecular modelling

2.2.1. The ECEPPGA program

The central conformational searching procedure employed in this study is a GA. It was designed by Stephens[120] for the specific purpose of investigating protein folding. In conjunction with the ECEPP/3 force field and a gradient minimiser, it constitutes the ECEPPGA program. The program appears to be suitable for the conformational prediction of small polypeptides and was validated by successfully locating the global minimum conformation of Met^5-Enkephalin.

For instructions on the use of ECEPPGA, the program manual presented in Stephens's dissertation[120] can be consulted. Details are provided there concerning installation, compilation and execution of the program. A description of the various input and output files is also supplied. Algorithms were developed in a *UNIX* environment: the ECEPP/3 algorithm[l] was written in Fortran 77, and the GA extension in C. Thus, portability of the program is warranted for most UNIX-compatible machines. A distributed version of ECEPPGA is included with Stephens's dissertation.[m][120]

2.2.1.1. ECEPP/3

The ECEPP/3 algorithm[51] is a subset of a larger MM programming package called ECEPPAK.[234] Some of the routines available in ECEPPAK are energy evaluation and minimisation (for single and multiple input conformations), conformational searching (systematic and MC), energy mapping, and calculation of root mean square deviations.

[l] ECEPP/3 is a program that reads in a single conformation, generates co-ordinates for the conformation, and then evaluates or minimises its energy using the ECEPP/3 force field.

[m] The source code for ECEPPGA is available upon request from the Chemistry Department at the University of the Witwatersrand.

Modelling options such as the application of distance constraints, the specification of variable dihedral angles, and the selection of sampling regions are also supplied.

UCEPP/3[51] is the most recent version of the ECEPP (Empirical Conformational Energy Program for Peptides) program[49] which has been widely used for computational energy calculations on polypeptides and proteins. The ECEPP atomic force field parameters are derived from structural data and CNDO/2 (ON) molecular orbital calculations,[n] and have been appropriately modified[50,250] as new experimental information has become available. The latest update[51] incorporates a revised geometry for the proline residue.

In ECEPP, the total conformational potential energy, E_T, is defined as the sum of the electrostatic energy, E_{ES}, nonbonded energy, E_{NB} (Lennard-Jones 6-12 potential), hydrogen bond energy, E_{HB} (10-12 potential), and torsional energy, E_{TOR}. Rigid geometry is used, *i.e.*, bond lengths and bond angles are fixed at experimental values, while dihedral angles in the backbone (ϕ, ψ, ω) and in the side-chains (χ's) are variable. The ECEPP formulation utilises the concept that intraresidue interactions play a dominant, but not exclusive, role in determining the conformation of a polypeptide.

2.2.1.2. The genetic algorithm

The standard GA method for global optimisation has been described in section 1.1.3.3 of this thesis. To develop an appropriate GA for the optimisation of small polypeptides, Stephens[120] modified the standard GA. Among the strategies added were the generation gap (ensures the survival of the best individuals from generation to generation), crowding (reduces the number of similar breeding individuals, thereby ensuring genetic diversity), distance-biased breeding (biases breeding towards, or away from, conformational similarity), family or niche breeding (simulates biological speciation by permitting breeding only within a family), template-based breeding (simulates speciation by permitting breeding only amongst "compatible" individuals), steady state population (allows newly-born individuals into the current breeding stock), generational population (forbids newly-born individuals from reproducing immediately) and crossover and mutation restrictions (limits sampling to specific regions of conformational

[n] CNDO/2 (Complete Neglect of Differential Overlap)[249] is a semi-empirical self-consistent field molecular orbital method. In ECEPP, it is used to determine the overlap normalised (ON) partial atomic charges of each amino acid residue.

space). A unique modification is the characterisation of chromosomes or genomes[o] as series of torsional angles instead of bit strings. Single gene units are extended to accommodate several torsional angles.

Parameter settings that dictate strategy choice and thus behaviour of the GA are contained in a main input file called `param.in`. An example file is reproduced in Appendix A. Specifications in this obligatory file control the use of optional input files which provide supporting data for operation of the GA.

For further details on the strategies available in the modified GA, as well as an overview of the methodology, Stephens's dissertation[120] should be consulted.

2.2.1.3. Interface with ECEPP/3

The GA was written as an add-in module to the ECEPP/3 program.[51] It can thus be considered as a supplementary option available to users of the ECEPPAK[234] package. It has been closely integrated with the main routines of ECEPPAK and can be invoked in the same way that the other peptide modelling options, such as local minimisation and MC searching, are invoked. The flow diagram in Figure 2.1 shows the interface between the GA global minimiser and ECEPP/3. SUMSL (Secant Unconstrained Minimisation Solver)[251] is one of the local minimisers incorporated into the standard ECEPP program.[p]

[o] The terms "chromosome" and "genome" are considered equivalent for the present purposes.

[p] The other available local minimiser is SMSNO which uses numerical gradients as opposed to the analytical gradients used by SUMSL.

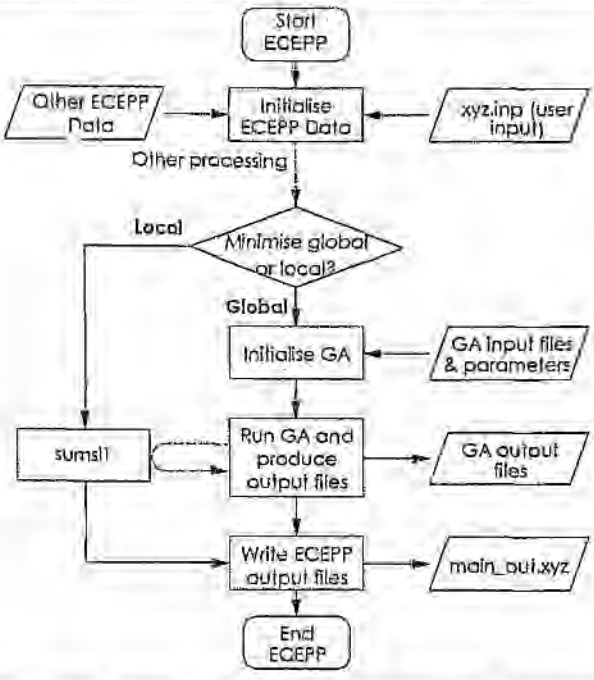


Figure 2.1: Flow diagram of the interface between the GA and ECEPP/3[120]

2.2.1.4. Optimisation of the genetic algorithm

Since the GA was designed to be highly customisable, the algorithm comprises a large set of input parameters that require optimisation with every new application. To evaluate the performance of the GA when searching the conformational space of the pentapeptide Met^5-Enkephalin, Stephens[120] sought optimal values for the following parameters: crossover rate (the probability of a crossover occurring at each gene on the genome during mating), mutation rate (the probability of a point mutation occurring at each gene), selective pressure (*via* the biased breeding rate[q]), population size (the number of offspring generated during each generation) and breeding strategy.[r] It was established that, of these parameters, population size and breeding strategy exert the most influence on performance.

[q] The "biased breeding rate" or "fitness to breeding rate relationship" biases random selection of a breeding pair from a population according to rank or fitness. The relationship can be uniform, linear, or logarithmic. The slope of the breeding rate is controlled by the "breeding rate parameter".

[r] Four breeding or mating strategies are available as options in the program: simple breeding, distance-biased breeding, family breeding, and template- and mask-controlled breeding.

2.2.2. Distance-constraints modelling

A partial solution to the problem of combinatorial explosion is to reduce the scope of a conformational search by imposing constraints on the sequence under investigation (*cf.* section 1.1.3.3). However, an alternative use for distance-constraints has been proposed by Tobias and Brooks.[110] In their investigation of the folding and unfolding of one turn of an α-helix in tripeptides of one type of amino acid, they subjected the peptides to a series of MD simulations in water with an end-to-end hydrogen bond distance constrained to a sequence of different values. This distance was defined as the distance between two atoms which form a particular hydrogen bond and was called the one-dimensional "reaction co-ordinate". It was discovered that, at small values of this co-ordinate, the peptides form α-helices while at large values the peptides are extended, as might be expected. To evaluate conformational equilibria, continuous free energy surfaces of the peptides as a function of the reaction co-ordinate were computed with the aid of the "umbrella sampling"[9] technique. This modelling strategy, encompassing both distance-constraints and "umbrella sampling", has been further employed in MD calculations of peptides in various solvents.[253,254]

The technique of applying distance-constraints to peptide sequences as a means of detecting conformational preferences was developed independently in the present work. Energetically optimal signal sequence conformations were sought as a function of oligomer compactness, the latter being controlled by distance-constraints. Since SP conformation *in vivo* is influenced by environmental (aqueous and lipidic) and steric conditions, it was thought that the imposition of distance-constraints within a SP might simulate these conditions where the preference for a particular atom-to-atom distance by the hydrophobic region of a signal sequence might enable it to be favourably located within a lipid membrane, as evidenced by a significant energy minimum. The technique also mimics the behaviour of signal sequences as they are in nature, *viz.*, with the one end restrained and attached to its nascent chain.

Distance-constraints modelling may thus be capable of performing several functions: it places restrictions on conformational space and thereby provides a computationally manageable number of structures; it can relate peptide conformational preference to molecule compactness; and it may simulate *in vivo* peptide behaviour.

[9] "Umbrella sampling" is a specialised conformational sampling technique that biases sampling toward a desired range of the reaction co-ordinate by adding an auxiliary harmonic potential to the energy of the system.[252]

A constraints option available in the ECEPPAK modelling package was used to model distance-constraints in this study. Constraints in the ECEPPAK program are enforced by a pseudo-potential which is included in E_T and which is conformationally dependent. Distance-constraint sets are defined in the main ECEPP/3 set-up file (`xyz.inp`),[1] along with other user-specified inputs such as calculation type, force field features, minimisation details, peptide sequence, and initial conformation variables (dihedral angles). Another file, `bounds.xyz` must also be generated. This file contains information pertaining to each constraint, *viz.*, the atoms defining the constraint, upper and lower bounds, and a weighting factor for the energy term.

[1] "xyz" can be replaced by any valid file name. An example set-up file is reproduced in Appendix A.

CHAPTER 3
EXPERIMENTAL METHODS

3.1. Signal peptide (SP) systems studied

One way of examining the correlation between SP structure and function is to study the effect that alterations in SP sequence have on translocation abilities. Thus, several SP systems which comprise a wild-type (WT) peptide and at least one mutated peptide have been reported in the literature. The reports often contain useful sequence information such as α-helical content, hydrophobicity and secondary structure prediction. Some of the more recently reported systems are those belonging to the secretory proteins of LamB,[173,192] CPY,[204] OmpA,[179,255] human lysozyme,[198] PhoE,[256] gC[257] and PhoA.[194]

The SP systems belonging to LamB, CPY and gC were investigated in this study. The criteria for selecting these particular systems were the availability of quantitative results regarding translocation efficiencies of system peptides and the diversity in efficiencies exhibited by a wide range of system mutants.

A list of abbreviations for amino acids pertinent to this thesis can be found on page xi.

3.1.1. LamB

LamB, also known as λ phage receptor or maltoporin, is an *Escherichia coli* outer membrane protein. It facilitates the passage of maltose and maltodextrins through the bacterial outer membrane and serves as the receptor for bacteriophage λ.[258] LamB has been well characterised by genetic research and its signal sequence is the most extensively used model for protein export. Export and conformational studies of the LamB wild-type signal sequence and mutant strains have been detailed in Chapter 1. To summarise, there exists a relationship between translocation efficiency, sequence hydrophobicity and secondary structure conformation. Furthermore, α-helix formation appears to be necessary for the proper functioning of these signal sequences.

The WT and mutant LamB signal sequences chosen for this study, together with their translocation efficiencies, are listed in Figure 3.1. Mutant Δ78 contains a deletion of four consecutive residues in the

h-core when compared to the WT, and does not export LamB. Export is restored in two revertants of this mutation, Δ78r1 and Δ78r2. In the revertant Δ78r1, the glycine residue is mutated to cysteine at position 13, while in revertant Δ78r2, the proline residue in position 9 is replaced by leucine. The fourth sequence, A13D, is mutated from the WT; it contains a replacement of alanine by aspartate at position 13. The hydrophobic regions in the WT and A13D peptides are seven residues long, while those in the deletion mutants are reduced to three residues in length.[173]

	1 5 10 15 20 25	trans[a] %
WT	met-met-ile-thr-leu-arg-lys-leu-pro-leu-ala-val-ala-val-ala-ala-gly-val-met-ser-ala-gln-ala-met-ala/	100
Δ78r2	met-met-ile-thr-leu-arg-lys-leu-leu – – – – val-ala-ala-gly-val-met-ser-ala-gln-ala-met-ala/	90
Δ78r1	met-met-ile-thr-leu-arg-lys-leu-pro – – – – val-ala-ala-cys-val-met-ser-ala-gln-ala-met-ala/	50
A13D	met-met-ile-thr-leu-arg-lys-leu-pro-leu-ala-val-asp-val-ala-ala-gly-val-met-ser-ala-gln-ala-met-ala/	10
Δ78	met-met-ile-thr-leu-arg-lys-leu-pro – – – – val-ala-ala-gly-val-met-ser-ala-gln-ala-met-ala/	0

[a] translocation efficiency or *in vivo* activity; values from McKnight *et al.*[173]

Figure 3.1: Aligned amino acid sequences of the LamB system signal peptides.
***Numbers* indicate amino acid sequence number from the *N*-terminus; *dashes* represent deleted residues; *slashes* represent cleavage sites; *underlined residues* indicate the h-core.[173]**

3.1.2. The CPY system

The second SP system selected consists of the WT and mutated signal sequences of the vacuolar protein carboxypeptidase Y (CPY), which occurs in *Saccharomyces cerevisiae* (yeast). The WT signal sequence of CPY is an example of a SP that is not functionally equivalent in different organisms; it functions efficiently in yeast, but not in mammals.[259] After having subjected the WT to various point mutations, Bird and co-workers[204] discovered that the resultant mutants were able to translocate CPY across mammalian membranes. In addition to translocation efficiencies, they reported hydrophobicities and predicted secondary structures of the mutants using statistical procedures. Their results implied a direct relation between hydrophobicity and export activity, and suggested that the functional mutant peptides are more structured than the non-functional CPY signal sequence around the h-core. All sequences were predicted to adopt either an α-helical or a β-strand conformation, the functional peptides displaying higher tendencies to form such conformations.

The signal peptides of CPY which were considered in this analysis, together with their translocation efficiencies, are given in Figure 3.2. When the hydrophobicity of the WT sequence is increased by mutating the glycine residue in position 12 to leucine, resulting in mutant CPYm2, activity in mammalian cells is improved considerably. This level of activity is maintained when the length of the WT h-core is shortened by deleting glycine and when additionally, the remaining glycine in position 10 is replaced with leucine: CPYm6. However, the shortened h-region in CPYm12, which results from the deletion of glycine at position 10, displays a relatively lower export activity. A lower activity is also exhibited by CPYm8, in which the same glycine is substituted for alanine.[204]

		trans[a] %
	1 5 10 15	
CPYm6	met-lys-ala-phe-thr-ser-leu-leu-cys-leu-leu – leu-ser-thr-thr-leu-ala-lys-ala/	97
	1 5 10 15 20	
CPYm2	met-lys-ala-phe-thr-ser-leu-leu-cys-gly-leu-leu-leu-ser-thr-thr-leu-ala-lys-ala/	94
	1 5 10 15 20	
CPYm8	met-lys-ala-phe-thr-ser-leu-leu-cys-ala-leu-gly-leu-ser-thr-thr-leu-ala-lys-ala/	27
	1 5 10 15	
CPYm12	met-lys-ala-phe-thr-ser-leu-leu-cys – leu-gly-leu-ser-thr-thr-leu-ala-lys-ala/	22
	1 5 10 15 20	
WT	met-lys-ala-phe-thr-ser-leu-leu-cys-gly-leu-gly-leu-ser-thr-thr-leu-ala-lys-ala/	undetectable

[a] translocation efficiency or *in vivo* mammalian activity; values from Ford *et al.*[204]

Figure 3.2: Aligned amino acid sequences of the CPY system signal peptides.
***Numbers* indicate amino acid sequence number from the N-terminus; *dashes* represent deleted residues; *slashes* represent cleavage sites; *underlined residues* indicate the h-core.[204]**

3.1.3. The gC system

Glycoprotein C (gC) is a eukaryotic protein encoded by a swine herpesvirus. Its signal sequence has been genetically analysed by Ryan and Edwards[257] in an effort to delineate the conformational constraints experienced by a eukaryotic signal sequence. They systematically introduced proline, a known α-helix breaker, into different positions in the sequence's h-core and then evaluated the export competencies of the resulting mutants. It was concluded that proline affects SP function by both reducing overall hydrophobicity and interrupting secondary structure, and that some positions within the WT α-helix are more susceptible to proline disruption than others (functional asymmetry in the α-helix was detected). Figure 3.3 depicts the gC signal sequences examined in the current work.

It can be noted from Figure 3.3 that the deletion of a single residue in the WT h-core results in a mutant strain ΔA10 which is still functionally efficient and that, as proline is substituted further down the h-region from A10P to L14P, *in vivo* activity decreases significantly.

		trans[a] %
	1 5 10 15 20	
WT	met-ala-ser-leu-ala-arg-ala-met-leu-ala-leu-leu-ala-leu-tyr-ala-ala-ala-ile-ala-ala-ala/	100
	1 5 10 15 20	
ΔA10	met-ala-ser-leu-ala-arg-ala-met-leu - leu-leu-ala-leu-tyr-ala-ala-ala-ile-ala-ala-ala/	99
	1 5 10 15 20	
A10P	met-ala-ser-leu-ala-arg-ala-met-leu-pro-leu-leu-ala-leu-tyr-ala-ala-ala-ile-ala-ala-ala/	98
	1 5 10 15 20	
L12P	met-ala-ser-leu-ala-arg-ala-met-leu-ala-leu-pro-ala-leu-tyr-ala-ala-ala-ile-ala-ala-ala/	19
	1 5 10 15 20	
L14P	met-ala-ser-leu-ala-arg-ala-met-leu-ala-leu-leu-ala-pro-tyr-ala-ala-ala-ile-ala-ala-ala/	5

[a] translocation efficiency or *in vivo* activity; values from Ryan and Edwards[257]

Figure 3.3: Aligned amino acid sequences of the gC system signal peptides.
***Numbers* indicate amino acid sequence number from the N-terminus; *dashes* represent deleted residues; *slashes* represent cleavage sites; *underlined residues* indicate the h-core.[257]**

3.2. Knowledge-based modelling

Access to the knowledge-based prediction programs described in the preceding chapter was gained through the Netscape (ver. 3.01) browser. Chou-Fasman, hydropathy, TMpred and PSA plots were graphed with the MS EXCEL (ver. 5.0) program, belonging to the MS OFFICE (ver. 4.2) suite of programs.

Chou-Fasman conformation profiles and hydropathy profiles were computed using a window size of five amino acid residues with the following unequal linear weighting: 1 2 3 2 1. Since ProtScale (*cf.* Chapter 2) requires that a peptide sequence be at least 25 residues in length for calculation purposes, the first six amino acids of the N-terminus of the relevant mature proteins were included in the profiles. Otherwise, the entire lengths of the signal sequences, as listed in the figures above, were submitted to prediction servers for analysis.

3.3. Molecular modelling

MM calculations were conducted at various stages of this project on several UNIX-based computers: two Silicon Graphics workstations (R3000 Indigo and R4000 Iris); a Sun workstation (SPARCcenter 2000); and a Cray computer (CS-64100). Besides these machines, MS-DOS based personal computers were also used in analysing results.

The SYBYL (ver. 6.0)[233] and ECEPPAK[234] molecular modelling packages were employed for systematic conformational searching; and ECEPPGA[120] was employed for genetic algorithm computations. Ramachandran (ϕ, ψ) maps and other relevant plots were generated with MS EXCEL. HyperChem (ver. 4.5)[260] was used to view 3-D polypeptide structures and to produce α-helical wheel plots and stereoview diagrams.

3.3.1. Systematic conformational searches

In the initial stages of this project, the only conformational s[illegible] at our disposal were the systematic and grid (systematic search combined with [illegible] imisation) searches supplied with SYBYL.[233] Preliminary investigations of sign[illegible] ed both fragment-based (model building) and torsional angle-based approaches[a] on [illegible] Although various strategies were adopted to contend with the unavoidable combinato[illegible] sion problem, e.g., limiting the length of the SPs, imposing geometrical constraints (α-helix or β-[illegible]d) on the h-cores, discarding conformations that violate specified energetic and geometric criteria, and biasing the search towards low-energy regions of conformational space, the deterministic searches proved to be extremely time-consuming and consequently too crude for the present application. Indeed, it was found that proper searches could only be conducted for a dipeptide. Thus, a detailed discussion of this preliminary portion of the work will be excluded from the thesis.

Since ECEPPAK[234] also incorporates a systematic searching method for global optimisation, it was decided to conduct a comparative study to that performed with the SYBYL systematic search. The study was confined to two of the LamB system signal sequences: the active Δ78r2 and the inactive

[a] During conformational searching, only ϕ and φ torsional angle values were scanned; ω angles were fixed in the *trans* position, except for proline (both *trans* and *cis* assumed). Various grids were explored. Conformational energies were minimised using the TRIPOS force field and conjugate gradients. Charges were excluded and solvent effects ignored.

Δ78. As with the SYBYL systematic searches, several strategies were used to overcome the problem of combinatorial explosion. The focus of the experimental procedure was the application of distance-constraints. Each sequence was constrained between two chosen points (the N-atom of leu-8 and the C-atom of val-18, *cf.* Figure 3.1) by a defined distance, which was varied systematically during conformational searching. These points were either the two ends of the h-core, or were such that the h-core was centred. *Rotations of flanking residues beyond the fixed points were therefore permitted to be* flexible. The upper and lower bounds of these distances were selected from preliminary minimisation results where sequences were constrained in the extreme conformations of α-helical and extended.

Several different sets of experiments were performed, the sets following a generalised scheme:

1. Eleven residues from each sequence were chosen for simulation; blocking end groups capped the sequences during computations (ACE or acetyl at the N-terminus and NME or methyl amide at the C-terminus).
2. Only the ϕ and ψ torsional angles of the seven central residues were subjected to conformational searching (in accord with the MM calculations of Perez *et al.*[206]); all other torsions remained constant. Peptide bonds were held fixed at 180°, except for the one preceding the pyrrolidine ring of proline (0° and 180°). The ϕ dihedral angle of proline was also fixed at ECEPP default values appropriate to both puckered forms ('exo' or 'up', and 'endo' or 'down')[51] of the prolyl ring.
3. Values for the fixed side-chain torsions were determined either by extensive minimisation from initial ideal α-helix and β-strand conformations, or from the literature.[70]
4. The (ϕ, ψ) dihedral pairs were confined to either the A or the E region of Zimmerman conformational space.[v] An initial angular interval of 40° was used. At later stages of optimisation, the interval was refined in regions around selected low-energy conformers.
5. Simultaneous searching of all the variable torsions proved to be too compute-intensive and the number of torsions varied during a single search was limited to either six or eight.
6. Each conformational search was followed by energy minimisation of the fifty lowest-energy structures found. Gradient minimisation, limited to 1000 iterations, was executed with the ECEPP/3 force field and all torsions were allowed to vary. ECEPP default values, including a dielectric constant of 2, were used for all other relevant parameters.

[v] Zimmerman *et al.*[69] use conformational letter codes to define regions of the (ϕ, ψ) map. A ($-110° \le \phi < -40°$; $-70° \le \psi < -10°$) denotes the region for the right-handed α-helix; E ($-180° \le \phi < -110°$, $-180° \le \psi < -140°$ and $110° \le \psi < 180°$) denotes the extended conformational state.

7. After minimisation, the conformational space of the lowest-energy structure was searched further; different torsions to those used previously were scanned.
8. The process of searching and minimising was continued until all torsions (within the above limitations) of the seven residues were explored.

3.3.2. ECEPPGA

The following experimental conditions were applied to ECEPPGA computations which cover both optimisation of the GA and conformational searching. Conditions particular to either GA optimisation or searching will be described later.

The specific sequences of amino acids of the LamB, CPY and gC wild-type and mutated leader peptides which were selected for GA modelling are listed in Table 3.1. Residue numbering is as shown in Figures 3.1, 3.2 and 3.3. Only those lengths of the signal sequences which were considered essential for conformational analysis were probed in order to reduce the combinatorial problem. It was thus assumed that the remainder of a particular sequence would affect the conformations of the resultant low-energy structures in a common fashion. The choice of amino acids centralised the hydrophobic h-region within each peptide segment. End groups ACE (amino-$COCH_3$) and NME (carboxyl-$NHCH_3$) were again used to block the sequences.

The distance-constraint modelling strategy was again the focal point of the experiments, its application being analogous to that of the ECEPPAK systematic searches. Table 3.1 also records the end points in each peptide which defined distance-constraints.

In this GA implementation, the conformations of each SP portion were represented by genomes which consisted of all the torsional angles in the conformation, *viz.*, each gene on the genome represented an angle value for a single torsion. Conformational space was explored by random sampling of these torsional angles during the process of gene mutation. Peptide structure was described with two identity templates,[w] a residue identity template and an angle identity mask. The former defines the number of encoded residues of a genome, the type (corresponding to ECEPP residue identity numbers) and order of these residues, and the number of torsional angles in each.[x] The latter contains one character for

[w] A "template" or "mask" is either a character or a string of characters which encodes mating information. Bitstring and phenotype-linked (alphanumeric characters) templates and masks are supported in ECEPPGA.

[x] The residue identity template resides in the main GA input file, `param.in` and represents exactly the ECEPP/3 reference conformation described in the accompanying ECEPP/3 main input file, `xyz.inp`.

each torsional angle position on a genome and identifies torsional angle type. Characters are available for the following angles: arbitrary type, ϕ, ψ, ω, χ^1, χ^2, χ^3, χ^4, χ^5, χ^6, χ^7, χ^8.

Table 3.1: Sequences and distance-constraint end points selected for GA modelling

Signal peptide	Selected sequence	Number of residues in selected sequence	Distance-constraint end points[a]	Number of distance-constrained residues
LamB				
WT	arg-6 to ser-20	15	leu-10 to ala-16	7
Δ78r2	arg-6 to ser-16	11	leu-8 to val-14	7
Δ78r1	arg-6 to ser-16	11	leu-8 to val-14	7
A13D	arg-6 to ser-20	15	leu-10 to ala-16	7
Δ78	arg-6 to ser-16	11	leu-8 to val-14	7
CPY				
CPYm6	phe-4 to thr-15	12	ser-6 to ser-13	8
CPYm2	phe-4 to thr-16	13	ser-6 to ser-14	9
CPYm8	phe-4 to thr-16	13	ser-6 to ser-14	9
CPYm12	phe-4 to thr-15	12	ser-6 to ser-13	8
WT	phe-4 to thr-16	13	ser-6 to ser-14	9
gC				
WT	arg-6 to ala-16	11	ala-7 to tyr-15	9
ΔA10	arg-6 to ala-15	10	ala-7 to tyr-14	8
A10P	arg-6 to ala-16	11	ala-7 to tyr-15	9
L12P	arg-6 to ala-16	11	ala-7 to tyr-15	9
L14P	arg-6 to ala-16	11	ala-7 to tyr-15	9

[a] from the *N*-atom of the first residue to the C-atom of the second

Individual peptides were subjected to a series of runs dictated by a sequence of distance-constraint values. A run is defined as the creation of an initial population, followed by the GA process iterated over a number of user-defined generations, and ends with the attainment of a final population. Details of significant modelling strategies used during the runs are given below.

a. An initial population of conformations was generated in a random fashion by mutations of the genome of a reference conformation. These gene mutations were non-local;[y] they occurred in a uniform distribution over the entire conformational space.

b. Two breeding or mating strategies were used to select a mating pair from a population: simple (random) and template-based. The latter will be discussed later (see section 3.3.2.2).

c. Mating strategies were biased towards higher ranking individuals (see (i), below) with the aid of selective pressure, *i.e.*, the relative breeding rates of individuals were controlled according to fitness.

d. A new generation was produced *via* local[z] point mutation of genes on the parent genomes, random crossover at each gene during mating, and generation gaps.

e. Individual torsions were considered as sampling units[aa] during gene mutation. Although a completely random torsional angle sampling strategy was adopted, implying that all torsions of the parent genomes were eligible for sampling, there were some rotations that were forced to remain at fixed values. These were the rotations about all the peptide bonds (the *trans* form was assumed and the rotations were therefore kept constant at a value ω=180°), and the ϕ rotation of the proline residue (the 'down' puckered form of the pyrrolidine ring was assumed and the ϕ value was fixed at -68.780°). The GA distinguished between variable and fixed torsions *via* a fixed angle mask. The mask assigns either a fixed angle or a variable angle character to each torsional angle on the genome.

f. Restricted generation gaps were created, *i.e.*, defined numbers of best individuals were copied unaltered from each generation into the next generation.

g. A new individual was considered viable if its energy was below a certain limit and if it was not too genetically similar to other population individuals. The latter option encompassed a "crowding" or "space sharing" strategy where similari was assessed in terms of a minimum root mean square distance between superimposed conformations.

---

[y] "Non-local mutations" ignore current gene values. Torsional angle sampling is permitted anywhere within allowed regions of the conformational space. In this application, no regions were defined, *i.e.*, no lengths on the genomes were restricted. Thus, the entire (ϕ, ψ) space was explored uniformly.

[z] "Local" mutations are based on current torsional angle values. Random changes, in a Gaussian distribution about the current value are allowed for every torsion.

[aa] "Sampling units" are *n*-dimensional maps of conformational space and "samples" are randomly selected points on the maps. For the present application, a sampling unit is equivalent to a gene and can consist of either an individual torsional angle, a set of backbone angles of each residue with separate side-chain angle sampling units, or a set of all the torsional angles of each residue.

h. A population individual was evaluated according to the minimised, continuous ECEPP/3 potential energy function. Local minimisation with the built-in SUMSL gradient optimiser was implemented prior to each population energy evaluation. All genomes in a population were minimised, all torsions were allowed to vary, the ECEPP default effective dielectric constant of 2 was used (thereby simulating a hydrophobic environment), and the maximum number of iterations per minimisation was restricted. The combination of GA global minimisation with local minimisation throughout the experiments ensured effective exploration of the valleys present on the energy surface (*cf.* section 1.1.3.3).

i. The ranking of an individual was accomplished with a fitness coefficient which was minimised ECEPP/3 energy in this case (other contributing factors to fitness may be family membership, mask string length, *etc.*); conformations with lower energies were ranked higher with respect to fitness.

j. Progress of the GA was monitored by taking snapshots of populations and of population statistics at various generation intervals during the runs.

GA parameter settings for these modelling strategies and for those that that have not been mentioned here, besides those discussed in the succeeding sections 3.3.2.1 and 3.3.2.2, were kept constant throughout the experiments at values recommended by Stephens.[120] A complete GA parameter set, including a concise explanation of every parameter, can be found in the `param.in` file supplied in the Appendix.

3.3.2.1. Optimisation of the performance of ECEPPGA

In an attempt to optimise the performance of ECEPPGA with respect to the conformational analysis of SPs, parameters which were thought to affect GA efficiency were examined using the three deletion mutants of the LamB signal sequence with their very different translocation efficiencies. Parameters investigated were generation gap size, maximum number of iterations during local minimisation, number of generations in each run, population size, selective pressure and breeding strategy. The variables were altered systematically for each sequence and their effects on performance analysed. Remaining variables were held constant.

For each value of the variable parameter, three separate runs with differing random seed values were performed. These three runs constitute a job.

3.3.2.2. ECEPPGA conformational searches

In searching the potential energy surfaces of the signal sequences, no knowledge of their native structures or GMECs was assumed. The lowest-energy structure from each run was postulated to be the GMEC. At times, final conformational structures which were considered dubious because of their relatively high energies were obtained. In such cases, the seed value of the GA random number generator was changed, and the runs were repeated, generating a revised conformation.

Apart from the modelling strategies specified above, the following optimised parameter choices, obtained from the GA optimisation experiments, were adopted:

- a generation gap size of 120
- 50 iterations per local minimisation
- 15 generations per run
- a population size of 400
- a linearly biased breeding rate with a high selective pressure: the probability of a genome being selected as a parent was linearly related to fitness or rank, with the gradient set at 0.9 (0 represents the highest rank and 1 represents the lowest)
- the template- and mask-controlled breeding strategy. Bitstring templates and masks[bb] of 3 characters in length controlled the selection of mating pairs. A pair was selected if the *mask* string of one parent matched the template string of the other, and *vice versa*.

Although the majority of the conformational searching runs were carried out under the experimental conditions already described, five sets of runs adopting slightly altered conditions were also conducted. These were (1) runs with restricted torsional angle space, (2) runs with shorter calculated peptide lengths, (3) runs with differing proline configurations, (4) runs with varying dielectric constants, and (5) runs without distance-constraints. Motives for performing these additional calculations will be propounded in the following chapter. Experimental procedures for each set will also be described.

[bb] "Bitstring templates" and "bitstring masks" are assigned randomly to each individual of the initial population and then passed down the generations through inheritance. Each character is considered to be one gene. Each gene undergoes non-local mutation and complete crossover (the entire gene at the crossover site is inherited from the alternative parent).

Results from the GA conformational searches were graphed as minimised energy *versus* distance-constraint curves. The curves were analysed and compared in each SP system. Ramachandran (ϕ, ψ) plots were employed to assess the α-helical nature of the lowest-energy structures, and helical wheels were plotted and to assess the amphipathicity of the helices.

CHAPTER 4
RESULTS AND DISCUSSION

4.1. Knowledge-based modelling

4.1.1. Globular protein-based predictions

Attempts by several researchers[179,190,203,204] to relate signal sequence activity to secondary structure, as predicted using the Chou-Fasman method,[22] has prompted an equivalent investigation in this work. Predictions obtained for the three SP systems, utilising the C-F method and the Consensus procedure, are reported below. It was anticipated that these prediction techniques might not furnish useful results for this application since they are derived from the analyses of globular proteins.

4.1.1.1. Chou-Fasman

C-F profiles for signal sequences of the LamB, CPY and gC signal peptide systems are shown in Figures 4.1, 4.2 and 4.3, respectively. The vertical axis in each profile is the probability scale; a probability of 1.0 signifies a 50% possibility for either the α-helical or the β-sheet conformation. The horizontal axis denotes amino acid residue positions in the SPs. Those residues that constitute the central hydrophobic region of each sequence are the ones of interest in evaluating distinctions or similarities among the sequences of a specific system.

- ***LamB***

Because of differences in sequence length in the LamB system (Figure 3.1), only the WT and A13D peptides, and the Δ78r2, Δ78r1 and Δ78 peptides can be compared directly. Predictions indicate probabilities of the WT and A13D h-cores (residues 10 to 16) that suggest the formation of either an α-helix or a β-sheet. As mentioned in the introductory chapter (section 1.2.4.2), these conflicting probabilities are possibly due to the fact the C-F parameters were originally established for soluble proteins and not for short, hydrophobic peptides. Although the profiles of the deletion mutants, Δ78r2, Δ78r1 and Δ78, show that their h-cores (residues 10 to 12) favour the α-helix slightly, no definite conformational preferences can be deduced.

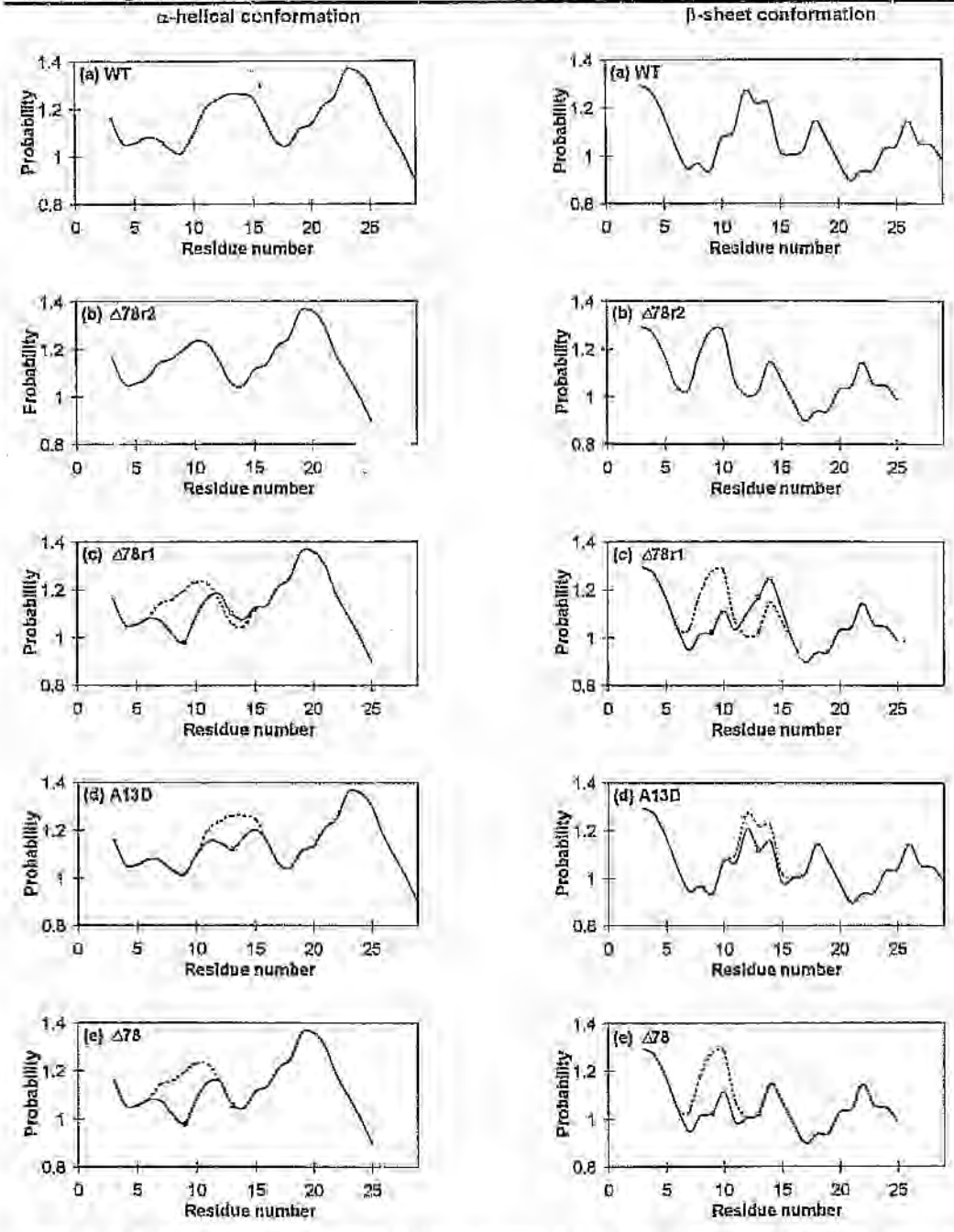


Figure 4.1: Probable secondary structure of the sequences of the LamB signal peptide system determined with the Chou-Fasman[25] prediction method. α-Helical probabilities for the sequences are shown on the left and β-sheet probabilities are shown on the right. A probability of 1.0 signifies a 50% probability for the conformation. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency. Dotted curves in (b) and (d) represent the conformational probabilities for the WT while those in (c) and (e) represent the conformational probabilities for Δ78r2. a = residue number 9, b = residue number 13.

On comparing the WT and A13D sequences in Figure 4.1(d), it can be seen that A13D has a decreased probability of occurring in both the α-helical and β-sheet conformations when compared to the WT, as evidenced by A13D's lower probability values for residues 11 to 15. As expected, the largest difference is at residue 13 (open circle in the figure), where alanine in the WT has been replaced with the less hydrophobic aspartate residue in A13D. Apart from its lower hydrophobic character, aspartate also has lower α-helical and β-sheet tendencies than alanine.

Graphs 4.1(d) and 4.1(e) show that the sequences of Δ78r1 and Δ78 display very similar conformational probabilities. Differences between the curves of these two SPs result from the replacement of cysteine in Δ78r1 at position 13 (open circles on the plots) by glycine; cysteine has slightly higher α-helical and β-sheet tendencies than glycine. The marked distinction between these curves and that of Δ78r2 results from the presence of proline at position 9 (dark circles on the plots) in Δ78r1 and Δ78. Proline has much lower tendencies than the leucine mutation in Δ78r2 to occur in α-helical and β-sheet conformations.

Observations in common with those of Emr and Silhavy[203] are: the α-helix in the WT occurs between proline-9 and glycine-17; the α-helix in Δ78r2 only experiences a disruption at glycine-13; and no distinct α-helix forms in the central region of Δ78 due to the close proximity of proline-9 and glycine-13. Discrepancies between results from these two sets of C-F calculations could arise from different window residue lengths and different decision functions, such as the weighting of values, that may have been used, and which are crucial to the accuracy of the predictions. Emr and Silhavy[203] calculated that the α-helix of the Δ78r1 revertant peptide only experiences a break at proline-9, whereas probability values here predict that no regular α-helix forms in its central region due to the close proximity of proline-9 and cysteine-13 (cysteine is a helix-indifferent residue[25]). The findings reiterate the importance in the C-F calculations of the presence of the helix-breaking residues, proline and glycine,[25] in a sequence.

If calculations for the α-helical conformation only are considered, it would appear that C-F predictions are able to correlate sequence with *in vivo* activity and α-helical content (see Table 4.1). The predictions suggest that the WT has a higher likelihood of forming an α-helix than A13D, and that Δ78r2 has a higher chance of occurring as an α-helix than do Δ78r1 and Δ78.

However, the fact that the C-F predictions for the β-sheet conformation also record high probability values undermines the reliability of this prediction procedure with regard to the current application.

Table 4.1: Summary of the activities and α-helical contents of the LamB signal peptides

Signal peptide	*In vivo* activity[a]	α-Helical content in SDS[b]	α-Helical content in aq. TFE[c]
	%	%	%
WT	100	70	55
Δ78r2	90	75	60
Δ78r1	50	40	40
A13D	10	60	not calculated
Δ78	0	35	30

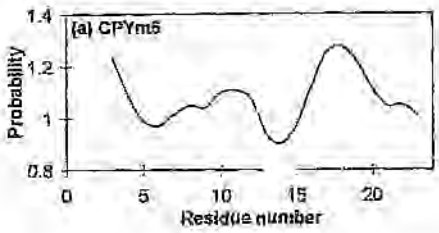
[a] values from McKnight *et al.*[173]
[b] calculated from CD spectra in a membrane-mimetic environment of sodium dodecyl sulphate micelles[173]
[c] calculated from CD spectra in a membrane-mimetic environment of aqueous trifluoroethanol[193]

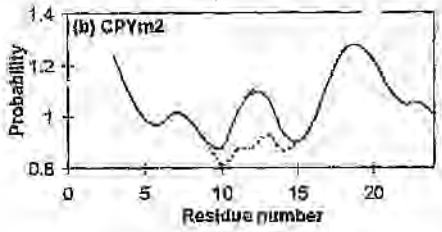
- ***CPY***

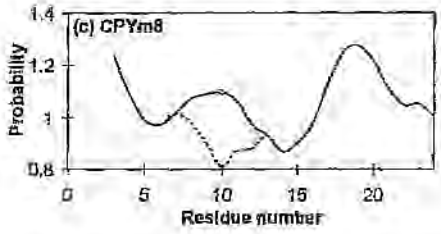
Because of differences in sequence length in the CPY system (Figure 3.2), CPYm6 and CPYm12 prediction curves were analysed separately (see Figure 4.2(d)) from the curves of the remaining sequences. As with the LamB system, the C-F algorithm predicts β-sheet probability values for the residues of the sequences which are comparable to their α-helical probability values. In fact, it would appear that, in general, the β-strand is slightly preferred. These findings concur with C-F predictions of this system calculated by Bird *et al.*[204] who demonstrated that all the sequences were likely to adopt either an α-helical or a β-strand conformation, with an unstructured region around residues 10 to 12 caused by the presence of one or more glycine residues.

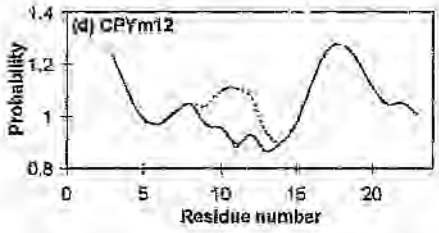
The residues of the CPYm6 h-core (residues 7 to 12) display a higher tendency for the α-helical and β-sheet conformations than those of the CPYm12 h-core (residues 7 to 12). This is shown in Figure 4.2(d). The difference between the CPYm6 and CPYm12 curves is due to the substitution of leucine-11 in CPYm6 with glycine (open circles on plots) in CPYm12; glycine has been proposed to be a strong α-helix breaker and a β-sheet breaker.[25]

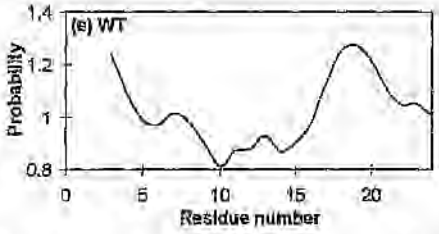
α-helical conformation



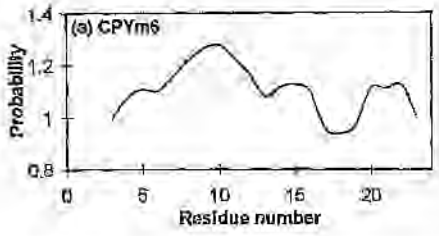


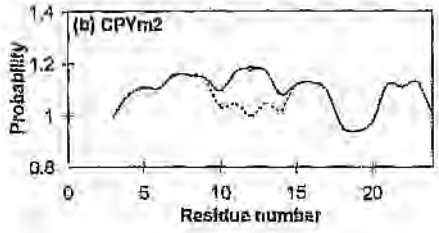


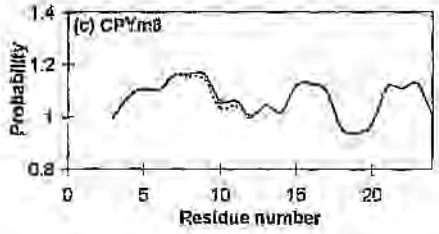


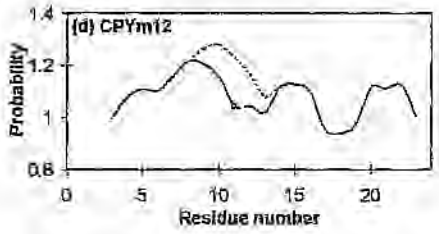


β-sheet conformation









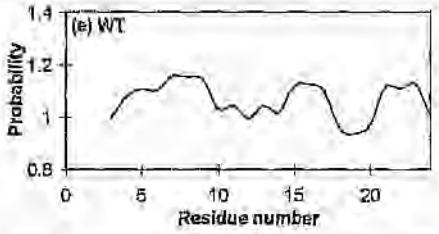


Figure 4.2: Probable secondary structure of the sequences of the CPY signal peptide system determined with the Chou-Fasman[25] prediction method. α-Helical probabilities for the sequences are shown on the left and β-sheet probabilities are shown on the right. A probability of 1.0 signifies a 50% probability for the conformation. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency. Dotted curves in (a), (b) and (c) represent the conformational probabilities for the WT while those in (d) represent the conformational probabilities for CPYm6. o = leu-12 in CPYm2, ala-10 in CPYm8, gly-11 in CPYm12.

A comparison of the CPYm2 (Figure 4.2(b)) and the CPYm8 (Figure 4.2(c)) α-helix probability plots with the WT plot indicates clearly that the two mutants have a higher possibility of forming helices in their hydrophobic regions (residues 7 to 13) than does the WT. The two glycine residues in the h-region of the WT are separated by only one residue, resulting in a low probability for α-helix formation. In CPYm2, the second glycine residue at position 12 of the WT has been replaced by leucine (open circles in figure), and in CPYm8, the first glycine at position 10 of the WT has been replaced by alanine (open circles in figure). Leucine and alanine are both presumed to be strong α-helix formers.[25] The presence of the helix indifferent cysteine-9 adjacent to glycine-10 in the CPYm2 sequence appears to cause a disruption in α-helix formation; this phenomenon is not observed in CPYm8, where cysteine-9 is adjacent to alanine-10.

Similarly to the C-F structural predictions of the Lamb system discussed earlier, α-helix predictions of the CPY system also correlate vaguely with *in vivo* activities (Figure 3.2). But, the high probability values calculated for β-structures once more weakens the reliability of the predictions.

- ***gC***

The effect on probable secondary structure formation induced by the systematic introduction of proline (open circles on the relevant plots) into a sequence can be noted in Figure 4.3. Proline is known to be both a strong α-helix breaker and a strong β-sheet breaker.[25] The probability that the gC sequences will form β-sheets is relatively lower than their probability of forming α-helices.

The h-core (residues 7 to 15) of the WT, whose curve is shown in Figure 4.3(e), displays a high probability of occurring in an α-helix. As proline is substituted for along the WT sequence, from position 10 to 12 to 14, disruption of this helix occurs in these positions. Although α-helix probability curves (a) and (b) can be related to corresponding *in vivo* SP activities (both curves demonstrate a high helix content, which is proposed to be analogous to a high export efficiency), curves (c) to (e) offer no correlation. Activity values can be referred to in Figure 3.3 of Chapter 3.

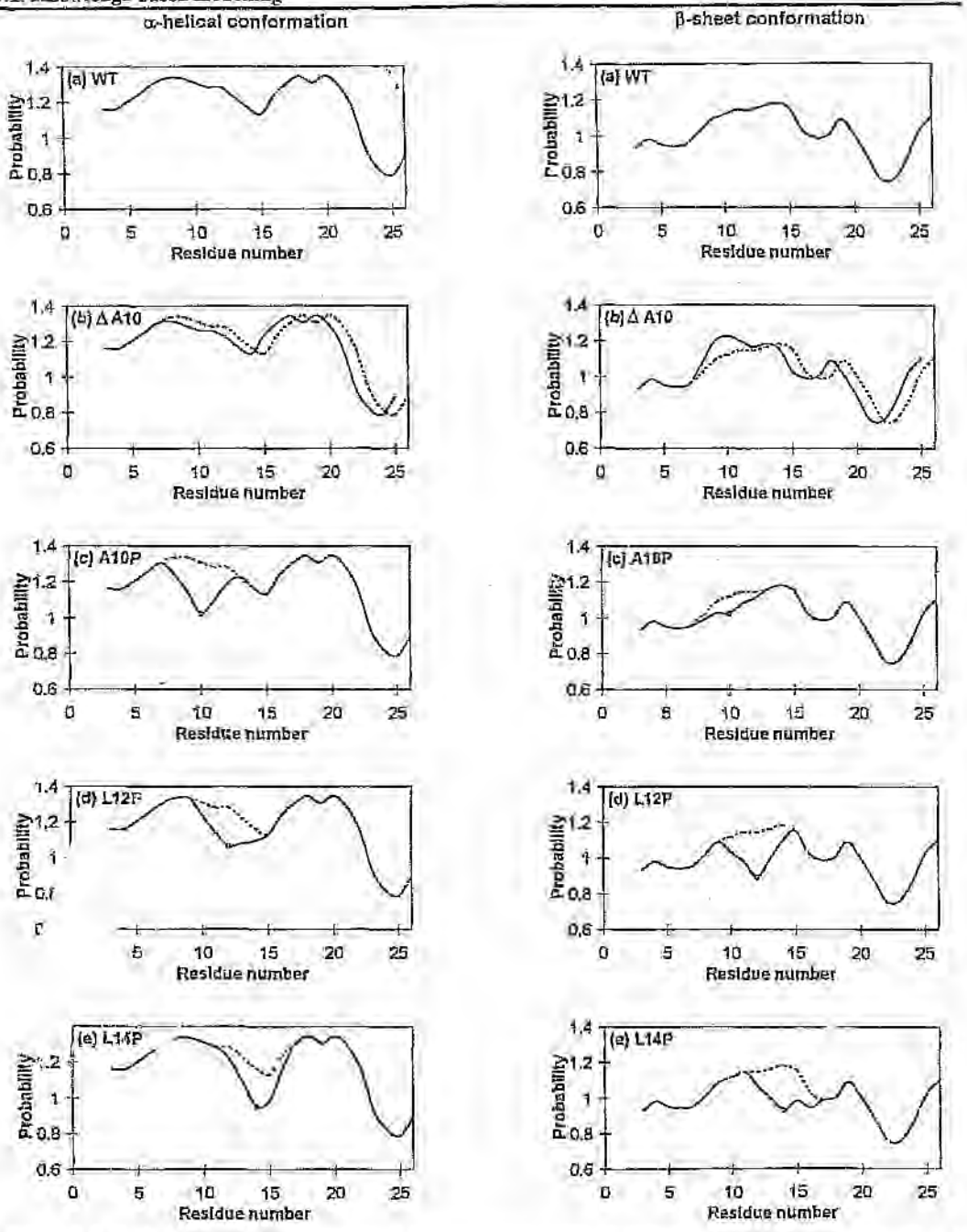


Figure 4.3: Probable secondary structure of the sequences of the gC signal peptide system determined with the Chou-Fasman[25] prediction method. α-Helical probabilities for the sequences are shown on the left and β-sheet probabilities are shown on the right. A probability of 1.0 signifies a 50% probability for the conformation. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency. Dotted curves represent the conformational probabilities for the WT. o = proline position.

4.1.1.2. Consensus procedure

Results from a consensus secondary structure prediction procedure for the LamB, CPY and gC signal peptide systems are given in Tables 4.2, 4.3 and 4.5, respectively. Predictions made by the individual methods are tabulated first, followed by joint predictions made with the consensus procedure. Discussion of the results will focus on the correlation of peptide translocation efficiency with predicted secondary structure. The hydrophobic regions in each sequence are in bold typeface.

- ***LamB***

The Gibrat, Levin and PHDsec methods predict α-helices for the h-regions of all the LamB sequences, with the exception of Δ78 for the Levin method. These methods falter in their ability to forecast functional efficiencies if α-helix content of the h-core is accepted as a measure of translocational efficiency. DPM results predict that residues in the h-regions of the LamB sequences adopt no regular structure; the regions consist of a mixture of α-helical, β-sheet and coil structures. Δ78 is again the exception here, its h-core assuming a completely random coil structure. Although this random coil may explain the functional inactivity of Δ78, the adoption of irregular conformations by the h-cores of the remaining sequences offers no correlation with activity. The failure of the SOPMA method to distinguish between signal sequences based on predicted h-core structure is evidenced by the similar predicted conformations for the h-regions of the WT and A13D sequences and the equivalent predicted conformations for the h-regions of the three deletion mutants.

If it is assumed that both the n-region (~ residues 5 to 9) and the h-region for each sequence are involved in secondary structure formation, the predictive accuracy of some of the above-mentioned methods appears to improve slightly. If the presence of an α-helix in a sequence is considered necessary for functional efficiency, and if conformations other than α-helical are presumed to cause functional inactivity, then the Gibrat and Levin predictions for the deletion mutants are able to account for relative activities. SOPMA seems to be the only method that is capable of assessing the comparative activities of the WT and A13D peptides.

Table 4.2: Secondary structure of the sequences of the LamB signal peptide system determined by joint prediction. Predictions were determined with the Gibrat,[238] Levin,[236] DPM,[237] SOPMA[238,239] and PHDsec[36,240,241] methods. The *Consensus* prediction combines these methods. H = α-helix, E = β-sheet, C = coil, T = turn, – = no prediction, *spaces* represent deleted residues.

Prediction method	Signal peptide[a]	Secondary structure prediction																							
	WT[b]	M	M	I	T	L	R	K	L	P	L	A	V	A	V	A	A	G	V	M	S	A	Q	A	M	A
		1				5					10					15					20					25
Gibrat	WT	H	H	H	H	H	H	H	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	Δ78r2	H	H	H	H	H	H	H	H	H					H	H	H	H	H	H	H	H	H	H	H	H
	Δ78r1	H	H	H	H	H	C	C	C	C					H	H	H	H	H	H	H	H	H	H	H	H
	A13D	H	H	H	H	H	H	H	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	Δ78	H	H	H	H	H	H	C	C	C					H	H	H	H	H	H	H	H	H	H	H	H
		1				5					10					15					20					25
Levin	WT	H	E	H	E	H	C	C	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	C	C	C
	Δ78r2	E	H	C	C	H	H	H	H	H					H	H	H	C	C	H	H	H	C	C	H	H
	Δ78r1	H	H	H	H	H	T	C	C	C					H	H	H	H	H	H	H	H	H	H	H	H
	A13D	H	H	H	H	H	C	C	C	C	H	H	H	H	H	H	H	H	C	H	H	H	H	C	C	H
	Δ78	H	H	H	H	H	H	C	C	C					T	H	T	H	C	C	H	H	H	H	H	H
		1				5					10					15					20					25
DPM	WT	C	C	E	E	E	H	H	C	C	H	H	E	H	E	H	H	C	E	H	H	H	H	H	C	C
	Δ78r2	C	C	E	E	E	E	H	H	H					E	H	H	C	E	H	C	H	H	H	C	C
	Δ78r1	C	C	E	E	E	C	C	C	C					E	H	H	E	E	H	C	H	H	H	C	C
	A13D	C	C	E	E	E	C	C	C	C	C	H	E	C	E	H	H	C	E	H	C	H	H	H	C	C
	Δ78	C	C	E	E	C	C	C	C	C					C	C	C	C	E	H	C	H	H	H	C	C
		1				5					10					15					20					25
SOPMA	WT	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	T	T	H	H	H	H	H	H	H	H
	Δ78r2	C	C	E	E	C	C	C	C	C					C	T	T	C	C	C	H	H	H	H	H	H
	Δ78r1	E	H	C	E	T	T	T	C	C					C	T	T	E	E	E	E	C	C	T	H	H
	A13D	H	H	H	E	E	C	C	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	Δ78	H	C	C	E	E	E	C	C	C					C	T	T	T	C	C	H	H	H	H	H	H
		1				5					10					15					20					25
PHDsec	WT	C	H	H	H	H	H	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	Δ78r2	C	H	H	H	H	H	H	H	H					H	H	H	H	H	H	H	H	H	H	H	C
	Δ78r1	C	H	H	H	H	H	C	C	C					H	H	H	H	H	H	H	H	H	H	H	C
	A13D	C	H	H	H	H	H	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	Δ78	C	H	H	H	H	H	C	C	H					H	H	H	H	H	H	H	H	H	H	H	C
		1				5					10					15					20					25
Consensus	WT	H	H	H	H	H	H	H	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	Δ78r2	C	H	–	–	H	H	H	H	H					H	H	H	C	–	H	H	H	H	H	H	H
	Δ78r1	–	H	H	H	H	–	C	C	C					H	H	H	H	H	H	H	H	H	H	H	H
	A13D	H	H	H	H	H	C	C	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	Δ78	H	H	H	H	H	H	C	C	C					–	H	–	H	–	H	H	H	H	H	H	H

[a] for each prediction method, the signal peptides are arranged in order of decreasing translocation efficiency

[b] the sequence of the WT is listed as a reference; a list of amino acid residue abbreviations is available on page xi of this thesis; residues in the h-core are underlined

Assuming that the α-helical secondary structure is a determinant of *in vivo* signal sequence translocation function, and assuming that it is the structure of the h-regions of the sequences that is significant, joint prediction, *i.e.*, the consensus procedure, fails to provide a correspondence between activity and structure; all h-regions, besides that of Δ78, are predicted to form α-helices. If both the n-region and the h-region are implicated in secondary structure formation, joint prediction seems to enhance predictive accuracy since a clear correspondence between activity and α-helix content can be noted; the α-helical form of the n- and h-regions decreases with decreasing activity. Thus, in this case, although the majority of the prediction methods individually yield inconclusive results when applied in isolation, joint analysis with the Consensus procedure is able to produce results that are more consistent with the literature.

- *CPY*

The CPY system was submitted for analysis at a later date than those of LamB and gC. This resulted in the unavailability of the PHDsec prediction method as the method had since been removed from the joint prediction procedure (due to its lengthy calculation time). Thus, the Consensus results reported in Table 4.3 incorporate only four predictive methods. A separate submission of the CPY system was made to PHDsec through the PredictProtein server. These results are given in Table 4.4.

Secondary structure predictions with the Gibrat method show a relationship between h-core α-helical content and translocation efficiency. As mammalian translocation efficiency decreases, the number of residues in a sequence that assume conformations other than α-helical increases. As with the Gibrat predictions, the Levin method proposes an α-helix for the h-region of the active CPYm6 peptide and a mixed coil and β-strand structure for the h-region of the non-functional WT peptide. However, inspection of the h-core residue conformations of the remaining peptides emphasises the unreliability of the Levin method for this application; there appears to be no direct correlation between structure and functional activity. Although the DPM and SOPMA methods also predict a mixed coil and β-strand structure for the h-region of the WT, their predictions for the h-core of CPYm6 are contrary to those of Gibrat and Levin; β-sheet structures are proposed.

Table 4.3: Secondary structure of the sequences of the CPY signal peptide system determined by joint prediction. Predictions were determined with the Gibrat,[235] Levin,[236] *DPM*[237] and *SOPMA*[238,239] methods. The *Consensus* prediction combines these methods. H = α-helix, E = β-sheet, C = coil, T = turn, S = bend, – = no prediction, *spaces* represent deleted residues.

Prediction method	Signal peptide[a]	Secondary structure prediction																	
	WT[b]	M	K	A	F	T	S	L	L	C	G	L	G	L	S	T	T	L	A	K	A
		1				5					10					15					20
Gibrat	CPYm6	H	H	H	H	H	H	H	H	H	H	H		H	H	H	H	H	H	H	H
	CPYm2	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	CPYm8	H	H	H	H	H	H	H	H	H	H	H	C	H	H	H	H	H	H	H	H
	CPYm12	H	H	H	H	H	H	H	H	H		H	C	C	H	H	H	H	H	H	H
	WT	H	H	H	H	H	H	E	E	E	C	C	C	C	C	H	H	H	H	H	H
		1				5					10					15					20
Levin	CPYm6	H	H	H	C	H	H	H	H	H	H	H		H	H	H	H	H	H	H	H
	CPYm2	H	T	H	C	C	H	E	H	H	H	C	C	H	H	T	T	C	C	H	H
	CPYm8	C	H	C	H	H	E	E	E	E	C	C	C	C	C	C	E	C	C	C	H
	CPYm12	C	H	T	C	H	H	H	H	H		H	H	H	H	H	C	H	H	H	H
	WT	C	T	T	C	C	E	E	E	E	H	C	C	C	C	E	E	E	–	H	T
		1				5					10					15					20
DPM	CPYm6	C	C	H	H	E	E	E	E	E	E	E		E	E	E	E	H	H	C	C
	CPYm2	C	C	H	H	E	E	E	E	E	C	E	E	E	E	E	E	H	H	C	C
	CPYm8	C	C	E	H	E	E	E	E	E	H	E	C	E	E	E	E	H	H	C	C
	CPYm12	C	C	H	H	E	E	E	E	E		E	C	E	E	E	E	H	H	C	C
	WT	C	C	H	H	E	E	E	E	E	C	C	C	E	C	E	E	H	H	C	C
		1				5					10					15					20
SOPMA	CPYm6	C	C	C	H	H	H	H	E	E	E	E		E	E	E	E	E	H	H	C
	CPYm2	C	E	E	E	E	E	E	E	C	C	C	T	T	T	C	H	H	H	H	H
	CPYm8	C	C	C	C	E	E	E	E	E	E	T	T	C	E	E	C	C	C	C	H
	CPYm12	C	C	E	H	H	H	E	E	H		T	T	T	T	T	C	H	H	H	H
	WT	C	C	C	C	E	E	E	E	C	C	C	C	T	E	C	C	E	E	H	H
		1				5					10					15					20
Consensus	CPYm6	C	C	H	H	H	H	H	H	H	H	H		H	H	H	H	H	H	H	C
	CPYm2	C	–	H	H	E	H	E	H	H	C	C	–	H	H	–	H	H	H	H	H
	CPYm8	C	C	C	H	H	E	E	E	E	H	–	C	C	E	E	E	C	C	C	H
	CPYm12	C	C	H	H	H	H	H	H	H		H	C	–	H	H	C	H	H	H	H
	WT	C	C	H	C	E	E	E	E	E	C	C	C	C	C	C	E	H	H	H	H

[a] for each prediction method, the signal peptides are arranged in order of decreasing translocation efficiency

[b] the sequence of the WT is listed as a reference; a list of amino acid residue abbreviations is available on page xv of this thesis; residues in the h-core are underlined

Although consensus prediction appears to improve predictive accuracy, with the exception of CPYm12, with regard to relating secondary structure to translocational efficiency, the results cannot be relied upon since they are derived from prediction methods which propose conflicting structure probabilities for functionally active SPs.

Secondary structure predictions calculated with the PHDsec method (Table 4.4) are somewhat surprising since the method seems incapable of differentiating between the different peptides of the system. It predicts that all the peptides will form complete α-helices; only the first and last residues of the sequences are not α-helical. It would be expected that the presence of the helix-breaking glycine residue in some of the peptides would cause breaks in the helices. One possible reason for this contrary result is the availability of too few sequence homologues for the purposes of multiple sequence alignment, *i.e.*, three for the LamB WT, one for the CPY WT, and four for the gC WT. This would lead to an expected decrease in prediction accuracy (< 72%).

Table 4.4: Secondary structure of the sequences of the CPY signal peptide system determined with the PHDsec[36,240,241] prediction method. H = α-helix, L = loop or coil, *spaces* represent deleted residues.

Signal peptide[a]	Secondary structure prediction																			
WT[b]	M	K	A	F	T	S	L	L	C	G	L	G	L	S	T	T	L	A	K	A
	1				5					10					15					20
CPYm6	L	H	H	H	H	H	H	H	H	H	H		H	H	H	H	H	H	H	L
CPYm2	L	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	L
CPYm8	L	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	L
CPYm12	L	H	H	H	H	H	H	H	H		H	H	H	H	H	H	H	H	H	L
WT	L	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	L

[a] the signal peptides are arranged in order of decreasing translocation efficiency
[b] the sequence of the WT is listed as a reference; a list of amino acid residue abbreviations is available on page xi of this thesis; residues in the h-core are underlined

- *gC*

Predictions for the gC system (Table 4.5) are also puzzling since all the methods predict that all the sequences of the system will form complete or almost complete α-helical structures. Unexpectedly, the occurrence of the helix-breaking proline residue in some of the sequences does not induce interruptions in the helical structures. Only the DPM method offers some differences in h-core secondary structure, where the less active peptides contain a few residues that prefer coil structures.

Table 4.5: Secondary structure of the sequences of the gC signal peptide system determined by joint prediction. Predictions were determined with the Gibrat,[235] Levin,[236] DPM,[237] SOPMA[238,239] and PHDsec[36,240,241] methods. The *Consensus* prediction combines these methods. H = α-helix, C = coil, T = turn, – = no prediction, *spaces* represent deleted residues.

Prediction method	Signal peptide[a]	Secondary structure prediction																					
	WT[b]	M	A	S	L	A	R	A	M	L	A	L	L	A	L	Y	A	A	A	I	A	A	A
		1				5					10					15					20		
Gibrat	WT	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	ΔA10	H	H	H	H	H	H	H	H	H		H	H	H	H	H	H	H	H	H	H	H	H
	A10P	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	L12P	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	L14P	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
		1				5					10					15					20		
Levin	WT	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	ΔA10	C	C	C	H	H	H	H	H	H		H	H	H	H	H	H	H	H	H	H	H	H
	A10P	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	L12P	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	L14P	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C	H	H	H	H	H	H	H
		1				5					10					15					20		
DPM	WT	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C	C
	ΔA10	C	C	H	H	H	H	H	H	H		H	H	H	H	H	H	H	H	H	H	C	C
	A10P	C	C	H	H	H	H	H	H	H	C	H	H	H	H	H	H	H	H	H	H	C	C
	L12P	C	C	H	H	H	H	H	H	H	H	H	C	H	H	H	H	H	H	H	H	C	C
	L14P	C	C	H	H	H	H	H	H	H	H	H	H	H	C	C	H	H	H	H	H	C	C
		1				5					10					15					20		
SOPMA	WT	C	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	A10	H	H	H	H	H	H	H	H	H		H	H	H	H	H	H	H	H	H	H	H	T
	A10P	T	T	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
	L12P	T	C	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	T
	L14P	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
		1				5					10					15					20		
PHDsec	WT	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	ΔA10	H	H	H	H	H	H	H	H	H		H	H	H	H	H	H	H	H	H	H	H	C
	A10P	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	L12P	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	L14P	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
		1				5					10					15					20		
Consensus	WT	C	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	ΔA10	H	H	H	H	H	H	H	H	H		H	H	H	H	H	H	H	H	H	H	H	–
	A10P	-	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	C
	L12P	–	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	.	–
	L14P	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H

[a] for each prediction method, the signal peptides are arranged in order of decreasing translocation efficiency

[b] the sequence of the WT is listed as a reference; a list of amino acid residue abbreviations is available on page xi of this thesis; residues in the h-core are underlined

It was initially surmised by us that the short length of the sequences was the cause of these similar prediction results. However, this may be ruled out as the peptides of the CPY system that were submitted for calculation are even shorter. As a consequence of the results of the individual prediction methods, joint analysis yields completely α-helical structures, allowing for no differentiation between sequences based on structure.

To summarise, the use of secondary structure prediction methods based on globular, water-soluble proteins to predict the conformations and hence the functional efficiencies of signal sequences is questionable. Both Chou-Fasman predictions and a Consensus prediction procedure are unable to propose structures for peptides of a system that correlate explicitly with the *in vivo* activity trend of the system. In the case of C-F calculations, the high β-structure probability values obtained for the peptides of LamB and CPY lead to inconclusive findings, and even though α-helical probabilities are higher than those of the β-structure for the peptides of the gC system, activity trends cannot be explained. In the case of Consensus prediction, only the activities of the LamB peptides can be somewhat rationalised. Thus, although globular protein-based methods have been established to be valid for some signal sequences, the inconsistency of predictions makes the methods unsuitable for the conformational analysis of hydrophobic peptides in general.

4.1.2. Membrane protein-based predictions

Hydrophobicity is postulated to be an important factor affecting SP function. Therefore, hydropathy plots of sequences investigated in the present study have been studied. The hydrophobic nature of signal sequences and their presumed adoption of a common conformation upon integration into membranes also prompted the use of secondary structure prediction techniques based on transmembrane sequences. PHDhtm, TMpred and PSA are such techniques that have been employed here.

4.1.2.1. Hydrophobicity

Figures 4.4, 4.5 and 4.6 are hydrophobicity plots for the signal peptides of the LamB, CPY and gC systems, respectively. The plots were calculated from three different hydropathy scales: the Hopp and Woods[242] hydrophilicity scale,[∞] the Rose *et al.*[243] scale which computes mean fractional area losses, and the "optimal matching hydrophobicity" (OMH) scale of Sweet and

[∞] Values obtained from this scale were assigned opposite signs and then plotted as hydrophobicity values.

Eisenberg.[244] Values of the scales have been normalised by the ProtScale program (*cf.* section 2.1.1) to fit the 0 to 1 range, or the -1 to 0 range for Hopp and Woods, for the purposes of comparison. For each SP system, sequences of the same length are plotted together.

- *LamB*

In Figure 4.4, the graphs of WT and A13D are on the left of the page, while those for the deletion mutants, Δ78r2, Δ78r1 and Δ78, are on the right.

The hydrophobic regions of the WT and A13D signal sequences stretch from residue 10 to residue 16 (*cf.* Figure 3.1), and it can be clearly seen from Figure 4.4(a), (b) and (c) that this particular region is more hydrophobic for the WT than for the inactive A13D. This is due to the replacement of alanine at position 13 in the WT with the highly hydrophilic residue of aspartate. The plot derived from the Sweet and Eisenberg scale for the WT suggests an h-core that is shifted down by two residues from the residue-10 to residue-16 range, and stretching from residue 8 to residue 14.

As with the WT and A13D, differences between the curves of Δ78r1 and Δ78 occur in the residue-10 to residue-16 range. All the scales compute Δ78r1 (50% active) to be more hydrophobic than Δ78 (inactive) in this region. The less hydrophobic nature of the Δ78 region results from the substitution of cysteine at position 13 in Δ78r1 by the more hydrophilic glycine residue. The absence of proline at position 9 in the active Δ78r2 sequence produces a peak in its hydropathy curve at residue 10, whereas the curves of proline-containing Δ78r1 and Δ78 peak at residue 13 or 14. Consequently, the hydrophobicity of Δ78r2 cannot be directly compared with the hydrophobicities of Δ78r1 and Δ78. These results contradict the assumption of McKnight *et al.*[173] that the h-cores of these deletion mutants consist of residues 10 to 12. The plots infer that the cores should be lengthened to include the residues from position 8 to position 16.

Separately, the plots of the WT and A13D and of Δ78r1 and Δ78 correlate with their respective sequence activities. However, no correlation with activity can be distinguished between the longer and shorter sequences. These findings are confirmed by values calculated for the average hydrophobicities of the h-regions of the peptides; the values are tabulated in Table 4.6. Note that the h-regions of Δ78r2, Δ78r1 and Δ78 have been lengthened as discussed above.

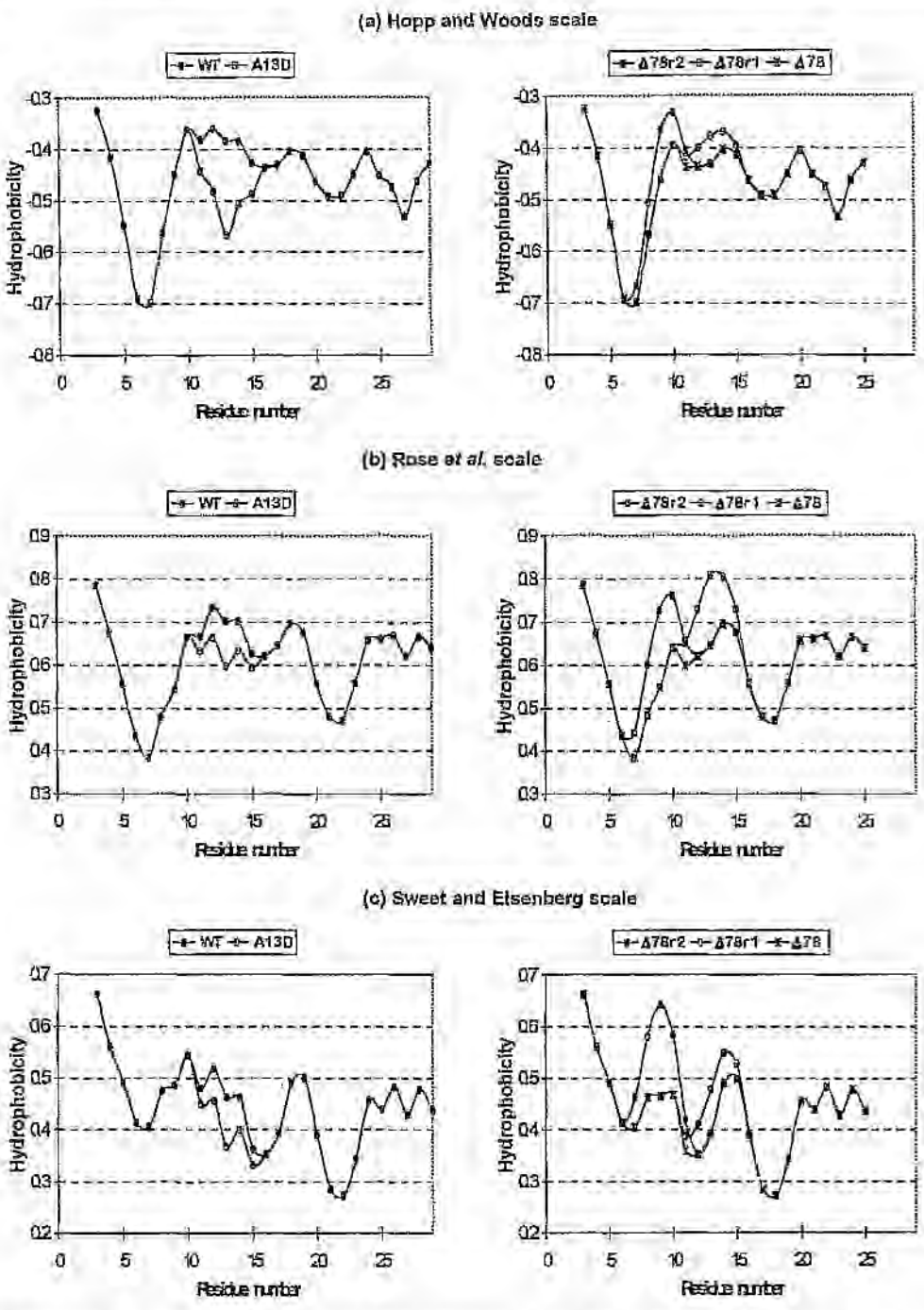


Figure 4.4: Hydrophobicity plots of the signal sequences of the LamB signal peptide system determined from the normalised hydropathy scales of (a) Hopp and Woods,[242] (b) Rose *et al.*[243] and (c) Sweet and Eisenberg[244]

Table 4.6: Summary of the translocation efficiencies and hydrophobicities of the signal peptides in the LamB system

Signal peptide	Translocation efficiency[a] /%	h-Region[b]	Average hydrophobicity of h-region[c] Hopp	Rose	Sweet
WT	100	LAVAVAA	0.971	0.790	0.206
Δ78r2	90	LLVAAGVMS	0.860	0.713	0.326
Δ78r1	50	LPVAACVMS	0.780	0.711	0.239
A13D	10	LAVDVAA	0.471	0.773	0.0757
Δ78	0	LPVAAGVMS	0.680	0.692	0.155

[a] values from McKnight *et al.*[173]

[b] refer to Figure 3.1; h-regions for Δ78r2, Δ78r1 and Δ78 have been adjusted according to hydropathy plot results

[c] calculated by adding the residue hydrophobicity values assigned by the Hopp and Woods,[242] Rose *et al.*[243] and Sweet and Eisenberg[244] scales, and dividing by the number of residues

- *CPY*

In Figure 4.5, the graphs for the shorter sequences of CPYm6 and CPYm12 are on the left side of the page, while those for the longer sequences of CPYm2, CPYm8 and the WT are on the right side.

The hydrophobic regions of CPYm6 and CPYm12 stretch from residue 7 to residue 12 (*cf.* Figure 3.2). Figure 4.5 shows that the h-region of CPYm6 is more hydrophobic than that of the less active CPYm12. This is attributed to the replacement of leucine at position 11 in CPYm6 with the more hydrophilic residue of glycine in CPYm12.

The substitution of glycine-10 in the WT by the slightly more hydrophobic alanine residue in CPYm8 leads to similar hydrophobicity plots; CPYm8 possesses a slightly more hydrophobic h-core (residues 7 to 13) than the WT. The area below the h-core (residues 7 to 13) curve of CPYm2 is larger than the areas below the corresponding curves of CPYm8 and the WT. The CPYm2 curve displays a break at position 10, which is occupied by glycine.

By combining the results from the shorter sequence plots with those from the longer sequence plots, it can be observed that, apart from CPYm12, results correlate to some extent with mammalian translocational efficiencies. However, the results cannot rationalise the large variation in activity between CPYm2 and CPYm8. These observations are confirmed by calculated values for the average hydrophobicities of the h-regions of the peptides, tabulated in

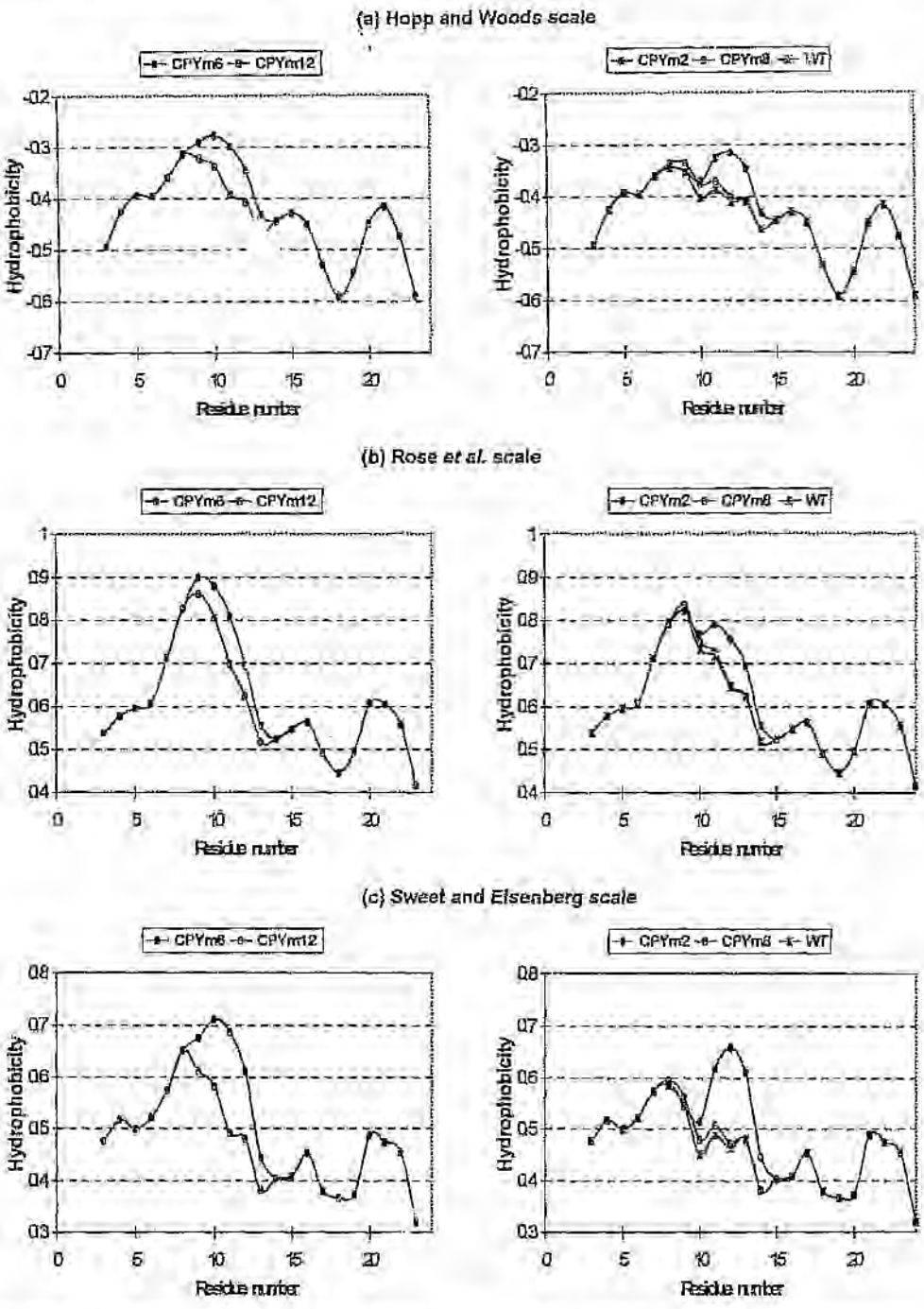


Figure 4.5: Hydrophobicity plots of the signal sequences of the CPY signal peptide system determined from the normalised hydropathy scales of (a) Hopp and Woods,[243] (b) Rose *et al.*[243] and (c) Sweet and Eisenberg[244]

Table 4.7. It would appear that, relative to CPYm6 and CPYm8, the average hydrophobicity value of the h-core of CPYm2 is too low.

Table 4.7: Summary of the mammalian translocation efficiencies and hydrophobicities of the signal peptides in the CPY system

Signal peptide	Translocation efficiency[a] /%	h-Region[b]	Average hydrophobicity of h-region[c] Hopp	Rose	Sweet
CPYm6	97	LLCLLL	1.667	0.860	1.05
CPYm2	94	LLCGLLL	1.429	0.840	0.800
CPYm8	27	LLCALGL	1.243	0.824	0.569
CPYm12	22	LLCLGL	1.367	0.838	0.730
CPY	undetectable	LLCGLGL	1.171	0.821	0.530

[a] values from Bird *et al.*[204]
[b] refer to Figure 3.2
[c] calculated by adding the residue hydrophobicity values assigned by the Hopp and Woods,[242] Rose *et al.*[243] and Sweet and Eisenberg[244] scales, and dividing by the number of residues

- *gC*

The graphs on the left side of Figure 4.6 compare the WT peptide with the shorter ΔA10 sequence. The first five residues of the hydrophobic region of ΔA10, *i.e.*, residues 8 to 12, have higher hydrophobicity values than the corresponding residues of the WT, but the h-core curve of the latter peptide (residues 8 to 16) is wider than that of ΔA10. Thus, their translocation efficiencies are expected to be similar. This is indeed the case, as can be noted in Table 4.8.

The graphs on the right side of Figure 4.6 compare the gC sequences which contain the proline residue. The insertion of proline into the WT sequence causes valleys to occur in the h-cores (residues 8 to 16) of these peptides at positions corresponding to that of proline. The shallower valley in the A10P curve seems to explain its greater ability to translocate secretory proteins when compared with L12P and L14P. However, the curves of L12P and L14P do not offer an immediate explanation concerning their relative translocation abilities.

Calculated values for the average hydrophobicities of the h-regions of the gC peptides are tabulated in Table 4.8. Trends followed by the values agree with the observations gathered from the hydropathy plots: the WT, ΔA10 and A10P behave similarly regarding their export abilities, and are more translocationally efficient than L12P and L14P; no distinction in export ability between L12P and L14P is perceived.

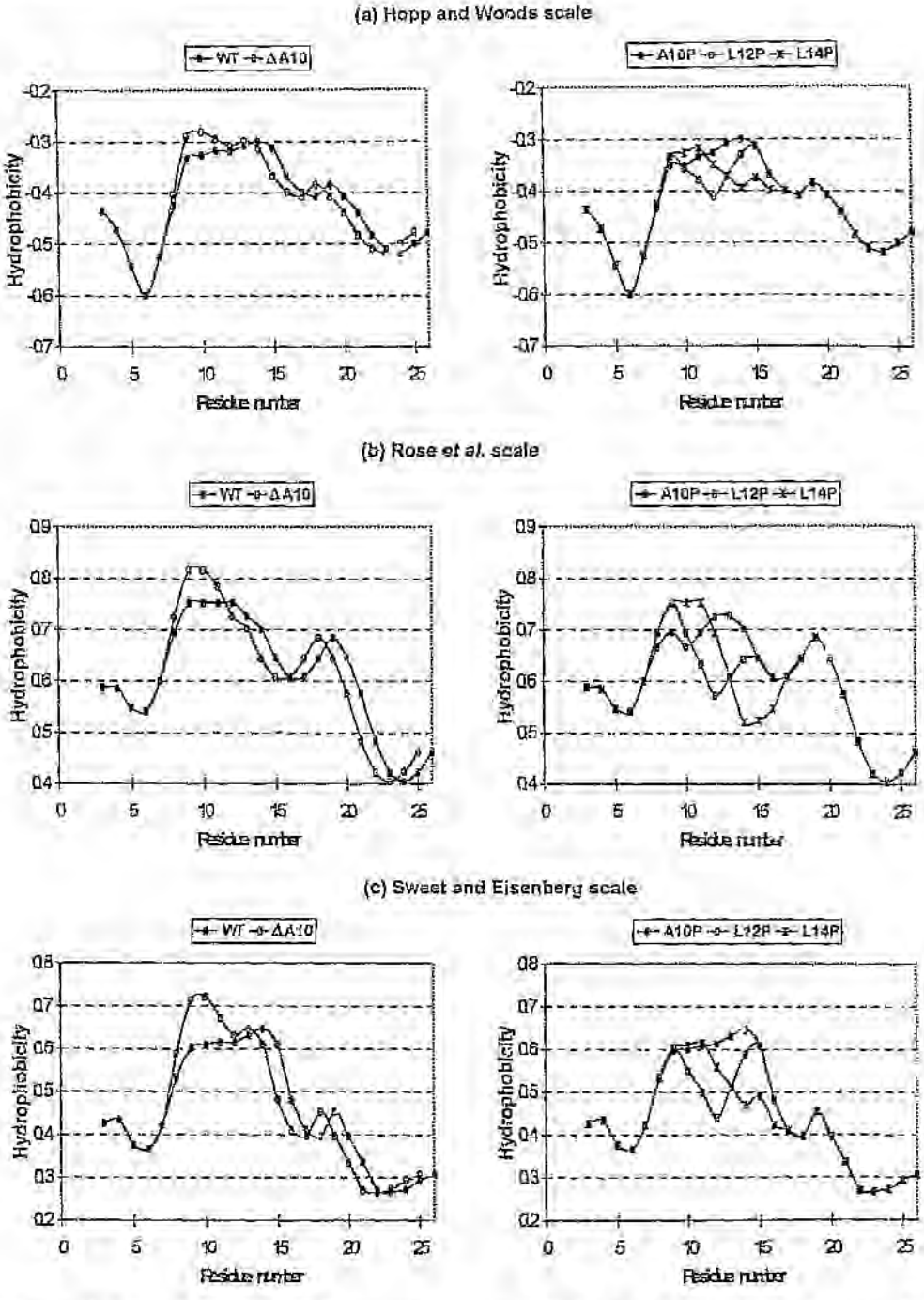


Figure 4.6: Hydrophobicity plots of the signal sequences of the gC signal peptide system determined from the normalised hydropathy scales of (a) Hopp and Woods,[242] (b) Rose *et al.*[243] and (c) Sweet and Eisenberg[244]

Table 4.8: Summary of the translocation efficiencies and hydrophobicities of the signal peptides in the gC system

Signal peptide	Translocation efficiency[a] /%	h-Region[b]	Average hydrophobicity of h-region[c] Hopp	Rose	Sweet
WT	100	AMLALLALY	1.37	0.803	0.708
ΔA10	99	AMLLLALY	1.48	0.811	0.846
A10P	98	AMLPLLALY	1.31	0.792	0.698
L12P	19	AMLALPALY	1.17	0.780	0.518
L14P	5	AMLALLAPY	1.17	0.780	0.518

[a] values from Ryan and Edwards[237]
[b] refer to Figure 3.3
[c] calculated by adding the residue hydrophobicity values assigned by the Hopp and Woods,[242] Rose *et al.*[243] and Sweet and Eisenberg[244] scales, and dividing by the number of residues

4.1.2.2. PHD methods

A summary of prediction findings using the PHDhtm[224] method for the recognition of α-helical transmembrane regions in a peptide is given in Table 4.9. The program predicts two states: helical transmembrane and non-transmembrane (loop). With the exclusion of LamB's WT and A13D, results appear to be illogical. The only explanation is that the length of sequence submitted for analysis is crucial to predictive accuracy. Indeed, membrane-spanning domains of proteins generally consist of 22 to 23 amino acids. In addition, the hydrophobic regions of these membrane-spanning segments are longer than the hydrophobic regions of SPs.

Solvent accessibilities for the signal sequences were calculated with PHDacc[245] (see Table 4.10). Data is provided in respect of three accessibility states: buried; intermediate; exposed. States listed in the table in bold typeface occur in the hydrophobic regions of the sequences. Predictions confirm that the entire sequences are highly hydrophobic since the majority of the sequence residues are buried or inaccessible to solvent (71% to 80% for LamB, 65% to 68% for CPY, 71% to 91% for gC). Buried residues implies small surface areas for the sequences which may, in turn, imply α-helical structure arrangement. No other useful information could be deduced from the results.

Table 4.9: Secondary structure prediction of the sequences of the LamB, CPY and gC systems determined with the PHDhtm[224] prediction method

System	Signal peptide	Secondary structure prediction		Length of submitted sequence[b]
		% α-helix	% loop (coil)	
WT	WT	100	0	25
	A13D	84	16	25
	Δ78r2, Δ78r1, Δ78	0	100	21
CPY	all[a]	0	100	19 or 20
gC	all[a]	100	0	21 or 22

[a] all the sequences of the system
[b] number of residues submitted for analysis

Table 4.10: Solvent accessibility prediction of the sequences of the LamB, CPY and gC systems determined with the PHDacc[245] prediction method. b = buried in interior, i = intermediate, e = exposed to solvent, *spaces* represent deleted residues.

System	Signal peptide[a]	Solvent accessibility prediction																							
LamB	WT[b]	M	M	I	T	L	R	K	L	P	L	A	V	A	V	A	A	G	V	M	S	A	Q	A	M	A
		1				5					10					15					20					25
	WT	b	b	b	b	b	i	e	b	e	b	b	b	b	b	b	b	b	b	b	b	b	e	b	b	e
	Δ78r2	b	b	b	b	b	i	e	e	b					b	b	b	b	b	b	b	b	e	b	b	e
	Δ78r1	b	b	b	b	b	e	e	b	e					e	b	b	b	b	b	b	b	e	b	b	e
	A13D	b	b	b	b	b	i	e	b	e	b	b	b	b	b	b	b	b	b	b	b	b	e	b	b	e
	Δ78	b	b	b	b	b	e	e	b	e					e	b	b	b	b	b	b	b	e	b	b	e
CPY	WT[b]	M	K	A	F	T	S	L	L	C	G	L	G	L	S	T	T	L	A	K	A					
		1				5					10					15					20					
	CPYm6	b	e	b	b	b	b	b	b	b	e	e		b	b	e	b	b	b	e	e					
	CPYm2	e	e	b	b	b	b	b	b	b	b	b	e	b	b	e	e	b	b	e	e					
	CPYm8	e	e	b	b	b	b	b	b	b	b	b	e	b	b	e	e	b	b	e	e					
	CPYm12	e	e	b	b	b	b	b	b	b		b	b	b	b	e	e	b	b	e	e					
	WT	e	e	b	b	b	b	b	b	b	b	b	e	b	b	e	e	b	b	e	e					
gC	WT[b]	M	A	S	L	A	R	A	M	L	A	L	L	A	L	V	A	A	A	I	A	A	A			
		1				5					10					15					20					
	WT	e	b	b	b	b	i	b	b	b	b	b	b	b	b	e	b	b	b	b	b	b	c			
	ΔA10	e	b	e	b	b	b	b	b	b		b	b	e	e	i	b	b	b	b	b	b	e			
	A10P	b	b	b	b	b	e	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	e			
	L12P	e	b	b	b	b	e	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	e			
	L14P	b	b	b	b	b	b	b	b	b	b	b	b	b	b	i	b	b	b	b	b	b	e			

[a] for each system, the signal peptides are arranged in order of decreasing translocation efficiency
[b] the sequence of the WT is listed as a reference; a list of amino acid residue abbreviations is available on page *xi* of this thesis; residues in the h-core are underlined

It must be emphasised that these PHDacc predictions should be treated with circumspection as the method, based on neural networks, was trained on a set of globular proteins, and hydrophobic segments in globular proteins (6 to 8 residues in length) are generally shorter than the hydrophobic regions of signal sequences (10 to 15 residues long).

4.1.2.3. TMpred

The TMpred program selected the central region of each signal sequence and predicted its transmembrane α-helical orientation (inside → outside or outside → inside). Predictions for the sequences of the three SP systems are plotted in Figures 4.7, 4.8 and 4.9. Both orientation models are depicted in the figures.

The majority of the sequences (LamB WT and A13D, CPYm6 and CPYm12, gC system) appear to prefer the inside → outside orientation, where the N-terminal of the peptide enters the lumen of the membrane from the cytoplasm, as opposed to the outside → inside orientation, where the N-terminal exits the lumen to the cytoplasm. This finding suggests that SPs may enter the membrane in an α-helical conformation. On exiting the membrane, the SPs may adopt an extended conformation.

Separately, the plots of the longer sequences (left of the figures) and of the shorter sequences (right of the figures) of the LamB and the CPY systems correspond with their relative signal sequence translocation efficiencies, *e.g.*, the LamB WT has a higher probability of occurring as a transmembrane helix than the less active A13D, and transmembrane probabilities decrease from the active Δ78r2 to the inactive Δ78. However, the longer sequence plots cannot be directly compared with the shorter sequence plots as a means of evaluating sequence activity. For example, CPYm6 and CPYm12 are postulated to have less tendency to contain membrane-spanning regions than CPYm2, CPYm8 and the WT.

Although predictions for the WT and ΔA10 sequences of the gC system (left of Figure 4.9) correlate with *in vivo* activities, predictions for the proline-containing sequences, A10P, L12P and L14P (right of Figure 4.9) do not.

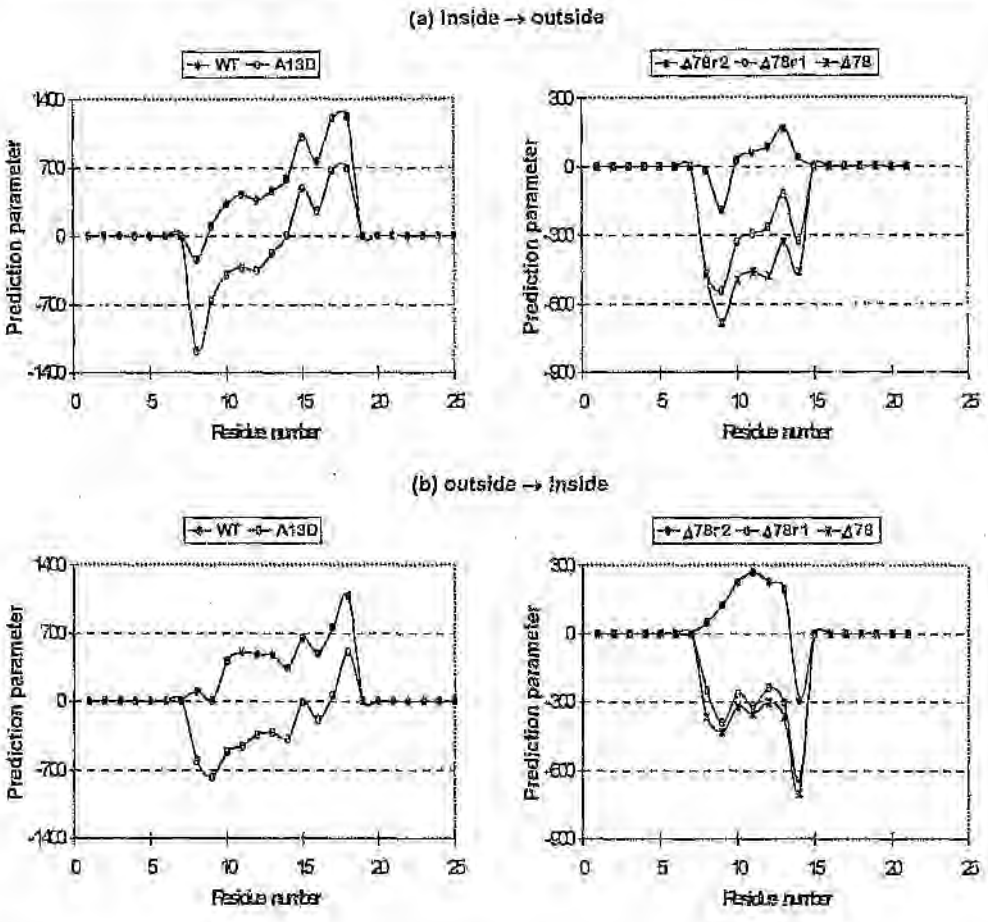


Figure 4.7: TMpred prediction of transmembrane helical regions of the sequences of the LamB signal peptide system. (a) inside → outside and (b) outside → inside orientations of the transmembrane regions are shown.

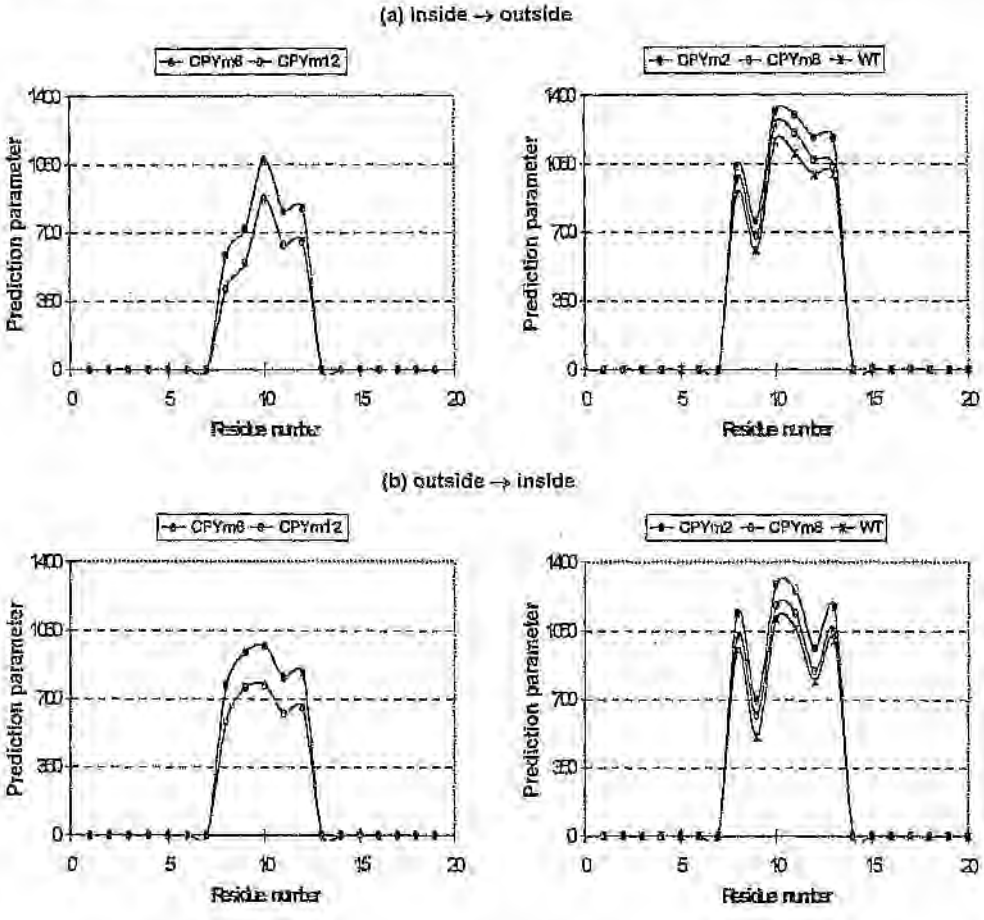


Figure 4.8: TMpred prediction of transmembrane helical regions of the sequences of the CPY signal peptide system. (a) inside → outside and (b) outside → inside orientations of the transmembrane regions are shown.

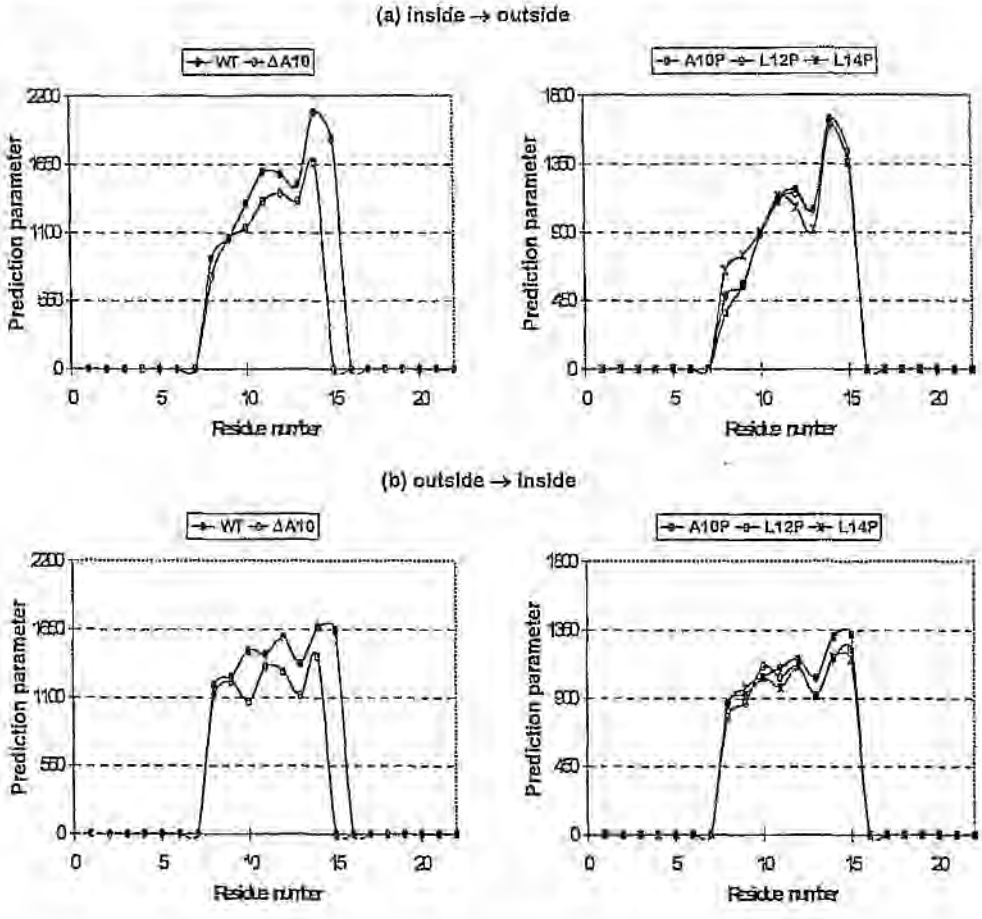


Figure 4.9: TMpred prediction of helical transmembrane regions of the sequences of the gC signal peptide system. (a) inside → outside and (b) outside → inside orientations of the transmembrane regions are shown.

4.1.2.4. PSA

Secondary structure probability trends obtained from PSA[247,248] calculations are graphed in Figures 4.10, 4.11 and 4.12, and the probabilities of the sequences being membrane-spanning regions are listed in Table 4.11. PSA analyses peptides for four structures: α-helices; loops; turns; β-strands. The combined probabilities of these structures is unity for each sequence.

Table 4.11: Translocation efficiencies and membrane-spanning probabilities of the sequences of the LamB, CPY and gC systems determined with the PSA[247,248] prediction method

System	Signal peptide	Translocation efficiency /%	Membrane-spanning probability
LamB	WT	100	0.9630
	Δ78r2	90	0.9390
	Δ78r1	50	0.9080
	A13D	10	0.9071
	Δ78	0	0.8742
CPY	CPYm6	97	0.8829
	CPYm2	94	0.8739
	CPYm8	27	0.8651
	CPYm12	22	0.8312
	CPY	undetectable	0.8034
gC	WT	100	0.9922
	ΔA10	99	0.9910
	A10P	9[illegible]	0.9812
	L12P	19	0.9785
	L14P	5	0.9792

The signal sequences in this study were predicted to form predominantly α-helices, with the loop or coil structure favoured next. Thus, only these structure probabilities are shown in the graphs. Probability values for the turn and β-strand structures were relatively low (< 0.16 for the LamB peptides, < 0.10 for the CPY peptides, < 0.053 for the gC peptides).

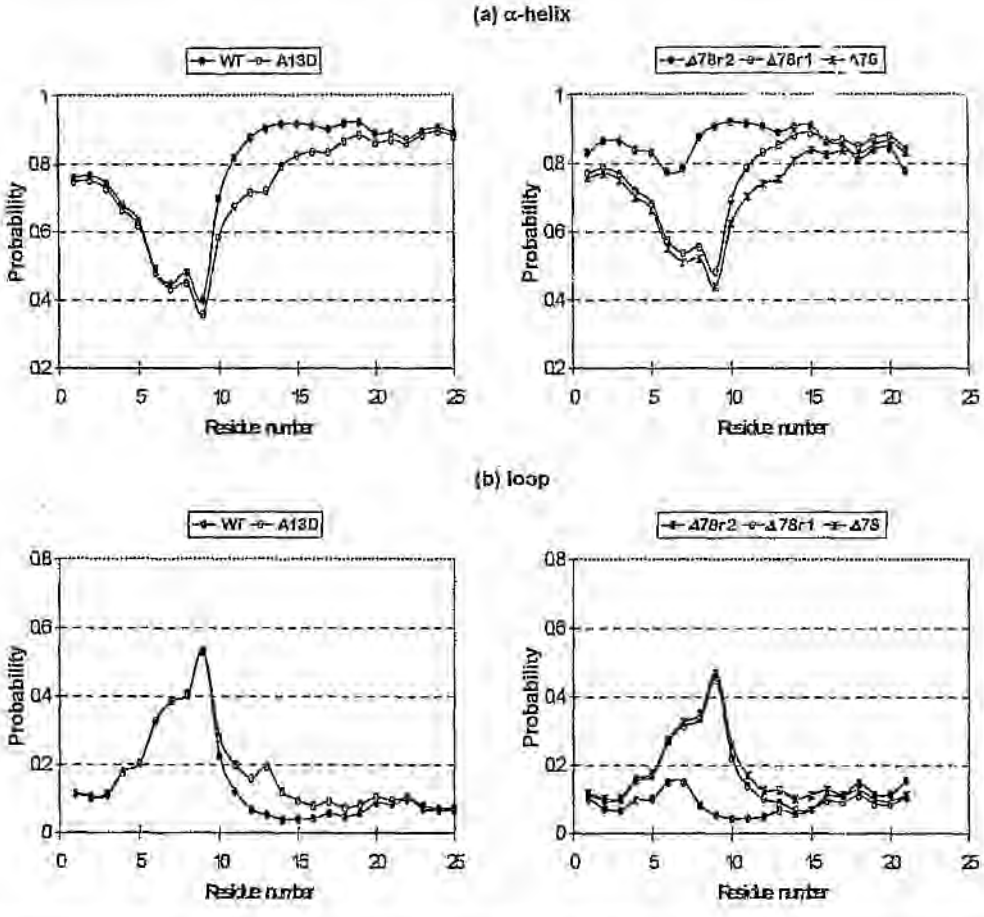


Figure 4.10: Probable secondary structure of the sequences of the LamB signal peptide system determined with the PSA[247,248] prediction method. (a) α-helical and (b) loop structures are shown.

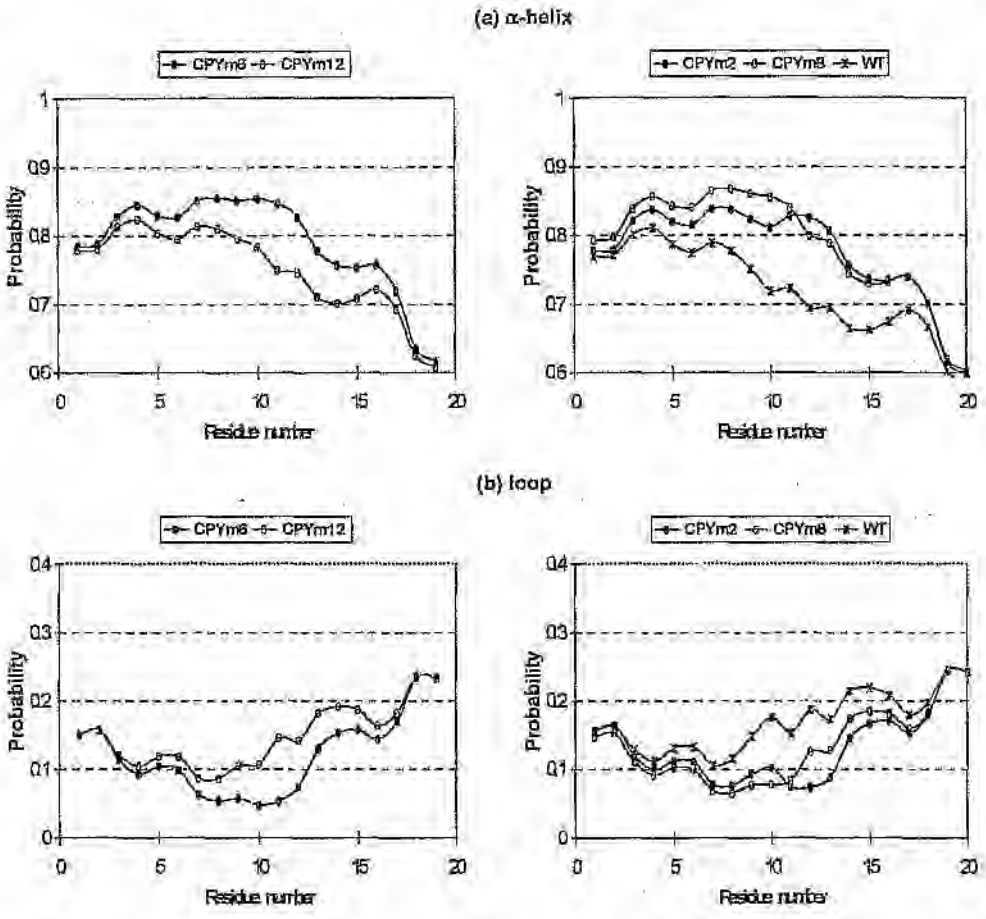


Figure 4.11: Probable secondary structure of the sequences of the CPY signal peptide system determined with the PSA[247,248] prediction method. (a) α-helical and (b) loop structures are shown.

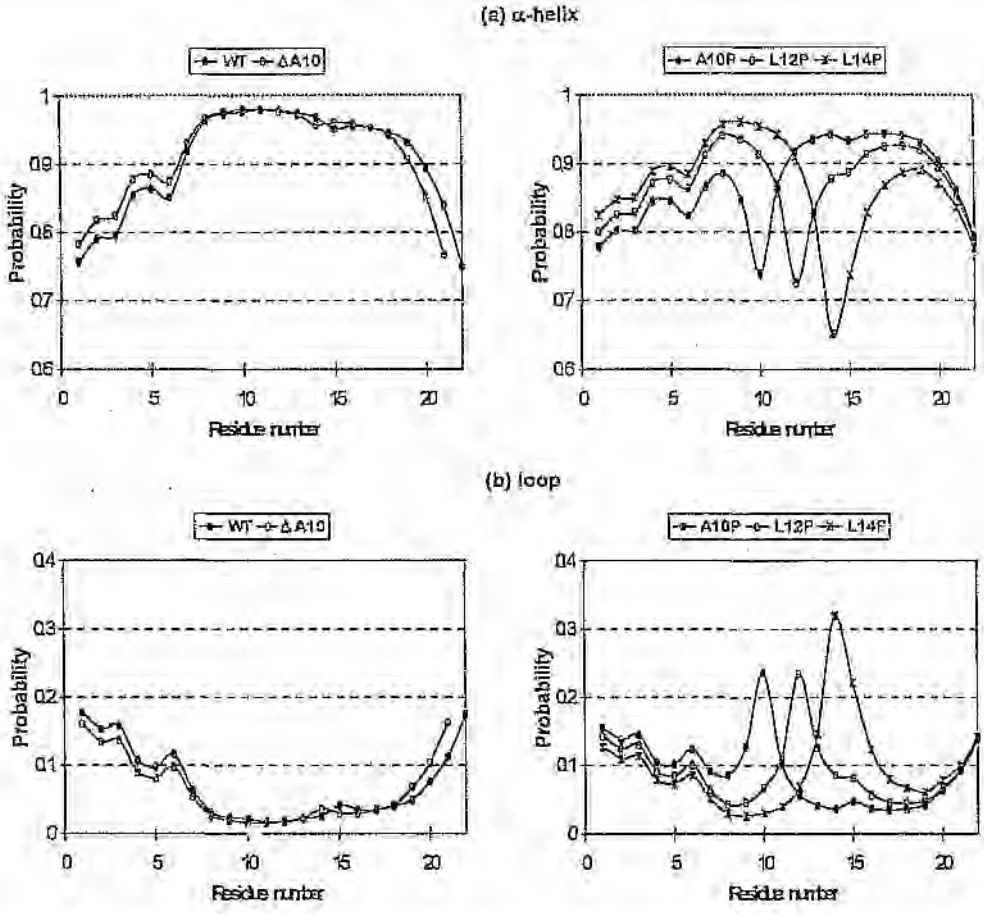


Figure 4.12: Probable secondary structure of the sequences of the gC signal peptide system determined with the PSA[247249] prediction method. (a) α-helical and (b) loop structures are shown.

Membrane-spanning probabilities correlate well with the ability of the SPs to secrete proteins, while α-helical structure probabilities correlate less well. The effect of the hydrophilic, helix-breaking proline residue on secondary structure formation can be distinctly discerned from Figures 4.10 and 4.12 (minima on the α-helix plots; maxima on the loop plots).

To summarise, secondary structure prediction methods based on membrane proteins appear to yield more useful information than methods based on globular proteins. Most of the membrane-based programs are able to propose structures for peptides of a system that correlate explicitly with *in vivo* activities. The inconsistencies in predictions that still occur are probably due to the fact that hydrophobic regions of the signal peptides are shorter than membrane-spanning segments, resulting in erroneous calculations. PSA is the only prediction method that seems to deliver results which correlate well with observations from the literature. Nevertheless, PSA and other knowledge-based modelling algorithms are only able to partially predict SP conformation as other factors, such as peptide interactions, may be pertinent.

4.2. Molecular modelling

4.2.1. Systematic conformational searches

4.2.1.1. SYBYL

Although systematic searches executed with SYBYL were too crude to yield any meaningful results, the searches did produce two general findings which are illustrated in Figure 4.13. In the figure, conformations attained from two searches of the Δ78 signal sequence, one with applied α- and one with applied β-constraints, are plotted against their respective calculated energy values. The findings are: (1) signal sequences are considerably more stable with their h-cores constrained in the α-helical geometry than in the β-sheet (the average conformational energy for the α-helical conformations is 7.443 kcal mol^{-1}, while the average energy for the β-sheet conformations is higher at 14.699 kcal mol^{-1}); (2) local minima energy wells for the β-constrained conformations are much shallower than those for the α-constrained conformations (the plot for the β-constrained conformations is relatively smoother than that for the α-constrained conformations). It would appear that the α-helix secondary structure is energetically favoured, while that of the β-sheet is statistically favoured.

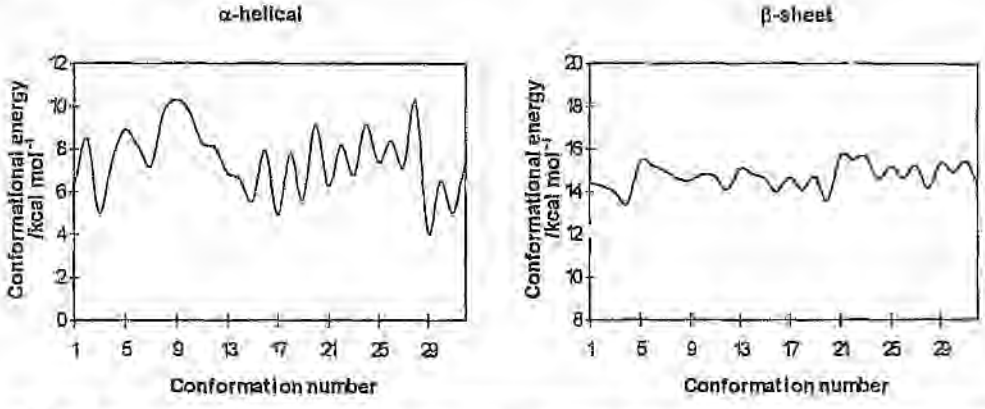


Figure 4.13: Comparison of local minima energy wells of the LamB Δ78 signal sequence constrained in the α-helical and β-sheet secondary structures during systematic searching with SYBYL. Proline ω = 180° and the prolyl ring is puckered 'down'. For comparison purposes, equal numbers of conformations for the two structures are graphed. Note the different energy scales on the ordinates.

Ramachandran plots of the lowest energy-minimised structures for the LamB signal sequences of active Δ78r2 and inactive Δ78 are given in Figure 4.14. Numbers drawn beside points on the plots indicate the sequence numbers of the amino acids of the peptides (*cf.* Figure 3.1). Since α-helical conformational constraints were imposed on the h-cores (residues 10 to 12) of the sequences during conformational searching, h-core (ϕ, ψ) values reside in the Zimmerman α-helix region. The majority of residues flanking the h-core assume other conformations.

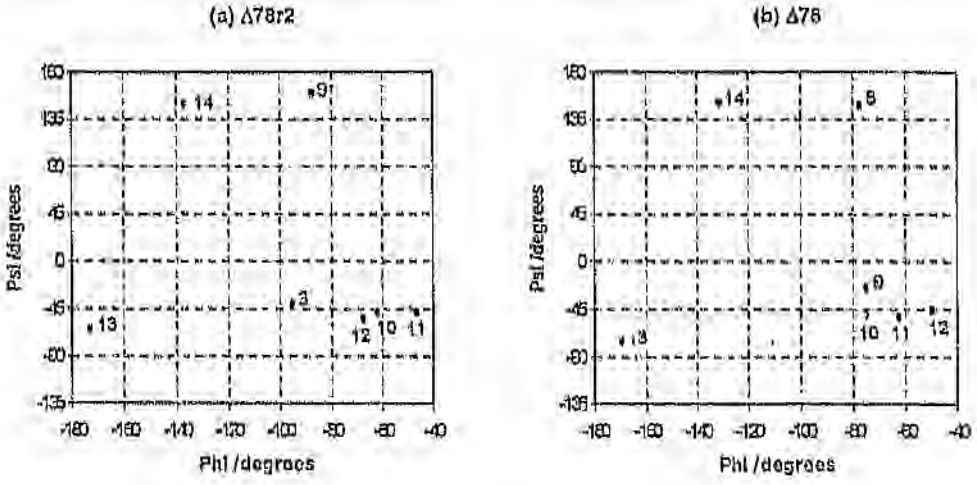


Figure 4.14: Ramachandran plots of the lowest energy-minimised structures for the (a) Δ78r2 and (b) Δ78 signal sequences of the LamB system determined with SYBYL systematic searches. For Δ78, proline ω = 180° and the prolyl ring is puckered 'down'. The h-cores of the sequences are constrained in the α-helix conformation.

As mentioned in section 3.3.1, both fragment-based and torsional angle-based approaches were explored during systematic conformational searches. Nevertheless, torsional-angle searching is the preferred technique because, firstly, it cannot be assumed that the conformation of each residue is independent of the entire peptide conformation, and secondly, individual residue conformations may change when placed in a peptide chain.

4.2.1.2. ECEPPAK

The distance-constraints modelling strategy was first employed in systematic searches performed with ECEPPAK. Since the ECEPPAK program imposes distance-constraints *via* the application of a pseudo-potential, the energy associated with this potential (low values which average 10^{-3} kcal mol^{-1}) was subtracted from the total minimum potential energy obtained for each resultant conformation. Although one cannot in principle apply calculations for isolated, single residues to globular proteins because medium- and long-range interactions are able to shift torsional angle values away from their pre-determined minima locations, side-chain torsional values acquired from the literature were nonetheless used here.

Minimised energies obtained from these sets of experiments (which were conducted with a dielectric constant of 2) are plotted as a function of distance-constraint in Figure 4.15 for α- and β-conformations. Only the lowest-energy values per distance-constraint are shown. Both plots (a) and (b) indicate that sequences constrained to be α-helical yield lower energies than those constrained to be in the β-sheet. This is an anticipated result as β-structures need to be stabilised relative to their environment *via* hydrogen bonding, which cannot happen in a hydrophobic medium. The searches may thus validate the assumption that SPs exist in highly hydrophobic media as α-helices. The α-helical energies also tend to form parabolic curves, whereas the β-sheet energies appear to increase linearly with distance-constraint.

Ramachandran plots of the lowest energy-minimised structures for the active Δ78r2 and the inactive Δ78 are given in Figure 4.16. Only those residues with relatively divergent (ϕ, ψ) values are numbered in the figure. All eleven residues per sequence chosen for MM simulation have (ϕ, ψ) values which correspond to the Zimmerman α-helical region. A comparison with the Ramachandran plots in Figure 4.14 infers that the optimum structures computed with ECEPPAK are probably closer to the GMEC than those computed with SYBYL since they display more α-helical character. However, this conclusion cannot be ascribed to the performance of the respective force fields as thorough searching was unrealised and different experimental conditions were exercised.

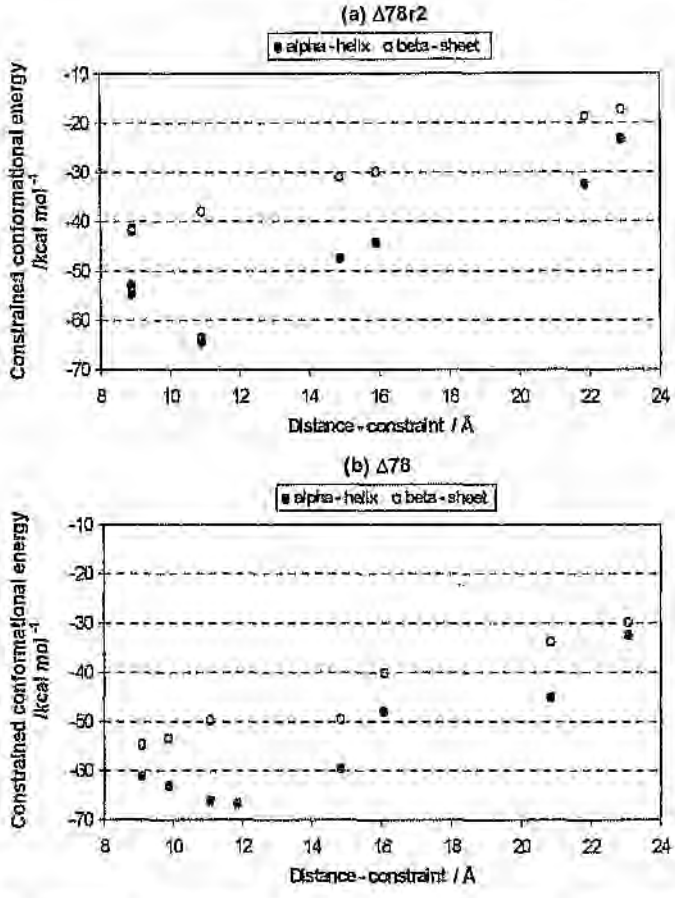


Figure 4.15: Minimised energy *versus* distance-constraint curves for the (a) Δ78r2 and (b) Δ78 signa' sequences of the LamB system determined with ECEPPAK systematic searches. For Δ78, proline ω = 180° and the prolyl ring is puckered 'down'. The central se.en residues of the sequences are constrained in the α-helix and β-sheet conformations.

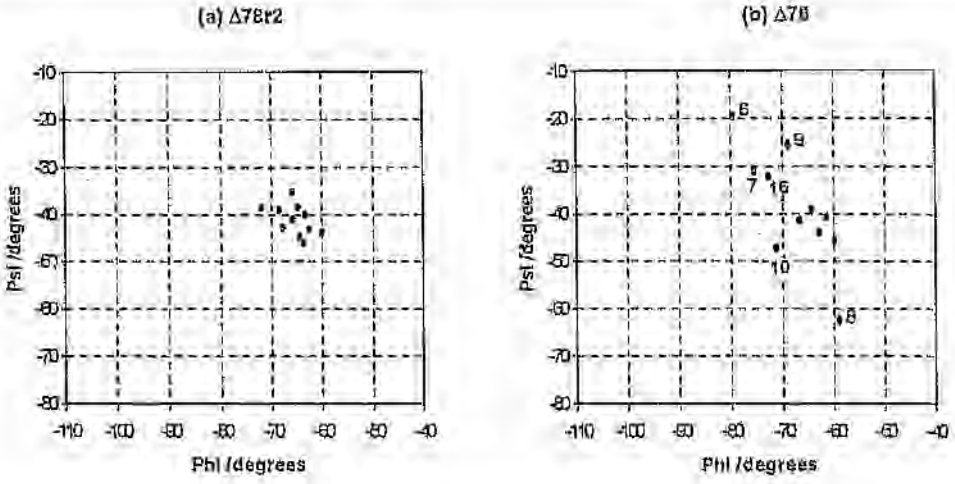


Figure 4.16: Ramachandran plots of the lowest energy-minimised structures for the (a) Δ78r2 and (b) Δ78 signal sequences of the LamB system determined with ECEPPAK systematic searches. For Δ78, proline ω = 180° and the prolyl ring is puckered 'down'. The central seven residues of the sequences are constrained in the α-helix conformation.

The method of deterministic searching using both SYBYL and ECEPPAK was found to be impractical in terms of required computer time and ineffective in locating local minimum energy structures; many more points than those shown in Figures 4.13 and 4.15 were attained, but these coincided with relatively high energies. Conclusions drawn from these results may therefore be suspect.

4.2.2. Optimisation of the performance of ECEPPGA

ECEPPGA parameters optimised for this application and parameter values employed during optimisation are listed in Table 4.12. The remaining parameters were fixed at values recommended by Stephens;[120] these values can be seen in the `param.in` file in the Appendix. The mutant sequences Δ78r2, Δ78r1 and Δ78 of LamB were subjected to investigation. Jobs, each consisting of three runs, were conducted with and without distance-constraints for Δ78r2. Only jobs with distance-constraints were conducted for Δ78r1 and Δ78.

Results presented for each job are derived from population energy statistics gathered during the experimental run which yielded the lowest energy. The progression of a job is plotted as a function of the number of generations passed. The plots illustrate the rate of minimisation achieved under different

experimental conditions. The counting of generations is dependent on the elected generation counting strategy. In this instance, generations were counted since the last improvement in best solution or structure.

Table 4.12: The variation of relevant ECEPPGA parameters during GA optimisation runs

Name of job	ECEPPGA parameter			
	Generation gap size: population size	Iterations per minimisation	Generations per run	Selective pressure[a] and breeding strategy
A	0.6	100	30	uniform and distance bias
B	0.3	100	30	uniform and distance bias
C	0.3	50	30	uniform and distance bias
D	0.3	50	15	uniform and distance bias
E	0.3	50	15	linear and template breeding

[a] fitness to breeding-rate relationship

Results for job B have been omitted from Figure 4.17(a) as the job required significantly more generations (210) than the others for completion of the GA. Plotted points for job D in the same figure cannot be distinguished due to the fact that it afforded the same results as the first 70 generations of job C. Job D was completed after 72 generations had been cycled.

Perusal of Figures 4.17, 4.18 and 4.19 shows a moderate degree of consistency in results when the optimisation procedure is applied to the three peptides. Thus, certain general observations may be made: halving of the generation gap size to population size ratio from 0.6 (job A) to 0.3 (job B) produces very similar minimisation rates; halving the maximum number of iterations allowed per minimisation (from job B to job C) generates conformations that are lower in energy than those generated by job B; halving the number of counted generations per run (from job C to job D) improves slightly on results from job C; changing the selective pressure and breeding strategy options (from job D to job E) yields the most noticeable effect by increasing the rate of minimisation and producing the lowest-energy conformations. Some of the inferior performances of jobs A to D may be ascribed to over-exploration, *i.e.*, recombination occurred too often. This is evidenced by the greater number of actual generations used to complete the runs.

The goal of optimising the GA was to reach the lowest-energy conformation using the least possible amount of generations and the least amount of computer time. This goal appears to have been achieved with the parameter values employed in job E. The success of the change in selective pressure from uniform, where the probability of a conformation being selected as a parent is independent of its

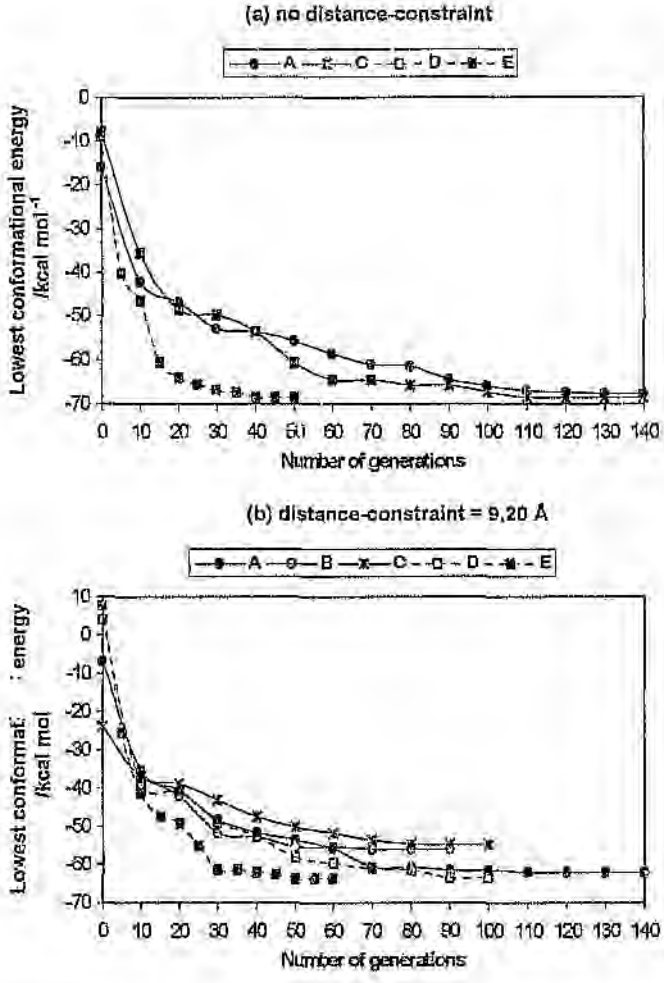


Figure 4.17 Effect of parameter variation on the minimisation achieved at various generations for the Δ78r2 signal sequence of LamB. (a): without distance-constraints; (b): with distance-constraint ($N_{Ile\text{-}8}$ to $C_{Val\text{-}14}$) = 9.20 Å.

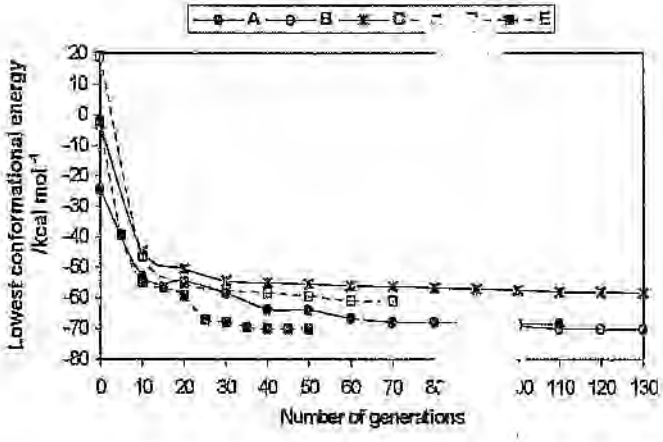


Figure 4.18: Effect of parameter variation on the minimisation achieved at various generations for the Δ78r1 signal sequence of LamB. The sequence was distance-constrained (N_{leu-8} to C_{val-14}) at 9.20 Å.

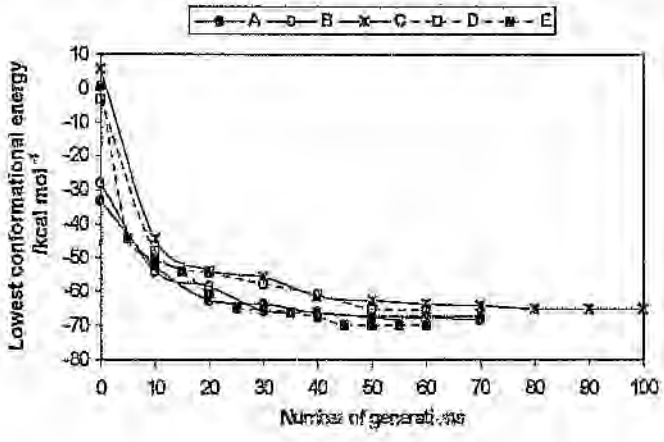


Figure 4.19: Effect of parameter variation on the minimisation achieved at various generations for the Δ78 signal sequence of LamB. The sequence was distance-constrained (N_{leu-8} to C_{val-14}) at 9.20 Å.

fitness, to linear, where the probability of a conformation being selected as a parent is linearly related to its fitness, and in breeding strategy from distance bias to template breeding concurs with observations made by Stephens.[120] Stephens found that a strongly biased breeding rate results in high selective pressure and thus faster optimisation, and that template breeding is highly effective in significantly improving ECEPPGA efficiency.

Although the information gleaned from this optimisation procedure prompted the adoption of parameter values corresponding to job E for ECEPPGA searches, the information must be interpreted with caution for several reasons. Firstly, reliability of the data will be very low because only three runs exist for every parameter setting, and more statistical data than this is required to infer reliable results. Secondly, only the effect of single parameters in isolation is considered, and it is possible that combinations of parameters exist that would yield better success rates. Nevertheless, the analysis has permitted us to select a serviceable procedure.

As optimisation of the GA was not one of the primary aims of this study, and as the final optimised parameters were judged to be satisfactory for our purposes, no further work concerning optimisation was deemed necessary.

4.2.3. Comparison of search methods

To assess the performance of the genetic algorithm, the lowest conformational energies found with the ECEPPAK systematic search (where residues were constrained in the α-helix) are compared to those found with ECEPPGA. Figure 4.20 shows minimised energies of the active LamB Δ78r2 and the inactive Δ78 signal sequences plotted as a function of distance-constraint. Fewer conformers were calculated with the systematic search than with the ECEPPGA search. It can be clearly seen that ECEPPGA consistently succeeded in finding lower-energy conformers.

The genetic algorithm therefore performs well relative to extensive systematic searching. However, the latter technique is only efficient for the conformational analysis of small peptides, and proper evaluation of the genetic algorithm performance requires contrasting it with other methods which are more competent than the systematic search. Still, ECEPPGA has succeeded in locating SP minimum-energy conformations that compare favourably with experimental observations from the literature; it has thus been proven to be a highly efficient conformational searching technique.

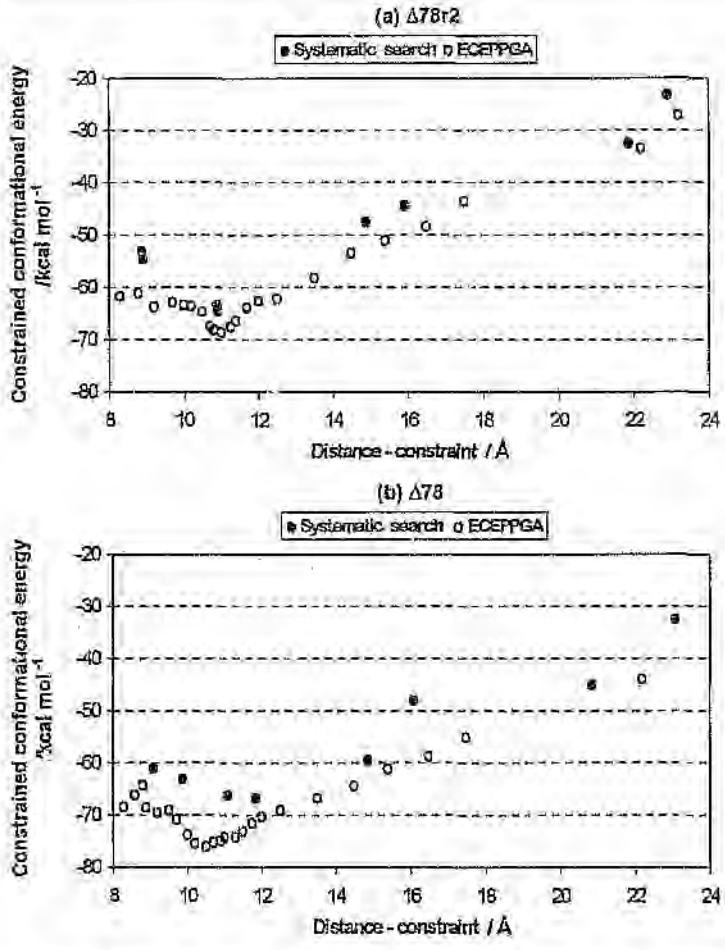


Figure 4.20: A comparison of minimised energy *versus* distance-constraint curves for the (a) Δ78r2 and (b) Δ78 signal sequences of the LamB system determined with the ECEPPAK systematic search and ECEPPGA. For Δ78, proline ω = 180° and the prolyl ring is puckered 'down'. The central seven residues of the sequences are constrained in the α-helix conformation for the systematic search.

4.2.4. ECEPPGA conformational searches

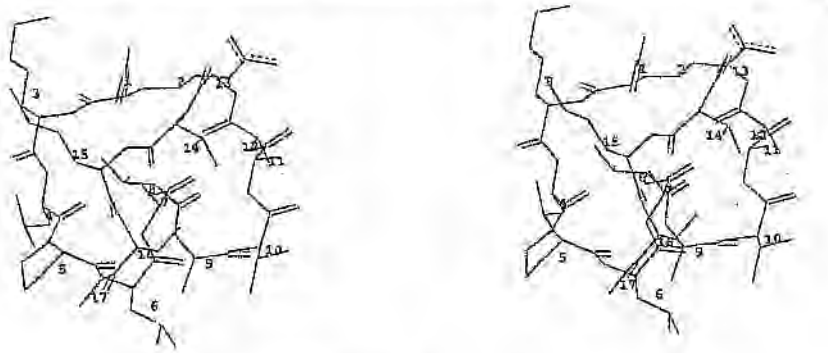
Results reported in this section were obtained from runs conducted with an effective dielectric constant of 2. Resultant ECEPPGA minimised energy *versus* distance-constraint curves for the three investigated systems are shown in Figures 4.22, 4.25 and 4.28.

Approximately parabolic-shaped curves were produced which vary in well depth, steepness of well sides and minimum position relative to distance-constraint. Although some curves display local minima, a principal well was obtained in most cases, suggesting that the global minimum had been reached. As the distance-constraint value increases from the principal well position, the conformation of the sequence changes from the low-energy α-helical form to the less stable extended structure. This is illustrated for the WT signal sequence of LamB in Figure 4.21. Generally, there is an increase in uniformity of the curve as the distance-constraint increases beyond the value at which the energy of the structure is a minimum. Conversely, as the distance-constraint decreases from its value at the energy minimum, the residues in the sequence are forced to be in close contact with each other, resulting in an irregular series of structures (Figure 4.21). Changes in peptide conformation during MM calculations have been witnessed by other researchers[110,253,254] who also use distance-constraints as a modelling strategy: alanine tripeptide and valine tripeptide unfold from an α-helix to a turn structure to an extended conformation;[110,253] an MeA (α-methylalanine) oligopeptide undergoes a transition from an α-helix to a 3_{10} helix.[254]

Quantitative information about the lowest-energy conformation calculated for each SP is provided in Tables 4.13, 4.14 and 4.15; data for both constrained and unconstrained runs are shown. The last column in each table is pertinent to both sets of runs. Conformational searching of sequences that were not distance-constrained attempted to confirm whether their lowest-energy structures were equivalent to those obtained from runs on constrained sequences. As expected, if the global minimum had been attained, minimum energy values from runs with the unconstrained sequences are similar to those acquired with the constrained sequences. Torsional (ϕ, ψ) values of residues belonging to lowest-energy conformations at distance-constraint values for which the WT sequence of each SP system is a minimum, and whose (ϕ, ψ) values reside in the Zimmerman-defined α-helical region, are presented as Ramachandran plots in Figures 4.23, 4.26 and 4.29. Helical wheel plots of these same minimum-energy conformations are supplied in Figures 4.24, 4.27 and 4.30.

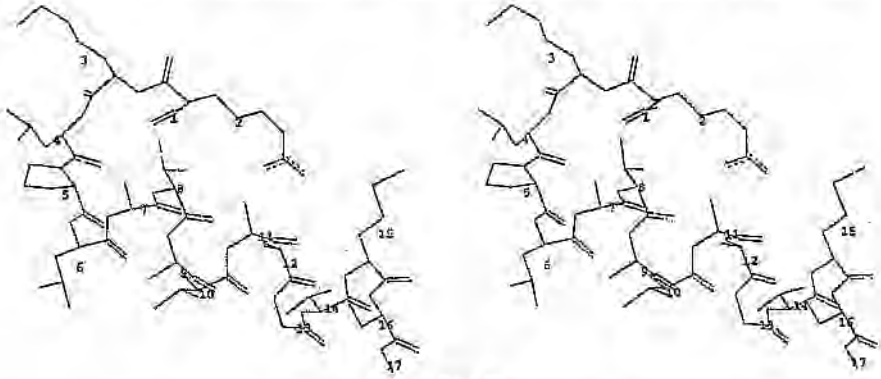
(a) 9.00 Å

E_{conf} = -74.6964 kcal mol^{-1}; residues 5 to 10 in the stereoview are α-helical



(b) 11.00 Å

E_{conf} = -86.4877 kcal mol^{-1}; residues 5 to 16 in the stereoview are α-helical



(c) 13.00 Å

E_{conf} = -78.4723 kcal mol^{-1}; residues 8 to 16 in the stereoview are α-helical

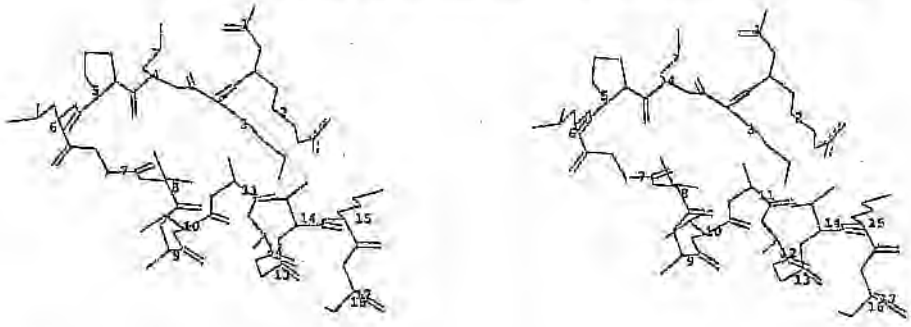


Figure 4.21: Stereoviews of minimised energy structures for the WT signal sequence of the LamB system determined with ECEPPGA at various distance-constraint values. Calculations were performed with a dielectric constant of 2.

4.2.4.1. LamB

The α-helical secondary structure is adopted by the lowest-energy conformations of the five peptide sequences of the LamB system (see Table 4.13). Of the fifteen residues (arg-6 to ser-20) in the calculated sequences of the WT and A13D, twelve (pro-9 to ser-20) conform to the α-helix. Of the 11 residues (arg-6 to ser-16) in Δ78r2, Δ78r1 and Δ78, Δ78r2 adopts a fully α-helical structure, while Δ78r1 and Δ78 adopt a partly α-helical structure from pro-9 to ser-16. The initiation of helices in the sequences which contain proline occurs at the proline position.

Table 4.13: A comparison of the lowest-energy minimised structures for the LamB signal peptide system determined with ECEPPGA, with and without distance-constraints. Calculations were performed with a dielectric constant of 2.

Signal peptide	Distance-constrained		Unconstrained		Residues in α-helical conformation
	Distance-constraint /Å	E_{conf} /kcal mol^{-1}	End to end distance[a] /Å	E_{conf} /kcal mol^{-1}	
WT	11.00	-86.4877	10.8805	-86.4601	pro-9 to ser-20
Δ78r2	11.00	-68.6256	10.9630	-68.6445	arg-6 to ser-16
Δ78r1	10.50	-77.1797	10.4162	-77.1498	pro-9 to ser-16
A13D	11.00	-101.610	10.8976	-100.773	pro-9 to ser-20
Δ78	10.50	-75.9978	10.4370	-76.0534	pro-9 to ser-16

[a] measured distance between end points described in Table 3.1

A comparison of the graphs from the calculations on the LamB sequences, Figure 4.22, shows that the moderately active Δ78r1 and inactive Δ78 sequences are most stable when the distance between the defined end points is approximately 10.50 Å, whereas this distance is about 11.00 Å for the WT, A13D and Δ78r2. If ideal α-helical conformation is assumed for all the constrained residues in each peptide, *i.e.*, a mean rise per residue of 1.5 Å, the minimum energy distance-constraint would be expected to be 10.50 Å, since seven residues were placed under constraint. Although all seven constrained residues in the WT (from N_{leu-10} to C_{ala-16}), A13D(from N_{leu-10} to C_{ala-16}) and Δ78r2 (from N_{leu-8} to C_{val-18}) display α-helical character, the observed distances (values in Table 4.13) are greater than 10.50 Å, so that these helices are somewhat extended. Of the seven constrained residues in Δ78r1 (from N_{leu-8} to C_{val-18}) and Δ78 (from N_{leu-8} to C_{val-18}), only six are α-helical. For both peptides, the seventh residue is leucine-8, which conforms to the β-sheet and which precedes proline in the sequence.

A comparison of energy-well depths in Figure.4.22 appears to indicate that the deletion mutant without proline, Δ78r2, exhibits a slightly narrower and deeper well than the two proline-containing deletion mutants, Δ78r1 and Δ78. This suggests that Δ78r2 has a higher α-helical conformational specificity than do Δ78r1 and Δ78. The longer calculated helix of Δ78r2, relative to Δ78r1 and Δ78, seems to corroborate this suggestion. Although the energy-well of the WT seems to be narrower and deeper than that of A13D, inferring that the WT is more likely to form an α-helix, results in Table 4.13 do not offer any differentiating information concerning structure.

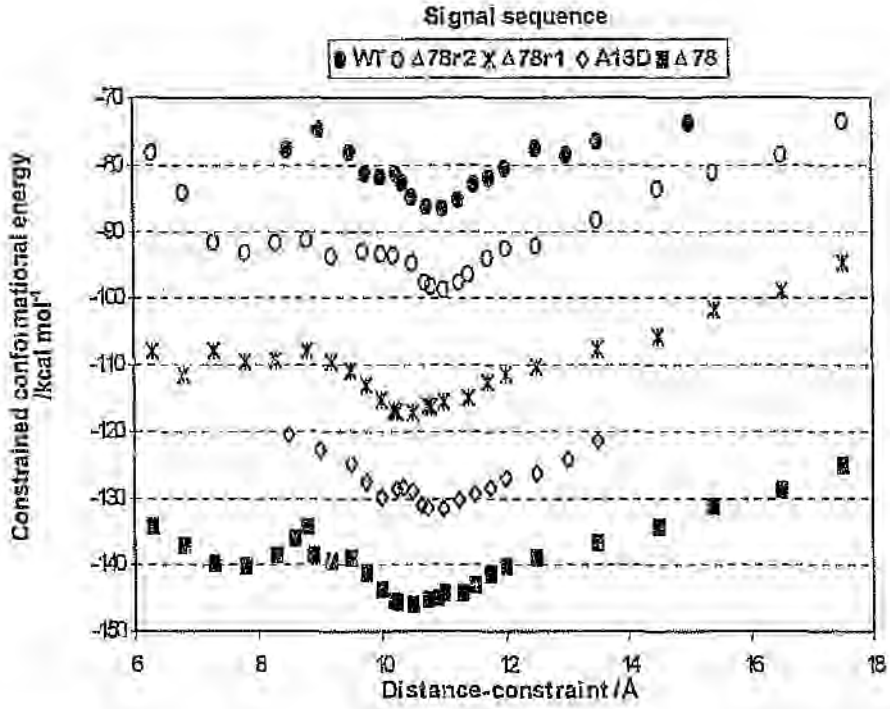


Figure 4.22: Minimised energy *versus* distance-constraint curves for the signal sequences of the LamB system determined with ECEPPGA with a dielectric constant of 2. The scale on the energy axis is correct for the WT curve. Energy values for the remaining curves have been shifted to accommodate this scale. The curves are arranged in order of translocation efficiency: WT (unshifted); Δ78r2 (shifted by -30 kcal mol^{-1}); Δ78r1 (shifted by -40 kcal mol^{-1}); A13D (shifted by -30 kcal mol^{-1}); Δ78 (shifted by -70 kcal mol^{-1}).

Ramachandran plots in Figure 4.23 show that at a distance-constraint of 11.00 Å, the (ϕ, ψ) values for all the amino acids in the computed Δ78r2 structure fall in a limited α-helical range of torsions, while the α-helices of the Δ78r1 and Δ78 structures contain amino acids with (ϕ, ψ) values which fall outside this limited range. The proline-9 residue, in particular, is removed from the range. These plots thus also infer that Δ78r2 has a greater tendency to form α-helices than do Δ78r1 and Δ78. However, the plots do not illuminate any conformational differences between Δ78r1 and Δ78, and between the WT and A13D.

If it is assumed that formation of an α-helix is necessary for peptide export, a higher degree of specificity for the α-helix would result in a higher efficiency of translocation across the membrane. According to translocation efficiencies (Figure 3.1), Δ78r2 is transported across the membrane with more success than are Δ78r1 and Δ78. Results from ECEPPGA searches agree with this finding but do not, however, explain the differences in activity between Δ78 and Δ78r1; Δ78r1 is 50% active, while Δ78 is completely inactive. On inspection of the α-helical contents calculated from CD spectra (in SDS micelles) in Table 4.1, Δ78r1 and Δ78 exhibit a 5% difference in content with a mean value of 37.5%, while Δ78r2 shows a relatively higher content at 75%. The computational procedure that we have used may thus be sufficient to differentiate among signal peptides on the basis of α-helical content if large differences in that content exist, but cannot discriminate between signal peptides with similar helical tendencies. It thus cannot discriminate among translocation efficiencies if α-helical content is to be the determining factor. This point is reinforced when data for the WT and A13D sequences are compared. The curves for the peptides, depicted in Figure 4.22, are similar and their α-helical contents differ by only 10%, but their effective activities (Table 3.1) are very different.

Since the structural properties of signal peptides, as reported here, do not appear to be determinative with regard to translocation efficiency, the physical properties of these peptides are probably also relevant. Indeed, it has already been concluded in an earlier section of this chapter that the hydrophobicity of each sequence correlates well with efficiency of translocation. For example, one alanine residue in the WT signal peptide has been replaced by the charged, less hydrophobic residue, aspartate, to form the A13D signal peptide. This replacement lowers the overall hydrophobicity of the sequence, which correlates with a reduced functional ability.

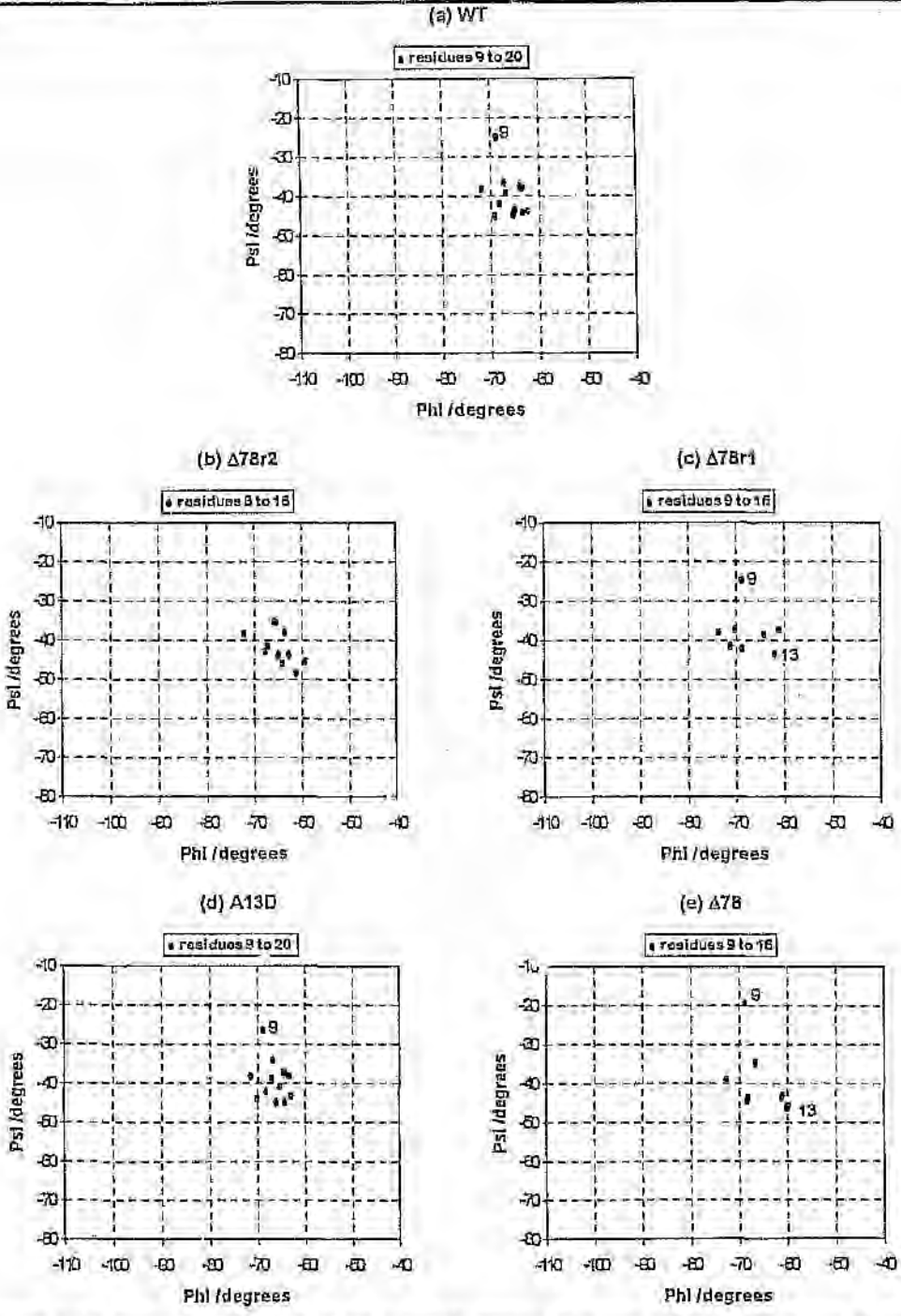


Figure 4.23: Ramachandran plots of the α-helical regions of energy-minimised structures of the signal sequences of the LamB system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 11.00 Å. Calculations were performed with a dielectric constant of 2. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency.

Amphipathicity of the α-helices is another physical property that may contribute to the functional differences exhibited by the signal sequences. However, apart from the helical plot for A13D, the helical plots in Figure 4.24 do not reveal definite divisions of the helical cylinders into polar and nonpolar sides.

Because neither secondary structure nor hydrophobicity nor amphipathicity appears to predominate in the determination of export ability, it must be presumed that the determinant is a combination of these factors. It has been previously suggested by Perez *et al.*[205], in their molecular mechanics study of the LamB system, that both α-helix formation and hydrophobicity of the signal peptide are the characteristics needed for successful protein export. It must be borne in mind, however, that the conformational searching strategies employed by them, *viz.*, build-up, random placement of points on the hypersurface, and rigid-geometry contouring, are not as thorough and extensive as the GA method used in this work, and their results must therefore be viewed with discretion.

The effect of the α-helix breaking residues, proline and glycine, on helix propagation was highlighted by Emr and Silhavy[203] after performing Chou-Fasman secondary structure calculations on the LamB signal peptides. They predicted that proline and glycine would cause α-helix disruption which would, in turn, decrease peptide translocation ability, and that protein export would be restored if the proline and glycine residues were replaced by residues that are commonly found in α-helices, leucine and cysteine in this instance. In our modelling work, the replacement of proline with leucine (converting Δ78 to Δ78r2) results in a fully α-helical structure, but the replacement of glycine with cysteine (converting Δ78 to Δ78r1) yields very similar results to those obtained with the original mutant, Δ78. Since glycine is considered to prefer the α-helix less than cysteine, it would be expected that Δ78r1 would show a higher helical content than Δ78.

It would appear from the data presented here that the presence of proline in sequences causes breaks in the α-helical forms of the entire sequences, and that the potential of glycine to be a helix-breaker is not as definitive in this molecular modelling approach as in the Chou-Fasman secondary structure prediction approach;[203] our observed minimum-energy structures indicate the propagation of α-helices from the proline residue and do not indicate any breaks in the helices due to glycine. These findings are in accord with the conclusions of Bruch and

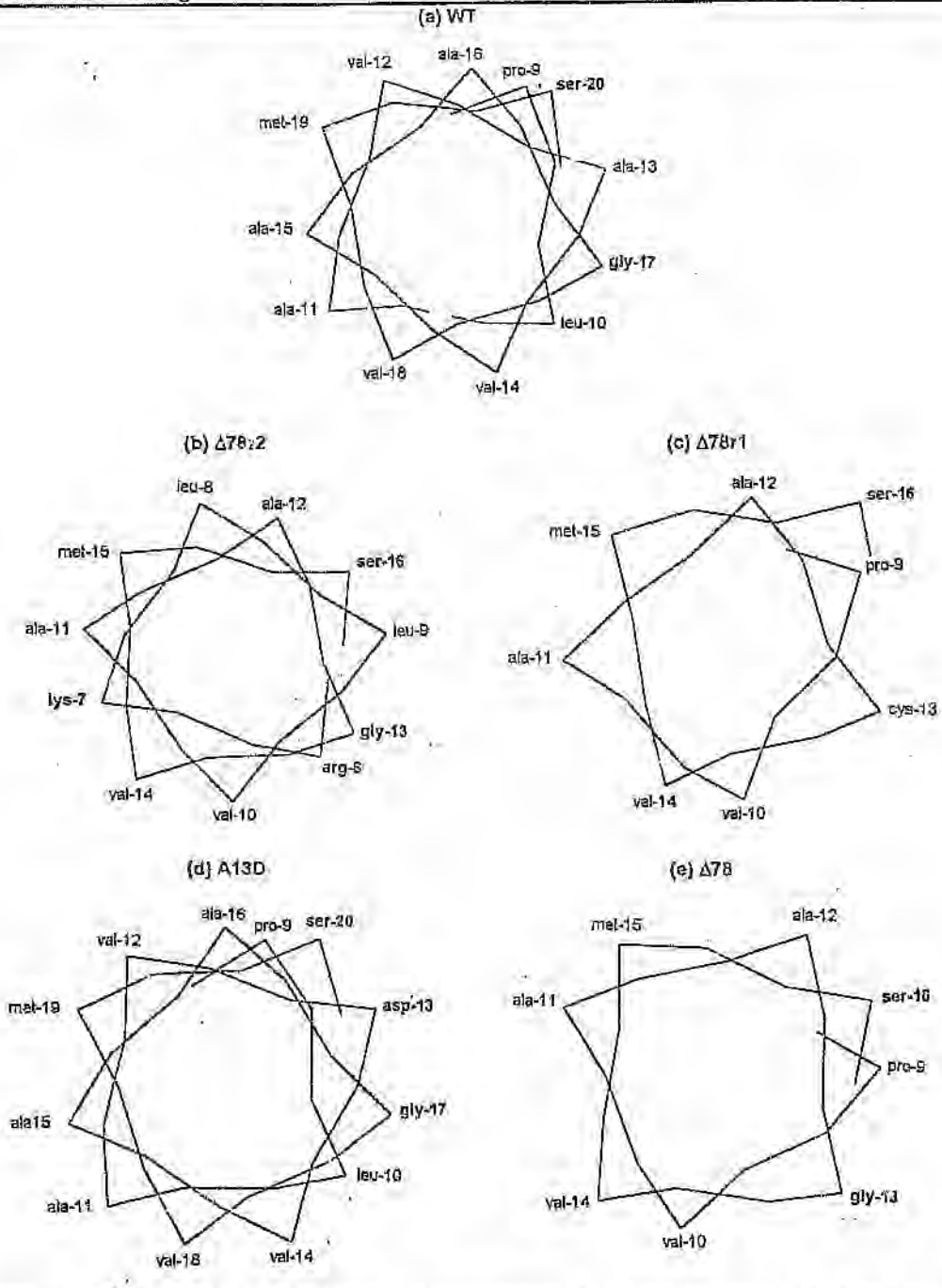


Figure 4.24: Helical wheel plots of energy-minimised structures of the signal sequences of the LamB system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 11.00 Å. Calculations were performed with a dielectric constant of 2. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency. Polar amino acids are shown in ***bold***.

Gierasch[193] who found that proline has a larger disruptive effect than does glycine, and that glycine does not have a strong helix-disrupting influence in the LamB mutant peptides.

4.2.4.2. CPY

The minimised structures of the CPY system display α-helical character throughout their investigated sequences (Table 4.14), resulting in the ...arly shaped curves in Figure 4.25 and the similar Ramachandran plots in Figure 4.26. In accord with the GA results for the LamB peptides, no direct correlation is found between translocation efficiency and extent of α-helical structure adopted by the lowest energy minimised structures of each sequence.

Table 4.14: A comparison of the lowest-energy minimised structures for the CPY signal peptide system determined with ECEPPGA, with and without distance-constraints. Calculations were performed with a dielectric constant of 2.

Signal peptide	Distance-constrained		Unconstrained		Residues in α-helical conformation
	Distance-constraint /Å	E_{cons} /kcal mol^{-1}	End to end distance[a] /Å	E_{cons} /kcal mol^{-1}	
CPYm6	12.00	-70.2973	12.0974	-70.4209	phe-4 to thr-15
CPYm2	13.75	-72.0831	13.8565	-72.2025	phe-4 to thr-16
CPYm8	13.75	-68.9508	13.8634	-69.1528	phe-4 to thr-16
CPYm12	12.00	-66.4533	12.0910	-66.5627	phe-4 to thr-15
WT	13.75	-68.5230	13.8530	-68.6361	phe-4 to thr-16

[a] measured distance between end points described in Table 3.1

The curves for the CPYm2, CPYm8 and the WT signal peptides in Figure 4.25 indicate that the peptides are most stable at a distance-constraint of approximately 13.75 Å. The minimum energy distance-constraint value observed for the CPYm6 and CPYm12 mutant peptides was 12.00 Å (see Table 4.14). This difference in values arises from the deletion of one residue in the sequences of both CPYm6 and CPYm12, causing them to be shorter in length than the other peptides. The CPYm6 and CPYm12 curves in Figure 4.25 have, therefore, been shifted by 1.50 Å, the mean rise per residue for an ideal helix, in order to eliminate this mutational artefact. The eight residues placed under constraint (from N_{ser-6} to C_{ser-13}) in each of these sequences appear to form ideal α-helical conformations, while the nine constrained residues (from N_{ser-6} to C_{ser-14}) in each of the CPYm2, CPYm8 and WT sequences appear to form α-helices that are slightly extended. The relatively lower minimum distance-constraint values obtained with the LamB sequences (Figure 4.22) result from the smaller number of residues constrained between the end points of the latter.

A comparison of energy-well depths in Figure 4.25 appears to indicate a slightly narrower and deeper well for CPYm6 than for CPYm12, thus suggesting that CPYm6 has a higher α-helical conformational specificity. This correlates with the translocational efficiencies recorded in Figure 3.2. However, energy-well depths of CPYm2, CPYm8 and the WT do not seem to correlate with efficiency. The sole use of these energy *versus* distance-constraint curves to predict translocational abilities is therefore questionable, and as suggested above, other factors such as hydrophobicity must play a role.

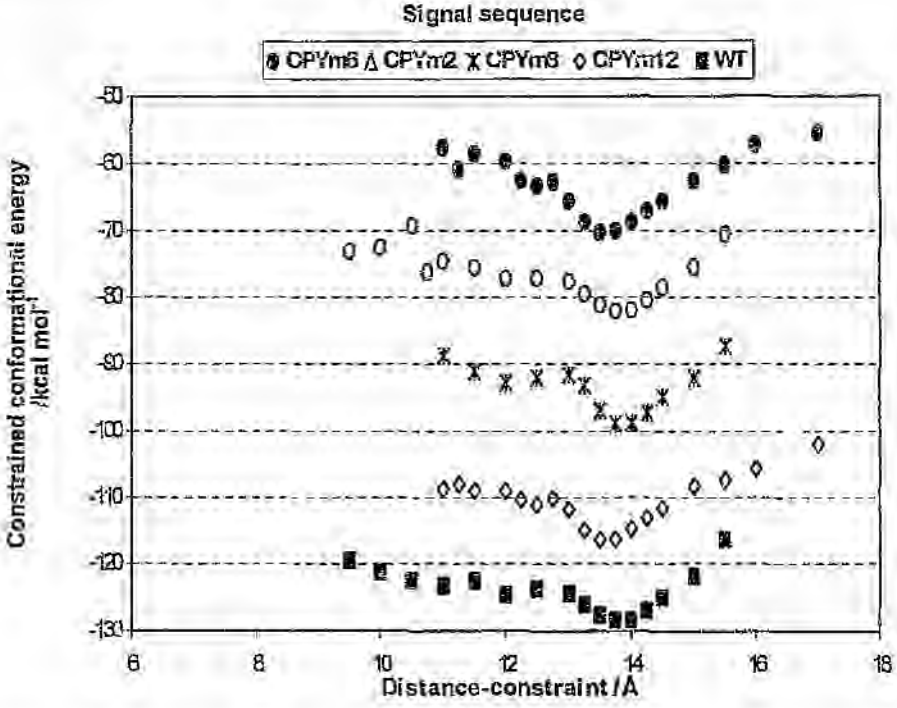


Figure 4.25: Minimised energy *versus* distance-constraint curves for the signal sequences of the CPY system determined with ECEPPGA with a dielectric constant of 2. The scale on the energy axis is correct for the WT curve. Energy values for the remaining curves have been shifted to accommodate this scale. Plots for CPYm6 and CPYm12 have also been shifted right by 1.50 Å along the distance-constraint axis, because one fewer residue of these sequences was constrained during modelling. The curves are arranged in order of translocation efficiency: CPYm6 (unshifted); CPYm2 (shifted by -10 kcal mol^{-1}); CPYm8 (shifted by -30 kcal mol^{-1}); CPYm12 (shifted by -50 kcal mol^{-1}); WT (shifted by -60 kcal mol^{-1}).

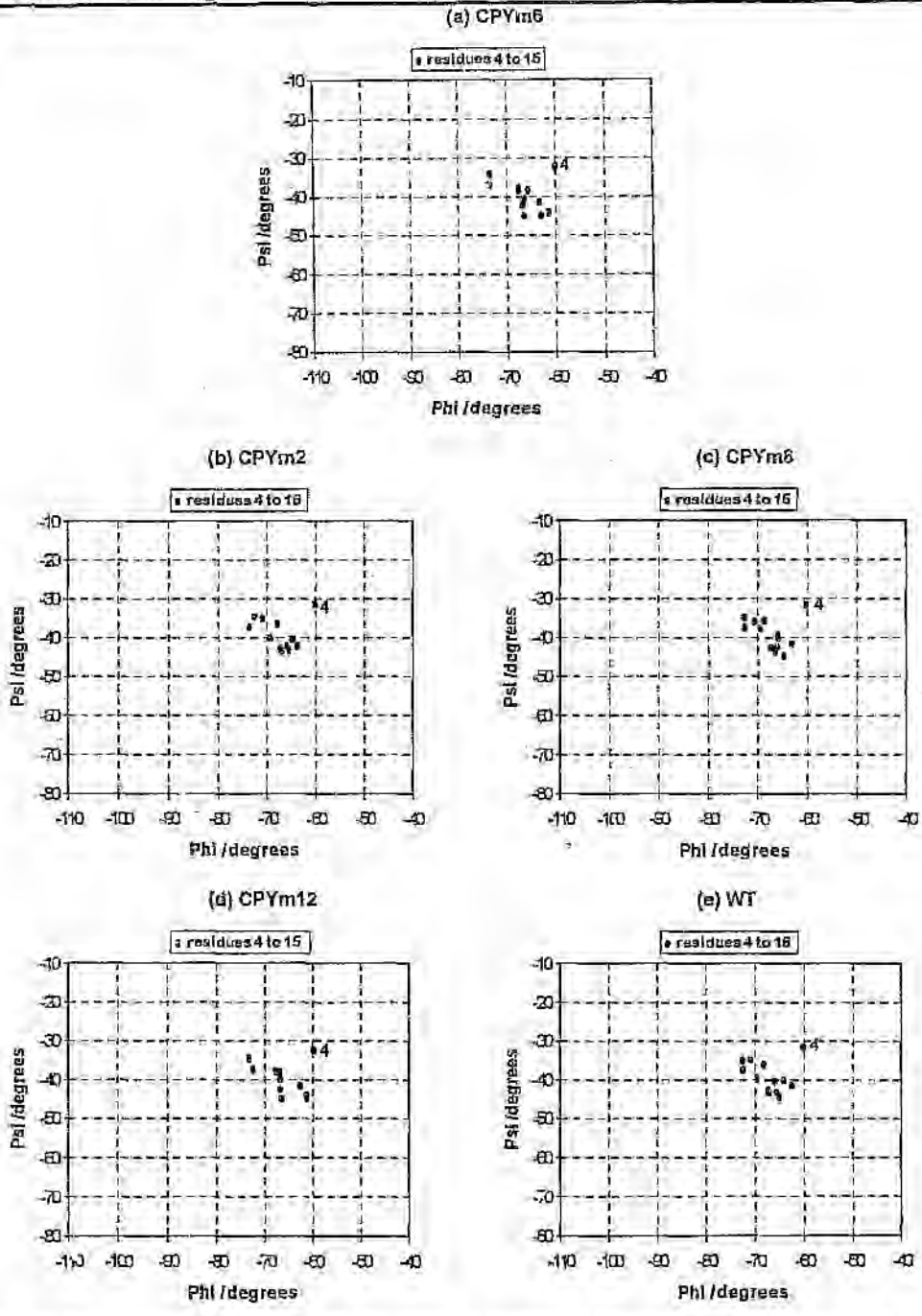


Figure 4.26: Ramachandran plots of the α-helical regions of energy-minimised structures of the signal sequences of the CPY system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 13.75 Å for CPYm2, CPYm8 and WT, and 12.25 Å for CPYm6 and CPYm12. Calculations were performed with a dielectric constant of 2. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency.

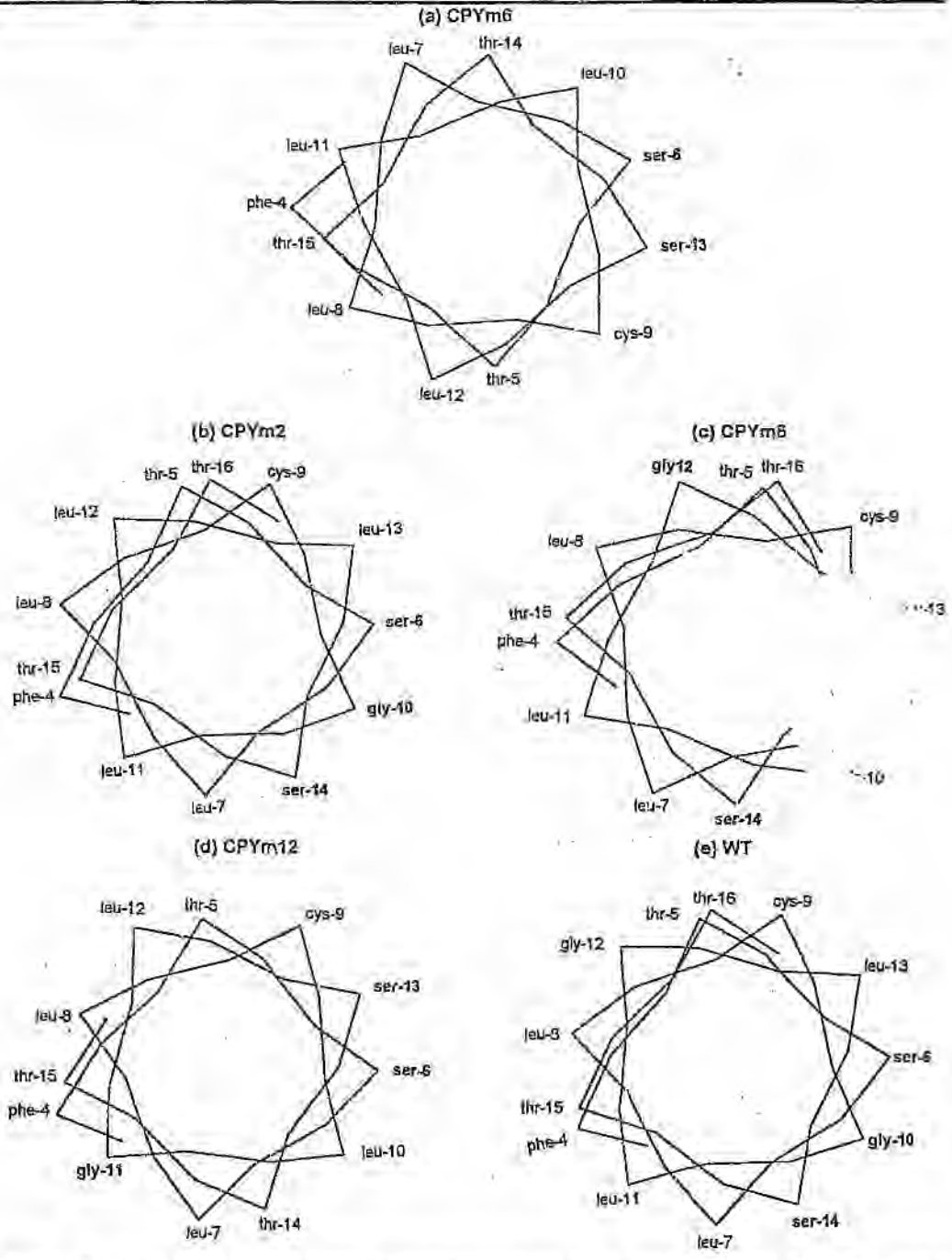


Figure 4.27: Helical wheel plots of energy-minimised structures of the signal sequences of the CPY system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 13.75 Å for CPYm2, CPYm8 and WT, and 12.25 Å for CPYm6 and CPYm12. Calculations were performed with a dielectric constant of 2. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency. Polar amino acids are shown in *bold*.

Similarly to the LamB system, the α-helices of the minimum-energy conformations for the CPY peptides, sketched in Figure 4.27, do not exhibit strong amphipathic character.

Chou-Fasman calculations performed by Bird *et al.*[204] predict breaks in the α-helices of CPYm2, CPYm8, CPYm12 and the WT due to the presence of glycine. However, the results here indicate that all the residues in these sequences form one α-helical chain. Thus, as with the LamB system, the study shows that glycine is not a helix-breaker of sufficient strength to cause a disruption in the secondary structures of these signal peptides.

4.2.4.3. gC

The influence exerted by the strongly helix-breaking proline residue on secondary structure formation is clearly discernible from the results presented here. In Figure 4.28, the curves for the signal peptides which do not contain proline, the WT and ΔA10, show definite parabolic-like character, whereas the curves for the peptides which do contain proline, A10P, A12P and L14P, are much flatter by comparison. Disruptions in the α-helices of the proline-containing sequences are reported in Table 4.15. These breaks occur at residues positioned just before proline.

The curve for the WT signal peptide in Figure 4.28 indicates that the peptide is most stable at a distance-constraint of approximately 14.00 Å. The minimum energy distance-constraint value observed for the shorter ΔA10 mutant peptide was 12.00 Å (see Table 4.15). Since ΔA10 is one residue shorter in length than the WT, its curve has been shifted by 1.50 Å. The eight residues placed under constraint (from $N_{ala\text{-}7}$ to $C_{tyr\text{-}14}$) in ΔA10) appear to form an ideal α-helix, while the nine constrained residues (from $N_{ala\text{-}7}$ to $C_{tyr\text{-}15}$) in the WT form an extended α-helix. Although conformations at minimum distance-constraint values are reported for A10P, L12P and L14P in Table 4.15, the conformations along their curves are so similar in energy, that these choices of lowest-energy conformations should be viewed as somewhat arbitrary. Their minimum distance-constraint values and minimum energy values are therefore shown in parenthesis in the table.

Table 4.15: A comparison of the lowest-energy minimised structures for the gC signal peptide system determined with ECEPPGA, with and without distance-constraints. Calculations were performed with a dielectric constant of 2.

Signal peptide	Distance-constrained		Unconstrained		Residues in α-helical conformation
	Distance-constraint /Å	E_{conf} /kcal mol^{-1}	End to end distance[a] /Å	E_{conf} /kcal mol^{-1}	
WT	14.00	-68.6478	13.9000	-68.7878	arg-6 to ala-16
ΔA10	12.00	-67.4761	12.1043	-67.6108	arg-6 to ala-15
A10P	(12.25)	(-78.7202)	12.2215	-78.7248	arg-6 to met-8, pro-10 to tyr-15
L12P	(9.75)	(-75.5688)	8.9126	-76.0517	arg-6 to ala-10, pro-12 to ala-16
L14P	(12.50)	(-76.9027)	11.4780	-76.5282	arg-6 to leu-12, pro-14 to ala-16

[a] measured distance between end points described in Table 3.1

Ramachandran plots in Figure 4.29 show that at a distance-constraint of 14.00 Å (12.50 Å for ΔA10 because of its shorter length), the WT structure consists of an α-helix that is more uniform than that of ΔA10. It can be clearly seen from plots (c), (d) and (e) that the proline residue in positions 10, 12 and 14 respectively distorts the α-helical structure.

These results verify once more that no distinct correlation exists between translocation efficiency and extent of α-helical structure adopted by each sequence. The transport abilities of the gC WT and ΔA10 can be accounted for, but those of the proline-containing peptides cannot. Thus, in agreement with the observations made by Ryan and Edwards,[257] whose work is the source for this gC system study, it must be concluded that a disturbance of secondary structure and a reduction of sequence hydrophobicity are jointly responsible for signal sequence dysfunction. Weak α-helix amphipathicities of the low-energy gC conformations (Figure 4.30) lead to the conclusion that amphipathicity has limited influence on functional efficiency.

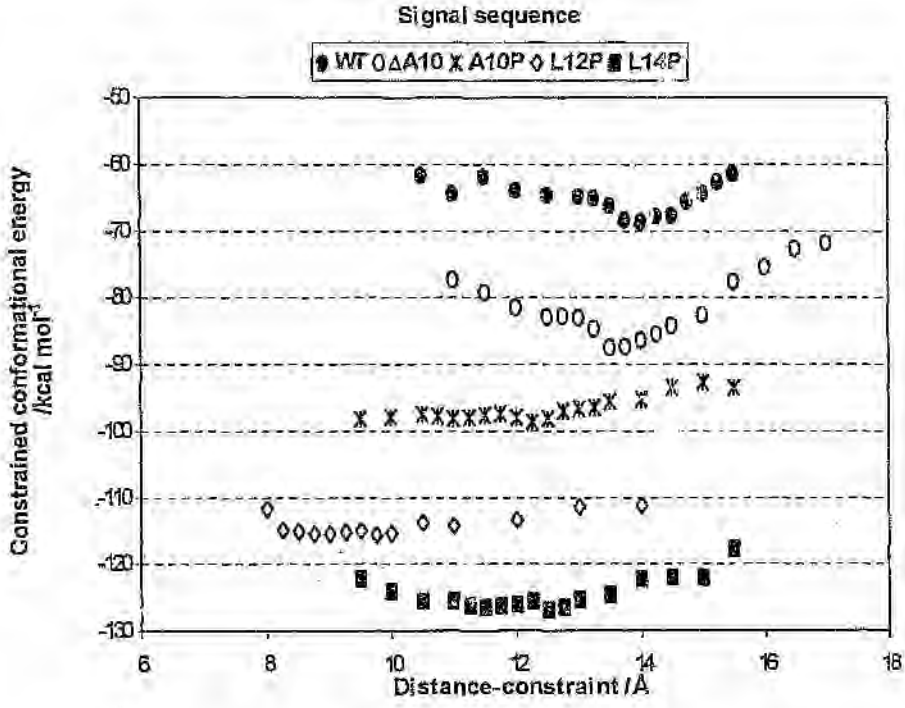


Figure 4.28: Minimised energy *versus* distance-constraint curves for the signal sequences of the gC system determined with ECEPPGA with a dielectric constant of 2. The scale on the energy axis is correct for the WT curve. Energy values for the remaining curves have been shifted to accommodate this scale. The plot for ΔA10 has also been shifted right by 1.50 Å along the distance-constraint axis, because one fewer residue of this sequence was constrained during modelling. The curves are arranged in order of translocation efficiency: WT (unshifted); ΔA10 (shifted by -20 kcal mol^{-1}); A10P (shifted by -20 kcal mol^{-1}); L12P (shifted by -40 kcal mol^{-1}); L14P (shifted by -50 kcal mol^{-1}).

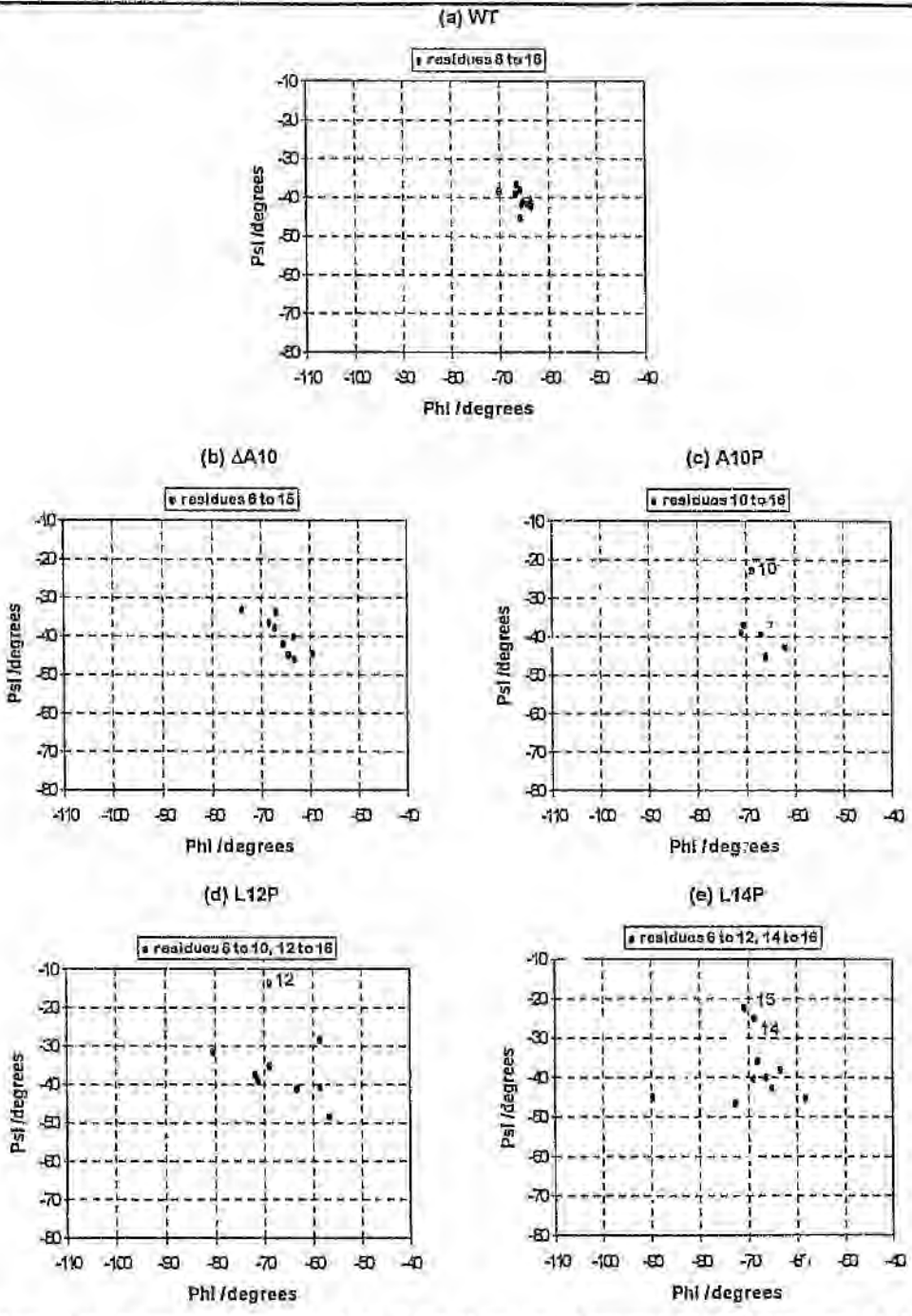


Figure 4.29: Ramachandran plots of the α-helical regions of energy-minimised structures of the signal sequences of the gC system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 14.00 Å for WT, A10P, L12P and L14P, and 12.50 Å for ΔA10. Calculations were performed with a dielectric constant of 2. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency.

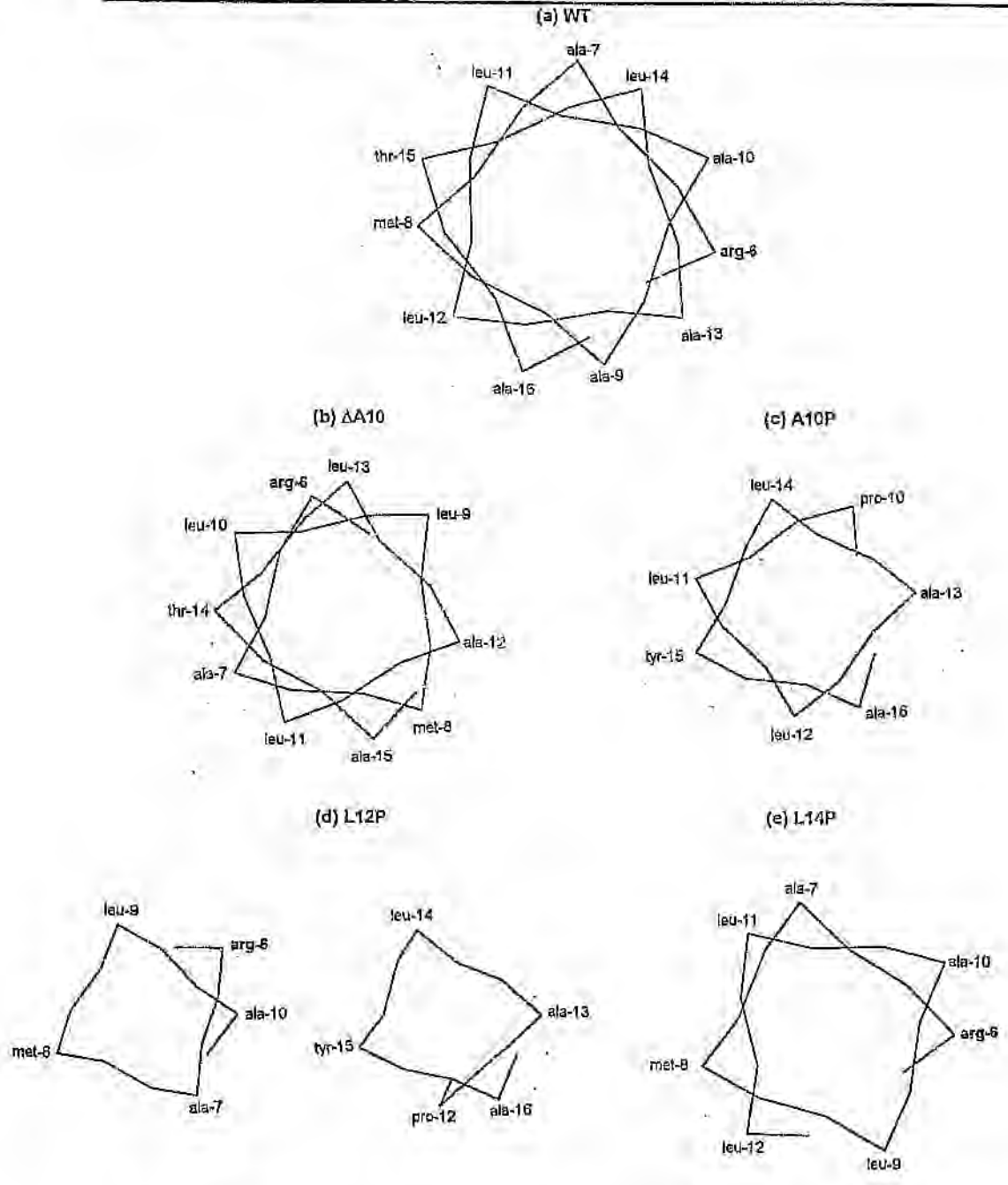


Figure 4.30: Helical wheel plots of energy-minimised structures of the signal sequences of the gC system determined with ECEPPGA at a distance-constraint for which the WT is a minimum, *i.e.*, 14.00 Å for WT, A10P, L12P and L14P, and 12.50 Å for ΔA10. Calculations were performed with a dielectric constant of 2. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency. Polar amino acids are shown in ***bold***.

4.2.5. ECEPPGA conformational searches with modified modelling conditions

Apart from a dependence on force field, molecular modelling calculations are also dependent on experimental conditions such as hypervolume of conformational space to search, length of calculated sequence, adopted proline configuration (where proline is present in a sequence), and dielectric permitivitty. These afore-mentioned conditions were explored, each to varying extents; results thereof are presented below.

4.2.5.1. Restricted conformational space

The ECEPPGA program includes an option which allows torsional angle sampling in user-defined conformational regions. Modelling in restricted space was carried out in this work firstly, to examine the capability of the genetic algorithm to find GMECs on selection of this option, and secondly, to locate all regions necessary to sufficiently explore the conformational surfaces of signal sequences. The technique of torsional angle sampling in restricted space is anticipated to be particularly beneficial for large peptides for which adequate searching of unrestricted space is computationally difficult.

The three deletion mutants of the LamB system, namely Δ78r2, Δ78r1 and Δ78, were investigated in these runs, which were also executed without distance-constraints. The Zimmerman[69] regions corresponding to the right-handed α-helix (A region) and the β-strand (C, D, E, F regions) were individually searched. GA parameters whose values were modified from those used in runs described in section 4.2.3 are width of local mutation[dd] (doubled), torsional angle sampling strategy, number of generations counted per run (increased to 20), and population size (halved). The ratio of generation gap size to population size was maintained at 0.3. Gaussian distributions within selected backbone conformational regions as opposed to selected residue conformational regions was the chosen sampling strategy.

In Tables 4.16 and 4.17, results from runs conducted in limited space corresponding to the α-helical conformation and to the β-strand conformation are compared with those obtained from runs conducted in unlimited space. It can be noted that higher conformational energies and larger end-to-end distances are attained when the moderately active Δ78r1 and inactive Δ78

[dd] "Width of local mutation" is the fraction of the total allowed torsional space that is available during a local mutation.

sequences are compelled to form α-helical secondary structures. Runs in unrestricted and restricted space for the active Δ78r2 sequence yield equivalent conformations with the restricted run requiring fewer generations for completion. The technique of limiting torsional angle space can therefore be used to reduce computation time.

Table 4.16: The effect of α-helical restricted conformational space on unconstrained conformational energies of the deletion mutants of the LamB signal peptide system determined with ECEPPGA. Calculations were performed with a dielectric constant of 2.

Signal peptide	Unrestricted space				α-restricted space			
	E_{conf} /kcal mol^{-1}	End to end distance[a] /Å	Residues in α-helical conformation	Number of gen.[b]	E_{conf} /kcal mol^{-1}	End to end distance[a] /Å	Residues in α-helical conformation	Number of gen.[b]
Δ78r2	-68.6445	10.9630	arg-6 to ser-16	52	-68.6454	10.9630	arg-6 to ser-16	37
Δ78r1	-77.1498	10.4162	pro-9 to ser-16	50	-72.8305	11.0690	arg-6 to ser-16	45
Δ78	-76.0534	10.4370	pro-9 to ser-16	53	-71.7770	11.0890	arg-6 to ser-16	40

[a] measured distance between end points described in Table 3.1
[b] the number of actual generations cycled during the run

As expected, higher energies for β-constrained structures are obtained (see Table 4.17). These results can be correlated with functional ability if differences in energy between unlimited space conformations and β-limited space conformations are considered. As translocation ability decreases from Δ78r2 to Δ78r1 to Δ78, this energy difference also decreases; this emphasises the fact that α-structures of functional signal peptides are much more stable than their β-structures, while α-structures of dysfunctional peptides are closer in energy to their β-structure counterparts.

Table 4.17: The effect of β-strand restricted conformational space on unconstrained conformational energies of the deletion mutants of the LamB signal peptide system determined with ECEPPGA. Calculations were performed with a dielectric constant of 2.

Signal peptide	E_{conf} /kcal mol^{-1}		
	Unrestricted space	β-restricted space	Δ(β-restricted – unrestricted)
Δ78r2	-68.6445	-48.1534	20.4911
Δ78r1	-77.1498	-61.6004	15.5494
Δ78	-76.0334	-62.3873	13.6661

The problem with imposing limits on torsional angle space is that one has to ensure that all the appropriate regions are searched and that no regions are excluded that may be needed. For example, the appropriate region (Zimmerman A region) for Δ78r2 was searched, yet this region was too restrictive for Δ78r1 and Δ78 because their lowest-energy conformations are not completely α-helical. Although it was assumed that the results reported here are those of the lowest-energy conformations of the sequences subject to experimental conditions, ECEPPGA parameter settings still require optimisation for minimisation in restricted space.

4.2.3.2. Varying sequence length

It was mentioned in section 3.3.2 that only those lengths of the signal sequences that were considered essential for analysis were used in the calculations. To investigate the validity of our choice of lengths, shorter sequences of the LamB WT and A13D, and of the CPY WT and CPYm2 were subjected to computation. A comparison of modelling conditions is tabulated in Table 4.18. ECEPPGA parameter values were unchanged.

Table 4.18: The varying sequences and distance-constraint end points selected for GA modelling

Signal peptide	Selected sequence	Number of residues in selected sequence	Distance-constraint end points[a]	Number of distance-constrained residues
LamB				
WT	arg-6 to ser-20	15	leu-10 to ala-16	7
short_WT	leu-8 to val-18	11	leu-10 to ala-16	7
A13D	arg-6 to ser-20	15	leu-10 to ala-16	7
short_A13D	leu-8 to val-18	11	leu-10 to ala-16	7
CPY				
CPYm2	phe-4 to thr-16	13	ser-6 to ser-14	9
short_CPYm2	thr-5 to thr-15	11	leu-7 to leu-13	7
WT	phe-4 to thr-16	13	ser-6 to ser-14	9
short_WT	thr-5 to thr-15	11	leu-7 to leu-13	7

[a] from the *N*-atom of the first residue to the *C*-atom of the second

Minimised energy *versus* distance-constraint curves for the LamB WT and A13D sequences can be contrasted with the curves of their shorter variants in Figure 4.31. No prominent differences between the shapes of the curves are observed, and similar distance-constraint values for which the energy of the sequences are a minimum are noted. Hence, a decrease in calculated sequence length, specifically for the LamB system peptides, would probably have been warranted in this study.

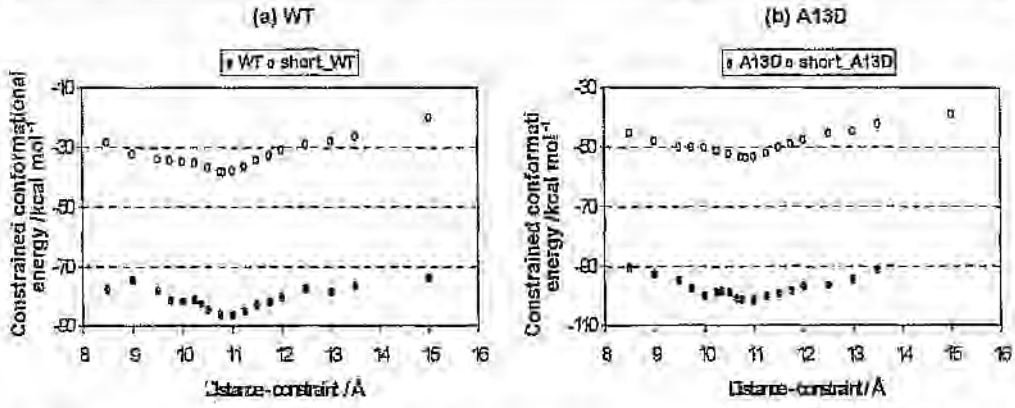


Figure 4.31: Minimised energy *versus* distance-constraint curves for the WT and A13D signal sequences of the LamB system determined with ECEPPGA, using varying selected sequence lengths

The curves for the CPYm2 and WT signal peptides in Figure 4.32 indicate that the peptides are most stable at a distance-constraint of approximately 13.75 Å. The minimum energy distance-constraint value observed for the short_CPYm2 and short_WT peptides was approximately 10.75 Å. This shift in the energy potential well experienced by the shorter sequences arises from the fewer number of residues that were chosen to be distance-constrained. The difference in minimum energy distance-constraint values concurs with the length of two residues in an ideal α-helix. The short_CPYm2 and short_WT curves in Figure 4.32 have, therefore, been shifted by 3.00 Å. Since the shapes of the curves are not markedly dissimilar, a decrease in calculated sequence length, specific to this system, would also have been justified.

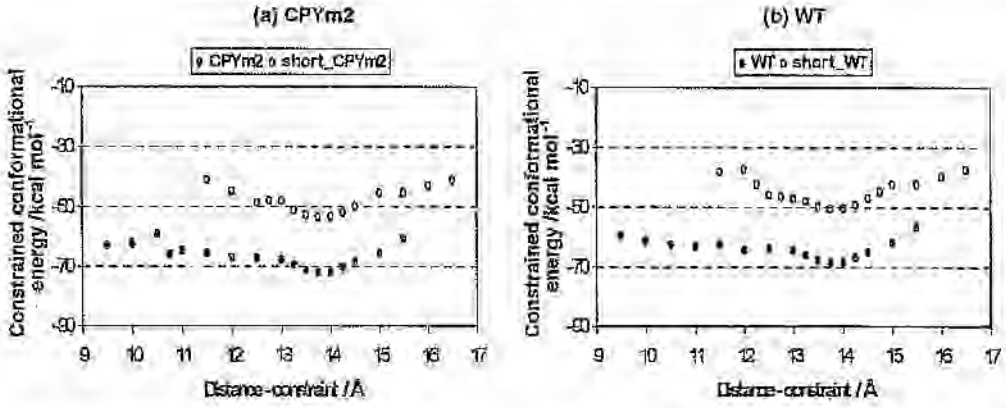


Figure 4.32: Minimised energy *versus* distance-constraint curves for the CPYm2 and WT signal sequences of the CPY system determined with ECEPPGA, using varying selected sequence lengths. Plots for short_CPYm2 and short_WT have been shifted right by 3.00 Å along the distance-constraint axis, because two fewer residues of the sequence was constrained during modelling.

4.2.5.3. Varying dielectric permittivity

Up till now, solvation has been unaccounted for in this molecular modelling study. For reasons already expounded in the Introduction (section 1.1.3.4), and owing to the speculated behaviour of SPs *in vivo* (the signal hypothesis postulates that nascent peptides cross cell membranes *via* aqueous pores, and the membrane trigger hypothesis postulates that nascent peptides undergo conformational changes on moving from the cytoplasm into the lipid membrane layer), the issue of solvation needs to be addressed.

The membrane trigger hypothesis was initially assumed in the present study, prompting the use of the ECEPP distance-dependent default dielectric constant (ε) of 2, which simulates the hydrophobic environment of the membrane bilayer. This "effective" ε of 2 actually corresponds to a medium with a ε of ~4 (estimated ε of polypeptide crystals).[89] To examine the influence of hydration on SP structure, ε was augmented (from 2 to 10, 40 and 80) to reflect increasingly hydrophilic environments. This approach was considered appropriate since the effect of environment on signal peptides is non-specific, *i.e.*, there are no specific hydrogen bonds between the signal peptides and their surroundings. Apart from changes in dielectric permittivity, no other modifications to modelling conditions were made.

Quantitative information about the lowest-energy conformation calculated for each of the signal sequences studied in this portion of the work, *i.e.*, the LamB system, CPYm2, the WT of CPY, the WT of gC, and L14P, as a function of ε is supplied in Tables 4.19, 4.20 and 4.21; data for both distance-constrained and unconstrained runs are shown (except for the CPY peptides, where only unconstrained runs were conducted). The last column in each table is appropriate to both sets of runs. Minimised energy *versus* distance-constraint plots, also as a function of ε, are presented in Figures 4.33 and 4.34.

An increase in ε does not appear to impact greatly on peptide secondary structure. Evidence for this is seen from the similar shapes of the curves in the figures, specific to each peptide, from the last column in each table (the α-helical structure for each peptide remains constant with a variation in ε), and from the almost constant end-to-end distances measured for each peptide. The exception in the latter case is L14P, a mutant of the gC signal peptide; the inconsistency in its end-to-end distances results from the peptide's lack of specificity for a regular structure, *viz.* the flat graphs in Figure 4.34.

Minimum-energy values decrease with an increase in ε; higher dielectric constants effectively mean suppression or removal of charges on the peptide, thereby inducing a reduction in repulsions and attractions and a subsequent lowering in energy. For each of the studied signal sequences (besides the LamB deletion mutants for which no calculations with $\varepsilon = 40$ were performed), there are nominal differences in energy between the $\varepsilon = 40$ and the $\varepsilon = 80$ conformations. This observation is probably a consequence of the force field, with $\varepsilon = 40$ simulating the water environment.

On comparing the conformational energy curves of the LamB WT with those of A13D in Figure 4.33, it can be noted that the energy difference between conformations calculated with the two dielectric constant extremes (totally hydrophobic, $\varepsilon = 2$ and totally hydrophilic, $\varepsilon = 80$) is larger for the WT than yields the same trend as functional activity decreases from Δ78r2 to Δ78r1 to Δ78, the difference in energy between the $\varepsilon = 2$ and the $\varepsilon = 80$ conformations decreases as well. This same phenomenon occurs for the gC WT and L14P (Figure 4.34), but not for the peptides belonging to the CPY system (Table 4.20). Results obtained for the LamB and gC sequences endorse the membrane trigger hypothesis; the larger energy differences imply that highly active signal peptides are much more stable in a hydrophobic medium than in aqueous, while the

smaller energy differences imply that the stabilities of less active peptides in hydrophobic and hydrophilic media are comparable.

Table 4.19: The effect of dielectric constant on the lowest-energy minimised structures for the LamB signal peptide system determined with ECEPPC/A, with and without distance-constraints. – = not calculated.

Signal peptide	Dielectric constant	Distance-constrained		Unconstrained		Residues in α-helical conformation
		Distance-constraint /Å	E_{conf} /kcal mol^{-1}	End to end distance[a] /Å	E_{conf} /kcal mol^{-1}	
WT	2	11.00	-86.4877	10.8805	-86.4601	pro-9 to ser-20
	10	10.75	-100.719	10.8806	-100.912	pro-9 to ser-20
	40	11.00	-106.350	10.8768	-106.502	pro-9 to ser-20
	80	11.00	-108.360	10.8795	-108.357	pro-9 to ser-20
Δ78r2	2	11.00	-68.6256	10.9630	-68.6445	arg-6 to ser-16
	10	–	–	10.9724	-79.2867	arg-6 to ser-16
	80	11.00	-84.9244	10.9771	-84.9321	arg-6 to ser-16
Δ78r1	2	10.50	-77.1797	10.4162	-77.1498	pro-9 to ser-16
	10	–	–	10.4014	-82.4256	pro-9 to ser-16
	80	10.50	-85.8220	10.4117	-85.9002	pro-9 to ser-16
A13D	2	11.00	-101.610	10.8976	-100.773	pro-9 to ser-20
	10	–	–	10.8723	-110.922	pro-9 to ser-20
	40	11.25	-112.176	10.8852	-111.087	pro-9 to ser-20
	80	11.00	-112.586	11.0059	-113.768	pro-9 to ser-20
Δ78	2	10.50	-75.9978	10.4370	-76.0534	pro-9 to ser-16
	10	–	–	10.4749	-80.3297	pro-9 to ser-16
	80	10.50	-83.0187	10.4380	-83.0577	pro-9 to ser-16

[a] measured distance between end points described in Table 3.1

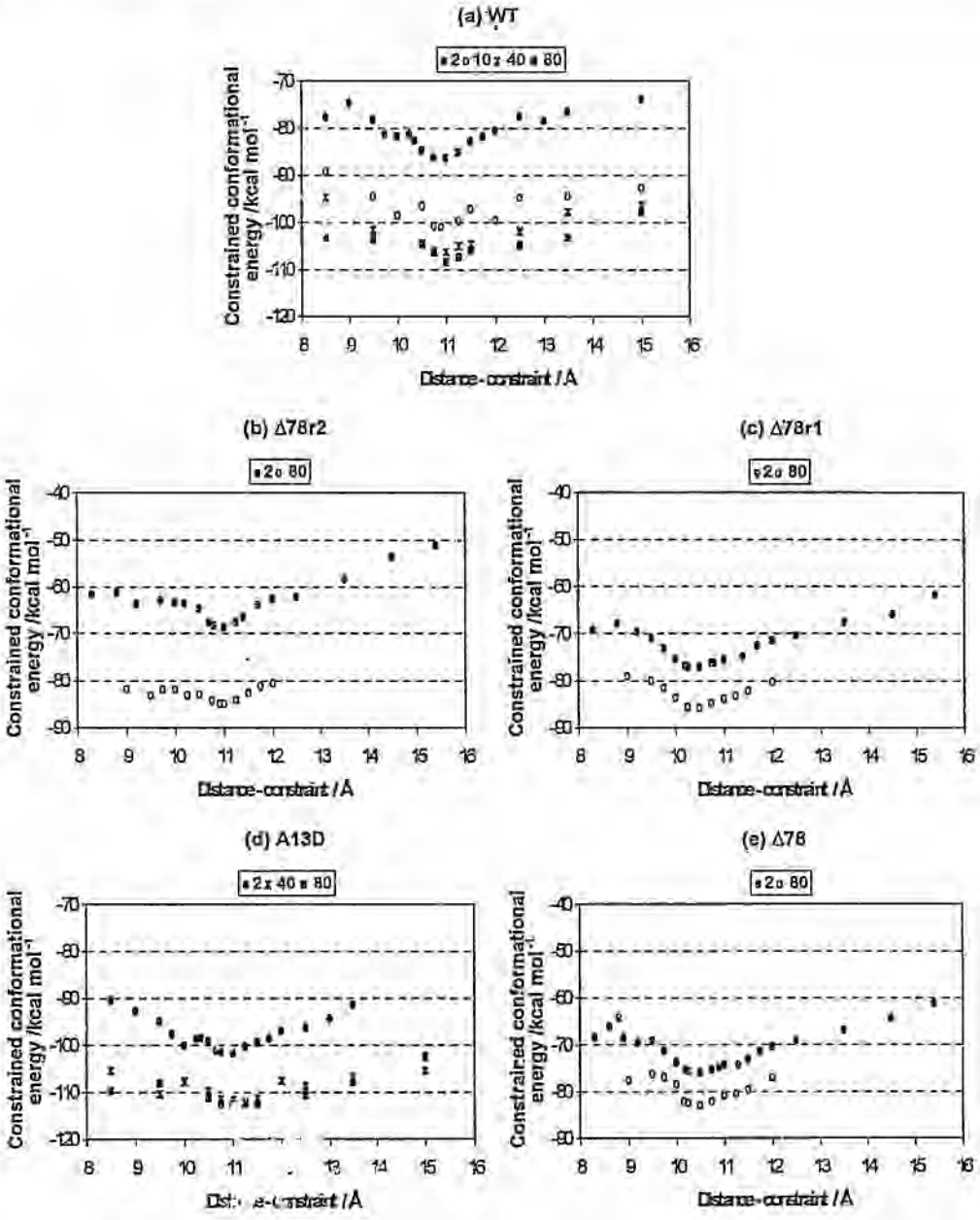


Figure 4.33: Minimised energy *versus* distance-constraint curves for the signal sequences of the LamB system determined with ECEPPGA with varying dielectric constants. The graphs are arranged (a) to (e) in order of decreasing translocation efficiency.

Table 4.20: The effect of dielectric constant on the lowest-energy minimised structures for the CPYm2 and WT signal sequences of the CPY system determined with ECEPPGA, without distance-constraints

Signal peptide	Dielectric constant	End to end distance[a] /Å	E_{conf} /kcal mol^{-1}	Residues in α-helical conformation
CPYm2	2	13.8563	-72.2025	phe-4 to thr-16
	10	13.8651	-96.5760	phe-4 to thr-16
	40	13.8684	-106.514	phe-4 to thr-16
	80	13.8696	-109.434	phe-4 to thr-16
WT	2	13.8530	-68.6361	phe-4 to thr-16
	10	13.8634	-92.5397	phe-4 to thr-16
	40	13.8683	-102.288	phe-4 to thr-16
	80	13.8697	-105.152	phe-4 to thr-16

[a] measured distance between end points described in Table 3.1

Table 4.21: The effect of dielectric constant on the lowest-energy minimised structures for the WT and L14P signal sequences of the gC system determined with ECEPPGA, with and without distance-constraints. – = not calculated.

Signal peptide	Dielectric constant	Distance-constrained Distance-point /Å	Distance-constrained E_{conf} /kcal mol^{-1}	Unconstrained End to end distance[a] /Å	Unconstrained E_{conf} /kcal mol^{-1}	Residues in α-helical conformation
WT	2	7.00	-68.6478	13.9000	-68.7878	arg-6 to ala-16
	10	–	–	13.9003	-81.3237	arg-6 to ala-16
	40	–	–	13.9111	-85.5669	arg-6 to ala-16
	80	14.00	-86.7999	13.9113	-87.0131	arg-6 to ala-16
L14P	2	12.50	-76.9027	11.4780	-76.5282	arg-6 to leu-12, pro-14 to tyr-15
	10	–	–	11.4551	-81.6901	arg-6 to leu-12, pro-14 to ala-16
	40	–	–	11.5873	-83.8537	arg-6 to leu-12, pro-14 to ala-16
	80	11.50	-84.4948	10.9896	-84.3952	arg-6 to leu-12, pro-14 to tyr-15

[a] measured distance between end points described in Table 3.1

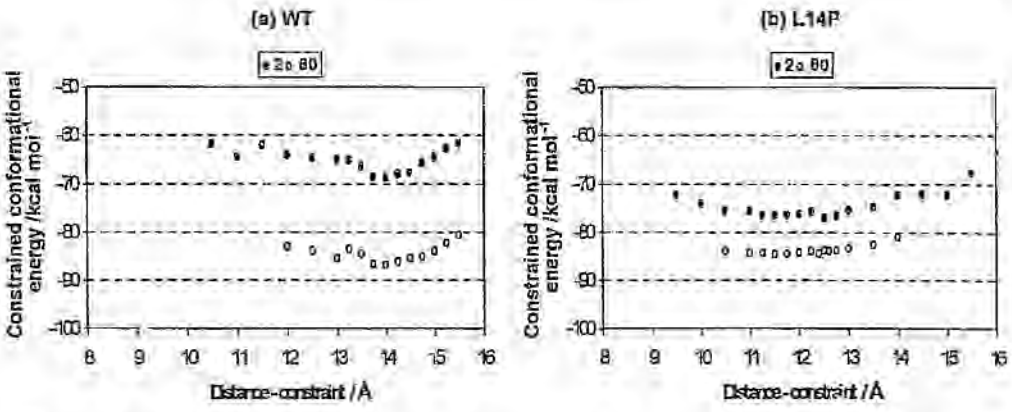


Figure 4.34: Minimised energy *versus* distance-constraint curves for the WT and L14P signal sequences of the gC system determined with ECEPPGA with dielectric constants of 2 and 80. The WT is the most translocationally efficient, while L14P is the least translocationally efficient.

Results obtained from conformational analyses of signal peptides *in vacuo* ($\varepsilon = 2$) were presented earlier. It was hoped that this hydrophobic-simulated environment (in which membrane proteins would be expected to operate) would be able to distinguish between functionally active and inactive peptides. Although this obj. e was achieved in some instances, overall results are unfortunately too inconsistent to provide conclusive information. An equivalent conformational analysis of the signal peptides of the LamB system, conducted with $\varepsilon = 80$, yielded similar findings (see Table 4.19 above) to those obtained for the $\varepsilon = 2$ runs.

Minimum-energy *versus* distance-constraint curves for the high dielectric situation are shown in Figure 4.35. Unlike the low dielectric curves (Figure 4.22), a comparison of the deletion mutant energy-well depths does not suggest a higher α-helical conformational specificity for the active Δ78r2 than for the moderately active Δ78r1 and the inactive Δ78. However, the curves do infer that the WT is more likely to form an α-helix than the inactive A13D (the energy-well of the WT is narrower and deeper than that of A13D). Hence, conflicting results are observed which offer no tangible conclusions.

The lack of useful information gleaned from this study to determine the effect of dielectric constant on structure could be attributed to the fact that the macroenvironment (external surrounding) and microenvironment (internal environment) of the peptides are not accounted for

separately. Internally, the peptides require a dielectric constant of 2, but when this is effected, the peptide is modelled in isolation – which is unrealistic. When a higher dielectric constant is applied, the peptide is modelled artificially and any coulombic interactions that may occur are effectively ignored. The use of a distance-dependent dielectric constant could also have led to erroneous electrostatics. Moreover, biological membranes are highly inhomogenous and their local dielectric constants may vary from place to place. The fact that all the signal sequences form α-helices here suggests that implicit solvation, as used in ECEPP, may be forcing the sequences into these structures: if there is no environment for the sequences to interact with, they only have the option of forming internal hydrogen bonds.

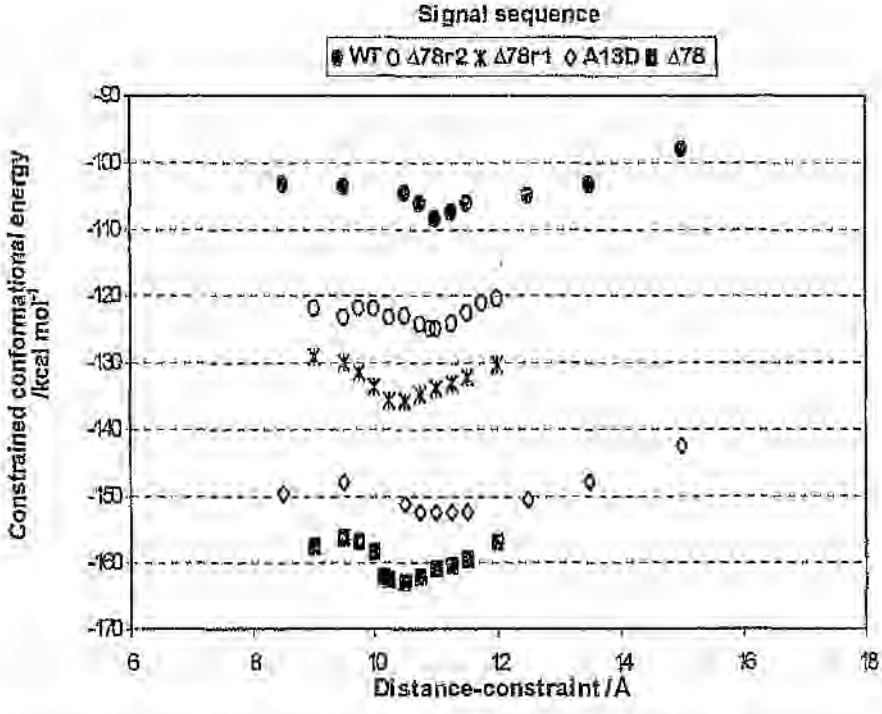


Figure 4.35: Minimised energy versus distance-constraint curves for the signal sequences of the LamB system determined with ECEPPGA with a dielectric constant of 80. The scale on the energy axis is correct for the WT curve. Energy values for the remaining curves have been shifted to accommodate this scale. The curves are arranged in order of translocation efficiency: WT (unshifted); Δ78r2 (shifted by -40 kcal mol^{-1}); Δ78r1 (shifted by -50 kcal mol^{-1}); A13D (shifted by -60 kcal mol^{-1}); Δ78 (shifted by -80 kcal mol^{-1}).

One possible way of overcoming the problem of solvent is to include it explicitly in the models (*cf.* 1.1 3.4). This entails molecular dynamics simulations of fairly large systems which would greatly increase computation time, but which is becoming more feasible with the increase in computer power.

4.2.5.4. Varying proline configuration

Crystal structure analyses, NMR measurements and molecular mechanics computations have ascertained that the prolyl ring of the proline residue occurs in two definite puckered conformations: the 'exo' or 'up' form and the 'endo' or 'down' form.[51] The 'down' puckering of the prolyl peptide has been calculated to be more energetically stable than the 'up' puckering.[51,261] Two forms of the peptide CO-NH bond preceding proline are also possible: the *trans* form and the *cis* form. *Cis-trans* isomerism about a peptide bond results from its partial double bond character (due to resonance). The *trans* form is favoured energetically due to fewer repulsions between nonbonded atoms surrounding the bond. However, when the peptide bond is followed by a proline residue, the cyclic chain of the proline diminishes these repulsions, resulting in comparable stabilities for the *trans* and *cis* isomers. In this instance, the *trans* isomer is only slightly favoured.

The existence of the two proline ring conformations ('up' and 'down'), and the two peptide bond configurations (*trans* and *cis*) prompted an investigation to determine the effect of these variations in form on minimised structure. Results derived from this investigation are given in Table 4.22.

The inactive deletion mutant of the LamB system, Δ78, was modelled. During the GA runs, the proline phi torsion was either fixed at -68.780° for the 'up' puckered conformation of proline, or at -53.040° for the 'down' puckered conformation. ECEPP incorporates a constant energy term, E_{pro}, which is specific to each puckered form and which is independent of the conformation of the sequence. Rotation about the proline peptide bond was set at $\omega = 180°$ for the *trans* form and $\omega = 0°$ for the *cis* form. The proline psi torsion was permitted to vary. No distance-constraints were assigned and no modifications of ECEPPGA parameter values were made. Dielectric constants of 2 and 80 were applied. The dielectric constant of 80 was employed in an effort to research the findings of Choi *et al.*[261a] who determined that solvation causes the 'down' conformation of the proline ring to change to the 'up' conformation.

Table 4.22: The effect of proline conformation on unconstrained conformational energies of the Δ78 signal sequence of the LamB system determined with ECEPPGA. Calculations were performed with dielectric constants of 2 and 80.

Dielectric constant	Prolyl ring conformation[a]	Proline peptide bond[b]	Unconstrained conformational energy /kcal mol^{-1}	Residues in α-helical conformation
2	down	trans	-76.0534	pro-9 to ser-16
		cis	-68.9490	ala-12 to ser-16
	up	trans	-76.3768	pro-9 to ser-16
		cis	-70.2392	val-10 to ser-16
		cis to trans[c]	-76.9861	pro-9 to ser-16
80	down	trans	-83.0469	pro-9 to ser-16
		cis	-79.4974	val-10 to ser-16
	up	trans	-82.2291	pro-9 to ser-16
		cis	-78.4709	val-10 to ser-16
		cis to trans[c]	-82.8384	pro-9 to ser-16

[a] proline ϕ was set at either -53.040° for the 'up' puckered conformation or at -68.780° for the 'down' puckered conformation

[b] proline ω was set at either 180° for the *trans* configuration or at 0° for the *cis* configuration

[c] during the course of the run, the peptide bond changed from *cis* to *trans*

Lower conformational energies are obtained for the *cis* isomers than for the *trans* isomers (see Table 4.22). The instability of the *cis* form is accented by the flipping of the peptide bond from the *cis* to the *trans* form during the course of some of the runs. Similar minimum energy values are calculated for the *trans* isomers of the 'up' and 'down' proline ring conformations. Calculations performed with a dielectric constant of 80 yield analogous results to those performed with a dielectric constant of 2. These observations partially validate the use of the *cis* and 'down' configurations in the current study. However, to fully examine the effect of variations in proline conformation and peptide bond configuration on peptide structure, a more extensive study is required.

4.2.6. Calculation of vibrational entropy

This work would be incomplete if no calculations of conformational free energy and subsequently, vibrational entropy were attempted. As outlined in section 1.1.3.5 of this thesis, the free energy of a single conformation can be computed by associating a statistical weight to the local conformation with the aid of an harmonic approximation. The normalised statistical weight w_i of the ith conformation is expressed as:[132,133]

$$w_i = (1/Z)(2\pi RT)^{k/2}(\det F_i)^{-1/2}\exp(-\Delta U_i/RT), \qquad (5)$$

where ΔU_i is the ith conformational energy relative to the lowest energy, R is the gas constant, T is the temperature, k is the number of variable torsions, F_i is the Hessian matrix (matrix of second derivatives) of the ith conformational energy, and Z is the partition function given by:

$$Z = (2\pi RT)^{k/2}\sum_{i=1}^{n}(\det F_i)^{-1/2}\exp(-\Delta U_i/RT), \qquad (6)$$

where n is the number of local energy minima in an accessible energy region.

The free energy G_i of the ith conformation:

$$G_i = -RT\ln w_i \qquad (7)$$

and the relative free energy ΔG_i:

$$\Delta G_i = G_i - G_0, \qquad (8)$$

where G_0 is th. free energy of the lowest energy conformation, can then be used to calculate the relative entropy ΔS_i of the signal peptide conformation:

$$\Delta S_i = (1/T)(\Delta U_i - \Delta G_i). \qquad (9)$$

It is virtually impossible to identify all of the local energy minima within an accessible energy region (equation *(6)*) and some approximations such as assuming the same value for all i for all determinants in equations *(5)* and *(6)*, *viz.*, approximating the free energy as the sum of the exponential terms, and estimating the number of accessible local equilibrium states for the system, must be made. The first approximation is equivalent to assuming a common librational or harmonic behaviour for all local minima within a potential well, *viz.*, ignoring the effect of entropy. It was decided to follow this route and instead of comparing conformational entropies of the SPs, statistical weights corresponding to their lowest-energy conformations were analysed.

Values calculated for the lowest-energy conformation of each SP are shown in Table 4.23. The value of n is the number of identified local energy minima within 10RT or 5.0 kcal mol^{-1} of the lowest energy. Translocation efficiencies are also provided. It would appear from the w_i values in the table that the above-mentioned estimations proved too crude for an accurate analysis against the translocation efficiencies; no definite trend in the values is discerned.

Table 4.23: Values calculated for the lowest-energy conformations of the signal sequences of the LamB, CPY and gC systems determined with ECEPPGA. ECEPPGA runs were conducted with distance-constraints and with a dielectric constant of 2.

Signal peptide	Distance-constraint /Å	n	U_i /kcal mol^{-1}	ΔU_i	w_i	G_i /kcal mol^{-1}	ΔG_i	ΔS_i	Translocation efficiency[a] /%
LamB									
WT	11.00	9	-86.4877	0	0.5144	0.3939	0	0	100
Δ78r2	11.00	11	-68.6256	0	0.2601	0.7978	0	0	90
Δ78r1	10.50	13	-77.1797	0	0.3154	0.6837	0	0	50
A13D	11.00	12	-101.610	0	0.4010	0.5414	0	0	10
Δ78	10.50	13	-75.9978	0	0.3355	0.6471	0	0	0
CPY									
CPYm6	12.00	7	-70.2573	0	0.5173	0.3906	0	0	97
CPYm2	13.75	9	-72.0831	0	0.4798	0.4352	0	0	94
CPYm8	13.75	5	-68.9578	0	0.5028	0.4073	0	0	27
CPYm1 2	12.00	7	-66.4533	0	0.5342	0.3715	0	0	22
WT	13.75	9	-68.5230	0	0.4804	0.4344	0	0	undetectable
gC									
WT	14.00	11	-68.6478	0	0.4541	0.4678	0	0	100
ΔA10	12.00	10	-67.4761	0	0.4908	0.4217	0	0	99
A10P	12.25	16	-78.7202	0	0.2543	0.8812	0	0	98
L12P	9.75	14	-75.5088	0	0.1795	1.0176	0	0	19
L14P	12.50	18	-76.9027	0	0.2647	0.7875	0	0	5

[a] values from McKnight *et al.*[173] for LamB, from Bird *et al.*[204] for CPY, from Ryan and Edwards[237] for gC

CHAPTER 5
CONCLUSIONS

5.1. The performance of knowledge-based modelling techniques

Various knowledge-based (statistical) conformational analysis methods have been tested against signal peptide functional activity to establish their validities as predictive tools in this study.

Methods based on globular proteins (Chou-Fasman and a consensus procedure) were found to be unreliable; in concordance with the literature, the Chou-Fasman algorithm calculated high probabilities for the signal sequences to occur in both α-helix and β-sheet conformations. Consensus prediction, however, proved to be an improvement over predictions made by the individual methods comprising the consensus procedure – but was still unable to render conclusive results; this was especially the case with the gC system.

In general, predictive methods based on membrane proteins (PHDhtm, TMpred and PSA) yielded more meaningful information than those based on globular proteins. However, only the PSA method calculated membrane-spanning probabilities for the signal sequences that seem to correlate well with translocation efficiencies.

The importance of hydrophobicity for signal sequence function has been highlighted in this work. Solvent accessibility predictions (PHDacc) indicate that the sequences are highly hydrophobic, while hydrophobicity plots, specifically of the h-cores, correspond to some extent with sequence functional activity. However, helical wheel plots did not produce results which could be interpreted.

Discrepancies in results obtained with the knowledge-based modelling techniques employed in this work are probably a consequence of not only the manner in which the methods are derived, but also of the various decision functions (window length, sequence length, homology, *etc.*) which are involved in the submission of sequences for analysis. Prediction techniques are usually empirically tailored for specific functional classes of either globular proteins or membrane proteins, and must therefore be used with caution for general application.

5.2. The performance of molecular modelling techniques

We appear to have developed a successful means of searching for low-energy signal peptide conformations. The genetic algorithm used has proved to be an efficient technique in overcoming the problem of multiple minima. (Pedersen and Moult[262] have recently confirmed the effectiveness of the genetic algorithm method for searching polypeptide conformations.) Results from our ECEPPGA searches are a considerable improvement over those obtained with the computationally intensive systematic searches, and correlate reasonably well with experimental data (from CD and NMR spectra) in the literature. In addition, we appear to have been successful in improving molecular modelling procedures for the conformational analysis of signal peptides; the present analysis is both more extended and definitive than that of Perez *et al.*[206] who carried out a study of smaller portions of four of the five signal peptides of the LamB system only. ECEPPGA also performs better than a recently-developed genetic algorithm for peptide conformational analysis;[263] that algorithm found the relatively large size of the peptides studied here to be unmanageable.

Our calculations were performed with a series of distance-constraints applied to each peptide sequence in an attempt to explain experimental observations. Although it had been anticipated that the potential energy hypersurface of the constrained signal sequences would prove to be complex, in fact, results were obtained which enabled us to follow quite readily the conformation of the lowest-energy structure as a function of conformational compactness. A smooth trend in conformational change from extended to contracted *via* an energy well was generally observed for the sequences. Unfortunately, no distinct relationship between the shapes of the curves or the depths of the energy wells and peptide functional activities could be deduced.

Results suggest that all the signal peptides examined in this study, whether transport-effective or not, adopt stable α-helical structures (whether completely or partially α-helical) in both hydrophobic- and hydrophilic-simulated environments. The occurrence of proline in a sequence induces a break in the α-helix conformation (the amide nitrogen of proline cannot form a hydrogen bond, and the bulky ring disrupts the preceding turn of the helix), while no noticeable effects on conformation from the presence of glycine are apparent. The extent of α-helical secondary structure adopted by the lowest-energy conformations of the signal peptides appears to be insufficiently pronounced for direct correlation with their function; it would seem that conformational properties are not paramount in determining transport efficiency in a hydrophobic environment. Hence, α-helical content, as calculated here, can only give a partial estimate, at best, of translocation ability. There exists an ill-defined relation between hydrophobicity and translocation

efficiency (as mentioned above), but no or little relation between helix amphiphilicity and efficiency. It is thus clear that other factors besides secondary structure conformation, such as the hydrophobicity of the sequences, must also be considered when analysing a signal peptide system for its translocation efficiency.

For obvious reasons, the ECEPPGA option of restricting the conformational space to be searched for a local minimum proved to be successful only for those peptides whose torsional angle values resided within that restricted space. The use of this option must therefore be limited to sequences of known conformation, and is only valid for assessing small structural differences between sequences displaying similar conformations.

The use of shorter sequence lengths during computation did not cause a marked change in results. It is believed that the use of longer sequence lengths (equal to the signal peptide length or longer with the adjacent mature protein region) than those employed here would also not affect results. Thus, it can be concluded that only the shorter sequences are required in modelling procedures.

The primary objective of this study has been to detect differences between native signal sequences and their mutants which would help to explain differences in sequence translocational efficiencies. An exploration of molecular modelling possibilities has led to the conclusion that, in conjunction with certain knowledge-based modelling techniques, *e.g.*, PSA, and sequence hydrophobicity analysis, a clearer picture of the structural differences between export-effective sequences and those mutated sequences which do not facilitate translocation can be gained. The effectiveness of these procedures is clouded by the complexity of the relations among the various properties examined.

5.3. Signal peptide conformation during protein secretion

We have attempted to formulate a more informed description of the translocation structure and process in this study. However, the only conclusion that can be drawn from these predictions is that functionally-active signal peptides tend to favour the α-helical secondary structure conformation slightly more than do functionally-inactive signal peptides. It would thus appear that the assumption that such a conformation is adopted during effective translocation is appropriate, as has been proposed by several other researchers; the differences in calculated energy values between the lowest-energy α-helical conformations and extended conformations are believed to be too large to suggest that extended structures will form during membrane crossing.

We were unable to substantiate the postulated change in peptide secondary structure on moving from an aqueous to a lipidic environment which arises in the membrane trigger hypothesis. If it is rather assumed that signal peptides traverse the membrane *via* an aqueous channel (and recent studies[264] indicate that this is a highly probable scenario), no environment-induced alteration in secondary structure would be expected to occur, and it is then highly probable that hydrophobicity differences between peptides is the principal determinant of translocational efficiency. Further studies in which the aqueous medium is better simulated than simply by a change in dielectric constant alone may produce more informative results.

5.4. Suggestions for future work

ECEPPGA was subjected to a measure of optimisation with respect to this particular application, and proved to be efficient in locating local energy minima. Nevertheless, further optimisation may improve the performance of the GA. It is recommended that the use of the program in other applications be accompanied by optimisation procedures specific to each application.

To ensure the attainment of the GMEC for each calculated peptide sequence, a number of low-energy conformations could be sampled in the form of the Boltzmann distribution. This would inform us as to whether the lowest-energy conformation obtained is also the conformation that would occur most often, *i.e.*, is the most stable. Another way to ensure that the GMEC is reached is to perform multiple runs (> 3) for each sequence. Of course, it must first be evaluated whether or not the ultimate attainment of the GMEC is required. (A protein may not necessarily occur as its global minimum in nature; it may form a specific conformation and the energy barrier to transition may be too high to allow it to reach its minimum-energy structure.)

To better simulate the behaviour of signal sequences in different media, molecular dynamics computations with solvent (aqueous and lipid bilayers) are necessary. The lipid bilayer could be modelled by setting up a "membrane" across a periodic box, the box simulating a larger sheet structure of lipids. Molecular mechanics calculations involving solvent will become more practical as advances in computer hardware are made. Progress in the computer industry will not only speed up computations, but will also make possible the application of conformational searching programs such as ECEPPGA to larger, more complex molecular systems.

REFERENCES

1 Anfinsen, C.B. (1973) *Science* **181**, 223-239.

2 Leach, A.R. (1991) in *Reviews in Computational emistry*, Vol. 2, (Lipkowitz, K.B., Boyd D.B., Eds.) VCH Publishers, New York, 1-55.

3 Epstein, C.J., Goldberger, R.F., Anfinsen, C.B. (1963) *Cold Spring Harbor Symp. Quant. Biol.* **27**, 439-449.

4 Wetlaufer, D.B., Ristow, S. (1973) *Annu. Rev. Biochem.* **42**, 135-158.

5 Dobson, C.M., Evans, P.A., Radford, S.E. (1994) *Trends in Biochem. Sci.* **19**, 31-37.

6 Schulz, G.E., Schirmer, R.H. (1979) in *Principles of Protein Structure*, (Cantor, C.R., Ed.) Springer-Verlag, New York.

7 Kim, P.S., Baldwin, R.L. (1982) *Annu. Rev. Biochem.* **51**, 459-489.

8 Wolynes, P.G., Onuchic, J.N., Thirumalai, D. (1995) *Science* **267**, 1619-1620.

9 Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z., Socci, N.D. (1995) *Proc. Natl. Acad. Sci., USA* **92**, 3626-362?

10 Rose ,1996) *The Sciences* Jan/Feb, 26-31.

11 Rost, B., Sander, C. (1996) *Annu. Rev. Biophys. Biomol. Struct.* **25**, 113-136.

12 Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M. (1987) *Nature* **326**, 347-352.

13 Eisenhaber, F., Persson, B., Argos, P. (1995) *Crit. Rev. Biochem. Mol. Biol.* **30**(1), 1-94.

14 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977) **112**, 535-542.

15 Bairoch, A., Boeckmann, B. (1994) *Nucleic Acids Res.* **22**, 3578-3580.

16 Needleman, S.B., Wunsch, C.D. (1970) *J. Mol. Biol.* **48**, 443-453.

17 Smith, T.F., Waterman, M.S. (1981) *J. Mol. Biol.* **147**, 195-197.

18 Kocher, J.P.A., Rooman, M.J., Wodak, S.J. (1994) *J. Mol. Biol.* **235**, 1598-1613.

19 Sippl, M.J. (1990) *J. Mol. Biol.* **213**, 859-883.

20 Maiorov, V.N., Crippen, G.M. (1992) *J. Mol. Biol* **227**, 876-888.

21 Creighton, T.E. (1993) in *Proteins: Structures and Molecular Properties* (2nd ed.), W.H. Freeman, New York, 255-257.

22 Chou, P.Y., Fasman, G.D. (1974) *Biochem.* **13**, 222-245.

23 Garnier, J., Osguthorpe, D.J., Robson, B. (1978) *J. Mol. Biol.* **120**, 97-120.

24 Lim, V.I., (1974) *J. Mol. Biol.* **88**, 857-872, 873-894.

25 Chou, P.Y., Fasman, G.D. (1978) *Adv. Enzymol.* **47**, 45-148.

26 Chou, K.C., Zhang, C.T. (1994) *J. Biol.Chem.* **269**, 22014-22020.

27 Eisenhaber, F., Imperiale, F., Argos, P., Frömmel, C. (1996) *Proteins: Struct. Funct. Genet.* **25**, 157-168.

28 Eisenhaber, F., Frömmel, C., Argos, P. (1996) *Proteins: Struct. Funct. Genet.* **25**, 169-179.

29 Levitt, M., Chothia, C. (1976) *Nature* **261**, 552-558.

[30] Cornette, J.L., Cease, K.B., Margalit, J.H., Spouge, J.L., Berzofsky, J.A., DeLisi, C. (1987) *J. Mol. Biol.* 195, 659-685.

[31] Benner, S.A., Badcoe, I., Cohen, M.A., Gerloff, D.L. (1994) *J. Mol. Biol.* 235, 926-958.

[32] Wako, H., Blundell, T.L. (1994) *J. Mol. Biol.* 238, 682-692, 693-708.

[33] Kabsch, W., Sander, C. (1984) *Proc. Natl. Acad. Sci., USA* 81, 1075-1078.
Dill, K.A. (1990) *Biochem.* 29, 7133-7155.

[35] Hunt, N.G., Gregoret, L.M., Cohen, F.E. (1994) *J. Mol. Biol.* 241, 214-225.

[36] Rost, B., Sander, C. (1993) *J. Mol. Biol.* 232, 584-599.

[37] Bamborough, P. (1995) *Chem. Des. Automation News* 10(1,2), 1.

[38] Scheraga, H.A. (1992) in *Reviews in Computational Chemistry*, Vol. 3, (Lipkowitz, K.B., Boyd, D.B., Eds.) VCH Publishers, New York, 73-142.

[39] Hehre, W.J., Radom, L., Schleyer, P., Pople, J.A. (1986) in *Ab Initio Molecular Orbital Theory*, John Wiley & Sons, New York.

[40] Pople, J.A., Beveridge, D.L. (1970) in *Approximate Molecular Orbital Theory*, McGraw-Hill, New York.

[41] Binkley, J.S., Frisch, M.J., de Frees, J., Raghavachari, K., Whitside, R.A., Schlegel, H.B., Fluder, E.M., Pople, J.A. (1982) *GAUSSIAN82*, Carnegie-Mellon University, Pittsburgh, USA.

[42] Dewar, M.J.S., Thiel, W. (1977) *J. Amer. Chem. Soc.* 99, 4899-4907.

[43] Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., Stewart, J.J.P. (1985) *J. Amer. Chem. Soc.* 107, 3902-3909.

[44] Stewart, J.J.P. (1989) *J. Comput. Chem.* 10, 209-220.

[45] Burkert, U., Allinger, N.L. (1982) *Molecular Mechanics*, ACS Monograph 177, American Chemical Society, Washington, DC.

[46] Cohen, N.C., Blaney, J.M., Humblet, C., Gund, P., Barry, D.C. (1990) *J. Med. Chem.* 33, 883-894.

[47] Allinger, N.L. (1977) *J. Amer. Chem. Soc.* 99, 8127-8134.

[48] Allinger, N.L., Yuh, Y.H., Lii, J.-H. (1989) *J. Amer. Chem. Soc.* 111, 8551-8566.

[49] Momany, F.A., McGuire, R.F., Burgess, A.W., Scheraga, H.A. (1975) *J. Phys. Chem.* 79, 2361-2381.

[50] Némethy, G., Pottle, M.S., Scheraga, H.A. (1983) *J. Phys. Chem.* 87, 1883-1887.

[51] Némethy, G., Gibson, K.D., Palmer, K.A., No Yoon, C., Paterlini, G., Zagari, A., Rumsey, S., Scheraga, H.A. (1992) *J. Phys. Chem.* 96, 6472-6484.

[52] Weiner, P.K., Kollman, P.A., Nguyen, L ., Case, D.A.(1986) *J. Comput. Chem.* 7, 230-252.

[53] Brooks, B.R., Bruccoleri, R.E., Olafson, L D., States, D.J., Swaminathan, S., Karplus, M. (1983) *J. Comput. Chem.* 4, 187-217.

[54] White, D.N.J. (1977) in *Computers & Chemistry*, Vol. 1, Pergamon Press, United Kingdom, 225-233.

[55] Rosenbrock, H.H. (1961) *Comput. J.* 6, 163.

[56] Wiberg, K.B. (1965) *J. Amer. Chem. Soc.* 87, 1070-1078.

[57] Fletcher, R., Reeves, C.M. (1964) *Comput. J.* 7, 149.

[58] Schlick, T. (1992) in *Reviews in Computational Chemistry*, Vol. 3, (Lipkowitz, K.B., Boyd, D.B., Eds.) VCH Publishers, New York, 1-71.

[59] Vásquez, M., Némethy, G., Scheraga, H.A. (1994) *Chem. Rev.* **94**, 2183-2239.

[60] Muñoz, V., Serrano, L. (1994) *Proteins: Struct. Funct. Genet.* **20** 301-311.

[61] Kamimura, M., Takahashi, Y. (1994) *CABIOS* **10**, 163-169.

[62] Vajda, S. (1993) *J. Mol. Biol.* **229**, 125-145.

[63] Piela, L., Olszewski, K.A., Pillardy, J. (1994) *J. Mol. Struct. (Theochem)* **308**, 229-239.

[64] Scheraga, H.A. (1996) *Biophys. Chem.* 59, 329-339.

[65] Judson, R.S., Jaeger, E.P., Treasurywala, A.M., Peterson, M.L. (1993) *J. Comput. Chem.* **14**, 1407-1414.

[66] Bruccoleri, R.E. (1993) *Mol. Simul.* **10**, 151-174.

[67] Meirovitch, H., Meirovitch, E., Michel, A.G., Vásquez, M. (1994) *J. Phys. Chem.* **98**, 6241-6243.

[68] Desmet, J., de Maeyer, M., Hazes, B., Lasters, I. (1992) *Nature* **356**, 539-542.

[69] Zimmerman, S.S., Pottle, M.S., Némethy, G., Scheraga, H.A. (1977) *Macromol.* **10**, 1-9.

[70] Vásquez, M., Némethy, G., Scheraga, H.A. (1983) *Macromol.* **16**, 1043-1049.

[71] Koča, J., Kříž, Z., Carlsen, P.H.J. (1994) *J. Mol. Struct. (Theochem)* **306**, 157-164.

[72] Vásquez, M., Scheraga, H.A. (1985) *Biopolym.* **24**, 1437-1447.

[73] Troyer, J.M., Cohen, F.E. (1991) in *Reviews in Computational Chemistry*, Vol. 2, (Lipkowitz, K.B., Boyd, D.B., Eds.) VCH Publishers, New York, 57-80.

[74] Koča, J., Carlsen, P.H.J. (1992) *J. Mol. Struct. (Theochem)* **257**, 131-141.

[75] Lambert, M.H., Scheraga, H.A. (1989) *J. Comput. Chem.* **10**, 770-797, 798-816, 817-831.

[76] Davidson, R.B., Zimmerman, S.S. (1994) *J. Chem. Inf. Comput. Sci* **34**, 1009-1013.

[77] Purisima, E.J., Smith, T.F., Vajda, S. (1994) *J. Comput. Chem.* **15**, 300-312.

[78] Ra ow, A.A., Scheraga, H.A (1993) *J. Mol. Biol.* **232**, 1157-1168.
Lau, K.F., Dill, K.A. (1989) *Macromol.* **22**, 3986-3997.

[80] Covell, D.G., Jernigan, R.L. (1990) *Biochem.* **29**, 3287-3294.

[81] Sikorski, A., Skolnick, J. (1990) *J. Mol. Biol.* **212**, 819-836.

[82] Chan, H.S., Dill, K.A. (1990) *Proc. Natl. Acad. Sci., USA* **87**, 6388-6392.

[83] Gō, N., Taketomi, H. (1978) *Proc. Natl. Acad. Sci., USA* **75**, 559-563.

[84] Abkevich, V.I., Gutin, A.M., Shakhnovich, E.I. (1994) *Biochem.* **33**, 10026-10036.

[85] Covell, D.G. (1994) *J. Mol. Biol.* **235**, 1032-1043.

[86] Levitt, M., Warshel, A. (1975) *Nature* **253**, 694.

[87] Tanaka, S., Scheraga, H.A. (1977) *Proc. Natl. Acad. Sci., USA* **74**, 1320-1323.

[88] Shakhnovich, E.I., Gutin, A.M. (1990) *Nature* **346**, 773-775.

[89] Sali, A., Shakhnovich, E.I., Karplus, M. (1994) *J. Mol. Biol.* **235**, 1614-1636.

[90] Ghose, A.K., Jaeger, E.P., Kowalczyk, P.J., Peterson, M.L., Treasurywala, A.M. (1993) *J. Comput. Chem.* **14**, 1050-1065.

[91] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953) *J. Chem. Phys.* **21**, 1087-1092.

[92] Li, Z., Scheraga, H.A. (1987) *Proc. Natl. Acad. Sci., USA* 84, 6611-6615.

[93] Ripoll, D.R., Scheraga, H.A. (1988) *Biopolym.* 27, 1283-1303.

[94] Piela, L., Scheraga, H.A. (1987) *Biopolym.* 26, S33-S58.

[95] Abagyan, R., Totrov, M. (1994) *J. Mol. Biol.* 235, 983-1002.

[96] Kirkpatrick, S., Gelatt, C.D., Jr., Vecchi, M.P. (1983) *Science* 220, 671-680.

[97] Goldberg, D.E. (1989) in *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 89-145.

[98] Maranas, C.D., Floudas, C.A. (1994) *J. Chem. Phys.* 100, 1247-1261.

[99] Crippen, G.M. (1978) *J. Comput. Phys.* 26, 449.

[100] Crippen, G.M. (1989) *J. Comput. Chem.* 10, 896-902.

[101] Purisima, E.O., Scheraga, H.A. (1986) *Proc. Natl. Acad. Sci., USA* 83, 2782-2786.

[102] Braun, W., Gō, N. (1985) *J. Mol. Biol.* 186, 611-626.

[103] Head-Gordon, T., Stillinger, F.H., Arrecis, J. (1991) *Proc. Natl. Acad. Sci., USA* 88, 11076-11080.

[104] Piela, L., Kostrowicki, J., Scheraga, H.A. (1989) *J. Phys. Chem.* 93, 3339-3346.

[105] Lybrand, T.P. (1990) in *Reviews in Computational Chemistry*, Vol. 1, (Lipkowitz, K.B., Boyd, D.B., Eds.) VCH Publishers, New York, 295-320.

[106] van Gunsteren, W.F., Berendsen, H.J.C. (1977) *Mol. Phys.* 34, 1311-1327.

[107] Dorofeev, V.E., Mazur, A.K. (1993) *J. Biomol. Struct. Dyn.* 10, 143-167.

[108] Scully, J.L., Hermans, J. (1993) *Mol. Simul.* 11, 67-77.

[109] Brünger, A.T., Karplus, M. (1991) *Acc. Chem. Res.* 24, 54-61.

[110] Tobias, D.J., Brooks, C.L., III (1991) *Biochem.* 30, 6059-6070.

[111] Brooks, C.L., III, Case, D.A. (1993) *Chem. Rev.* 93, 2487-2502.

[112] Kollman, P.A. (1993) *Chem. Rev.* 93, 2395-2417.

[113] Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI.

[114] Garduño-Juárez, R., Romero, D. (1994) *J. Mol. Struct. (Theochem)* 308, 115-123.

[115] Mestres, J., Scuseria, G.E. (1995) *J. Comput. Chem.* 16, 729-742.

[116] Brodmeier, T., Pretsch, E. (1994) *J. Comput. Chem.* 15, 588-595.

[117] Shaffer, R.E., Small, G.W. (1997) *Anal. Chem.* 69, 236A-242A.

[118] McGarrah, D.B., Judson, R.S. (1993) *J. Comput. Chem.* 14, 1385-1395.

[119] Meza, J.C., Judson, R.S., Faulkner, T.R., Treasurywala, A.M. (1996) *J. Comput. Chem.* 17, 1142-1151.

[120] Stephens, P.P. (1995) in *A Novel Genetic Algorithm for Peptide Conformational Searching*, M.Sc. dissertation, University of the Witwatersrand, Johannesburg.

[120a] Pedersen, J.T., Moult, J. (1996) *Curr. Opin. Struct. Biol.* 6, 227-231.

[121] Unger, R., Moult, J. (1993) *J. Mol. Biol.* 231, 75-81.

[122] Dandekar, T., Argos, P. (1994) *J. Mol. Biol.* 236, 844-861.

[122a] Sun, S. (1993) *Protein Sci.* **2**, 762-785.

[122b] Khimasia, M.M., Coveney, P.V. (1997) *Mol. Simul.* **19**, 205-226.

[123] Bowie, J.U., Eisenberg, D. (1994) *Proc. Natl. Acad. Sci., USA* **91**, 4436-4440.

[124] Seibel, G.L., Kollman, P.A. (1990) in *Comprehensive Medicinal Chemistry*, Vol. 4, Pergamon Press, New York, 125-138.

[125] Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L. (1983) *J. Chem. Phys.* **79**, 926-935.

[126] Berendsen, H.J.C., Grigera, J.R., Straatsma, T.P. (1987) *J. Phys. Chem.* **91**, 6269-6271.

[127] Eisenberg, D., McLachlan, A.D. (1986) *Nature* **319**, 199-203.

[128] Marlow, G.E., Perkyns, J.S., Pettitt, B.M. (1993) *Chem. Rev.* **93**, 2503-2521.

[129] Gilson, M.K., Honig, B.H. (1988) *Proteins: Struct. Funct. Genet.* **3**, 32.

[130] Sharp, K.A., Honig, B. (1990) *Annu. Rev. Biophys. Biophys. Chem.* **19**, 301-332.

[131] Lipkowitz, K.B. (1995) *J. Chem. Educ.* **72**, 1070-1075.

[132] Gō, N., Scheraga, H.A. (1969) *J. Chem. Phys.* **51**, 4751-4767.

[133] Gō, N., Scheraga, H.A. (1976) *Macromol.* **9**, 535-542.

[134] Creighton, T.E. (1993) in *Proteins: Structures and Molecular Properties* (2nd ed.), W.H. Freeman, New York, 64-78.

[135] Arkowitz, R.A., Bassilana, M. (1994) *Biochim. Biophys. Acta* **1197**, 311-343.

[136] von Heijne, G. (1990) *J. Membrane Biol.* **115**, 195-201.

[137] Prabhakaran, M. (1990) *Biochem. J.* **269**, 691-696.

[138] Schneider, G., Wrede, P. (1993) *Protein Seq. Data Anal.* **5**, 227-236.

[139] Schatz, G., Dobberstein, B. (1996) *Science* **271**, 1519-1526.

[140] Randall, L.L., Hardy, S.J.S. (1989) *Science* **243**, 1156-1159.

[141] Briggs, M.S., Gierasch, L.M. (1986) *Advances in Protein Chem.* **38**, 109-180.

[142] von Heijne, G., Steppuhn, J., Herrmann, R.G. (1989) *Eur. J. Biochem.* **180**, 535-545.

[143] Engelman, D.M., Steitz, T.A. (1981) *Cell* **23**, 411-422.

[144] Pagés, J.-M., Piovant, M., Varenne, S., Lazdunski, C. (1978) *Eur. J. Biochem.* **86**, 589-602.

[145] Chen, L., Tai, P.C. (1985) *Proc. Natl. Acad. Sci., USA* **82**, 4384-4388.

[145] Simon, S.M., Peskin, C.S., Oster, G.F. (1992) *Proc. Natl. Acad. Sci., USA* **89**, 3770-3774.

[147] Hartl, F-U., Hlodan, R., Langer, T. (1994) *Trends in Biochem. Sci.* **19**, 20-25.

[148] Randall, L.L., Hardy, S.J.S. (1995) *Trends in Biochem. Sci.* **20**, 65-69.

[149] Bochkareva, E.S., Lissin, N.M., Girshovich, A.S. (1988) *Nature* **336**, 254-257.

[150] Kumamoto, C.A. (1989) *Proc. Natl. Acad. Sci., USA* **86**, 5320-5324.

[151] Landry, S.J., Gierasch, L.M. (1991) *Trends in Biochem. Sci.* **16**, 159-163.

[152] Weiss, J.B., Bassford, P.J., Jr. (1990) *J. Bacteriol.* **172**, 3023-3029.

[153] Liu, G., Topping, T.B., Randall, L.L. (1989) *Proc. Natl. Acad. Sci., USA* **86**, 9213-9217.

[154] Blobel, G., Dobberstein, B. (1975) *J. Cell Biol.* 67, 835-851, 852-862.

[155] Wickner, W. (1979) *Annu. Rev. Biochem.* 48, 23-45.

[156] Engelhard, V.H. (1994) *Sc. Amer.* August, 44-51.

[157] Simon, S.M., Blobel, G. (1991) *Cell* 65, 371-380.

[158] Crowley, K.S., Reinhart, G.D., Johnson, A.E. (1993) *Cell* 73, 1101-1115.

[159] Crowley, K.S., Liao, S., Worrell, V.E., Reinhart, G.D., Johnson, A.E. (1994) *Cell* 78, 461-471.

[160] Matlack, K.E.S., Walter, P. (1995) *J. Mol. Biol.* 270, 6170-6180.

[161] Krieg, U.C., Johnson, A.E., Walter, P. (1989) *J. Cell. Biol.* 109, 2033-2043.

[162] Wiedmann, M., Görlich, D., Hartmann, E., Kurzchalia, T.V., Rapoport, T.A. (1989) *FEBS Lett.* 257, 263-268.

[163] Görlich, D., Prehn, S., Hartmann, E., Kalies, K.-U., Rapoport, T.A. (1992) *Cell* 71, 489-503.

[164] Görlich, D., Rapoport, T.A. (1993) *Cell* 75, 615-630.

[165] Mothes, W., Prehn, S., Rapoport, T.A. (1994) *EMBO J.* 13, 3973-3982.

[165a] Beckmann, R., Bubeck, D., Grassucci, R., Penczek, P., Verschoor, A., Blobel, G., Frank, J. (1997) *Science* 278, 2123-2125.

[166] Martoglio, B., Hofmann, M.W., Brunner, J., Dobberstein, B. (1995) *Cell* 81, 207-214.

[167] Rapoport, T.A. (1985) *FEBS Lett.* 187, 1-10.

[168] Miller, J.D., Bernstein, H.D., Walter, P. (1994) *Nature* 367, 657-659.

[169] Shimnar, A.E., Kaiser, E.T. (1984) *J. Amer. Chem. Soc.* 106, 5006-5007.

[170] Nagaraj, R. (1984) *FEBS Lett.* 165, 79-82.

[171] Briggs, M.S., Gierasch, L.M., Zlotnick, A., Lear, J.D., DeGrado, W.F. (1985) *Science* 228, 1096-1099.

[172] Batenburg, A.M., Demel, R.A., Verkleij, A.J., de Kruijff, B. (1988) *Biochem.* 27, 5768-5685.

[173] McKnight, C.J., Briggs, M.S., Gierasch, L.M. (1989) *J. Biol. Chem.* 264, 17293-17297.

[174] Sankaram, M.B., Jones, J.D. (1994) *J. Biol. Chem.* 269, 23477-23483.

[175] Moll, T.S., Thompson, T.E. (1994) *Biochem.* 33, 15469-15482.

[176] Soekarjo, M., Eisenhawer, M., Kuhn, A., Vogel, H. (1996) *Biochem.* 35, 1232-1241.

[177] Cornell, D.G., Dluhy, R.A., Briggs, M.S., McKnight, C.J., Gierasch, L.M. (1989) *Biochem.* 28, 2789-2797.

[178] Kusters, R., Dowhan, W., de Kruijff, B. (1991) *J. Biol. Chem.* 266, 8659-8662.

[179] Goldstein, J., Lehnhardt, S., Inouye, M. (1990) *J. Bacteriol.* 172, 1225-1321.

[180] Izard, J.W., Kendall, D.A. (1994) *Mol. Microbiol.* 13, 765-773.

[181] Inouye, M., Halegoua, S. (1980) *CRC Crit. Rev. Biochem.* 7, 339-371.

[182] Shaw, A.S., Rottier, P.J.M., Rose, J.K. (1988) *Proc. Natl. Acad. Sci., USA* 85, 7592-7596.

[183] von Heijne, G., Blomberg, C. (1979) *Eur. J. Biochem.* 97, 175-181.

[184] Randall, L.L., Hardy, S.J.S. (1984) *Microbiol. Rev.* 48, 290-298.

[185] Nesmayanova, M.A. (1982) *FEBS Lett.* 142, 189-193.

[186] Killian, J.A., de Jong, A.M.P., Bijvelt, J., Verkleij, A.J., de Kruijff, B. (1990) *EMBO J.* 9, 815-819.

[187] Rietveld, A.G., Koorengevel, M.C., de Kruijff, B. (1995) *EMBO J.* 14, 5506-5513.

[188] Keller, R.C.A., Killian, J.A., de Kruijff, B. (1992) *Biochem.* **31**, 1672-1677.

[189] McKnight, C.J., Stradley, S.J., Jones, J.D., Gierasch, L.M. (1991) *Proc. Natl. Acad. Sci., USA* **88**, 5799-5803.

[190] Rosenblatt, M., Beaudette, N.V., Fasman, G.D. (1980) *Proc. Natl. Acad. Sci., USA* **77**, 3983-3987.

[191] Briggs, M.S., Gierasch, L.M. (1984) *Biochem.* **23**, 3111-3114.

[192] Bruch, M.D., McKnight, C.J., Gierasch, L.M. (1989) *Biochem.* **28**, 8554-8561.

[193] Bruch, M.D., Gierasch, L.M. (1990) *J. Biol. Chem.* **265**, 3851-3858.

[194] Izard, J.W., Doughty, M.B., Kendall, D.A. (1995) *Biochem.* **34**, 9904-9912.

[195] Briggs, M.S., Cornell, D.G., Dluhy, R.A., Gierasch, L.M. (1986) *Science* **233**, 206-208.

[196] Sui, S.-F., Wu, H., Sheng, J., Guo, Y. (1994) *J. Biochem.* **115**, 1053-1057.

[197] de Vrije, T., Batenburg, A.M., Killian, J.A., de Kruijff, B. (1990) *Mol. Microbiol.* **4**, 143-150.

[198] YamamotoY., Ohkubo, T., Kohara, A., Tanaka, T., Tanaka, T., Kikuchi, M. (1990) *Biochem.* **29**, 8998- 9006.

[199] Rizo, J., Blanco, F.J., Kobe, B., Bruch, M.D., Gierasch, L.M. (1993) *Biochem.* **32**, 4881-4894.

[200] Chupin, V., Killian, J.A., Breg, J., de Jongh, H.H.J., Boelens, R., Kaptein, R., de Kruijff, B. (1995) *Biochem.* **34**, 11617-11624.

[201] Kaiser, E.T., Kézdy, F.J. (1987) *Annu. Rev. Biophys. Biophys. Chem.* **16**, 561-581.

[202] Williams, R.W., Chang, A., Juretic, D., Loughran, S. (1987) *Biochim. Biophys. Acta* **916**, 200-204.

[203] Emr, S.D., Silhavy, T.J. (1983) *Proc. Natl. Acad. Sci., USA* **80**, 4599-4603.

[204] Bird, P., Gething, M.-J., Sambrook, J. (1990) *J. Biol. Chem.* **265**, 8420-8425.

[205] Pincus, M.R., Klausner, R.D. (1982) *Proc. Natl. Acad. Sci., USA* **79**, 3413-3417.

[206] Perez, J.J., Ricart, J.M., Masip, J. (1991) *Int. J. Biol. Macromol.* **13**, 241-246.

[207] Arêas, E.P.G., Pascutti, P.G., Schreier, S., Mundim, K.C., Bisch, P.M. (1995) *J. Phys. Chem.* **99**, 14885-14892.

[208] Powell, M.J.D. (1964) *Comput. J.* **7**, 155-162.

[209] van Gunsteren, W.F., Berendsen, H.J.C. (1990) *Angew. Chem., Int. Ed. Engl.* **29**, 992.

[210] Cowan, S.W., Rosenbusch, J.P. (1994) *Science* **264**, 914-916.

[211] Hinde, U.P. (1990) *Biosci. Rept.***10**, 537-546.

[212] Kyte, J., Doolittle, R.F. (1982) *J. Mol. Biol.* **157**, 105-132.

[213] Schiffer, M., Edmundson, A.B. (1967) *Biophys. J.* **7**, 121-135.

[214] Eisenberg, D., Schwarz, E., Komaromy, M., Wall, R. (1984) *J. Mol. Biol.* **179**, 125-142.

[215] Rao, J.K., Argos, P. (1986) *Biochim. Biophys. Acta* **869**, 197-214.

[216] von Heijne, G. (1986) *EMBO J.* **5**, 3021-3027.

[217] Persson, B., Argos, P. (1994) *J. Mol. Biol.* **237**, 182-192.

[218] Parodi, L.A., Granatir, C.A., Maggiora, G.M. (1994) *CABIOS* **10**, 527-535.

[219] von Heijne, G. (1992) *J. Mol. Biol.* **225**, 487-494.

[220] Claros, M.G., von Heijne, G. (1994) *CABIOS* **10**, 685-686.

[221] Jones, D.T., Taylor, W.R., Thornton, J.M. (1994) *Biochem.* **33**, 3038-3049.

[222] Persson, B., Argos, P. (1996) *Protein Sci.* **5**, 363-371.

[223] Lohmann, R., Schneider, G., Behrens, D., Wrede, P. (1994) *Protein Sci.* 3, 1597-1601.

[224] Rost, B., Casadio, R., Fariselli, P., Sander, C. (1995) *Protein Sci.* 4, 521-533.

[225] Fariselli, P., Casadio, R. (1996) *CABIOS* 12, 41-48.

[226] Pincus, M.R., Klausner, R.D. Scheraga, H.A. (1982) *Proc. Natl. Acad. Sci., USA* 79, 5107-5110.

[227] Terwilliger, T.C., Weissman, L., Eisenberg, D. (1982) *Biophys. J.* 37, 353-361.

[228] Head-Gordon, T., Stillinger, F.H. (1993) *Biopolym.* 33, 293-303.

[229] Tobias, D.J., Gesell, J., Klein, M.L., Opella, S.J. (1995) *J. Mol. Biol.* 253, 391-395.

[230] Efremov, R.G., Vergroten, G. (1995) *J. Phys. Chem.* 99, 10658-10666.

[231] Kovacs, H., Mark, A.E., nsson, J., van Gunsteren, W.F. (1995) *J. Mol. Biol.* 247, 808-822.

[232] Tuffery, P., Etchebest, C., Lavery, R. (1995) in *Modelling of Biomolecular Structures and Mechanisms*, (A. Pullman *et al.*, Eds.) Kluwer Academic Publishers, Netherlands, 1-9.

[233] TRIPOS Ass., 1699 S. Hanley Road, St. Louis, MO.

[234] The source code for ECEPPAK (and generous help with information thereof) was provided by Dr. Daniel Ripoll of The Theory Center, Cornell University, Ithaca, NY.

[235] Gibrat, J.-F., Garnier, J., Robson, B. (1987) *J. Mol. Biol.* 198, 425-443.

[236] Levin, J.M., Robson, B., Garnier, J. (1986) *FEBS Lett.* 205, 303-308.

[237] Deléage, G., Roux, B. (1987) *Prot. Eng.* 1, 289-294.

[238] Geourjon, C., Deléage, G. (1994) *Prot. Eng.* 7, 157-164.

[239] Geourjon, C., Deléage, G. (1995) *CABIOS* 11, 681-684.

[240] Rost, B., Sander, C. (1994) *Proteins* 19, 55-72.

[241] Rost, B., Sander, C., Schneider, R. (1994) *CABIOS* 10, 53-60.

[242] Hopp, T.P., Woods, K.R., (1981) *Proc. Natl. Acad. Sci., USA* 78, 3824-3828.

[243] Rose, G.D., Geselowitz, A.R., Glenn, J.L., Lee, R.H., Zehfus, M.H. (1985) *Science* 229, 834-838.

[244] Sweet, R.M., Eisenberg, D. (1983) *J. Mol. Biol.* 171, 479-488.

[245] Rost, B., Sander, C. (1994) *Proteins* 20, 216-226.

[246] Hofmann, K., Stoffel, W. (1993) *Biol. Chem. Hoppe-Seyler* 374, 166.

[247] Stultz, C.M., White, J.V., Smith, T.F. (1993) *Protein Sci.* 2, 305-314.

[248] White, J.V., Stultz, C.M., Smith, T.F. (1994) *Math. Biosci.* 119, 35-75.

[249] Pople, J.A., Segal, G.A. (1966) *J. Chem. Phys.* 44, 3289.

[250] Sippl, M.J., Némethy, G., Scheraga, H.A. (1984) *J. Phys. Chem.* 88, 6231-6233.

[251] Gay, D.M. (1983) *ACM Trans. Math. Software* 9, 503-524.

[252] Tobias, D.J., Sneddon, S.F., Brooks, C.L., III (1991) *J. Mol. Biol.* 216, 783-796.

[253] Brooks, C.L., III, Nilsson, L. (1993) *J. Amer. Chem. Soc.* 115, 11034-11035.

[254] Smythe, M.L., Huston, S.E., M rs, G.R. (1993) *J. Amer. Chem. Soc.* 115, 11594-11595.

[255] Rizo, J., Blanco, F.J., Kobe, B., Bruch, M.D., Gierasch, L.M. (1993) *Biochem.* 32, 4881-4894.

[256] Nouwen, N., Tommassen, J., de Kruijff, B. (1994) *J. Biol. Chem.* 269, 16029-16033.

[257] Ryan, P., Edwards, C.O. (1995) *J. Biol. Chem.* **270**, 27876-27879.

[258] Szmelcman, S., Schwartz, M., Silhavy, T.J., Boos, W. (1976) *Eur. J. Biochem.* **65**, 13-19.

[259] Bird, P., Gething, M.-J., Sambrook, J. (1987) *J. Cell Biol.* **105**, 2905-2914.

[260] HyperChem, Hypercube, Inc., Waterloo, Canada.

[261] Tanaka, S., Scheraga, H.A. (1974) *Macromol.* **7**, 698.

[261a] Choi, S.H., Yu, J.Y., Shin, J.K., Jhon, M.S. (1994) *J. Mol. Struct.* **323**, 233-242.

[262] Pedersen, J.T., Moult, J. (1997) *J. Mol. Biol.* **269**, 240-259.

[263] Herrmann, F. *Personal communication*, University of Stuttgart, Germany.

[264] Powers, T., Walter, P. (1997) *Science* **278**, 2072-2073.

[265] Talmud, P., Lins, L., Brasseur, R. (1996) *Prot. Eng.* **9**, 317-321.

[266] Brasseur, R., Vanloo, B., Deleys, R., Lins, L., Labeur, C., Taveirne, J., Ruysschaert, J-M., Rosseneu, M. (1993) *Biochim. et Biophys. Acta* **1170**, 1-7.

[267] Brasseur, R., Pillot, T., Lins, L., Vandekerckhove, J., Rosseneu, M. (1997) *Trends in Biochem. Sci.* **22**, 167-171.

[268] Brasseur, R., Ruysschaert, J-M. (1986) *Biochem. J.* **238**, 1-11.

[269] Brasseur, R. (1991) *J. Biol. Chem.* **266**, 16120-16127.

APPENDIX A

EXAMPLES OF ECEPPGA INPUT FILES

A.1. Example of an ECEPPGA main input file (param.in)

```
/* rparm - real valued parameters */

/* 0 - p_mutate, probability of a point mutation at each site (gene)
   on each genome per generation */
r0=0.3

/* 1 - p_xover, probability of a crossover occuring at each site
   (gene) during mating */
r1=0.1

/* 2 - Energy cutoff maximum = maximum difference between lowest
   energy and energy of new offspring */
r2=5e1

/* 3 - Space sharing strategy: initial RMS distance resource limit -
   minimum distance between genomes in the set */
r3=10.0

/* 4 - Space sharing strategy: final RMS distance resource limit -
   minimum distance between genomes in the set */
r4=5.0

/* 5 - Space sharing strategy: resource limit gradient - change in
   limit per generation */
r5=-0.1

/* 9 - Random seed (use a negative integer) */
r9=-123456

/* 10 - Width of log curve for parent selection probability - OR
   gradient of linear curve for same */
r10=0.9

/* 11 - Probability of point mutation in template string */
r11=0.1

/* 12 - Probability of crossover at each site in template string */
r12=0.1

/* 13 - Probability of point mutation in mask string */
r13=0.1
```

```
/* 14 - Probability of crossover at each site in mask string */
r14=0.1

/* 15 - Probability of finding phenotype character/s in mask */
r15=0.9

/* 16 - Probability of finding non-phenotype character/s in mask */
r16=0.3

/* 17 - Probability of a '1' in a bitstring template */
r17=0.5

/* 18 - Probability of a '1' in a bitstring mask */
r18=0.45

/* 19 - Probability of a '#' in a bitstring mask */
r19=0.1

/* 20 - 29: These parameters control the relative probabilities of
   using different types of mutation operators in template strings.
   Probabilities must add up to unity */

/* 20 - Duplicate gene from following or previous site in genome -
   whichever exists, select at random if both exist */
r20=0.05
/* 21 - Add confs from following or previous as for 20 */
r21=0.05
/* 22 - Delete confs from following or previous (as for 20) from set
   of confs in current gene */
r22=0.05
/* 23 - Add a random conf to set */
r23=0.05
/* 24 - Delete one of the confs from the set */
r24=0.05
/* 25 - Delete all confs from set, i.e. assign "don't care" */
r25=0.05
/* 26 - Assign one (random) conf. Uses probabilities in r15-r16 if
   i16 is 1 */
r26=0.05
/* 27 - Assign a random length set of randomly selected confs.  Uses
   the probabilities in r15-r16 if i16 is 1 */
r27=0.05
/* 28 - Invert set, i.e. assign all unassigned confs and unassign all
   assigned confs */
r28=0.6

/* 30 - 39: These parameters control the relative probabilities of
   using different types of crossover operators in template strings.
   Probabilities must add up to unity */
```

```
/* 30 - logical AND: assign confs common to both parents */
r30=0.1
/* 31 - logical OR: assign confs occurring in either parent */
r31=0.1
/* 32 - A NOT B: assign confs occurring in parent1, not duplicated in
   parent2 */
r32=0.05
/* 33 - B NOT A: assign confs occurring in parent2, not duplicated in
   parent1 */
r33=0.05
/* 34 - logical XOR: assign confs occurring in either parent except
   those that occur in both parents */
r34=0.1
/* 35 - Complete crossover, inherit gene from parent2 */
r35=0.6

/* 40 - Probability of crossbreed during interbreed cycle */
r40=0.05

/* 41 - Probability of interbreed during crossbreed cycle */
r41=0.05

/* 42 - Probability of gene mutation being local, i.e. gaussian
   distr. around current value. Otherwise mutation is non-local,
   i.e. random according to strategy in i0 */
r42=0.9

/* 43 - Width of local mutation (Variance of gaussian distr). As a
   fraction of the total torsional space available, i.e. fraction of
   width between boundaries for defined confs and fraction of 360
   degrees divided by the RMS adjustment factor for unrestricted
   angles. */
r43=0.05

/* 44 - Cutoff percentage for gaussian distribution, i.e. percentage
   of population rejected by boundary limits of defined backbone and
   residue conformations - determines relative width of gaussian
   curve within such boundaries. */
r44=0.01      /* = 1% rejected */

/* 45 - Clumping parameter for demarcation of families */
r45=0.0

/* 47 - Limiting parameter for E difference between generations (det.
   end of run)*/
r47=0.3

/* 48 - Limiting parameter for E difference between generations (det.
   swap inter- with cross-breed*/
r48=0.3
```

```
/* iparm - integer valued parameters */

/* 0 - Torsional angle sampling strategy (for mutations and assigning
   random angles in initial population) */
   /* 0: Completely random - all angles between -180 and +180 */
   /* 1: Gaussian or normal distributions within selected backbone
      conformations - read data from file: bb_confs.dat.  A backbone
      angle mask is required (cparm2). A conformations mask
      (avail_cnfs.mask) can also be used to limit the conformations
      available to each monomer unit */
   /* 2: Gaussian or normal distributions within selected monomer
      conformations - read data from file: res_confs.dat.  A residue
      identity string is required i  addition to an angle identity
      string (cparm3 and cparm4).     conformations mask
      (avail_cnfs.mask) can also be used. */
i0=0

/* 1 - Space sharing strategy flag */
   /* 0: None - all genomes below the energy cutoff are accepted */
   /* 1: Space sharing - new genomes must be some minimum RMS
      distance removed from all other genomes in the set. i28
      determines whether user has supplied distance adjustment
      factors */
i1=1

/* 2 - Fitness to breeding rate relationship */
   /* 0: Uniform: - Probability of genome being selected as a parent
      does not depend on its fitness */
   /* 1: Linear - Probability of genome being selected as a parent is
      linearly related to fitness (Position in set). Line gradient
      given by r10 */
   /* 2: Power Law - Probability of genome being selected as a parent
      is logarithmicaly related to fitness (Position in set). Curve
      width given by r10 */
i2=1

/* 3 - Mating strategy */
   /* 0: Random - random parents are chosen for mating */
   /* 1: Closeness bias - same as random, except that for each 1st
      parent chosen, N (i4) other possible parents are randomly
      selected, the one within the smallest or largest (depending on
      the interbreed/crossbreed criterion - i5) RMS distance is then
      used as the 2nd parent. */
   /* 2: Family breeding based on user defined classes of backbone or
      residue conformation - family definitions are read from
      "families.def" and conformational states are read from
      "bb_confs.dat" or "res_confs.dat". Minimum population of a
      family is given by i12 */
   /* 3: Family breeding based on dynamically generated families.
      Families are updated every N (i6) generations and coarseness
```

```
        for demarcation of families is controlled by r15. No families
        are assigned for the first N (i7) generations. Minimum
        population of a family is given by i12.  Penalty/Reward for
        belonging to a family is given by i13, i14 controls whether it
        is added to each genome or shared between all members of the
        family */
    /* 4: Template breeding - template strings are attached to the
        genomes to control fit mates. Length of the string is given by
        i8 and nature of the template is controlled by i9. Mutation and
        crossover rates in the template and mask are controlled by r11
        through r14. Relative probabilities of using different mutation
        and crossover operators are controlled by r20 through 29 and
        r30 through 39 respectively */
i3=4

/* 4 - Number of random samples of prospective parents */
i4=50

/* 6 - Number of generations between automatic family updates */
i6=0

/* 7 - Number of initial generations before demarcation of families /
i7=0

/* 8 - Length of mating template bitstring */
i8=3

/* 9 - Template string type */
    /* 0: Bit string, randomly assigned to genomes. Separate template
        and mask strings are used */
    /* 1: Phenotype linked mask string. Only mask string is used,
        genome serves as template by comparing to conformational zones
        in mask. Depending on the torsional angle sample strategy (i0),
        the mask string will contain either backbone or monomer
        conformational characters from the files in which they are
        defined. Length of string must be equal to number of residues.
        */
i9=0

/* 10 - Generation gap strategy flag */
    /* 0: No generation gap */
    /* 1: Select N (i11) best genomes from entire set to copy
        unaltered into next generation */
    /* 2: Select N (i11) best genomes from each family to copy
        unaltered into next generation */
    /* 3: Continuous generations - no fixed generation boundaries,
        parents and children co-exist and compete in the same space
        Arbitrarily end of generation processing is done after every N
        children - where N is the population size */
i10=1
```

```
/* 11 - Generation gap size */
i11=120

/* 12 - Minimum population of family. Members of families with fewer
   members than this minimum are added to the "general" family, i.e.
   remain unassigned to a family */
i12=0

/* 13 - Penalty/Reward for belonging to a family (not the "general"
   family) */
   /* +ve: Reward */
   /* -ve: Penalty */
i13=0

/* 14 - Penalty/Reward sharing
   /* 0: Penalty/Reward is a                    genomes fitness */
   /* 1: Penalty/Reward is sh                    ily members */
i14=0

/* 15 - Number of genomes to min                    neration */
i15=1000

/* 16 - Assignment of initial mask strings i    phenotype dependent
   template breeding */
   /* 0: Assign random mask chars from available set */
   /* 1: Assign mask chars according to probabilities in r15-r16 */
i16=1

/* 17 - Crossover and mutation control for phenotype dependent
template
   breeding */
   /* 0: Crossover and mutation in mask string is separate from same
in
   genome */
   /* 1: Crossover and mutation occur at the same site in the mask as
in
   the genome.  Only works when whole residues are treated as gene
   units on the genome */
i17=0

/* 18 - Number of generations counted before ending GA run */
i18=15

/* 19 - Number of generations counting strategy flag */
   /* 0: Count all generations */
   /* 1: Count all generations in which best solution doesn't improve
      */
   /* 2: Count all generations in which average energy doesn't
      improve */
   /* 3: Count all generations in which neither improves */
   /* 4: Count generations since last improvement in best solution */
```

```
   /* 5: Count generations since last improvement in average energy
      */
   /* 6: Count generations since last improvement in either */
i19=4

/* 20 - Population size */
i20=400

/* 21 - Minimization strategy flag */
   /* 0: No local minimization */
   /* 1: Locally minimize N (i15) random genomes */
   /* 2: Locally minimize the N (i15) worst genomes */
   /* 3: Locally minimize the N (i15) best genomes */
i21=1

/* 22 - Number of generations passed before start of I/C swapping */
i22=1

/* 23 - Number of interbreed generations between I/C swaps */
i23=3

/* 24 - Number of crossbreed generations between I/C swaps */
i24=3

/* 25 - Interbreed/Crossbreed (I/C) strategy flag */
   /* 0: No I/C: fitness alone determines breeding pattern */
   /* 1: Fixed interbreed or crossbreed according to i26 */
   /* 2: Swap interbreed/crossbreed according to parameters, i26
      through i27. */
i25=2

/* 26 - I/C starting strategy */
   /* 0: Start run with interbreeding strategy */
   /* 1: Start run with crossbreeding strategy */
i26=0

/* 27 - Number of generations (for i22-i24) counting strategy flag */
   /* 0: Count all generations */
   /* 1: Count all generations in which best solution doesn't improve
      */
   /* 2: Count all generations in which average energy doesn't
      improve */
   /* 3: Count all generations in which neither improves */
   /* 4: Count generations since last improvement in best solution */
   /* 5: Count generations since last improvement in average energy
      */
   /* 6: Count generations since last improvement in either */
i27=4

/* 28 - Use RMS distance distance adjustment factors */
   /* 0: Don't use adjustment factors */
```

```
    /* 1: Read RMS distance adjustment factors from file "RMSD.adj",
       which contains a series of floating point values (one for each
       torsional angle).  RMS distances for angles are then multiplied
       by these factors during RMS calculations.  The factors are
       stored in the array PC_RMS_fact_at_pos */
i28=0

/* 29 - Snapshot interval in generations */
i29=5

/* 30 - Use program's default parameters. If this value is not 0, all
   data in this file will be ignored! */
i30=0

/* 31 - Maximum number of overlapping regions allowed per region per
   type of residue or in the entire set of backbone conformations.
   i.e. each defined region may not have more than this number of
   regions overlapping it. */
i31=10

/* 32 - Statistics interval in generations */
i32=1

/* 35 - Maximum number of error messages to be generated */
i35=100

/* 36 - Compress minimum square distance matrix */
i36=0

/* 40 - Fitness range (ex 10000) of energy values */
i40=1000

/* 41 - Tolerance of fitness range of energy values - i.e. How far
   can fitness (due to energy) venture outside the specified range
   before a re-calibration is ordered */
i41=50

/* 42 - Fitness penalty for each # in bit mask string */
i42=5

/* 43 - Fitness penalty per 5% unit of available confs present in
   template mask string */
i43=0

/* 44 - Lower limit percentage for inducing penalty in i43 - i.e.
   genome is only penalized if the mask contains a larger percentage
   of the available confs than that specified by this limit */
i44=100
```

```
   /* string parameters */

   /* cparm1 - Fixed angle mask */
      /* 0: Angle remains fixed at initial position */
      /* 1: Angle can be varied during GA run */

c1=10110111111111101111111011111011011111101110111111011101111101110111011
   0111110111111110111

   /* cparm3 - angle identity mask */
      /* 0: unknown angle type */
      /* 1: phi angles */
      /* 2: psi angles */
      /* 3: omega angles */
      /* 4: chi1 angles */
      /* 5: chi2 angles */
      /* 6: chi3 angles */
      /* 7: chi4 angles */
      /* 8: chi5 angles */
      /* 9: chi6 angles */
      /*    chi7 angles */
      /*    chi8 angles */

c3=23123456789;12345678123456723123456712341234561234123456123412341231
   34561234567123451

   /* cparm4 - residue identity template */
      /* pos 0: number of residues encoded in template */
      /* pos odd:  residue identity: a number from 1-255, corresponding
         to the position of the residue in the "res_confs.mask" file.
         Program startup will only make assignments correctly if these
         positions correspond to ECEPP residue identities */
      /* pos even: number of genes belonging to current residue */

   c4=17 132 2 15 10 9 8 10 7 13 2 10 7 1 4 18 6 1 4 18 6 1 4 1 4 6 2 18
      6 11 7 16 5 143 1
```

A.2. Example of an ECEPP/3 main input file (xyz.inp)

```
$cntrl
runtyp=ga
res_code=ecepp
var_angles=all
$end

$dist_const
n1pair=0
n2pair=1
$end

$minim
minimizer=sumsl    ! Used SUMSL as the minimizer
maxit= 50          ! maximum no. of iterations during minimization.
$end

$ffield
force_field=ecepp
constr_mov
$end

$geom
  58.527-179.657
 -68.502 -18.096 173.398-165.030 169.655 176.723  80.286   0.246 179.490
-1.265
 -54.864 -39.380 177.562-176.441 175.873-179.572 -64.485 176.979
-119.426  75.618-175.544 -57.488 167.946  60.330 -52.236
 -68.780 157.198 177.711
 -69.058 -40.287-179.127 177.733  62.442  52.269 179.050
 -59.920 -44.186-179.088 -60.587
 -64.245 -45.470-178.612 165.195  50.758 172.534
 -59.920 -44.186-179.088 -60.587
 -64.245 -45.470-178.612 165.195  50.758 172.534
 -59.920 -44.186-179.088 -60.587
 -64.135 -38.426 179.432-179.541
 -62.981 -42.921-178.380
 -67.922 -42.643-178.758 165.835 -69.129 173.085
 -64.299 -38.517 179.915 -74.372 167.979-179.791  57.781
 -73.654 -37.884 178.350 -57.565  82.440
-179.370
$end

$seq
 4
   15  9 10 13 10  1 18  1 18  1  1  6 18 11 16
15
$end
```

```
$ga
directory=./11.0/
$end
```

A.3. Example of a distance-constraints input file (bounds.xyz)

```
1          0        154        15.000
            1   6 N        1  12 C      11.000  11.000  1000.000
```

Author: Chantson, Tracy Elizabeth.
Name of thesis: A conformational analysis of signal peptides.

PUBLISHER:
University of the Witwatersrand, Johannesburg
©2015

LEGALNOTICES:

Copyright Notice: All materials on the University of the Witwatersrand, Johannesburg Library website are protected by South African copyright law and may not be distributed, transmitted, displayed or otherwise published in any format, without the prior written permission of the copyright owner.

Disclaimer and Terms of Use: Provided that you maintain all copyright and other notices contained therein, you may download material (one machine readable copy and one print copy per page)for your personal and/or educational non-commercial use only.

The University of the Witwatersrand, Johannesburg, is not responsible for any errors or omissions and excludes any and all liability for any errors in or omissions from the information on the Library website.