Optimisation of Biochemical Systems

by

Evangelos Simeonidis

A thesis submitted for the degree of

Doctor of Philosophy

of the University of London

Department of Chemical Engineering

University College London

London, 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

The main interest of this work is the application of mathematical programming and optimisation methodologies to problems of biological nature. Biological data forms the basis for modelling, simulation and optimisation - techniques developed and matured successfully within the Process Systems Engineering community - to be carried out in systems like biological networks, metabolic pathways or proteins. Mathematical programming techniques have not yet extensively been applied to such systems.

In the first part of the thesis, optimisation methods for the analysis of biological networks are developed. Metabolic pathway distances and their correlations with genome distance and enzyme function for *E. coli* small molecule metabolism are examined through the use of a linear programming algorithm. The same technique is also applied to the study of the robustness of the p53 cell cycle and apoptotic control network. The p53 network is found to be robust against mutational perturbation, but vulnerable to directed assault against its hubs from tumour-inducing viruses, which act as "biological hackers" to attack the system.

The second part studies protein folding using lattice models. A mixed integer linear programming framework is developed, to implement a successful three-step solution strategy for reading the 3D structure of proteins from only the knowledge of the amino acid sequence and the contact energies among amino acids. The methodology is validated by its application on model proteins designed to fold in a cubical lattice.

Finally, the third part presents mathematical models for the concurrent synthesis of optimal peptide tags and purification steps for protein downstream processing in biochemical processes. In particular, a mixed integer non-linear programming model for the solution of the problem is proposed. A mixed integer linear programming model is then developed by modifying the process synthesis constraints and applying linear approximations of the non-linear functions. The applicability of the models is demonstrated by examples that rely on experimental data.

Acknowledgments

I take this opportunity to thank everyone who made the completion of this thesis possible.

Special thanks go to my supervisor, Dr Lazaros Papageorgiou, for introducing me to the fascinating world of optimisation theory and mathematical programming. His excellent supervision, contribution, support and guidance throughout my PhD studies are more than appreciated.

Financial support from EPSRC, iCPSE, Royal Academy of Engineering and UCL Graduate School is gratefully acknowledged.

Jose Pinto, for an excellent cooperation and for accommodating my research visit to New York. Sophia Tsoka, Maria Elena Lienqueo and Paul Dalby for offering their invaluable insight. David Bogle and Eric Fraga, for comments and discussions. Peter Karp and Pedro Romero from EcoCyc, for kindly generating the *E. coli* dataset used in Chapter 2. Stewart Rison and Lewis Dartnell, for fruitful collaborations.

All past and present members of the Department of Chemical Engineering at UCL and particularly my friends and colleagues, Aris and Dimitris.

My team-mates on UCL's basketball team, for sweating and bleeding with me on the court and for being a constant source of encouragement with their enthusiasm and passion for life. Remember, once we were Warriors.

All my friends in Greece, the UK and around the world. Thank you for being the wonderful people you are and for helping me be a better person.

My partner and best friend, Ilia. Thank you for standing by me, your love and understanding are my most prized possessions.

Finally, I thank my parents, Dimitris and Koula, for providing me with their unconditional love and continuous financial and moral support all these years as well as my brother, Pantelis, for always being there for me.

στους γονείς μου

ευχαριστώ για την υπομονή σας

στους παππούδες μου Βαγγέλη και Παντελή και τις γιαγιάδες μου Χρύσα και Κατίνα

που με προσέχουν από ψηλά

Table of contents

PART II: PROTEIN STRUCTURE PREDICTION

PART III: CHROMATOGRAPHIC PURIFICATION OF PROTEINS USING PEPTIDE TAGS

List of figures

List of tables

Publications related to this thesis

The chapter studying pathway distances and their correlations with genome distance and enzyme function (Chapter 2) contains work that was partly published as an article in Metabolic Engineering (Simeonidis *et al.*, 2003). The results have also been presented at two conferences (Intelligent Systems for Molecular Biology, Copenhagen, Denmark, 2001; Metabolic Engineering IV: Applied Systems Biology, Il Ciocco, Italy, 2002).

The chapter on the robustness of the p53 apoptotic control network (Chapter 3) is based on an article published in FEBS Letters (Dartnell *et al.*, 2005). The work has also been accepted for presentation at a conference (7^{th} World Congress of Chemical Engineering, 2005, Glasgow, UK).

The chapter describing the MINLP model for the design of optimal peptide tags and the synthesis of downstream protein processing (Chapter 5) is based on an article published in Biotechnology Progress (Simeonidis *et al.*, 2005); a related summary article was also published in the Proceedings of the European Symposium on Computer-Aided Process Engineering-14 (Simeonidis *et al.*, 2004).

A short summary of Chapter 6, which presents a linearised MILP model for the design of optimal peptide tags and the synthesis of downstream protein processing, will be published in the Proceedings of the European Symposium on Computer-Aided Process Engineering-15 (Simeonidis *et al.*, 2005).

In all cases, the chapters contain data and analyses not present in the published work.

Chapter 1

Introduction

Mathematical programming techniques have long been recognised as a fertile environment to sustain a wide range of applications. Numerous practitioners in the academia and industry have made vital contributions in the area of process systems engineering with the most celebrated being the pioneering work of Prof. Roger Sargent (1977). Over the years, as computational power became cheaper and widely available, mathematical programming started to realise its true potential by harnessing the number-crunching capabilities of modern computers coupled with major advances in the field of process systems engineering.

1.1. Mathematical programming and biochemical systems

Recently, there has been increasing interest for the application of process systems methodologies to problems traditionally belonging to biology and biotechnology. Advances in sequencing, DNA replication and protein analysis and quantification

have made an unprecedented amount of biological data available, thus opening new horizons for Life Sciences in the near future. The sudden influx of information requires the combination of a large body of biological, biochemical and structural data, from very different levels of organisation, both inside the cell and across organisms. There is a need for comprehensive overviews of processes and systematic methodologies in order to study the biological information offered.

This wealth of data can form the basis for modelling, simulation and optimisation, techniques developed and matured successfully within the process systems engineering community, to be carried out in biological systems, where mathematical programming techniques have not extensively been applied. Working at the interface of process systems engineering, computational biology and bioinformatics will allow us to explore fundamental issues of cellular function and dynamics in ways that have not been possible before. According to Williams (1999), the motivation behind mathematical programming model building is three-fold:

- Gain insight into the problem. The actual exercise of building a mathematical model often reveals relationships that were not apparent previously. As a result greater understanding of the problem is achieved.

- Identify non-obvious solutions to the problem. Having built a model it is then possible to analyse it mathematically and help suggest course of actions that might not otherwise be obvious.

- Investigate extreme aspects of the problem. Computational experiments can be conducted when it is not possible or desirable to conduct an experiment in real-life, providing in this way useful information concerning the problem under investigation.

In the past, many researchers have endeavoured to incorporate optimisation methodologies in biological or biochemical studies. For instance, the potential of the application of pathway analysis for the purposes of metabolic engineering have been illustrated by initial attempts to synthesise metabolic pathways (Mavrovouniotis *et al.*, 1990). Also, kinetic models have been constructed from the stoichiometric and

rate equation for each reaction assumed present in a metabolic pathway. Biochemical Systems Theory used S-system representations (Savageau, 1969; Savageau, 1970) to characterise steady states with linear equations expressed in terms of the logarithms of the original variables, thus capturing some of the non-linear properties of metabolic networks. An extension of this approach was later attempted utilising the linearity of S-systems as a basis for linear programming and establishing a mathematical framework (Voit, 1992). Systems theory remains a popular and widely used technique for studying metabolic pathways; a recent study proposed the modelling of metabolic networks through classical optimisation formulations with an additional constraint to enforce stability of the system (Chang and Sahinidis, 2005). Mathematical programming was also applied on the basis of kinetic or steady state data in order to determine the controlling metabolites and enzymes of a metabolic system using the maximisation of a production rate or the minimisation of by-products as the objective (Regan *et al.*, 1993). This method defined a bound for the production and indicated an appropriate combination of variables.

A model that identifies changes in the regulatory characteristics of enzymes has been presented by Hatzimanikatis *et al.* (1996). The S-system representation was used to formulate the problem using integer programming techniques in order to build a regulatory superstructure that contained all alternative structures of the pathway. A log-linear kinetic model for the estimation of the performance of metabolic systems based on experimentally determined elasticities and control coefficients was also developed (Hatzimanikatis and Bailey, 1997). The latter model was extended in a subsequent paper and an optimisation-based method was applied for the identification of combinations of the metabolic characteristics of enzymes from yeast and bacteria, in order to maximise ethanol production (Hatzimanikatis *et al.*, 1998). A systematic quantitative framework for modelling the regulation of transitions in mammalian cell cycle was formulated (Hatzimanikatis *et al.*, 1999), which provided exceptional agreement with experimental observations. A computational framework applied to the tryptophan biosynthetic pathway (Li *et al.*, 2004) was developed for the construction and calculation of metabolic pathways, that creates new metabolic routes for known and novel metabolites in biological systems. A mathematical model of the critical steps in eukaryotic heat-shock response in yeast and humans was

presented (Rieger *et al.*, 2005) to understand how exposure to a stress signal is converted into specific molecular events for activation and/or repression of heat-shock transcription factors. Finally, a novel optimisation framework for the inference on gene regulatory networks from DNA array data using an S-systems modelling of gene expression was presented (Thomas *et al.*, 2004).

The study of metabolic pathways is a major area where mathematical programming methodologies have been applied broadly (Hatzimanikatis *et al.*, 1996; Pramanik and Keasling, 1997; Petkov and Maranas, 1997; Burgard and Maranas, 2001). Dissimilar to S-systems or kinetic methods, a class of simplified models only uses the stoichiometric data of a metabolic network to generate wider limits of flux distributions available to the cell. Palsson and his research group proposed Flux Balance Analysis (FBA) (Varma and Palsson, 1993), which provided a linear programming framework for modelling metabolism and studying the metabolic capabilities of an organism. This optimisation framework was later improved and extended to larger and more comprehensive datasets of *E. coli* metabolism (Pramanik and Keasling, 1997; Edwards and Palsson, 2000a). The computational predictions of the metabolic capabilities for growth rate of *E. coli* were shown to be consistent with experimental data (Edwards *et al.*, 2001). Over the years, FBA has been applied to the metabolic networks of systems other than *E. coli*, such as *Haemophilus influenzae* (Schilling and Palsson, 2000), mitochondrial metabolism (Ramakrishna *et al.*, 2001), *Helicobacter pylori* (Schilling *et al.*, 2002) and yeast (Forster *et al.*, 2003). The analysis capabilities of FBA have advanced to the degree that we are now able to evaluate the robustness of a metabolic network (Edwards and Palsson, 2000b), predict phenotypes from the genome of an organism and analyse high throughput microbial datasets (Reed and Palsson, 2004; Fong and Palsson, 2004).

Other research groups have also used FBA models, such as the work on *E. coli* by Pramanik and Keasling (1997) and a study on the hypothesis that knockout metabolic fluxes undergo a minimal redistribution with respect to the flux configuration of the wild type (Segre *et al.*, 2002), which introduced the method of minimisation of metabolic adjustment (MOMA). The incorporation of discrete decision variables in the FBA formulation offered the capability of gene additions and/or deletions to the

metabolic network and thus the assessment of the performance limits of metabolic networks (Burgard and Maranas, 2001). As FBA provides a framework for predicting metabolic flux distributions in the absence of kinetic data, optimisation-based models can be used for testing hypothesised objective functions for the metabolic fluxes are consistent with experimental data (Burgard and Maranas, 2003), for suggesting gene knockout strategies in order to enhance biochemical production (Burgard *et al.*, 2003; Pharkya *et al.*, 2003), or for genome-scale metabolic reconstructions (Burgard *et al.*, 2004; Nikolaev *et al.*, 2005). In addition to their work on FBA techniques, Maranas and co-workers have made a number of other important contributions to the field of systems biology. Predictive models for modelling and optimisation of DNA recombination in order to generate novel enzymes were proposed (Moore *et al.*, 2000). A systematic computational framework for designing DNA sequences through codon usage optimisation was presented (Moore and Maranas, 2002a). An approach that utilises thermodynamic and sequence information to calculate the frequency of out-of-sequence reassembly in DNA shuffling experiments was developed (Moore and Maranas, 2002b). An integer programming framework for calculating time delay in gene regulatory networks (Dasika *et al.*, 2004) was also presented. The large-scale inference of the transcriptional regulatory network of *B. subtilis* was addressed with two alternative methodologies, a linear and a power-law model (Gupta *et al.*, 2005).

There have been other attempts to apply mathematical programming to biological systems. For example, a recursive optimisation algorithm was proposed for rigorously finding all alternate optima in metabolic networks to allow for data interpretation or future experiment design (Lee *et al.*, 2000). A model that can enumerate all the ways fluxes can distribute in a metabolic network was described for generating alternative flux scenarios, forecasting responses to mutation, or designing different experiments (Phalakornule *et al.*, 2001). The proposed model was recently used in a different capacity: to provide bounds that help with the solution of the non-linear, non-convex problem of identifying metabolic fluxes from ^{13}C labelling experiments (Ghosh *et al.*, 2005).

There has also been considerable research in microarray technology, which allows the generation of large sets of time series data. A major challenge is to extract the biologically relevant information from the array experiments in an efficient and informative way. We have already reviewed relevant research from Hatzimanikatis and co-workers (Hatzimanikatis *et al.*, 1999; Thomas *et al.*, 2004). In other studies, the organisation and regulation of genetic pathways as dynamic systems were modelled and a mathematical framework for modelling genome expression and regulation was introduced (Wolkenhauer, 2002). Lin *et al.* (2003) combined microarray experiments with novel integer programming methods to define topologies of biological signal transduction pathways. A linear programming method for the classification of tumour samples based on microarray data was proposed (Antonov, 2004). A framework that integrates machine learning and optimisation methodologies for the selection of maximally informative genes in microarray expression experiments was also introduced (Androulakis, 2005).

Finally, an important group of contributions belongs to the area of protein folding. Because of its importance and complexity, the problem of protein folding has been investigated by many researchers, nevertheless a relatively small number of mathematical programming techniques have been applied in the field. Here, we summarise mathematical programming efforts to characterise the computational complexity of protein structure prediction with a few indicative examples. Backofen and Will (2003) presented a constraint programming approach based on the hydrophobic properties of amino acids. Wagner *et al.* (2004) demonstrated large-scale optimisation techniques for the solution of the protein folding problem. Xu *et al.* (2003 and 2004) used a mathematical programming approach to implement a protein threading program for protein structure prediction. Kingsford *et al.* (2005) proposed an integer programming formulation for side-chain positioning (a significant component of homology modelling and protein design). A comprehensive optimisation based approach for the prediction of three-dimensional structures of proteins from their amino acid sequence was also presented by Klepeis and Floudas (2003a).

1.2. Aims and objectives

Motivated by the promise of better understanding and enhanced problem-solving capabilities offered by mathematical programming, the aim of this work is *to facilitate biological studies by applying mathematical programming techniques to problems of biochemical nature*. More specifically, our goal is to develop a number of mathematical modelling frameworks in order to accommodate characteristic problems biologists are faced with today and play the role of a first step towards the establishment of mathematical programming techniques as valuable analysis tools in the fields of computational biology and bioinformatics.

In order to achieve these goals, the following areas are examined:

Biological networks: As availability of biological data grows with an exponential rate, it is becoming obvious that certain topological characteristics are common in otherwise diverse biological networks ranging from signalling and protein interaction networks, to gene networks and metabolism. At the same time, the function of many of these networks is not yet fully understood, or the collected data are incomplete. For this reason, researchers often turn to the study of network architecture instead of function, using graph-theoretical representations of biological networks, which can yield valuable information on network evolution, as well as on the importance of key nodes in the network.

Protein structure prediction: Predicting the three-dimensional structure of a protein can be characterised as the Holy Grail of modern biology. Each protein has a certain function, and it is the unique conformation into which its amino acid sequence folds that allows it to accomplish this function. The problem of protein structure prediction has its basis on the hypothesis that the folding of a protein only depends on the specific amino acids that comprise its sequence (Anfinsen, 1973). Therefore, one could predict the native structure of a protein by minimising a model of its free energy; hence the scope for application of optimisation methodologies.

Downstream protein purification: Proteins are often the products of biochemical production plants, produced through genetic manipulation of mutant bacterial strains

and collected after cell breakage and consequent processing of the mixture. Downstream purification is arguably the most important stage of the process, incurring a large part of the manufacturing and investment costs. It is therefore beneficial to reduce the size of the flowsheet, by decreasing the number of required purification steps. Since the predominant method for purification of the protein mixture is chromatography, it is possible to enhance the efficiency of the chromatographic techniques employed, by manipulating the physicochemical properties of the product protein. This can be accomplished by genetically fusing a small number of amino acids (a peptide tag) to the product protein. Even though modern technology offers the means for such genetic manipulation, it is still a daunting task to select the appropriate amino acids that will make up the optimal tag for the job for each particular protein.

The research areas that this thesis focuses on are clearly distinct, but there is a common theme running through all three of them: the problem is always formulated and solved as a mathematical programming model, facilitating in this way the investigation of the biological problem under question, and at the same time demonstrating the applicability of mathematical programming to the study of biochemical systems. In fact, there are additional commonalities between some of the objects of our research that are not immediately obvious. For example, in the case of downstream protein purification, one of the most important considerations related to the modification of the amino acid chain of the product protein when fusing a peptide tag is the need to avoid interference with protein structure. The influence the tag may have on the folding of the protein can be very substantial and largely depends on the properties of the amino acids selected, so this is a point where two of the research areas presented above (protein folding and use of purification tags for downstream processing) tend to converge.

The problems described in the thesis are formulated as Linear Programming (LP), Mixed Integer Linear Programming (MILP) or Mixed Integer Non-Linear Programming (MINLP) optimisation models. For their solution the General Algebraic Modeling System (GAMS; Brooke *et al.*, 1998) is used, which is a

specially-designed software with access to many different solvers for modelling and solving linear, non-linear and mixed integer optimisation problems.

1.3. Thesis outline

The rest of the thesis is structured in three parts. Part I explores the area of biological network analysis and comprises Chapter 2 and Chapter 3. Part II addresses the problem of protein structure prediction and consists of Chapter 4. Part III tackles the problem of synthesis of downstream protein processing with simultaneous optimal purification tag design and includes Chapters 5 and 6.

Chapter 2 studies metabolic pathway distances, which constitute an important step for evolutionary studies, metabolic reconstruction or selection of key nodes in a metabolic graph. Pathway distances are examined through the use of an LP algorithm. The applicability of the algorithm is illustrated by calculating the minimal pathway distances for *E. coli* small molecule metabolism enzymes, and considering their correlations with genome distance and enzyme function.

A second application of the LP technique for biochemical network analysis is presented in Chapter 3. The algorithm is applied to the study of the robustness of the p53 cell cycle and apoptotic control network. The diameter of the network (the average path length among all nodes) is calculated and used as a measure of network functionality, to study the response of the network to external attacks. The p53 network is found to be robust against mutational perturbation, but vulnerable to directed assault against its hubs from tumour-inducing viruses, which act as "biological hackers" to attack the apoptotic control system.

Chapter 4 presents an MILP framework for the prediction of the 3D structure of a protein based only on sequence data, which is typically an extremely complex task. Using small proteins, we applied a successful three-step strategy for reading the three-dimensional conformation of lattice model-designed proteins from only the knowledge of their amino acid sequence and the contact energies among the amino

acids. The methodology is validated by its application to model proteins designed to fold in a cubical lattice.

An MINLP framework, which integrates the selection of optimal peptide purification tags into an established approach for the synthesis of protein purification processes, is presented in Chapter 5. Considerable improvements in yields and costs of downstream protein purification processes can be achieved with the use of purification tags, which are comparatively short sequences of amino acids genetically fused on the product protein. The goal is to modify the physical properties of the desired product in a way that will enhance its separation from contaminants. The methodology is illustrated through its application on two example protein mixtures involving up to 13 contaminants and a set of 11 candidate chromatographic steps.

Chapter 6 describes the development of an MILP model for the synthesis of the most advantageous purification tags and purification steps for downstream processing in biochemical processes, by modifying the previous process synthesis constraints and reformulating the MINLP model through piecewise linear approximations of the non-linear functions. The results are indicative of the benefits resulting by the appropriate use of peptide tags in purification processes and provide a guideline for both optimal tag design and downstream process synthesis.

Finally, Chapter 7 summarises the main contributions of the thesis and provides recommendations for future research work.

PART I

BIOLOGICAL NETWORK ANALYSIS

Chapter 2

Analysis of metabolic pathways

An organism is an open system that maintains a continuous flow of energy and matter with the environment. Metabolism is the way in which the organism maintains this flux. Simply put, a metabolic pathway is a series of steps converting substrates into products. Several steps are required as generally each enzyme is only capable of performing simple chemical reactions and discrete steps are required to allow the energy released to be 'harnessed'. Cellular metabolism is however complex: pathways consist of hundreds of metabolic reactions and may be spatially organised or highly branched such that one substrate may be utilised in a variety of pathways. This enormous complexity is a fertile ground for the development of optimisation methodologies that will offer insight into the workings and the evolution of metabolic pathways.

This chapter presents the mathematical programming formulation of an algorithm designed to calculate minimal pathway distances in biochemical networks, based on LP techniques. Two graph-representations of small molecule metabolism are considered; both derived from the EcoCyc database (Karp *et al.*, 2002): one protein-centric, the other metabolite-centric. The derived model is applied to the *E. coli* metabolic network, and the correlations of minimal pathway distance with genome

distance (*i.e.* the number of base pairs separating two SMM genes on the *E. coli* chromosome), and minimal pathway distance with enzyme function, as described by Enzyme Commission (EC) number, are investigated.

2.1. Metabolic pathways

2.1.1. Small molecule metabolism

No strict definition of small molecule metabolism (SMM) exists, but the term usually describes the metabolism of all non-macromolecules (Teichmann *et al.*, 2001). Metabolism is divided into catabolism, during which the degradation of metabolites is performed, and anabolism, during which metabolites are biosynthesised. A definition of metabolism can be given as the sum of all physical and chemical processes by which living organised substance is produced and maintained (anabolism) and also the transformation by which energy is made available for the uses of the organism (catabolism). The catabolic breakdown of complex metabolites produces the free energy that is harnessed in high-energy compounds such as ATP and NADPH. In turn, theses molecules are sources of energy for anabolic pathways (Voet and Voet, 1995). Nearly all metabolic reactions require biological catalysts that are called enzymes.

2.1.2. Pathway evolution

There are a number of theories that attempt to explain the evolution of metabolic pathways. The two most prominent that have gathered the most support are the patchwork model and the retrograde model (Rison and Thornton, 2002).

The patchwork model proposes that metabolic pathways evolved by *ad hoc* recruitment of broad-specificity enzymes – capable of catalysing a variety of metabolic reactions (Jensen, 1976). These enzymes exhibited broad substrate specificities and catalyse classes of reactions. The patchwork model of evolution

would suggest that metabolically-close enzymes are no more likely to be functionally and evolutionarily similar that distant ones.

The retrograde model proposes that enzymes were recruited in a direction reverse to the metabolic flow from the preceding enzyme in the pathway (Horowitz, 1945). The model supposes the pre-existence of a chemical environment where both key metabolites and potential intermediates were available. An organism would use up environmental reserves of an essential metabolite A, to the point where falling availability limits growth. An organism capable of synthesising A from environmental precursors B and C would therefore have an evolutionary advantage. If metabolism evolved according to the retrograde model, it would mean that nearby enzymes are likely to be evolutionarily related, and share some functionality.

2.1.3. *Escherichia coli*

Escherichia coli (*E. coli*) was first isolated in 1885 by the German bacteriologist Theodor Escherich (Madigan *et al.*, 1997). *E. coli* is a rod-shaped bacterium approximately 1μm long (Margulis and Schwartz, 1998). The metabolic network of *E. coli* is very well characterised (Karp *et al.*, 2002), making it an ideal specimen for the study of metabolic systems. In fact, this bacterium is the primary model organism in biology, as it is easy to manipulate in the lab, it is capable of conjugation making it suitable for genetic experimentation, and it is able to support the growth of a range of bacterial viruses.

The genome of the non-pathogenic *E. coli K-12* was one of the first to be fully sequenced (Blattner *et al.*, 1997). This is the strain which is being used in this study as well, therefore the term "*E. coli*" in this chapter will refer to *E. coli K-12*. The *E. coli* bacterium is a free-living organism and, as such, has a set of the small molecular metabolic pathways sufficient for independent life. Similar sets are believed to exist in all bacteria and eukaryotes (Teichmann *et al.*, 2001). Because of the extensive experimentation with *E. coli*, knowledge of these pathways is probably close to complete. There exist numerous databases dedicated to these pathways, such as the EcoCyc database (Karp et al., 2002), which was exploited during this study.

2.1.4. Modelling metabolism

Much work has already been done on modelling metabolism (Edwards and Palsson, 2000a; Reed and Palsson, 2004) and analysing the possible mechanisms of pathway evolution (Teichmann *et al.*, 2001; Rison *et al.*, 2002; Rison and Thornton, 2002; Schmidt *et al.*, 2003; Light and Kraulis, 2004). The wealth of currently available data can be used in the creation of models that may also be applied for the simulation and optimisation of metabolic networks. Optimisation techniques have already been used in studies to meet objectives such as flux maximisation, optimal growth and studying the effect of gene deletions or additions to network robustness (Varma and Palsson, 1993; Regan *et al.*, 1993; Pramanik and Keasling, 1997; Schilling *et al.*, 1999; Edwards and Palsson, 2000b; Burgard and Maranas, 2001; Burgard and Maranas, 2003; Fong and Palsson, 2004; Nikolaev *et al.*, 2005).

Lately, there has been an increasing interest in metabolic pathways as an indicator of "connectivity" between genes (Marcotte *et al.*, 1999; Kolesov *et al.*, 2001; Rison *et al.*, 2002). The pathway distance metric can serve as such a measured descriptor of the relationship between two enzymes in the metabolic network. Minimal pathway distances are identified as the smallest number of metabolic steps separating two enzymes: the shortest path from one point in the network to another.

Metrics based on the application of shortest path algorithms in biochemical systems have been considered before. Graph-oriented representations of metabolism have been used to reconstruct metabolic pathways (Arita, 2000). The large-scale organisation of cellular networks has been addressed with a systematic comparative mathematical analysis based on a shortest path algorithm that examines the properties of the metabolic networks of different organisms (Jeong *et al.*, 2000; Almaas *et al.*, 2004). A quantitative basis for identifying a set of central metabolites defining the core of metabolism by calculating the shortest distances between substrates has also been established (Fell and Wagner, 2000).

Recently, the biochemical properties of the *E. coli* Small Molecule Metabolism (SMM) genes and enzymes were investigated using a simple but inefficient graph depth-first-traversal algorithm (Rison *et al.*, 2002). The work demonstrated that

propinquity of SMM genes on the *E. coli* chromosome was matched by propinquity of the encoded proteins in the metabolic network. Patterns of enzyme homologies and conservation of catalytic chemistry between homologues were suggestive of a patchwork model of pathway evolution, as opposed to the retrograde model of evolution (Rison *et al.*, 2002; Rison and Thornton, 2002). A network approach was also used to study the evolution of enzymes in metabolism (Alves *et al.*, 2002). Interestingly, the authors find that neighbouring enzymes (less than 3 steps apart) in the reaction network are more likely to be homologous than distant enzymes (more than 3 steps apart). The work also suggests that blocks of similar catalysis have evolved in metabolism.

Here, a mathematical programming formulation of an algorithm is demonstrated, designed to calculate the minimal pathway distances of the SMM of *E. coli*, based on LP techniques. The correlations of minimal pathway distance with genome distance, and enzyme function are investigated. Results demonstrate the effectiveness of the LP technique, and provide insight into the evolution of metabolic pathways.

2.2. Problem statement

Overall, the problem under examination can be stated as follows:

Given:

- the SMM network of *E. coli*, which consists of a set of metabolites taking part in a set of metabolic reactions, catalysed by a set of enzymes;
- the chromosomal location of all genes encoding the SMM enzymes investigated, which are used to derive genome distances for all gene pairs;
- the EC numbers of all enzymes in the dataset.

Determine:

- the minimal pathway distances among enzymes of the SMM of *E. coli.*

So as to investigate the correlations of minimal pathway distance with genome distance, and minimal pathway distance with enzyme function.

2.3. Algorithm

2.3.1. Shortest paths

The recognition of the shortest possible directed path from a specified source node to some other node of a weighted, directed graph is known as a shortest path problem. A variety of combinatorial problems can be formulated and solved as shortest path problems. In addition, a number of more complex problems can be solved by procedures, which call upon shortest path algorithms (Lawler, 1976). Instead of finding the shortest path from one specified source node to one specified destination node separately, it is more convenient to compute all shortest paths from a single source node to all other nodes in the network.

Next, we discuss the appropriate expressions to derive the length of shortest paths (D_i) from the source node to nodes i. First, assume a directed graph with n nodes, which is characterised by the following parameters.

a_{ij} = the (finite) weight of edge (i,j) if there is such an edge; $+\infty$ otherwise

The source node is numbered 1; the aim here is to calculate the shortest path distances from node 1 to all other nodes in the network. If there are no negative-weight cycles (*i.e.* no cycles for which the sum of edge weights is less than zero) reachable from the source node, then D_1 is equal to zero. Then, for each node j ($j>1$), there must be a final edge (k,j) in a shortest path from 1 to j, where k is the next to last node on the path. Thus, the shortest path lengths must satisfy the following equations, referred to as Bellman's equations (Bellman, 1958):

$$D_1 = 0 \tag{2.1}$$

$$D_j = \min_{k \neq j} \{D_k + a_{kj}\} \qquad \forall j = 2,...,n \tag{2.2}$$

Bellman's equations solve the single-source shortest path problem in the general case, in which edge weights may be negative, given a weighted, directed graph with no negative-weight cycles. Equation (2.2) implies a system of *n-1* inequalities and, for fixed j and $k \neq j$, we have:

$$D_j \leq D_k + a_{kj} \tag{2.3}$$

Also, for fixed values of D_k and $k \neq j$, the correct value of D_j can be determined by a simple LP model by maximising D_j subject to inequalities (2.3), thus satisfying equation (2.2):

$$\textit{maximise } D_j \tag{2.4}$$

subject to:

$$D_j \leq D_k + a_{kj} \qquad \forall k \neq j \tag{2.5}$$

2.3.2. LP model

From the above analysis, an LP model (Lawler, 1976; Cormen *et al.*, 2001) applied to metabolic networks is suggested, capable of finding in a single pass the minimal pathway distances (shortest path lengths) of all enzymes in a network that are reachable from a source enzyme (i^*). First, the notation used in the mathematical model is given:

Indices

i,j enzymes

Parameters

E_{ij} 1 if there is an edge (link) from i to j; *0* otherwise

Positive continuous variables

D_i distance from the i^* source enzyme to enzyme i

For each source enzyme (i^*) in the network, the algorithm finds the minimal pathway distances to all other enzymes by solving the following LP optimisation model:

$$maximise \sum_i D_i \tag{2.6}$$

subject to:

$$D_j \leq D_i + 1 \qquad \forall (i,j): E_{ij} = 1 \tag{2.7}$$

$$D_{i^*} = 0 \tag{2.8}$$

$$D_i \geq 0 \tag{2.9}$$

According to the analysis described in section 2.3.1, the above LP model can represent Bellman's equations. Constraints (2.7) incorporate pathway information related to reaction connectivity, circularity and reaction directionality, facilitated by the use of parameter E_{ij} (for reversible reactions $E_{ij} = E_{ji} = 1$, however for irreversible reactions $E_{ij} = 1$ and $E_{ji} = 0$). Constraint (2.8) assigns the initial value of zero to enzyme i^* to denote it as the source enzyme, while constraints (2.9) require all D_i variables to take positive values.

Finally, unbounded solutions can be avoided by adding:

$$D_i \leq T \qquad \forall i \tag{2.10}$$

where T is an appropriately large number. It should be noted that if D_i equals T at the final solution then it can be concluded that there is *no* path connecting the i^* source enzyme with enzyme i in the network under consideration. This feature of the algorithm is particularly useful to identify cases where the connectivity of part of the network is missing.

2.4. Methods

2.4.1. Generating the pathway dataset

Often, the metabolic network is subdivided into individual pathways, as commonly depicted in biochemistry textbooks (*e.g.* Glycolysis, TCA, fatty-acid biosynthesis) (Voet and Voet, 1995). However, whilst each individual pathway can be considered a separate entity, and distinction can be made between inter-pathway and intra-pathway properties (Teichmann *et al.*, 2001), metabolism is a complex and complete network. Thus, the division of metabolism into distinct pathways is arbitrary (Gerrard *et al.*, 2001). A possible way to deal with this issue is to ignore these divisions, and instead consider metabolism as a single network. Herein, such a network approach, similar to that of Alves *et al.* (2002), was adopted. When individual pathways are mentioned in the text, this is done in order to simplify the discussion; the analyses presented were performed on the whole network, not on a "per pathway" basis.

The SMM network used was obtained from the EcoCyc database (Karp *et al.*, 2002). EcoCyc is an organism-specific pathway/genome database implemented in Common Lisp, which describes the metabolic and signal-transduction pathways of *E. coli*, its enzymes, its transport proteins and its mechanisms of transcriptional control of gene expression (Karp *et al.*, 2002).

Two representations of the *E. coli* network were implemented. The metabolite-centric representation is the customary representation (Michal, 1998), with the metabolites as nodes and the enzymes catalysing the reactions between two metabolites/nodes as the edges. Even though metabolite-centric representations of metabolic networks are the most common, in this work a second, protein-centric representation was also adapted (Gerrard *et al.*, 2001). As illustrated in Figure 2.1, the enzymes are considered as the nodes of the graph, and the substrates are the edges. In both cases, the input to the LP algorithm is a list of node pairs connected by a single edge.

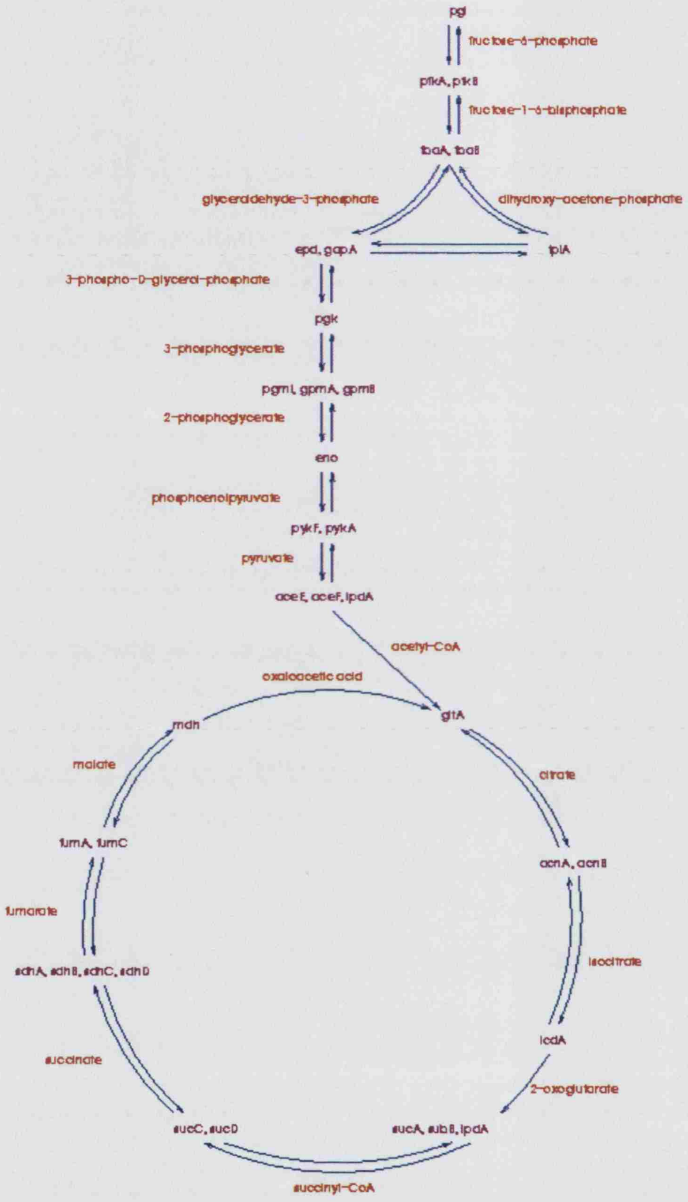

Figure 2.1: A protein-centric (Gerrard et al., 2001) view of Glycolysis and the tricarboxylic acid (TCA) cycle (adapted from EcoCyc; http://www.ecocyc.org/).

Key to Figure: pgi: phosphoglucose isomerase; pfkA and pfkB: 6-phosphofructokinase-1 and 2; fbaB and fbaA: fructose bisphosphate aldolase class I and II; tpiA: triose phosphate isomerase; epd: glyceraldehyde-3-phosphate dehydrogenase 2; gapA: glyceraldehyde-3-phosphate dehydrogenase-A; pgk: phosphoglycerate kinase; gpmA and gpmB: phosphoglycerate mutase 1 and 2; pgmI: phosphoglycerate mutase, cofactor independent; eno: enolase; pykF and pykA: pyruvate kinase I and II; aceE, aceF and lpdA: pyruvate dehydrogenase multienzyme complex; gltA: citrate syntahse; acnA and acnB: aconitase A and B; icdA: isocitrate dehydrogenase; subA, sucB and lpdA: 2-oxoglutarate dehydrogenase complex; sucC and sucD: succinyl-CoA synthase complex; sdhA, sdhB, sdhC and sdhD: succinate dehydrogenase complex; fumA and fumC: fumarase A and fumarase C; mdh: malate dehydrogenase.

In Figure 2.1, an example of a protein-centric representation of pathways is presented. Enzymes are in purple, substrates in red and only key metabolites are shown. The arrows can be read as *"produces a substrate for"*. Enolase (the gene product of *eno*) produces substrate '*phosphoenolpyruvate*' for PykF and PykA. Likewise, PykF and PykA produce '*phosphoenolpyruvate*' for Eno when catalysing the reverse direction reaction. Malate dehydrogenase (the gene product of *mdh*) produces substrate '*oxaloacetic acid*' for GltA, but GltA does not produce '*oxaloacetic acid*' for Mdh. The minimal pathway distance from GltA to Mdh is therefore 1 if directionality is not taken into account (all edges are assumed to be bi-directional), but 7 if directionality is considered (clockwise around the tricarboxylic acid (TCA) cycle).

When using the metabolite-centric representation, pathway distances for the *E. coli* SMM enzymes are derived by reversing the network first. To achieve the reversal, two enzymes are regarded as connected if they catalyse reactions in which the same metabolite appears. Therefore in both cases, the final input to the LP model is a protein-centric network.

Metabolite-centric representations tend to collapse around certain ubiquitous metabolites (*e.g.* ATP, NAD(P), O_2, water, etc.), therefore all metabolites appearing in more than 10 reactions were removed from the dataset. A similar strategy is implemented by Alves *et al.* (2002). The selection of which highly connected metabolites needed to be removed is significant, because of the effect it has on the results; for this reason we have extensively studied the behaviour of the metabolite-centric dataset and we illustrate the results in Figure 2.2. When all metabolites appearing in more than 100 reactions are removed, the average over all minimal pathway distances in the network is only 2.86, and the largest pathway distance observed is 10. As more of these ubiquitous metabolites are removed from the dataset, these values increase considerably, and the behaviour of the metabolite-centric network approximates that of the protein-centric one. It is clear from the graph that the progressive removal of highly connected metabolites decreases the connectivity, and brings the network in line with the protein-centric representation.

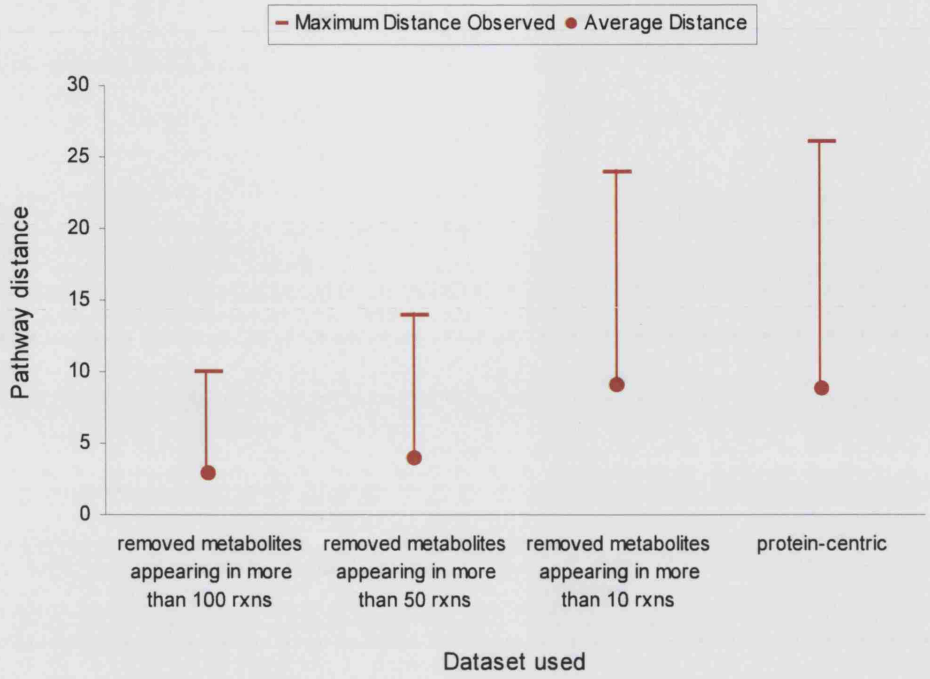

Figure 2.2: Effect of removal of promiscuous metabolites from metabolite-centric dataset.

2.4.2. Genome distance

Genes encoding the SMM enzymes investigated were assigned a chromosomal location by consulting the Gene Table for *E. coli* (Blattner *et al.*, 1997). These were used to derive genome distances for gene pairs, *i.e.* the smallest distance in base pairs (bp) separating the two genes on the chromosome. The genome of *E. coli* is constituted by a single double-stranded circular DNA chromosome; a graphical representation of the genome is presented in Figure 2.3. Since the *E. coli* chromosome is ~4.6Mbp and only the smallest genome distance is considered, two genes can, at most, be separated by ~2.3Mbp.

For the needs of the study, pairs are sorted into bins containing genes separated by: less than 100bp, 101-1,000bp, 1,001-10,000bp, 10,001-100,000bp, 100,001-1,000,000 and more than 1,000,000bp. The choice of bin sizes has a biological rationale. The first of these bins accounts for genes likely to belong to the same operon (Salgado *et al.*, 2000), the second bin size approximates to the average size of

a prokaryotic gene (Casjens, 1998). Subsequent bins were simply enlarged by an order of magnitude.

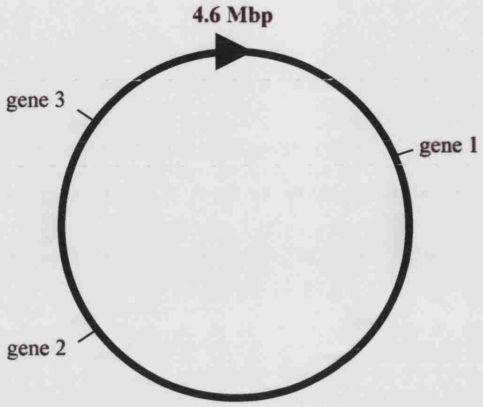

Figure 2.3: The circular chromosome of the E. coli genome.

2.4.3. Function similarity

Enzymes in the datasets were assigned an EC number by reference to the GenProtEC database (Riley, 1998), and following communications with the database curators. EC numbers classify reactions within a hierarchical 4-level scheme. For example, the reaction catalysed by the enzyme *glyceraldehyde-3-phosphate dehydrogenase* has EC number 1.2.1.12 (Enzyme Nomenclature, 1992), as demonstrated in Figure 2.4.

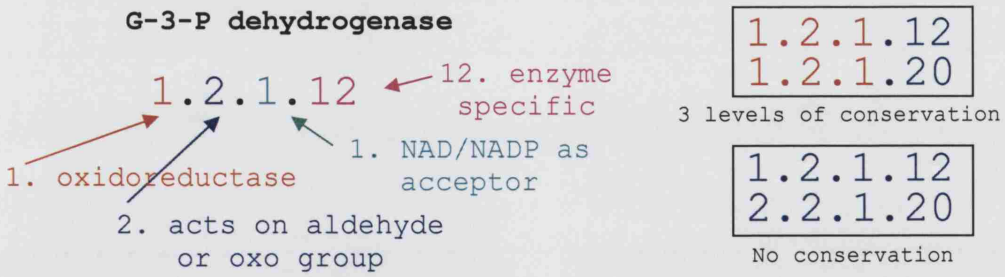

Figure 2.4: Example of classification with an EC number of a reaction catalysed by enzyme G-3-P dehydrogenase. On the right the concept of EC number conservation is demonstrated.

The level to which EC numbers assigned to two enzymes are identical can be used as a measure of the similarity of the function they perform (Martin *et al.*, 1998; Todd *et al.*, 2001). Enzymes assigned identical EC numbers perform the same biochemical function, enzymes with only the first EC level in common share only very generalised functional similarity (*e.g.* both *oxidoreductases*). Finally, enzymes assigned completely different EC numbers often share little or no functional commonalities. Therefore, the number of matching EC levels (none, 1, 2, 3 or 4) is used as the functional similarity metric.

2.5. Computational results

The algorithm was implemented within the GAMS software (Brooke *et al.*, 1998), using the CPLEX 6.5 LP solver for solving LP problems such as the one in hand. Post-processing calculations were incorporated in the algorithm to derive correlations of minimal pathway distance with genome distance and function similarity.

2.5.1. SMM datasets

The metabolite-centric dataset was composed of 973 metabolite pairs accounting for 795 distinct enzymes and 877 distinct metabolites, of which 71 promiscuous compounds were removed, leaving 806 metabolites for input into the LP model. A chromosomal localisation was identified for 740 of the genes encoding for the enzymes, and these were used to investigate the correlation between pathway distance and genome distance. For the study of the correlation between pathway distance and function similarity, an EC number was assigned to 634 genes. Pathway distances obtained by the solution of the algorithm ranged from 1 to 24.

The protein-centric dataset was composed of 599 enzyme pairs and 391 distinct metabolites. For 540 distinct enzymes a chromosomal localisation was identified, and 507 enzymes were assigned an EC number. Pathway distances obtained by the solution of the algorithm ranged from 1 to 26. Both datasets were kindly provided by

the curators of the EcoCyc database (Karp *et al.*, 2002). After a certain pathway distance, the results cease to be informative because: i) they do not deviate substantially from that found at the previous pathway distance; and/or ii) they are based on such a small number of pairs that their validity is questionable. Therefore, in all plots, only pathway distances up to 15 are considered.

2.5.2. Pathway distance and genome distance

The minimal pathway distances for all gene pairs in the SMM network were calculated for both representations. Then, for the established pairs, the base pair separation of the genes encoding the enzymes in the *E. coli* genome was determined, and the enzyme pairs assigned to the corresponding genome distance bin. For example, using the results from the protein-centric dataset, the enzymes glyceraldehyde-3-phosphate dehydrogenase 2 and phosphoglycerate kinase (respectively *epd* and *pgk* in Figure 2.1) have a pathway distance of 1, and are encoded by genes separated by only 50 base pairs. The pair therefore falls into the first bin (0-100bp). However, the enzymes phosphoglycerate kinase and phosphoglycerate mutase 1 (respectively *pgk* and *gpmA* in Figure 2.1), which also have a pathway distance of 1, are encoded by genes separated by 2,282,661bp. The results for the metabolite-centric representation are presented in Table 2.1 and the results for the protein-centric one in Table 2.2.

Table 2.1: Number of gene pairs in the six genome distance bins for each pathway distance for the metabolite-centric representation.

	Pathway Distance														
Genome Distance Bin	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0-100bp	217	29	11	6	5	3	5	2	4	2	1	1	0	0	0
101-1,000bp	91	21	13	8	10	7	5	5	2	2	4	2	1	1	0
1,001-10,000bp	231	78	98	64	84	85	57	52	70	69	53	53	16	23	7
10,001-100,000bp	94	196	284	375	516	750	893	848	823	706	526	502	453	336	316
100,001-1,000,000bp	928	1644	2403	3794	5019	6697	7719	8098	7995	7023	5901	5196	4162	3511	2263
1,000,001-10,000,000bp	1250	2265	3661	5381	7282	9710	11228	12111	11659	10169	8751	7692	5873	4735	3545
Totals	2811	4233	6470	9628	12916	17252	19907	21116	20553	17971	15236	13446	10505	8606	6131

Table 2.2: Number of gene pairs in the six genome distance bins for each pathway distance for the protein-centric representation.

	Pathway Distance														
Genome Distance Bin	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0-100bp	52	45	8	2	3	2	2	0	0	0	0	0	0	0	0
101-1,000bp	12	21	4	2	1	1	1	1	1	1	1	0	0	1	1
1,001-10,000bp	48	63	31	21	17	12	7	7	3	7	3	1	2	2	2
10,001-100,000bp	25	35	41	55	54	63	59	71	78	93	59	45	43	33	46
100,001-1,000,000bp	174	311	463	557	645	679	705	777	856	701	516	457	379	304	274
1,000,001-10,000,000bp	263	453	638	816	1015	1081	1062	1125	1108	1061	766	586	550	507	424
Totals	574	928	1185	1453	1735	1838	1836	1981	2046	1863	1345	1089	974	847	747

The percentages of gene pairs in the first four genome distance bins are plotted against pathway distance in Figure 2.5 for the metabolite-centric representation and in Figure 2.6 for the protein-centric representation. At each pathway distance (x-axis), the percentage of enzyme pairs within various genome distance bins is plotted.

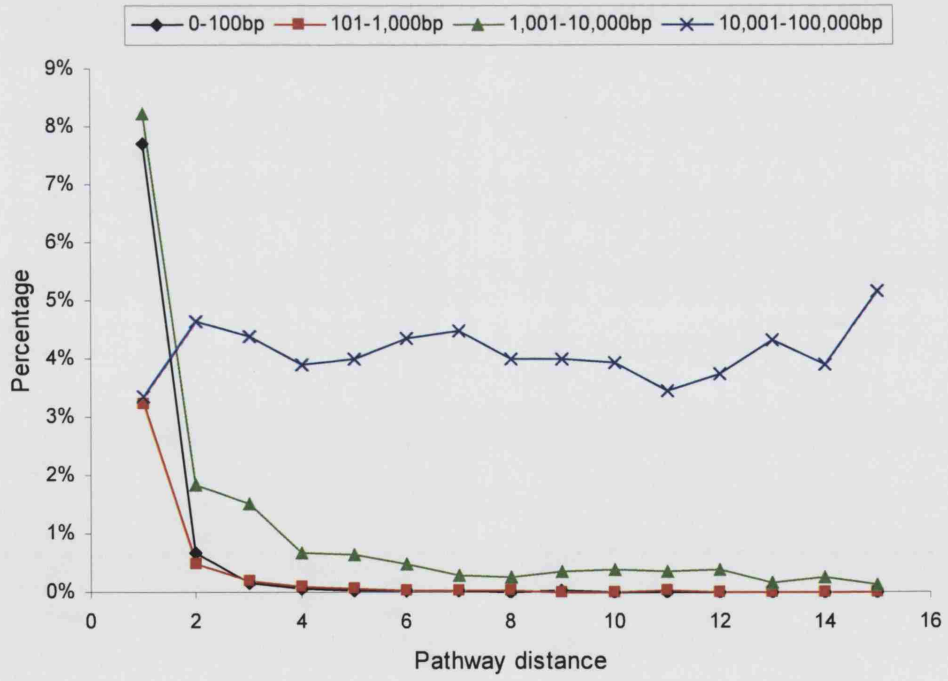

Figure 2.5: Pathway distance and genome distance (metabolite-centric dataset with metabolites involved in more than 10 reactions removed).

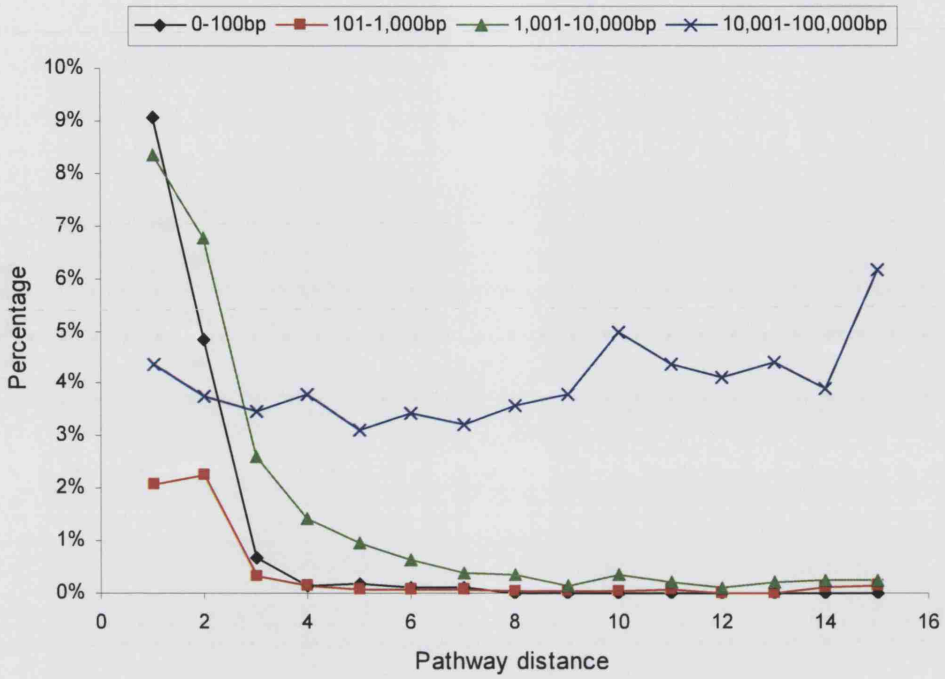

Figure 2.6: Pathway distance and genome distance (protein-centric dataset).

Interestingly, the protein-centric and metabolite-centric representations do not yield identical graphs. This demonstrates once again the problems with SMM representations. The metabolite-centric representations traditionally illustrated in biochemistry textbooks have implicitly considered the ancillary role of promiscuous compounds such as ATP. However, the input to the LP algorithm always consists of node-pairs. In the metabolite-pair representation, since all metabolites are represented, promiscuous compounds will "collapse" the network. In the process of generating the protein-centric representation, the promiscuous compounds are empirically accounted for and therefore they do not dramatically affect the results. This can be seen by contrasting Figure 2.5 and Figure 2.6; when comparing the percentages of homologies at various distances, there are more pairs detected at short distances in the more compact network derived from the metabolite-centric representation, but the actual number of homologous enzyme pairs remains constant, hence a lower percentage of homologous pairs, and the sharp drop between pathway distances 1 and 2. The problem of promiscuous compounds is also discussed by Alves *et al.* (2002).

There is a clear correlation between pathway distance and genome distance. As pathway distance increases, the percentage of genes separated by short genome distances drops. In the results produced with the protein centric dataset (Figure 2.6), for pathway distances of 1, 2, 3, and 4 steps, gene pairs separated by at most 10,000bp (*i.e.* bins 0-100bp, 101-1,000bp, and 1,001-10,000bp) account for 19.51%, 13.9%, 3.63% and 1.72% respectively of the pairs analysed. In the metabolite-centric representation (with all metabolites involved in more than 10 reactions omitted), the drop in chromosomally-close gene pairs with increasing distance is also observed, but the fall is much more sudden, with percentages of 19.17%, 3.02%, 1.89% and 0.81% (Figure 2.5). For the other three distance bins (101,000bp-1,000,000bp and 1,000,001bp and above, which are not plotted here), no clear trend is evident.

2.5.3. Statistical analysis

A statistical measure is applied to demonstrate that the results of the analysis are not due to chance. We are using the standard normal deviate, or Z-score, which measures the distance of a value from the mean of a distribution in standard deviation units. For the needs of this analysis, the mean and standard deviation used are those of randomised networks.

Random interconnected networks were created by arbitrarily pairing the enzymes of the *E. coli* small molecule metabolism, making sure that the same number of pairs was created for each distance as for the original protein-centric *E. coli* network (*i.e.* the connectivity of all the random networks was the same as for the protein-centric network): 574 enzyme pairs at pathway distance 1; 928 pairs at distance 2; 1185 pairs at distance 3; *etc.* Then, a mean and a standard deviation of the number of pairs in each genome distance bin were calculated, by averaging over the pairs produced for 100 random networks. The distance in standard deviation units of the mean of the distribution from the number of pairs of the protein-centric network existing in each bin and each pathway distance was calculated: there are 52 pairs with a pathway distance of 1 in the 0-100bp bin (X_{0-100}), but only 0.75 pairs appear on average in the same bin for randomised networks ($\bar{X}_{0-100}$). The standard deviation for this bin

for randomised networks (σ_{0-100}) is 0.78. Therefore, at a pathway distance of 1 and for the genome distance bin 0-100bp, we have:

$$Z_{0-100} = \frac{X_{0-100} - \bar{X}_{0-100}}{\sigma_{0-100}} = \frac{52 - 0.75}{0.78} = 65.75 \tag{2.11}$$

Figure 2.7 presents the Z-score results calculated for the protein-centric network only. The Z-scores indicate how far and in what direction each item deviates from the random mean, expressed in standard deviation units. Z-score values greater than 3 are usually considered to be significant. As observed in Figure 2.7, our results for the first 3 bins and the first 4 pathway distances deviate the most from the random estimations. After that, the network approaches a more or less random behaviour in the distribution of its enzyme pairs.

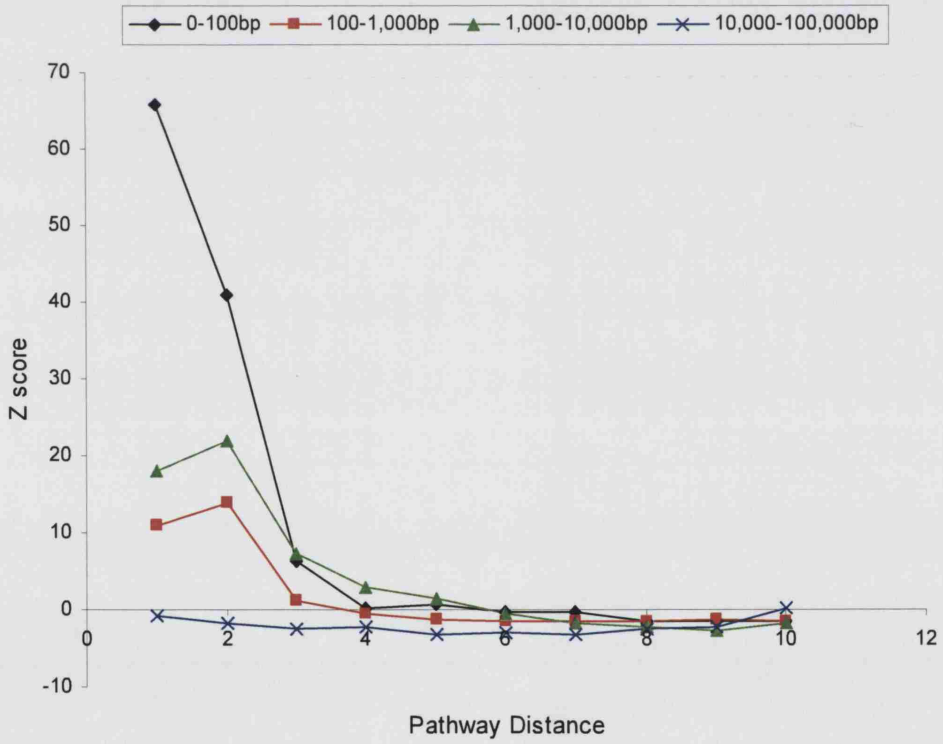

Figure 2.7: At each pathway distance, the Z-score of the number of enzyme pairs within various genome distance bins is plotted.

The observations made from the study of correlations between pathway distance and genome distance for *E. coli* and the statistical analysis performed indicate that SMM genes are "metabolically clustered" on the genome. Furthermore, the relatively high percentage of gene pairs found within 100bp (a very short distance in a ~4.6Mbp long chromosome) suggests that this clustering is the consequence of prokaryotic operon structures in which co-regulated genes are rarely separated by longer distances (Salgado *et al.*, 2000). The observation that short genome distances are often detected for functionally related genes has been made before (Tamames *et al.*, 1997; Overbeek *et al.*, 1999; Rison *et al.*, 2002). However, this relationship has not previously been quantitatively explored in depth and verified for the whole SMM of an organism. Here, we show that the observation holds true using co-participation in a metabolic pathway as an indication of shared function and "measuring" this relationship using our pathway and metabolic distance metrics.

An intriguing feature of these results is that the main "contributors" to the trend observed are the genes in the 0-100bp bin. The next chromosomal distance bin, 101-1,000bp, is nearly always the rarest. A possible explanation for this comes from assuming an average gene length of approximately 1,000bp; a length thought to be uniform in bacterial genomes (Casjens, 1998). Since the 101-1,000bp just reaches the average length of a gene, it represents an "impossible distance": two genes will either be contiguous (and hence separated by 100bp or less), or separated by at least one gene (so separated by at least 1,000bp) – thus avoiding the 101-1,000bp bin.

2.5.4. Pathway distance and function similarity

EC numbers were used as an indicator of shared function. The EC numbers assigned to each enzyme were compared, and the level of EC number conservation was determined. The metabolite-centric data are presented in Figure 2.8. The results for the protein-centric representation are plotted in Figure 2.9. At each pathway distance (x-axis), the percentage of enzyme pairs with all (L1+L2+L3+L4), 3 or more (L1+L2+L3), 2 or more (L1+L2) or 1 or more (L1) EC levels matching is plotted. L1+L2+L3+L4 is a subset of the L1+L2+L3 set (which in turn is a subset of L1+L2, etc.).

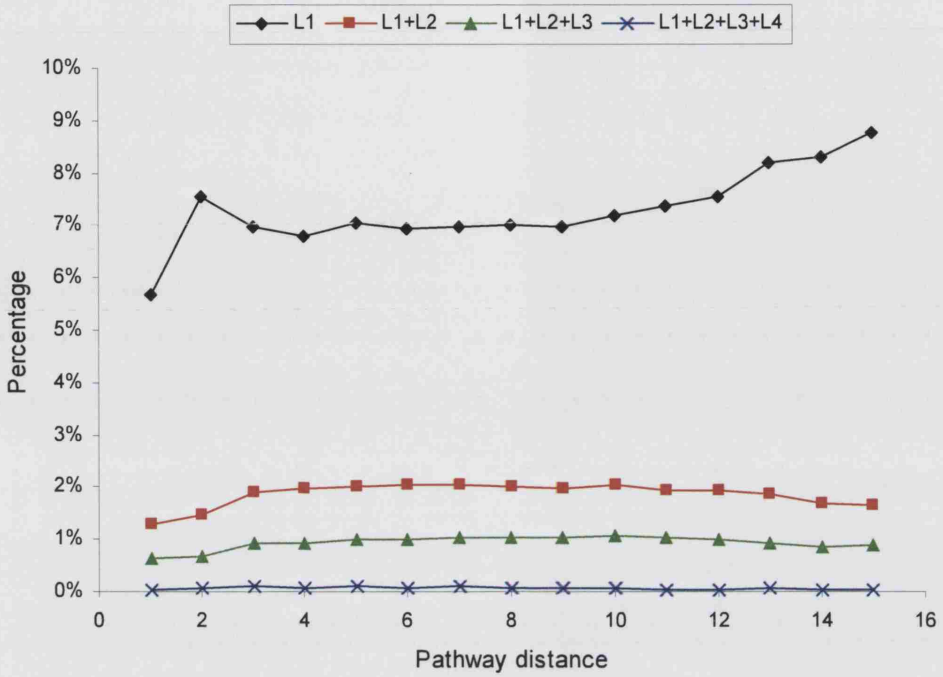

Figure 2.8: Pathway distance and function similarity (metabolite-centric dataset).

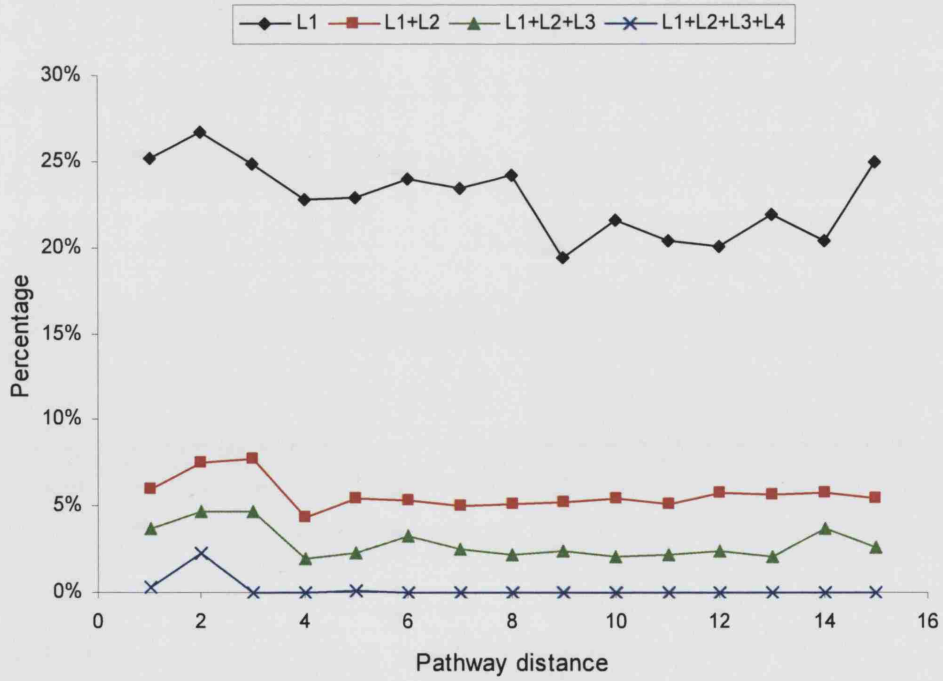

Figure 2.9: Pathway distance and function similarity (protein-centric dataset).

No obvious correlation between EC number and pathway distance could be established. Furthermore, the data show that conservation of EC number is relatively rare at all distances (the percentage of enzyme pairs with at least two EC levels is always under 8% for the protein-centric dataset, and under 3% for the metabolite-centric dataset). Even at short pathway distances, and for the protein-centric data which produce a larger number of first EC number conservations, enzyme pairs only catalyse the same type of reaction (as defined by an identical first EC number) approximately once out of 4. Moreover, this percentage is relatively constant at all distances, suggesting no particular bias for EC number conservation at shorter distances.

It is known that the relationship between EC numbers and pathways is complex, with pathways requiring a number of enzyme types to perform their task (Tsoka and Ouzounis, 2001). These data would suggest that enzymatic chemistries are varied along the substrate conversion routes. This contrasts with the recent work of Alves *et al.* (2002) who, when analysing the metabolic networks of 12 organisms derived from the metabolite-centric KEGG database (Kanehisa *et al.*, 2002), concluded there was often a clustering effect of enzymes belonging to the same class (*i.e.* sharing the same first EC number) in metabolic networks. In Alves' work, although levels of function conservation in enzyme less than 3 steps apart are significantly higher than that in enzyme pairs more than 3 steps apart regardless of homology, the correlation is substantially more pronounced when considering homologous pairs. In our work, we consider all pairs regardless of homology. It is hard to directly compare the two studies since they use different databases, and the latter study exploits pathway distance indirectly (comparing conservation of chemistry in pairs less than 3 steps apart and pairs 3 or more steps apart).

2.5.5. A case for patchwork evolution

The correlations between pathway distance and genome distance, and also between pathway distance and enzyme function for the SMM pathways of *E. coli* were investigated, in order to search for the existence of evolutionary relationships among the *E. coli* enzymes. As expected, there is a demonstrable relationship between

pathway distance and genome distance, with genes nearby in the genome far more likely to encode enzymes acting close-by in metabolism. This correlation can be attributed to operon structure. It can be surmised that, for *E. coli* SMM enzyme-encoding genes, operons cluster genes that are within a short pathway distance.

No clear trend was observed when examining the relationship between pathway distance and conservation of function. The lack of obvious correlation between pathway distance and EC numbers supports the notion of a "substrate-driven" evolution (as opposed to a "chemistry-driven" one), which is commonly associated with the patchwork model of pathway evolution: enzymes were almost randomly recruited on a need-only basis within the metabolic network of an organism. In conjunction with the results from investigations that have been performed by other researchers (Tsoka and Ouzounis, 2001; Teichmann *et al.*, 2002; Rison *et al.*, 2002; Rison, 2002), the data presented in this thesis can support the growing body of evidence suggesting patchwork evolution as the prevailing pathway evolution strategy.

2.6. Conclusions

This work has two salient conclusions: i) the LP technique presented is a fast and effective method of analysing certain properties of metabolic networks; and ii) pathway distance and genome distance correlate, but pathway distance and enzyme function do not, which offers insight into the likely model of pathway evolution.

The algorithm that has been presented in this chapter is a single-source shortest path algorithm formulated as an LP model. The algorithm is characterised by its simplicity and deals efficiently with network circularity (*i.e.* cycles within metabolic pathways). All the computational experiments were performed on an IBM RS6000 workstation. In the case of the study of correlations between minimal pathway distance and genome distance, experiments with the protein-centric dataset required 127s for the solution of 540 LPs and experiments with the metabolite-centric dataset required 1862s for the solution of 795 LPs. In the case of the study of correlations

between minimal pathway distance and enzyme function, the experiments required 124s for the solution of 507 LPs; experiments with the metabolite-centric dataset required 6498s for the solution of 634 LPs. It should be noted that these CPU times include pre- and post-processing of the data, a fairly time-consuming part of the process.

Minimal pathway distances between *E. coli* SMM enzymes have been studied using the algorithm. The issues related to the use of various conceptualisations of pathway can be seen by comparing the results obtained from the metabolite-centric and the protein-centric datasets. In the protein-centric dataset, human intervention has dealt with the issue of promiscuous compounds such as ATP, NAD(P) or water; in the metabolite-centric dataset, such compounds are included and "collapse" the network, giving it undesired properties (Alves *et al.*, 2002).

The correlations between minimal pathway distance and genome distance and enzyme function have been investigated. As expected, pathway distance correlated with genome distance with a higher probability of proximity on the genome for genes encoding enzymes involved in nearby metabolic reactions. However, pathway distance did not correlate with enzyme function as described by assigning EC numbers to SMM enzymes. These data, in conjunction with the result of previous analyses incorporating work concerning sequence and structural similarity of SMM enzymes (Teichmann *et al.*, 2002; Rison *et al.*, 2002), suggest a patchwork model of pathway evolution: the lack of obvious correlation between pathway distance and EC numbers is consistent with the *ad hoc* recruitment of enzymes where required within the metabolism of an organism (Jensen, 1976).

Chapter 3

Robustness of the p53 network and biological hackers

In this chapter, the LP model that was presented in chapter 2 is applied to a different biochemical problem. The p53 cell cycle and apoptosis control network is crucial in regulating the multicellular (metazoan) cell cycle and apoptosis. Here, the robustness of the p53 network is studied by analyzing its degeneration under two modes of attack. The LP algorithm is used to calculate average path lengths among proteins and the network diameter as measures of functionality. The significance of the results is considered with respect to mutational knockouts of proteins and the attacks mounted by tumour inducing viruses (TIVs).

3.1. Apoptosis and the p53 network

Cellular proliferation is tightly regulated in multicellular organisms. Cells accumulate in a coordinated manner during growth or repair, and undergo programmed cell death (apoptosis) when genetically damaged, when virally infected

or when the developmental program requires the death of the cell — a necessary sacrifice to save the organism as a whole. Proliferation is regulated via cyclical activation of different cyclin-dependent kinases (CDKs), which mediate the temporal activation of cell growth, DNA synthesis and cell division. Apoptosis is triggered in response to specific signals, and the cell is destroyed when a cascade of proteases (the caspases) are activated. Failure of this control system, leading to either unregulated proliferation or unnecessary apoptosis, is causative of both tumourigenesis and developmental diseases.

All organisms control progression through the cell cycle, and can respond to cellular stress by activating cell cycle checkpoints and by repairing damaged components if necessary. In addition, multicellular organisms have evolved the ability to trigger apoptosis in cases where risk to the organism is unacceptably high. In response to stress, a metazoan cell must decide between continued progression through the cell cycle or initiation of apoptosis. This decision is mediated by a protein-interaction network, at the centre of which lies p53.

The p53 protein is found only in metazoan cells, and combines protein-interaction domains, regulatory domains and a sequence specific DNA recognition domain that allow the integration of both intracellular and intercellular signals with gene transcription (Vogelstein *et al.*, 2000). Under normal conditions, p53 is turned over rapidly by proteolysis and is inactive. Cellular stress signals result in the stabilisation of p53 so that it rises in concentration to a level where it can activate transcription of its target genes. Depending on the circumstances (for example cell type and nature and strength of the stress signal), gene transcription resulting from elevated p53 levels produces responses including pausing of cell cycle, DNA repair, permanent arrest of replication, or apoptosis. The p53 protein is activated by both intrinsic and extrinsic stress signals, including DNA damage (*e.g.* from ionising radiation), mitotic spindle damage, aberrant growth signals, hypoxia, ribonucleotide depletion, and loss of cell adhesion (Robles *et al.*, 2002), many of which are characteristic indicators of tumourigenesis.

The p53 network is therefore crucial in pausing the cell cycle, in order to repair DNA damage or initiate apoptosis to destroy a tumorous cell. Such a vital cellular

mechanism should logically be relatively immune to attack; nevertheless there still exist certain threats against the p53 network, such as mutation, that can damage it. Mutations reducing p53 activity are present in over 50% of human tumours (Haupt *et al.*, 2003), either directly by knocking out the p53 gene, or indirectly by over-expressing inhibitors of the protein. For a study of the robustness of the network, one must be able to observe the response of the network to stimulation and/or perturbation, but the processes of the p53 network are unfortunately not yet understood adequately in order to model them. We must then turn to the structure of the network to find an acceptable measure of network functionality (Mahadevan and Palsson, 2005).

3.2. Network architectures

Graph theory is a branch of mathematics used to analyse complex networks of nodes and edges. Until recently, graphs were divided into two main categories; regular and random. The connections in a regular graph are very strictly ordered with all nodes of the network having the same degree or connectivity k, (the total number of connections from a node to other nodes), much like the chemical bonds in a crystal lattice. In a random network, connections are placed between any two nodes with a given probability.

The structure of a network with N nodes is often summarised by plotting k against the probability distribution function, $P(k)$ (the number of nodes with k connections, divided by the total number of nodes in the network). The plot of $P(k)$ for a random graph follows a Poisson distribution peaking at the average value of k. For a regular network, the plot is a spike as all nodes have the same connectivity. Counting the minimum number of connections that must be followed to traverse from one node, i, to another, j, yields the path length for that pair, l_{ij}. Within a random graph, connections can 'short-cut' across the entire network and so path lengths are typically much shorter than in a regular graph (Barabási and Oltvai, 2004). One global metric

of the structure of a network is its diameter, which represents the mean path length between all nodes. It is defined as:

$$D = \frac{2}{N \cdot (N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j \geq i+1}^{N} l_{ij} \tag{3.1}$$

where N is the total number of nodes in the graph.

Watts and Strogatz (1998) described the properties of a third class of graphs with an architecture in-between that of the regular and random extremes. This structure is similar to that of social networks, where people have close-knit circles of friends within a larger, but still inter-connected, social network. Millgram (1967) demonstrated the infamous 'small-world' nature of human acquaintances, and claimed that any two people in the world are no more than six 'degrees of separation' apart. Based on the same principle, Watts and Strogatz (1998) named their graphs 'small-world' networks.

'Small-world' networks combine the small diameter of random graphs with the high local connectivity of regular graphs. In addition to this small-world property, many networks with no pre-designed architecture that grow and evolve over time have a characteristic pattern of connectivity. $P(k)$ decays as a power-law – the vast majority of nodes have only a few connections, but there are several hubs that are very highly connected. Unlike regular or random graphs there is no characteristic degree of connectivity, and so such networks are termed 'scale-free'. Figure 3.1 demonstrates the characteristic topologies and $P(k)$ plots of random and scale-free networks.

In recent years, a great number of networks have been shown to be scale-free, including the Internet (Albert *et al.*, 1999), social interactions (Albert *et al.*, 2000), neural networks (Strogatz, 2001), ecological food webs (Strogatz, 2001), protein-protein interactions (Jeong *et al.*, 2001), gene transcription regulation networks (Barabási and Oltvai, 2004) and metabolism (Jeong *et al.*, 2000). Interestingly, a recent study (Arita, 2004) claims that the small-world properties of *E. coli* metabolism are not inherent in the network but are an artefact of the way that metabolic relationships are considered. In any case, "small-worldness" does seem to

be an inherent property of most, if not all biological networks. One explanation for the occurrence of this structure is that of network growth through preferential attachment of additional nodes. An initially small network grows by the connection of additional nodes, not to randomly-selected extant nodes, but with a probability proportional to each extant node's current connectivity (k), with hubs being created by a positive-feedback 'rich getting richer' process.

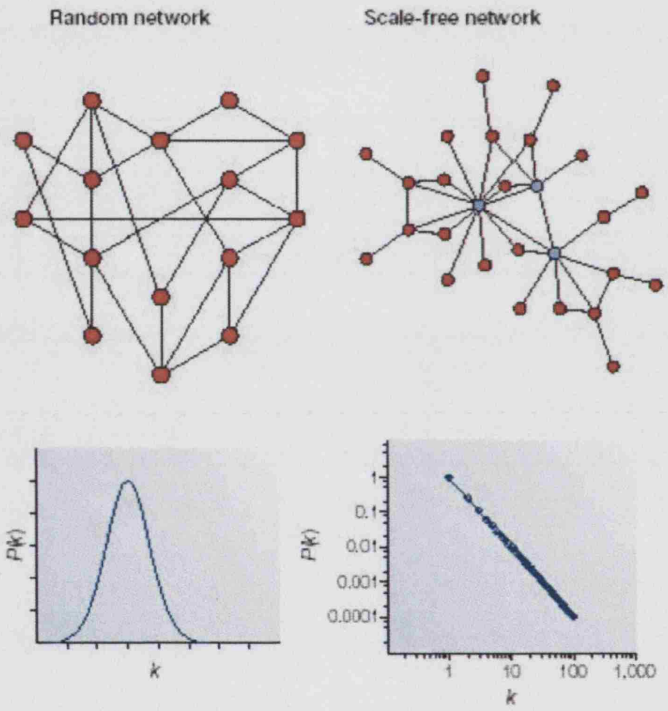

Figure 3.1: P(k) plots for random and scale free-networks (from Barabási and Oltvai, 2004).

Barabási and Oltvai (2004) explain how this process might operate in biological systems. Protein-protein interaction networks grow through gene duplication. When a randomly selected gene duplicates, the protein transcript of the copied gene retains all the interactions of the original. Thus, all proteins already linked to the original protein gain an extra connection. Highly connected proteins are more likely to link to the original protein, and this provides the mechanism for preferential attachment that results in the power law distribution and high clustering coefficient of small-world networks. A natural corollary of this mechanism is that the proteins that appeared earliest in the history of the growing network should become the most well-

connected. Indeed, in metabolic networks the hubs include coenzyme A, NAD and GTP, believed to be remnants of the RNA world. In protein interaction networks, the most evolutionarily ancient proteins, as revealed by comparative genomics, have the highest k values (Barabási and Oltvai, 2004).

3.2.1. Network robustness

Much research has been conducted into the robustness of networks, that is, their ability to remain relatively undisrupted in the face of perturbation. Robustness can be defined, in topological terms, as the remaining communication ability within a network as nodes or connections are removed. Real networks also perform one or more functions, be it electricity distribution or genetic regulation, but modelling a complex network's performance is often prohibitively difficult. Network navigability is a necessary (although not sufficient) prerequisite for adequate function, and so the diameter of the damaged network is taken as an acceptable proxy (Albert *et al.*, 2000). In order to study the robustness of the network, a series of knockouts (deletions) of nodes can be performed, and the response of the network diameter is observed.

Either individual connections, or entire nodes can be removed from a network, with the latter having a greater impact. There are two main modes of attack upon the nodes of a network – either removed at random, or the preferential targeting of the hubs. A random graph responds identically to both random and directed attacks as its connectivity is homogenous — the majority of its nodes have roughly the same number of links, approximately equal to the network's average degree. Upon successive deletion of nodes the diameter increases monotonically until a critical threshold fraction has been disabled, and the network undergoes a phase transition as it disintegrates into isolated fragments. In contrast, a scale-free network is relatively immune to random node failure, but extremely vulnerable to a targeted onslaught (Barabási and Oltvai, 2004).

3.3. Problem statement

Overall, the problem of the study of robustness of the p53 protein interaction network can be stated as follows:

Given:

- the list of proteins (nodes) and interactions amongst them (edges) of the p53 cell cycle and apoptosis control network.

Determine:

- the average path lengths (APLs) of all proteins in the p53 network;
- the diameter of the p53 network, defined as the mean path length among all nodes.
- the minimal pathway distances among enzymes of the SMM of *E. coli*.

So as to investigate the robustness of the p53 network against mutational perturbation and against directed attacks on its hubs.

3.4. Algorithm

The LP model used in Chapter 2 for the calculation of pathway distances in the *E. coli* small molecule metabolism is also applied here to calculate the average path lengths of the proteins in the p53 network. The notation used in the mathematical model is adapted to suit the needs of the study.

3.4.1. Nomenclature

Indices

i,j	proteins

Parameters

N	total number of proteins in the network
C_{ij}	1 if there is a connection from protein i to protein j; 0 otherwise
T	large number

Positive continuous variables

APL_{i^*}	average path length of source protein i^*
l_i	path length from protein i^* to protein i

3.4.2. Model constraints

The same algorithm used in chapter 2 is also applied here. The only difference is the objective function (3.2), which has been modified to better reflect the fact that we use the algorithm to identify the APLs of all proteins in the network. Each protein is systematically set as the source, i^*, and the algorithm finds the shortest path to all other proteins by solving the following LP optimisation model:

$$maximise\ APL_{i^*} = \frac{\sum_i l_i}{N-1} \tag{3.2}$$

subject to:

$$l_j \le l_i + 1 \qquad \forall (i,j) : C_{ij} = 1 \tag{3.3}$$

$$l_{i^*} = 0 \tag{3.4}$$

$$l_i \ge 0 \tag{3.5}$$

Constraint (3.3) incorporates pathway information related to network connectivity, facilitated by the use of parameter C_{ij}. Constraint (3.4) assigns the initial value of zero to protein i^* to denote it as the source protein, and constraint (3.5) requires all l_i variables to be positive. Unbounded solutions are avoided by adding:

$$l_i \le T \qquad \forall i \tag{3.6}$$

If l_i equals T then there is no path connecting the i^* source protein to protein i.

3.5. Methods

A model of the p53 network (a large module of the entire metazoan protein interaction network) was constructed. This model was then subjected to both random and directed modes of attack, and its changing diameter studied. The objective was to analyze how the network behaves in response to the stochastic protein knockouts from mutation during tumourigenesis, and also a targeted attack.

3.5.1. Generation of dataset

There are over 35,000 published articles relating to p53, its interactions, its functions, and the consequences of its inactivation. These articles describe the function of p53 in multiple organisms, multiple cell types within those organisms, and under a wide range of different situations within those cells. The result is a bewildering volume of information relating to p53 interactions, the relative merits of which can be extremely difficult to judge. Consequently, very few studies have attempted to fully connect the network in any meaningful way. The raw data therefore had to be created almost from scratch.

Kohn (1999) performed an extensive literature review and presented an annotated molecular interaction map of proteins involved in mammalian cell cycle progression and checkpoints, DNA repair, and apoptosis. The molecular maps of these processes all feature p53 prominently, and although the connections are incomplete and inevitably contain inaccuracies, the majority of the described interactions are experimentally validated and well understood. It is unlikely that a few false-positives or negatives would drastically alter the architecture of the network or calculations of its diameter.

Although there is a high degree of modularity within interaction networks, the dissociation of the "p53 network" follows largely arbitrary borders. This study therefore followed the same boundaries selected in the work of Kohn (1999), yielding a network containing 104 nodes and 226 unique connections. A representation of this was constructed as presented in Figure 3.2. It can be seen that

the vast majority of the nodes are poorly connected within the network, whereas a very small minority of the nodes are hubs (highlighted here in colour) with a high centrality. The list of the 104 nodes included in the p53 network is presented in Table 3.1. The list of the 226 interactions (connections) amongst those nodes can be seen in Table 3.2.

Table 3.1: The 104 nodes of the p53 protein interaction network.

14-3-3	C-TAK1	Gadd45	p19ARF	RHA
Abl	CycA	HBP1	p21	RPA
AP2	CycB	HDAC1	p27	Rpase_2
APC	CycD	Histones	p300	Skp1
ATM	CycE	HMG	p36MAT1	Skp2
Bax	CycH	HR23B	p53	SL1
BRCA1	DMP1	JNK	p57	Sp1
C-EBP	DNA-PK	Jun	p68	ssb
Casp3	DP1-2	Karp-1	PARP	ssDNA
Cdc25A	Dpase_a	Ku70	Paxillin	TAFII250
Cdc25C	Dpase_b	Ku80	pCAF	TBP
Cdk1	Dpase_d	Ligase_1	PCNA	TFIIH
Cdk2	dsDNA	Ligase_3	PKC	U-glyc
Cdk4-6	E2F1-2-3	MAPK	Plk1	Wee1
Cdk7	E2F4	Max	pRb	XPA
Chk1	E2F5	Mdm2	Rad51	XPB
CK1d-k	E2F6	Myc	Rad52	XPC
CK2	E-cad	Myt1	Raf1	XPD
Cks1	ERCC1	p107	Ras	XPF
Crk	FEN-1	p130	Rep_fork	XRCC1
CSB	Fos	p16	RF-C	

Several nodes, such as ssDNA (single stranded DNA), are not proteins, but are none the less crucial objects that interact within the p53 control network and therefore they are included in this study. All interactions were taken to be undirected, a valid assumption for mutual binding in a complex or phosphorylation of a target protein by a kinase. Several of the interactions, however, were transcription regulation events, such as the p53-activated expression of Bax (Kohn, 1999; Vogelstein *et al.*, 2000). These are directional, in that p53 affects Bax, but not visa versa. Directed interactions account for only around 5% of the total described by Kohn (1999). Although the LP algorithm used here is capable of dealing with directionality, the entire map was taken to be undirected to simplify the analysis.

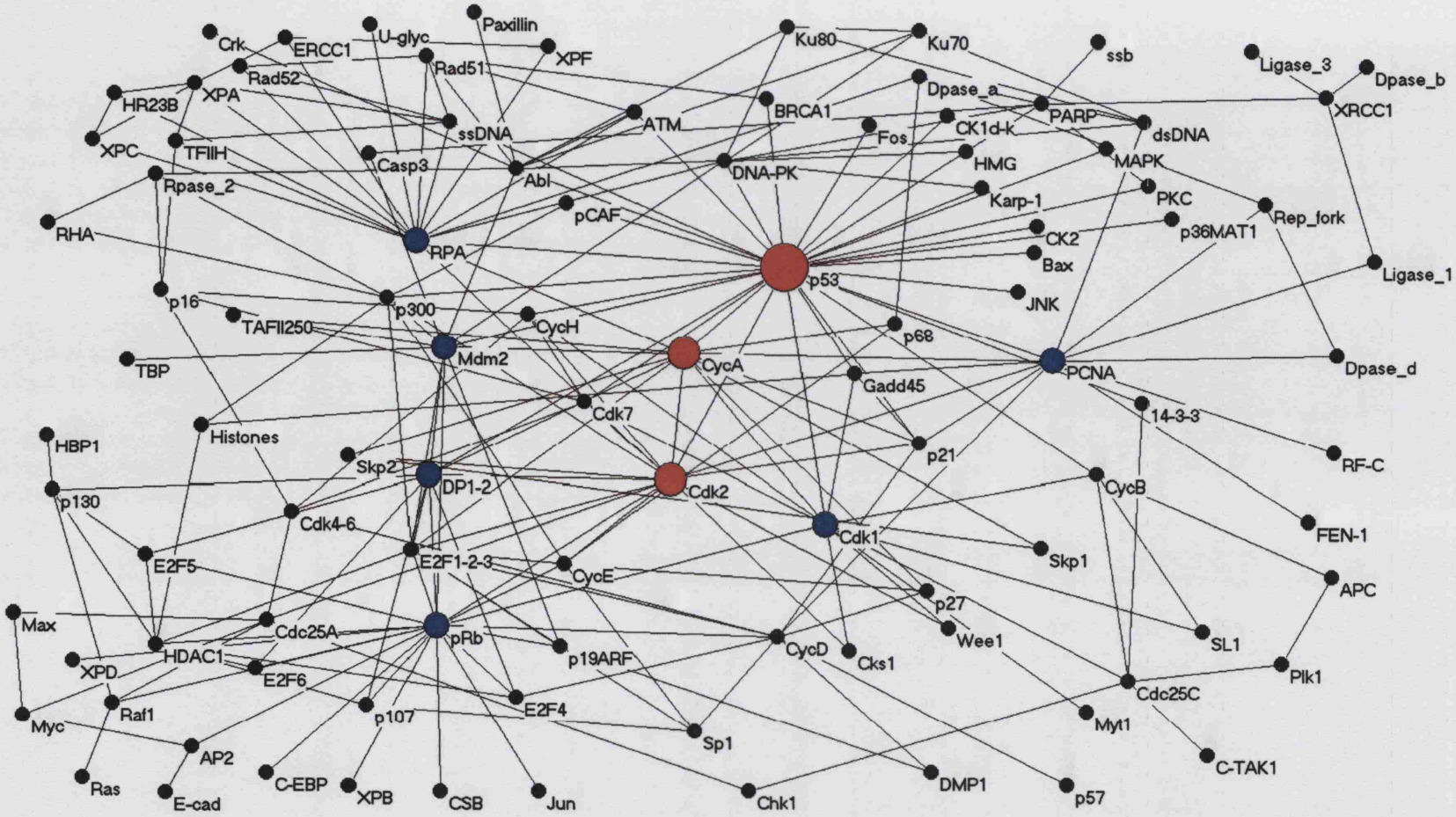

Figure 3.2: Visualisation of the p53 network, colour-coded with respect to centrality.

Table 3.2: List of interactions in the p53 network. All connections are considered to be undirected.

14-3-3	→ Cdc25C	Chk1	→ Cdc25A	E2F1-2-3	→ CycA	p300	→ Mdm2	PARP	→ ssb	RHA	→ p300
Abl	→ Crk	Chk1	→ Cdc25C	E2F1-2-3	→ CycD	p300	→ pCAF	PARP	→ XRCC1	RHA	→ Rpase_2
Abl	→ Ku70	Cks1	→ Cdk1	E2F1-2-3	→ CycE	p300	→ Rpase_2	PCNA	→ CycD	RPA	→ ATM
Abl	→ Ku80	Cks1	→ Cdk2	E2F1-2-3	→ Mdm2	p53	→ 14-3-3	PCNA	→ Dpase_d	RPA	→ DNA-PK
Abl	→ Mdm2	C-TAK1	→ Cdc25C	E2F1-2-3	→ p19ARF	p53	→ ATM	PCNA	→ dsDNA	RPA	→ Dpase_a
Abl	→ p53	CycA	→ p21	E2F4	→ CycD	p53	→ Bax	PCNA	→ FEN-1	RPA	→ ERCC1
Abl	→ Paxillin	CycA	→ p27	E2F5	→ p130	p53	→ BRCA1	PCNA	→ Gadd45	RPA	→ HR23B
Abl	→ Rpase_2	CycA	→ p68	Gadd45	→ Cdk1	p53	→ Cdk1	PCNA	→ Ligase_1	RPA	→ TFIIH
AP2	→ E-cad	CycA	→ PCNA	Gadd45	→ p21	p53	→ Cdk2	PCNA	→ RF-C	RPA	→ U-glyc
AP2	→ Myc	CycA	→ RPA	HBP1	→ p130	p53	→ Cdk7	pRb	→ C-EBP	RPA	→ XPF
AP2	→ pRb	CycA	→ TAFII250	HDAC1	→ E2F1-2-3	p53	→ CK1d-k	pRb	→ Cdk1	Skp1	→ CycA
APC	→ CycB	CycD	→ DMP1	HDAC1	→ E2F4	p53	→ CK2	pRb	→ Cdk2	Skp1	→ Skp2
APC	→ Plk1	CycD	→ p21	HDAC1	→ E2F5	p53	→ CycA	pRb	→ CSB	Skp2	→ Cdk2
ATM	→ Abl	CycD	→ p27	HDAC1	→ E2F6	p53	→ CycB	pRb	→ CycD	Skp2	→ CycA
ATM	→ Rad51	CycD	→ p57	HDAC1	→ p107	p53	→ CycH	pRb	→ CycE	SL1	→ Cdk1
Casp3	→ Mdm2	CycD	→ Sp1	HDAC1	→ p130	p53	→ DNA-PK	pRb	→ DP1-2	SL1	→ CycB
Casp3	→ PARP	CycE	→ p27	HDAC1	→ pRb	p53	→ DP1-2	pRb	→ E2F1-2-3	Sp1	→ CycE
Cdc25A	→ Cdk2	CycE	→ p68	Histones	→ Gadd45	p53	→ Fos	pRb	→ E2F4	Sp1	→ E2F1-2-3
Cdc25A	→ Cdk4-6	CycH	→ Cdk1	Histones	→ HDAC1	p53	→ Gadd45	pRb	→ E2F5	Sp1	→ p107
Cdc25A	→ Max	CycH	→ Cdk2	Histones	→ p300	p53	→ HMG	pRb	→ E2F6	ssDNA	→ RPA
Cdc25A	→ Myc	CycH	→ Cdk4-6	HMG	→ DNA-PK	p53	→ JNK	pRb	→ Jun	ssDNA	→ TFIIH
Cdc25A	→ Raf1	DMP1	→ p19ARF	Ku70	→ dsDNA	p53	→ Karp-1	pRb	→ Mdm2	ssDNA	→ XPA
Cdc25C	→ Cdk1	DNA-PK	→ Abl	Ku70	→ Ku80	p53	→ MAPK	pRb	→ p19ARF	TBP	→ Mdm2
Cdc25C	→ CycB	DNA-PK	→ dsDNA	Ku80	→ dsDNA	p53	→ Mdm2	pRb	→ XPB	Wee1	→ Cdk1
Cdc25C	→ Plk1	DNA-PK	→ Karp-1	Mdm2	→ CycA	p53	→ p21	pRb	→ XPD	Wee1	→ Cdk2
Cdk1	→ CycA	DNA-PK	→ Ku70	Mdm2	→ p19ARF	p53	→ p300	Rad51	→ Abl	XPA	→ ERCC1
Cdk1	→ CycB	DNA-PK	→ Ku80	Mdm2	→ TAFII250	p53	→ p36MAT1	Rad51	→ BRCA1	XPA	→ HR23B
Cdk2	→ CycA	DNA-PK	→ Mdm2	Myc	→ Max	p53	→ PARP	Rad51	→ Rad52	XPA	→ RPA
Cdk2	→ CycE	DP1-2	→ Cdk2	Myt1	→ Cdk1	p53	→ pCAF	Rad51	→ RPA	XPA	→ TFIIH
Cdk2	→ p21	DP1-2	→ CycA	p107	→ pRb	p53	→ PCNA	Rad51	→ ssDNA	XPC	→ HR23B
Cdk2	→ PCNA	DP1-2	→ E2F1-2-3	p16	→ Cdk7	p53	→ PKC	Rad52	→ RPA	XPC	→ RPA
Cdk2	→ RPA	DP1-2	→ E2F4	p16	→ CycH	p53	→ RPA	Rad52	→ ssDNA	XPC	→ XPA
Cdk4-6	→ CycD	DP1-2	→ E2F5	p16	→ Rpase_2	p53	→ ssDNA	Raf1	→ p130	XPF	→ ERCC1
Cdk4-6	→ p16	DP1-2	→ E2F6	p16	→ TFIIH	p68	→ Dpase_a	Raf1	→ pRb	XRCC1	→ Dpase_b
Cdk7	→ Cdk1	DP1-2	→ Mdm2	p21	→ PCNA	PARP	→ DNA-PK	Raf1	→ Ras	XRCC1	→ Ligase_1
Cdk7	→ Cdk2	DP1-2	→ p107	p300	→ Cdk2	PARP	→ Dpase_a	Rep_fork	→ Dpase_a	XRCC1	→ Ligase_3
Cdk7	→ Cdk4-6	DP1-2	→ p130	p300	→ CycE	PARP	→ dsDNA	Rep_fork	→ Dpase_d		
Cdk7	→ CycH	E2F1-2-3	→ Cdk2	p300	→ E2F1-2-3	PARP	→ PKC	Rep_fork	→ PCNA		

3.5.2. Centrality of network nodes

The average path length (APL) for a protein is the mean of the shortest paths between it and all other nodes in the network. Fell and Wagner (2000) have calculated this metric for metabolites in *E. coli* core metabolism. Some studies on network architecture use "connectivity" as the metric for a node's importance. It is a simple count of the number of other nodes it connects to, but it can often be biased. A node that links to many dead-end nodes, which are not themselves well integrated into the network, has a high connectivity, but the significance of the node within the network as a whole is greatly overemphasised. This study, however, calculates the average path length for a protein in the p53 network, summarising the propinquity of a node to every other — *i.e.* the centrality of the node within the graph. It is a global measure of a node's importance, and so the adoption of "centrality" rather than "connectivity" is preferred as the standard metric for node rankings. The network diameter was calculated by dividing the sum of all path lengths by the total number of protein pairings ($N \cdot (N - 1)$).

3.5.3. Network attacks

For studies on the attack tolerance of networks, either individual edges or entire nodes (and thus all involved connections as well) can be removed. The biological equivalents of these are simple. A mutation may occur in the regulatory DNA that creates a null mutant for that protein (absolutely no gene product created – node knockout) or else within the DNA coding for the binding domain to another protein that obliterates one specific interaction (edge knockout). Here the focus was on node eliminations, as the effects on network topology are more obvious.

The survivability of the p53 network in the face of both a directed or random attack against its nodes was examined. At each stage of the attack an additional protein was knocked-out and the diameter recalculated. For the directed attack regime the proteins were added to the exclusion list in rank order of centrality (as defined by their average path lengths). For the random attack a random permutation was applied, and the attack was repeated 100 times; the diameter at each step averaged across all runs.

Here, the removal of a node destroyed all the connections it possessed, but an attack against only one connection at a time is also possible.

A protein knockout decrements the number of nodes over which the diameter is calculated. Another possible consequence, especially from the loss of a hub, is that nodes may become isolated from the rest of the network. This produces nodes with no navigable route to each other, in which case constraint (3.6) sets the path lengths equal to the arbitrarily large number T, given the value of 100 in this study. We have studied the sensitivity of the results to the value of parameter T; the effect is illustrated in Figure 3.3, which demonstrates a directed attack against the hubs of the p53 network with different choices of T. It is obvious that the choice of the value of the parameter does not affect the outcomes of our study. There are no qualitative differences among the plots; the only variation is the scale of the y-axis, which does not influence at all any of the conclusions drawn in the chapter.

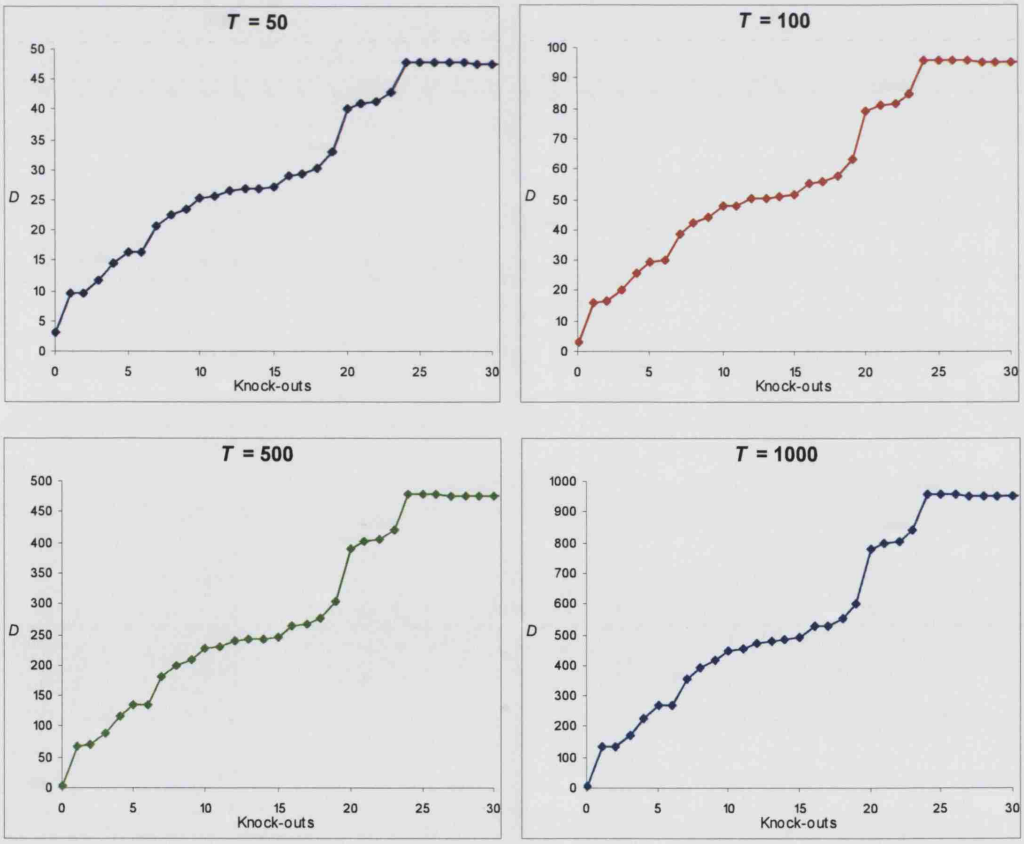

Figure 3.3: Sensitivity of results to the value of parameter T.

A path length of 100 means that there is no route between two proteins, for example if one has already been knocked out or has become isolated from the rest of the network due to the loss of a crucial hub. The diameter of the network at successive steps was calculated using all path lengths, including those with a value of 100. Although this method involves arbitrarily assigned values, it is at least consistent between different attack modes, and computationally simple to calculate. The alternative would have been to only calculate the diameter of the largest cluster, only including nodes that are known to remain connected. Aside from the difficulty of estimating which is the largest cluster (which can only be accomplished by redrawing the interaction map at every time step), this approach creates another problem: knocking out a hub removes many possible routes, causing a sharp increase in diameter. But hub removal also results in fragmentation into isolated sub-clusters, so that the largest cluster rapidly decreases in size. If only the diameter of the largest cluster was considered, these two effects would largely counteract each other so that the calculated diameter would barely change, grossly misrepresenting the actual degeneration of the network. Fragmented networks contain many path lengths of 100, and so the tactic used does capture the breakdown of the network.

3.6. Computational results

Figure 3.4 shows the relationship between connectivity, k, and the probability distribution function, $P(k)$. The number of nodes with a given connectivity can be seen to decay logarithmically as connectivity increases. This power-law relationship is a defining feature of a scale-free graph, and so it is possible that the p53 network also possesses such architecture.

When calculating the average path lengths and diameter of the network, the LP model presented in section 3.3 was solved with an algorithm implemented with GAMS (Brooke *et al.*, 1998), using the CPLEX 6.5 solver algorithm, and run on a RS6000 workstation. Table 3.3 shows the nodes with lower APLs, with hubs being defined as those proteins with the very lowest scores.

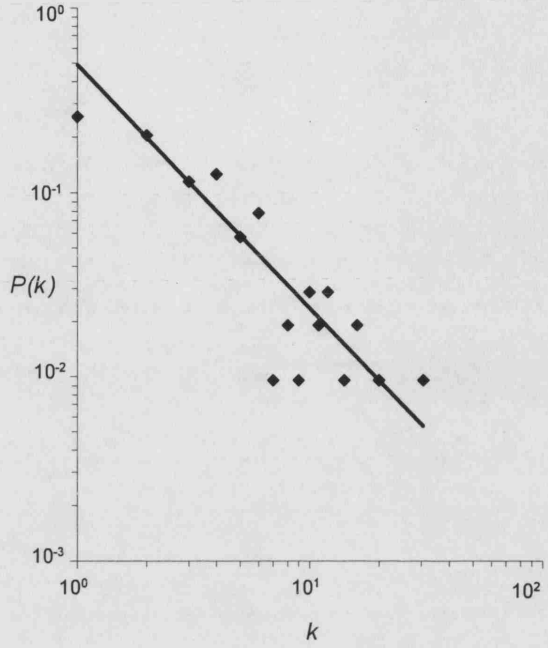

Figure 3.4: The power law relationship between k and P(k) for the p53 network.

Table 3.3: The 30 best-connected nodes in the p53 network, in order of ascending Average Path Length (APL).

APL	Proteins
1.9	p53
2.1	Cdk2
2.2	CycA
2.3	Cdk1, Mdm2, DP1-2, pRb
2.4	PCNA, RPA
2.5	DNA-PK, p21, p300, E2F1-2-3, Cdk7, CycH
2.6	Abl, Gadd45
2.7	CycB, CycD, CycE, PARP, ATM
2.8	ssDNA, Cdc25A, 14-3-3, pCAF, PKC
2.9	HMG, Karp-1, BRCA1

3.6.1. Protein knockouts

Figure 3.5 presents the plots of the diameter of the p53 network over the first 30 knockouts with nodes removed in either a random attack, or one directed against the hubs (nodes knocked-out in rank order of average path length). The diameter of the p53 network under a random attack increases very slowly. Thanks to its architecture, the majority of nodes in the network are poorly-connected and therefore their removal has a very small effect on network navigability. Hub nodes are uncommon and so they are rarely hit. The p53 network is shown here to be resistant to a random pattern of attack, which equates to robustness to mutational perturbation: mutations are commonly held to be randomly distributed events, and so assuming genes in the p53 network are of roughly equal length, mutations would knock out proteins in a random failure pattern.

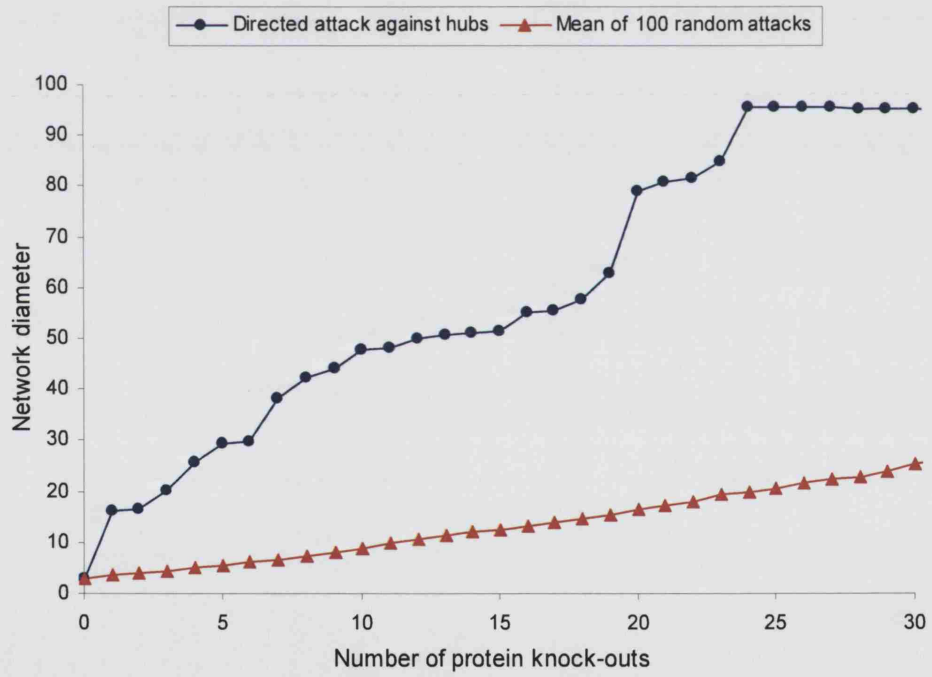

Figure 3.5: Degeneration of network diameter when nodes are knocked out in either a random pattern, or in a directed attack against the hubs.

Under a directed onslaught, however, network communication fails as the diameter of the network rapidly degenerates. The result of knocking out the first hub protein, p53,

is an increase in network diameter of over five-fold; from 3.1 to 16.1. The loss of no other protein has such a devastating effect: removal of the next four hubs produces only a further doubling of the diameter. After the 24th knockout the diameter levels off: the directed attack has removed all of the hubs, and consequently the majority of routes between proteins. The network has shattered into isolated subclusters, and most of the path lengths between protein pairs are designated as 100, therefore further knockouts can damage the network no further.

3.6.2. Statistical analysis

To assess the results statistically, the standard normal deviate or Z-score is used, which measures the distance of any value from the mean of a population in standard deviation units. The results of the directed attack performed on the p53 network (see Figure 3.5) are compared to the results of the same pattern of attack on 100 random networks. For each random network, 104 protein nodes were linked through random generation of 226 edges; the mean diameter ($\bar{D}$) and the standard deviation (σ) from 100 random networks, and the p53 network's diameter (D) were used for Z-score calculation.

$$Z = \frac{D - \bar{D}}{\sigma} \tag{3.7}$$

Figure 3.6 presents statistical analyses using Z-scores between the p53 network and random networks. Note that diameter values were first normalised to compensate for the fact that the p53 network has a lower initial diameter from any random network, due to its scale-free nature. Z-scores indicate how far and in what direction each item deviates from the random mean. Values greater than 3 are typically considered to be significant. As can be seen in Figure 3.6, our results differ considerably from random; thus indicating that trends observed cannot be attributed to chance.

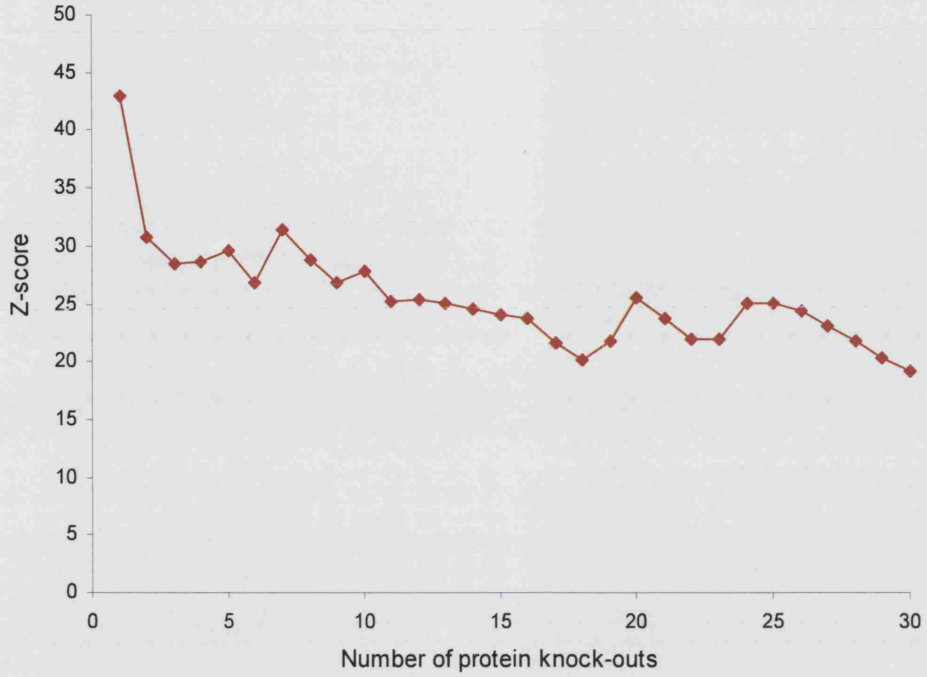

Figure 3.6: At each knock-out, the Z-score of the diameter of the network is plotted. The mean and standard deviation used for the estimation of the Z-score are those of random networks with 104 nodes and 226 unique connections.

The plot of network degeneration under directed attack (Figure 3.5) shows some interesting features. There are several small plateaux where diameter is temporarily stable. For example, navigability barely alters between the 2nd and 3rd knockouts (CDK2 and Cyclin A, respectively). These two proteins bind together to allow progression through a cell cycle checkpoint and activation of the DNA replication machinery. They thus bind to a large number of the same proteins, and so within this graph model are largely redundant — it is not until the second one is knocked out that routes between certain nodes are lost and the diameter jumps up. However, this behaviour under attack is an artefact of the nature of the model. In reality, CDK2 and Cyclin A bind together as a complex and so need each other in order to function. Knockout of either results in a loss of function of the other (and thus disappearance of its connections), in other words, such nodes in the p53 network are not independent. The model also assumes bidirectionality of connections, a non-dynamic presence of proteins in the cell (*i.e.* the temporal component of protein concentrations is ignored),

and all connections are given an equal weighting. When more complete data on protein dependencies, protein complexes, temporal fluctuations, and relative importance of interactions become available, these factors could be incorporated into models such as this one.

3.7. Biological hackers

The complete set of proteins involved within the p53 network and connections among them is still being mapped out, but the power-law relationship demonstrated in Figure 3.4 suggests that it may posses a scale-free structure. The p53 network shows a similar response under attack to scale-free networks such as the Internet; it is robust against a random attack as most of the protein knockouts will have negligible impact on the global integrity of the network. This reliance on highly-connected nodes, however, renders the network vulnerable to a directed attack. The most important nodes are selectively targeted and the diameter of the p53 network rapidly degenerates. A similar result was obtained on simulated attacks on the Internet, which was found to be robust to random server failures, but vulnerable to the activities of hackers deliberately targeting the hubs so as to wreak maximal havoc (Albert *et al.*, 2000). The question arises as to whether it is biologically possible to orchestrate a directed attack against the p53 network.

Such a threat does in fact exist in nature, operating not at a genetic level, but against the translated proteins. DNA tumour-inducing viruses (TIVs) increase their replication rate and survival chances with an armoury of proteins that suppress the normal apoptotic infection response, short-circuit the cell cycle into continually synthesizing viral DNA, and force the cell into a stealth mode to evade immune system surveillance. For example, DNA comprising the adenovirus is translated by the host cell into a variety of proteins (Burgert *et al.*, 2002). E1a reprograms the cell for continuous DNA synthesis, E1b/55Kd interrupts the induced intrinsic apoptotic response to viral infection by inhibiting p53, and E3 prevents apoptosis through the extrinsic pathway by interfering with the death receptors. Other viral proteins prevent

antigen presentation from the infected cell by inhibiting MHC molecules. Table 3.4 summarises the oncoproteins produced by adenoviruses, and the effect of their inhibition of target cellular proteins (data from Levine, 1992; Burgert *et al.*, 2002; Banks *et al.*, 2003).

Table 3.4: Tumour inducing viruses, the nodes in the p53 network their oncoproteins target, and the extent of damage inflicted on the network by those knockouts.

TIV	Viral protein	Host protein	Effect	Diameter after KO
Adenovirus	E1a	pRb	Apoptosis evasion	24.98
	E1b/55Kd	p53	Apoptosis evasion	
Coxsackie	unknown	cyclin D1	Transcription/reactivation	5.00
HCMV[a]	pp71	pRb, p107, p130	Transcription/reactivation	14.37
HPV[b] 16/18	E6	p53	Apoptosis evasion	27.07
	E7	pRb, p107, p130	Transcription/reactivation	
HSV[c]	ICP0[d]	DNA-PK	Transcription/reactivation	3.12
SV40[e]	Lg T Ag[f]	pRb, p53	Apoptosis evasion	24.98

[a] human cytomegalovirus
[b] human papillomavirus
[c] herpes simplex virus
[d] infected cell protein
[e] simian virus 40
[f] large T antigen

The two most common targets, p53 and pRb, are also two of the most central proteins in the network, with APLs of 1.9 and 2.3, respectively. Targeting these proteins allows the infected cell to escape apoptosis or halting of cell division. Knocking out multiple downstream effector proteins would have the same effect, but as this study suggests it is much more efficient to selectively remove the hubs. This is especially important for viruses, as their genome is often optimised for rapid replication and cannot afford the information cost of coding many oncoproteins. It is not advantageous for the TIVs to completely destroy the p53 network either (as then no

DNA replication or cell division would occur); they need only disable regions that halt cell cycle progression or trigger apoptosis. It is conceivably for this reason that not all of the target proteins are hubs (although none have an APL greater than 3.2), but their removal disables specific functions of the p53 network.

The final column in Table 3.4 shows the calculated diameter after the targeted proteins of the third column have been knocked out using the LP algorithm. The TIV directed strikes are effective at disrupting communication within the p53 network, but do not increase the diameter so much that the network shatters and function fails completely. TIVs thus behave like biological hackers, targeting their attack against some of the p53 network hubs and so exploiting the weakness in its architecture.

The attack pattern of tumour inducing viruses might be expected to exert a selective evolutionary pressure on the p53 network, pushing evolution towards architectures less vulnerable to this type of directed attack. The p53 network, as previously explained, is particularly crucial for multicellular animals, and so it might be interesting to speculate as to whether the network has developed through evolution ways to counteract the threat presented by TIVs. One possible way to protect the network from the ill effects of the deactivation or inhibition of its hubs is network redundancy.

Redundancy can exist on two levels in biological networks: backup links along crucial routes, such as the p19ARF-MDM2-p53 pathway (edge redundancy); or proteins with overlapping functions (node redundancy). Intriguingly, this later form of redundancy has recently been observed: p53 was originally assumed to be a unique gene, but two analogous genes, p63 and p73, have now been discovered (Irwin and Kaelin, 2001). They are more closely related to each other than either to p53, but they do share significant sequence homology in three domains, including activation and DNA binding. Although neither appear to be crucial tumour-suppressor genes (they are rarely mutated in cancer tissue and null mutants are not tumour-prone) they can perform many of p53's functions, including activating transcription of p53-responsive genes and inducing apoptosis. None of the viral oncoproteins listed in Table 3.4 can bind to and inactivate p63 or p73. This evidence supports the hypothesis that ancient gene duplications yielded backup copies of p53-like genes, which were selected due

to the extra robustness they imparted on this crucial network against random or even directed attacks.

3.8. Conclusions

The p53 cell cycle and apoptosis control network is inherently robust to random knockouts of its proteins, which signifies resilience against mutational perturbation provided by the structure of the network itself. This robustness against mutations, however, gives the network an Achilles Heel, as the reliance on highly-connected nodes makes it vulnerable to the loss of its hubs. Evolution has produced organisms that exploit this very weakness in order to disrupt the cell cycle and apoptosis system for their own ends: tumour inducing viruses target specific proteins to disrupt the p53 network, and this study has identified these same proteins as the network hubs. Although TIVs have previously been likened to 'biological hackers', here we show why the TIV attack is so effective – TIVs target a specific vulnerability of the network that can be explained in terms of network architecture. A Z-score analysis of the results has demonstrated that our findings differ considerably from random and cannot be attributed to chance.

From the computational perspective, we display the effectiveness of the algorithm in analysing the properties of a large protein interaction signalling network. Thus, the applicability of mathematical programming techniques in the analysis of biochemical networks is demonstrated. The algorithm has already been applied in the previous chapter in the study of the correlations between minimal pathway distances of the *E. coli* SMM enzymes and genome distance, and between *E. coli* minimal pathway distances and enzyme function. Here it is proven to be a valuable analysis tool for complex biological networks.

The connectionist network presented is a first-level model of the p53 cell cycle and apoptotic control network with a specific and clearly-defined function. The fact that we can represent and test the p53 network offers the future possibility to attach directions and strength values to the connections as more biological data become

available, in order to make accurate predictions about the importance of individual nodes and edges. This will allow frameworks like the one presented to be used in comparative analyses of how and why the variable dynamic network components operate under different evolutionary and cell type conditions.

The application of the LP algorithm has provided insight into the mode of attack utilised by tumour-inducing viruses upon the p53 apoptotic control network. As the p53 network is vital for multicellular animals, it might be possible that the prevalence of hubs or specific interactions has been adaptively tweaked by the evolutionary pressures caused from viral attacks to render the p53 network even more 'hardened' than other analogous networks.

PART II

PROTEIN STRUCTURE PREDICTION

Chapter 4

Protein folding using lattice models

All proteins are composed of the same building blocks; the 20 known amino acids (Table 4.1). Amino acids have the capacity to form long chains that fold into a unique three-dimensional structure. This structure is extremely important, since it allows the protein to perform its biochemical function. All the information required for the folding of a protein into its functional native conformation is contained in its amino acid sequence. Despite the fact that the number of conformations of an amino acid chain is too large to sample, peptide chains are able to fold extremely rapidly to the native state (Levinthal paradox; Levinthal, 1969). Yet, it is still an extremely complex task to extract this information in order to predict the 3D structure, especially for large proteins.

Anfinsen (1973) showed that protein folding is a process governed by physical constraints only, *i.e.* it only depends on the specific amino acids that comprise the sequence. Anfinsen's hypothesis suggests that we can theoretically predict the three-dimensional structure of a protein by minimising a model of its free energy. This is the basis of the protein folding problem; nevertheless the prediction of protein

structure remains extremely difficult, due to the very large number of possible conformations available to one amino acid sequence. Protein structure prediction is one of the most important unsolved problems of computational biology today.

Table 4.1: The 20 amino acids and their abbreviations.

Names	Abbreviations	
Alanine	ala	A
Arginine	arg	R
Asparagine	asn	N
Aspartic acid	asp	D
Cysteine	cys	C
Glutamine	gln	Q
Glutamic acid	glu	E
Glycine	gly	G
Histidine	his	H
Isoleucine	ile	I
Leucine	leu	L
Lysine	lys	K
Methionine	met	M
Phenylalanine	phe	F
Proline	pro	P
Serine	ser	S
Threonine	thr	T
Tryptophan	trp	W
Tyrosine	tyr	Y
Valine	val	V

4.1. Protein structure prediction

Over the years, a large number of techniques that attempt to predict the structure of proteins from their amino acid sequence have been developed. Comparative modelling methods (Altschul *et al.*, 1997; Karplus *et al.*, 1999; Notredame, 2002; Tramontano and Morea, 2003) predict the structure of a protein by comparing its amino acid sequence to other proteins, for which the 3D structure is known. The efficiency of these methods depends on the level of similarity between the two

sequences; if they share a large percentage of their sequence the prediction is likely to be of high quality (Kopp and Schwede, 2004). One advantage of comparative modelling is that its importance will continue to grow as more and more protein structures are identified, and therefore become available for comparison.

A second class of methods rely on fold recognition to predict the native structure of a protein. The basic idea behind this methodology is the fact that the number of different folds that exist in nature is extremely smaller than the number of different sequences. Fold recognition methods select a model fold for a given sequence among the known folds in protein databases, even when no sequence similarity is detected. Approaches of fold recognition include secondary structure prediction (Jones, 1999a; An and Friesner, 2002; Przybylski and Rost, 2004) and threading (Jones, 1999b; Xu and Xu, 2000; Kim *et al.*, 2003, Skolnick *et al.*, 2004).

Fragment based methods (Aloy *et al.*, 2003; Bradley *et al.*, 2003; Jones and Guffin, 2003; Zhang *et al.*, 2003; Lee *et al.*, 2004; Rohl *et al.*, 2004) compare short amino acid subsequences of the target protein to fragments of known proteins. When suitable fragments are identified, they are randomly put together to form a structure. The final conformation is selected with the use of scoring functions and optimisation algorithms.

First principles methods base predictions on physical models of the mechanisms and driving forces of protein folding (Pillardy *et al.*, 2001; Lee *et al.*, 2001; Liwo *et al.*, 2002; Czaplewski *et al.*, 2004a). Techniques belonging in this category attempt to identify the minimum of a free energy function for the structure of the protein. Such methods are computationally demanding, but they remain important approaches for prediction of protein folding, because in many cases even remotely related structural homologues for the knowledge-based methods are not available. Floudas and coworkers (Klepeis and Floudas, 2003a) introduced a first principles method that identifies helical regions through detailed free energy calculations and the application of global optimisation methodologies (Klepeis and Floudas, 2002). β-strands are solved using an MILP formulation to maximise hydrophobic interactions (Klepeis and Floudas, 2003b). Finally, tertiary structure is identified through hybrid global optimisation algorithms (Klepeis *et al.*, 2003a; Klepeis *et al.*, 2003b).

4.1.1. Lattice models

The protein folding problem is NP-hard (Berger and Leighton, 1998). This fact, combined with the ability of amino acids, as components of a protein, to actually fold to a virtually countless number of conformations, even further complicates the development of an efficient algorithm for the solution of the problem. Therefore, theoretical intervention is required through the use of simplified models for protein folding. Only some aspects of protein structure are modelled; for this reason these models are also known as low-resolution models.

Lattice models constitute the most important category of simplified models. Proteins are represented as self-avoiding random walks on a lattice where vertices indicate the possible positions of the amino acids. Lattice models lack atomic detail, but contain the fundamental microscopic attributes of proteins, like linear connectivity, excluded volume, chain flexibility and sequence dependent intra-chain interactions. The most stable structure is usually the one with the minimum energy (thus the scope for application of optimisation techniques). The apparent limitations of lattice proteins are the artificial discrete degrees of freedom and the short range of the interactions.

A general discussion of lattice models can be found in Dill *et al.* (1995). The following assumptions (Backofen and Will, 2003) are made in a lattice model: i) amino acids all have the same size; ii) bonds all have the same length; iii) the positions of the amino acids are restricted to lattice positions; and iv) a simplified energy function is applied. A number of lattice model methodologies have been suggested in the literature, with varying characteristics. Important aspects of lattice model methods are the kind of lattice that they employ to approximate real protein conformations, the energy function that is used to discriminate native from non-native structures, and the methodology for searching the space of possible configurations for identifying the minimal energy conformation (Backofen and Will, 2003). Examples for each of these three aspects are given below.

The first important decision is the kind of lattice that is going to be used. Two-dimensional approaches use a square lattice (Crippen, 1991; Fast and Istrail, 1996; Cui *et al.*, 2002). In 3D space, most commonly used is the cubic lattice (Shakhnovich

et al., 1990; Sali *et al.*, 1994a; Dinner *et al.*, 1996; Kussell *et al.*, 2003; Broglia *et al.*, 2004; Tiana *et al.*, 2004), but other, more complex lattice models have been proposed, *e.g.* the face-centred-cubic (FCC) lattice (Park and Levitt, 1995; Backofen and Will, 2003). While highly complex lattices can be used to closely approximate real proteins, square and cubic lattices are typically preferred to study basic principles of protein structure. An example of a cubic lattice is presented in Figure 4.1.

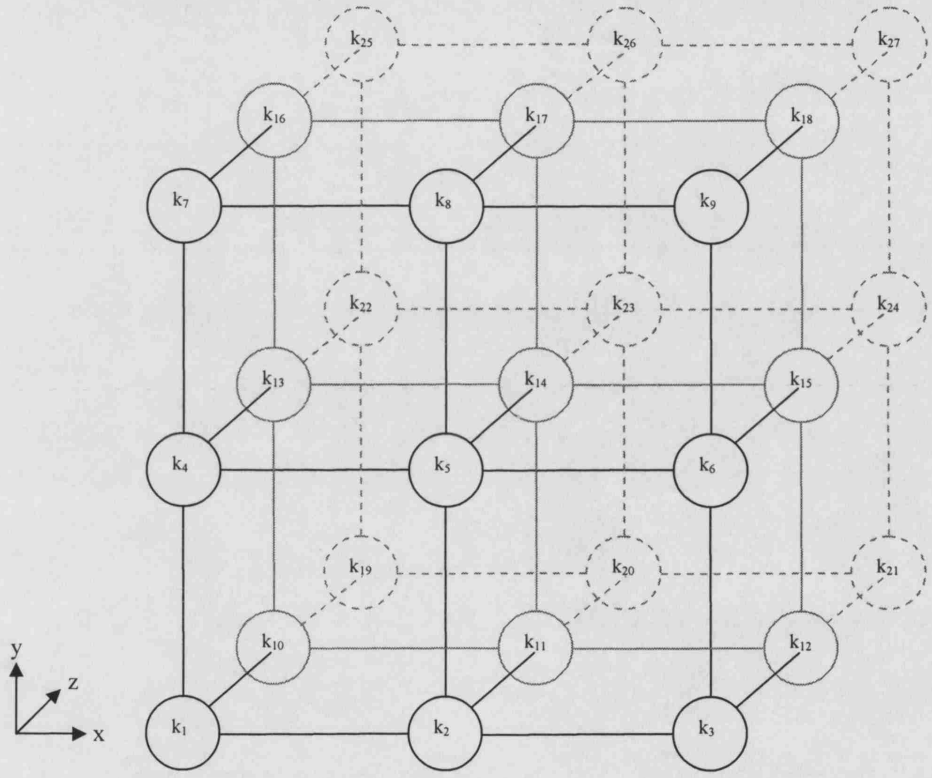

Figure 4.1: Example of a 3×3×3 cubic lattice.

The selection of an energy function is an equally important aspect. One of the most important representatives is the HP model (Yue and Dill, 1995; Backofen *et al.*, 1999, Cui *et al.*, 2002; Backofen and Will, 2003), first introduced by Lau and Dill (1989). The 20 amino acids are reduced to a two letter alphabet: H, representing hydrophobic amino acids and P, representing polar (hydrophilic) amino acids. The contribution of a contact to the energy function is -1 if both amino acids are hydrophobic and 0 otherwise. In order for a contact to be formed, two amino acids

have to be in neighbouring positions in the lattice and they must not be connected through a bond (*i.e.* they are not sequential in the amino acid sequence). The HP model has the problem that its degeneracy is large. For this reason, extended models, such as the HPNX-model* (Bornberg-Bauer, 1997) and the Go model (Ueda *et al.*, 1975; Dill and Chan, 1997; Du *et al.*, 1998) have been introduced. Other models use energy parameters derived from the random energy model (Kussell *et al.*, 2003) or from experimentally determined potentials, such as the Miyazawa-Jernigan (1985) potential. The results of these models are energetically comparable to real proteins and display a more realistic folding behaviour (Sali *et al.*, 1994a; Sali *et al.*, 1994b; Abkevich *et al.*, 1995; Dinner *et al.*, 1996; Unger and Moult, 1996; Govindarajan and Goldstein, 1997; Broglia and Tiana, 2001a; Broglia *et al.*, 2004).

Finally, a choice is needed for a technique that searches for the minimal energy conformation. Most examples from the literature use heuristic methods, ranging from Monte Carlo simulations (Dinner *et al.*, 1996; Broglia *et al.*, 1998), to Simulated Annealing (Kirkpatrick *et al.*, 1983; Brower *et al.*, 1993; MacDonald *et al.*, 2000), to genetic algorithms (Unger and Moult, 1996), purely heuristic methods (Dill *et al.*, 1993; Bornberg-Bauer, 1997), and even complete enumeration (Sali *et al.*, 1994a; Xia *et al.*, 2000).

Optimisation techniques are not yet widely employed for the identification of the structure with the minimum free energy, but have been applied before in the field of protein folding (Klepeis and Floudas, 2003a; Wagner *et al.*, 2004). Also, constraint programming has been employed for the solution of the protein structure prediction problem (Backofen, 1998; Backofen *et al.*, 1999; Krippahl and Barahona, 1999; Backofen and Will, 2003; Palu *et al.*, 2004) by applying a global search technique with an HP model.

In this chapter, a lattice model approach is presented that utilises a cubic lattice and an energy function based on a 20x20 energy matrix of experimentally determined potentials (in dimensionless *RT* units) for each pair of amino acids (*i.e.* a 20-letter alphabet of amino acids, compared to the 2-letter one applied in the HP model). The

* HPNX: Hydrophobic, Positive, Negative, Neutral

problem is formulated as an MILP model and, using the proposed model, a three-step solution strategy (Broglia *et al.*, 2004) is applied, which reads the 3D configuration (*i.e.* specifies the coordinates of all monomers in the native conformation) of lattice model-designed proteins from only the knowledge of their amino acid sequence and of the contact energies (Miyazawa and Jernigan, 1985; see Table 4.2) among the amino acids. In all following discussions, energies are represented in dimensionless *RT* units, where *R* is the gas constant and *T* is the absolute temperature.

Table 4.2: Contact energies among amino acids in RT units (from Miyazawa and Jernigan, 1985).

	cys																			
cys	-1.06	**met**																		
met	0.19	0.04	**phe**																	
phe	-0.23	-0.42	-0.44	**ile**																
ile	0.16	-0.28	-0.19	-0.22	**leu**															
leu	-0.08	-0.20	-0.30	-0.41	-0.27	**val**														
val	0.06	-0.14	-0.22	-0.25	-0.29	-0.29	**trp**													
trp	0.08	-0.67	-0.16	0.02	-0.09	-0.07	-0.12	**tyr**												
tyr	0.04	-0.13	0.00	0.11	0.24	0.02	-0.04	-0.06	**ala**											
ala	0.00	0.25	0.03	-0.22	-0.01	-0.10	-0.09	0.09	-0.13	**gly**										
gly	-0.08	0.19	0.38	0.25	0.23	0.16	0.18	0.14	-0.07	-0.38	**thr**									
thr	0.19	0.19	0.31	0.14	0.20	0.25	0.22	0.13	-0.09	-0.26	0.03	**ser**								
ser	-0.02	0.14	0.29	0.21	0.25	0.18	0.34	0.09	-0.06	-0.16	-0.08	-0.20	**gln**							
gln	0.05	0.46	0.49	0.36	0.26	0 24	0.08	-0.20	0.08	-0.06	-0.14	-0.14	0.29	**asn**						
asn	0.13	0.08	0.18	0.53	0.30	0.50	0.06	-0.20	0.28	-0.14	-0.11	-0.14	-0.25	-0.53	**glu**					
glu	0.69	0.44	0.27	0.35	0.43	0.34	0.29	-0.10	0.26	0.25	0.00	-0.26	-0.17	-0.32	-0.03	**asp**				
asp	0.03	0.65	0.39	0.59	0.67	0.58	0.24	0.00	0.12	-0.22	-0.29	-0.31	-0.17	-0.30	-0.15	0.04	**his**			
his	-0.19	0.99	-0.16	0.49	0.16	0.19	-0.12	-0.34	0.34	0.20	-0.19	-0.05	-0.02	-0.24	-0.45	-0.39	-0.29	**arg**		
arg	0.24	0.31	0.41	0.42	0.35	0.30	-0.16	-0.25	0.43	-0.04	-0.35	0.17	-0.52	-0.14	-0.74	-0.72	-0.12	0.11	**lys**	
lys	0.71	0.00	0.44	0.36	0.19	0.44	0.22	-0.21	0.14	0.11	-0.09	-0.13	-0.38	-0.33	-0.97	-0.76	0.22	0.75	0.25	**pro**
pro	0.00	-0.34	0.20	0.25	0.42	0.09	-0.28	-0.33	0.10	-0.11	-0.07	0.01	-0.42	-0.18	-0.10	0.04	-0.21	-0.38	0.11	0.26

4.2. Problem statement

The folding of a protein is determined by the native conformation being at an energy minimum (Shakhnovich and Gutin, 1993). Studies of protein aggregation suggest that proteins fold by forming partially folded intermediates (elementary structures),

which are controlled by local contacts among the most strongly interacting amino acids (Broglia *et al.*, 1998). These structures are formed early in the process, and determine the final folding conformation of the protein (Tiana and Broglia, 2001). Local contacts that are generally characterised as elementary structures are fast forming and stable bonds built by the most strongly interacting amino acids, while the structure of the rest of the amino acids in the sequence is formed later in the folding process. Elementary structures form a nucleus around which the protein's final structure is based.

Based on the above analysis of the way proteins fold, a three-step strategy (Broglia *et al.*, 2004) is adopted for the solution of the problem of protein structure prediction. The problem can be stated as follows:

Given:

- the amino acid sequence of a protein;
- the 20x20 contact energy matrix describing the interaction among the amino acids (Table 4.2);
- a finite cubic lattice.

Determine:

1. the possible elementary structures;
2. the possible folding nucleus, by allowing the elementary structures to interact amongst them;
3. the position s of the remaining amino acids in the sequence.

So as to identify the compact structure that displays the minimum energy, as defined by the summation of all the contact energies of the amino acids that are "in contact". A contact between two amino acids is considered to exist when they rest on vertices that are connected by a single lattice edge. Contacts are not considered for amino

acids that are sequential. The structure with the minimum energy must be the native conformation.

4.3. Mathematical formulation

The MILP model presented herein can be used to determine the 3D structure of a protein, by fixing its amino acids in a predetermined cubical lattice. Amino acids are viewed as spheres that assume rigid positions on the vertices of the lattice. First, the indices and parameters associated with the problem are listed.

4.3.1. Nomenclature

Indices

i,j	amino acids
k	position in lattice (vertex)

Sets

I_{fixed}	initially fixed amino acid
IJ_{fixed}	identified pairs of amino acids (i,j) that are in contact
K_{fixed}	lattice position where amino acids belonging to I_{fixed} are fixed

Parameters

CE_{ij}	contact energy between amino acids i and j
$\hat{E}_{ij}$	1 if the pair of amino acids i and j belongs to IJ_{fixed}; 0 otherwise
M	large number
N	total number of amino acids i in the sequence
$\hat{X}_k, \hat{Y}_k, \hat{Z}_k$	coordinates of vertex k

Positive continuous variables

X_i, Y_i, Z_i	coordinates of amino acid i
L_{ij}	relative distance in x direction from amino acid i to j, if i is left of j
R_{ij}	relative distance in x direction from amino acid i to j, if i is right of j

B_{ij} relative distance in y direction from amino acid i to j, if i is below j

A_{ij} relative distance in y direction from amino acid i to j, if i is above j

U_{ij} relative distance in z direction from amino acid i to j, if i is higher than j

D_{ij} relative distance in z direction from amino acid i to j, if i is lower than j

Binary variables

E_{ij} 1 if amino acid i is in contact with amino acid j; 0 otherwise

W_{ik} 1 if amino acid i is assigned to vertex k; 0 otherwise

4.3.2. Model constraints

4.3.2.1. Allocation constraints

Each amino acid has to be assigned to one unique position in the lattice. This is accomplished with equations (4.1)-(4.5). Sets I_{fixed} and K_{fixed} are used for excluding any amino acids for which their position in the lattice is fixed beforehand; usually this is one amino acid that is arbitrarily placed at a certain point in the lattice, so as to exclude symmetries in the search for the optimal conformation.

$$X_i = \sum_{k \notin K_{fixed}} \hat{X}_k \cdot W_{ik} \qquad \forall\, i \notin I_{fixed} \tag{4.1}$$

$$Y_i = \sum_{k \notin K_{fixed}} \hat{Y}_k \cdot W_{ik} \qquad \forall\, i \notin I_{fixed} \tag{4.2}$$

$$Z_i = \sum_{k \notin K_{fixed}} \hat{Z}_k \cdot W_{ik} \qquad \forall\, i \notin I_{fixed} \tag{4.3}$$

$$\sum_{k \notin K_{fixed}} W_{ik} = 1 \qquad \forall\, i \notin I_{fixed} \tag{4.4}$$

$$\sum_{i \notin I_{fixed}} W_{ik} \leq 1 \qquad \forall\, k \notin K_{fixed} \tag{4.5}$$

Binary variable W_{ik} decides whether amino acid i is allocated to position k. The coordinates $(\hat{X}_k, \hat{Y}_k, \hat{Z}_k)$ of vertex k are assigned to variables X_i, Y_i, Z_i with constraints (4.1), (4.2), (4.3), respectively. Equation (4.4) specifies that each amino

acid i must be assigned to exactly one position on the lattice. Equation (4.5) indicates that each position k must either be left empty or be occupied from at most one amino acid.

4.3.2.2. Continuity constraints

When $j = i + 1$, the two amino acids i and j are next to each other in the sequence. The amino acid chain (and therefore the identity) of the protein must be maintained; to this end relative distances are considered for sequential amino acids. Relative distance constraints are written for pairs of amino acids (i,j) and calculate their relative distance in each of the three dimensions.

$$L_{ij} - R_{ij} = X_i - X_j \qquad \forall\, i,j : j = i+1 \tag{4.6}$$

$$B_{ij} - A_{ij} = Y_i - Y_j \qquad \forall\, i,j : j = i+1 \tag{4.7}$$

$$D_{ij} - U_{ij} = Z_i - Z_j \qquad \forall\, i,j : j = i+1 \tag{4.8}$$

The total rectilinear distance between two sequential amino acids on the chain must be equal to one; this is accomplished with the following constraint:

$$L_{ij} + R_{ij} + B_{ij} + A_{ij} + D_{ij} + U_{ij} = 1 \qquad \forall\, i,j : (j = i+1) \vee IJ_{fixed} \tag{4.9}$$

Set IJ_{fixed} incorporates pairs of amino acids for which we already know that they are in contact; *i.e.* pairs of amino acids that have been identified as being in neighbouring positions in a previous step of the analysis.

4.3.2.3. Contact constraints

Relative distance constraints are also conditionally written for the rest of the amino acids in the sequence:

$$L_{ij} - R_{ij} = X_i - X_j \qquad \forall\, i,j : j \geq i+3; j - i \equiv odd; CE_{ij} < 0 \tag{4.10}$$

$$B_{ij} - A_{ij} = Y_i - Y_j \qquad \forall\, i,j : j \geq i+3; j - i \equiv odd; CE_{ij} < 0 \tag{4.11}$$

$$D_{ij} - U_{ij} = Z_i - Z_j \qquad \forall\, i,j : j \geq i+3; j-i \equiv odd; CE_{ij} < 0 \qquad (4.12)$$

We need to take a moment and explain the conditions imposed on these equations, as they appear in many of the constraints of this chapter. First, condition $j \geq i+3$ accomplishes two objectives: it guarantees that amino acid pairs appear only once (*e.g.* 1-4, but not 4-1); and additionally, it eliminates unnecessary variables, for example when $j = i+2$ there is no possible way for the two amino acids to fall in neighbouring positions in the lattice.

The second condition ($j - i \equiv odd$) is imposed because only when it holds true is it physically possible for two amino acids to fall in neighbouring positions in the lattice, *e.g.* the first amino acid in the sequence can be in contact with the fourth, sixth, eighth, etc. amino acids, but not with the fifth, seventh, ninth, etc.

Condition $CE_{ij} < 0$ is imposed in order to reduce the size of the model. Contact energies CE_{ij} are used for the calculation of the objective function. Because the energy function (4.17) is minimised, when $CE_{ij} < 0$ only one variable at most in each couple (L_{ij} and R_{ij} ; B_{ij} and A_{ij} ; U_{ij} and D_{ij}) is guaranteed to be non-zero at the optimal solution. If we also considered contacts with $CE_{ij} > 0$, additional constraints would be required to guarantee the correct estimation of the relative distances. Theoretically, contacts between amino acids with positive contact energy ($CE_{ij} > 0$) *are* possible, but in practice these contacts are so rare that the predicted structure of the protein is not influenced in the final result if we do not consider them.

Contact constraints identify which amino acids are in contact with each other. Only amino acids that are separated by an odd number of steps on the chain are considered, as explained above. Binary variable E_{ij} must be equal to 1 if amino acids i and j are in contact and 0 otherwise.

$$L_{ij} + R_{ij} + B_{ij} + A_{ij} + D_{ij} + U_{ij} \leq 1 + M \cdot (1 - E_{ij})$$
$$\forall\, i,j \notin IJ_{fixed} : j \geq i+3; j-i \equiv odd; CE_{ij} < 0 \qquad (4.13)$$

$$L_{ij} + R_{ij} + B_{ij} + A_{ij} + D_{ij} + U_{ij} \geq 2 - E_{ij}$$
$$\forall\, i,j \notin IJ_{fixed} : j \geq i+3; j-i \equiv odd; CE_{ij} < 0 \qquad (4.14)$$

Amino acids i and j are in contact when the sum of the relative distances in all directions is equal to 1. Constraint (4.13) forces E_{ij} to 0 when the sum of rectilinear distances is greater than 1, because the objective function is minimised. If the sum is equal to 1, constraint (4.14) forces E_{ij} to the value of 1.

4.3.2.4. Logical constraints

The nature of the lattice model only allows up to 6 neighbouring positions for each vertex and, because two of those positions have to be occupied by the next and the previous amino acid in the sequence, at most 4 contacts in total per amino acid are possible. The 1st and the N^{th} amino acids in the chain constitute an obvious exception; for them 5 contacts are possible.

$$\sum_{\substack{j\notin IJ_{fixed}\\ i\geq j+3\\ i-j\equiv odd\\ CE_{ij}<0}} E_{ji} + \sum_{\substack{j\notin IJ_{fixed}\\ j\geq i+3\\ j-i\equiv odd\\ CE_{ij}<0}} E_{ij} \leq 4 \qquad \forall\, i = 2,\ldots,N-1 \tag{4.15}$$

$$\sum_{\substack{j\notin IJ_{fixed}\\ i\geq j+3\\ i-j\equiv odd\\ CE_{ij}<0}} E_{ji} + \sum_{\substack{j\notin IJ_{fixed}\\ j\geq i+3\\ j-i\equiv odd\\ CE_{ij}<0}} E_{ij} \leq 5 \qquad i = 1 \vee i = N \tag{4.16}$$

4.3.3. Model summary

The overall objective function used is the sum of all the contact energies of the pairs of amino acids that are in contact with each other. Note that the second term of the objective function is fixed and representes the energy contributions from the pairs of amino acids belonging to set IJ_{fixed}. The entire model formulation described in section 4.3.2 can be summarised as the following MILP problem (problem S):

Problem S

$$\textit{minimise} \quad \sum_{i} \sum_{\substack{j\notin IJ_{fixed}\\ j\geq i+3\\ j-i\equiv odd\\ CE_{ij}<0}} CE_{ij}\cdot E_{ij} + \sum_{i} \sum_{j\in IJ_{fixed}} CE_{ij}\cdot \hat{E}_{ij} \tag{4.17}$$

subject to:

allocation constraints (4.1), (4.2), (4.3),(4 .4) and (4.5);

continuity constraints (4.6), (4.7), (4.8) and (4.9);

contact constraints (4.10), (4.11), (4.12), (4.13) and (4.14);

logical constraints (4.15) and (4.16);

$$E_{ij}, W_{ik} \in \{0,1\} \qquad \forall i \notin I_{fixed} \qquad (4.18)$$

$$X_i, Y_i, Z_i, R_{ij}, L_{ij}, A_{ij}, B_{ij}, U_{ij}, D_{ij} \geq 0 \qquad \forall i, j \qquad (4.19)$$

4.4. Solution procedure

A three-step solution strategy for reading the 3D structure of lattice-designed proteins from the knowledge of only their amino acid sequence and the contact energy matrix among the amino acids (Table 4.2) is implemented (Broglia *et al.*, 2004). The proposed methodology first looks for small elementary structures in the amino acid sequence and then creates a folding core from combining these structures. The final 3D conformation of the protein is identified by positioning the remaining amino acids around the nucleus. Each step of the solution strategy is presented in detail below.

4.4.1. Step 1: Search for elementary structures

The folding process has been found to follow a hierarchical sequence of events (Tiana and Broglia, 2001). The first stage of the procedure is to identify possible elementary structures, which are formed by a small number of residues relatively close to each otheralong the chain (2 to 10 monomers apart); these structures are created very quickly and are remarkably stable. The probability, P_{ij}, that the residue couple between the i^{th} and the j^{th} amino acid of the chain will bind together, depends essentially only on their distance and the contact energy of the pair, CE_{ij}, and can be fitted by the following function (Tiana and Broglia, 2001):

$$P_{ij} = (j-i)^{-1.68} \cdot e^{-\frac{CE_{ij}}{T_{eff}}} \qquad \forall\, i,j : i+3 \le j \le i+9; (j-i) \equiv odd \qquad (4.20)$$

where T_{eff} is an effective temperature set equal to the standard deviation of the interaction matrix. For the case of the contact energy matrix used here (Table 4.2; Miyazawa and Jernigan, 1985) the standard deviation σ is equal to 0.305. The elementary structures are selected among those pairs that display the highest values of probability, P_{ij}. Pairs of amino acids that must be in contact in order for the identified elementary structures to form properly are added to set IJ_{fixed}, which will be used in the next step of the process.

4.4.2. Step 2: Search for the folding core

We search for a folding nucleus that results from the docking of the elementary structures identified in the previous step. A cubic lattice is used, the size of which depends on the size of the protein under investigation, but its dimensions are selected to be as close to a cubic lattice as possible. For instance, for the two examples that are presented later in this chapter, a 3×3×3 cubic lattice is applied for a 27mer protein (see section 4.5), and in the case of a 36mer protein a 4×3×3 lattice is used. The three-dimensional conformation of the structural nucleus is identified by implementing problem S in order to optimally position the elementary structures relative to each other.

Only the amino acids belonging to the elementary structures are considered. One amino acid (usually an amino acid belonging to the pair with the highest binding probability (P_{ij}) from the previous step) is fixed to a randomly selected position (for practical reasons, the centre of the lattice, or a position as close to the centre as possible is selected). The fixing of the position of one amino acid drastically improves the performance of the model, because it eliminates some of the very large number of symmetrical solutions that exist. The fixed amino acid is the only one that belongs to set I_{fixed} and the lattice position where it is fixed belongs to set K_{fixed}. Set IJ_{fixed} incorporates pairs of amino acids that are known to be in contact, as identified in step 1. The rest of the amino acids in the elementary structures are allowed to take positions with the condition that the elementary structures stay intact. The potential

folding core is selected as the one that minimises the total energy of the structure. All pairs of amino acids that are in contact in the folding core are added to set IJ_{fixed} for step 3.

4.4.3. Step 3: Position remaining amino acids

To determine the final structure with the minimum energy, as with the folding core before, problem S is applied. Using the nucleus as a basis, the rest of the amino acids are placed around it in search of a conformation of the system that minimises its energy. Again, only one amino acid is fixed, selected from those belonging to the core. For reasons of eliminating symmetrical solutions from the search space, we pick the amino acid with the most identified contacts in the nucleus (*i.e.* amino acid i that is in contact with the most amino acids j, where $(i,j) \in IJ_{fixed}$). If there is more than one amino acid with the same number of identified contacts, then we pick the amino acid with the lowest sum of contact energies CE_{ij}, where $(i,j) \in IJ_{fixed}$. This is the only amino acid belonging to set I_{fixed}. The selected amino acid has to be fixed to an appropriate position in the lattice (the same lattice as the one used in step 2), which belongs to set K_{fixed}. A small subset of the available lattice positions need to be tested, because most positions are symmetrical. The rest of the amino acids that belong to the folding core are not fixed to a position, but they are forced to maintain the contacts that they are given by the solution of step 2. These constitute the pairs of amino acids belonging to set IJ_{fixed}.

4.5. Computational results

The applicability of the procedure presented in the previous section is demonstrated here with the help of two proteins: a 27mer and a 36mer peptide. Their sequences are presented in Table 4.3, using abbreviations for the names of amino acids (the full names can be found in Table 4.1). The two illustrative examples will serve to demonstrate and clarify the three-step solution strategy presented in section 4.4. The

sequences of Table 4.3 have been designed to fold fast into a unique 3D conformation.

Table 4.3: Peptide sequences of the two examples.

Protein	Lattice	Amino acid sequence
27mer	Cubic, 3×3×3	QFPHLKAPLVAILGMVCWANGIYTSRD[a]
36mer	Cubic, 4×3×3	SQKWLERGATRIADGDLPVNGTYFSCKIMENVHPLA[b]

[a] Abkevich *et al.*, 1994; Sali *et al.*, 1994a and 1994b

[b] Abkevich *et al.*, 1994; Klimov and Thirumalai, 1996

Both examples were implemented in GAMS (Brooke *et al.*, 1998), using the CPLEX 6.5 LP solver with a 5% margin of optimality. Solutions were obtained by running the models on an IBM RS6000 workstation with a maximum computational time limit of 10,000 seconds.

4.5.1. Example 1

4.5.1.1. Step 1: Elementary structures

The methodology is first applied to the 27mer sequence. The values of probability P_{ij} are calculated in order to predict the formation of possible elementary structures. It should be noted that the numbers in the figures and anywhere else below this point refer to the order in which an amino acid appears in the sequence of the protein under examination, for example **2** corresponds to Phenylalanine (F) and **5** corresponds to Leucine (L) for the 27mer sequence (Table 4.3).

Figure 4.2 displays the results of P_{ij} calculations using equation (4.20), associated with each couple of monomers. Two bonds have a value of P_{ij} much larger than all the other values and thus qualify as good candidates for elementary structures. Therefore, the elementary structures are selected with the condition that monomers **2–5** and **15–18** form the contacts of Figure 4.3 and the two pairs are added to set IJ_{fixed}.

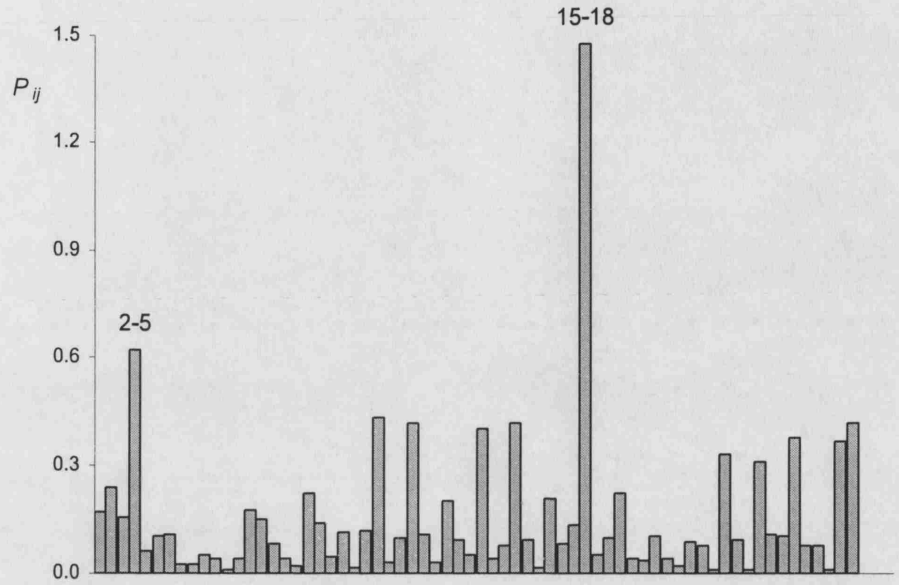

Figure 4.2: Binding probability for the 27mer. Pairs that demonstrate high binding probability form between amino acids 2-5 and 15-18.

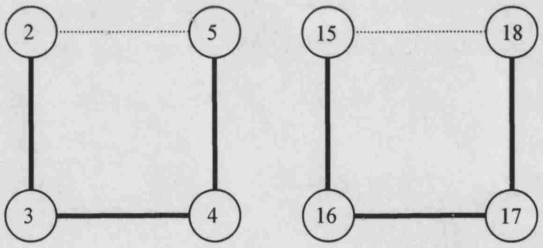

Figure 4.3: Elementary structures for 27mer sequence.

4.5.1.2. Step 2: Folding nucleus

The elementary structures of the 27mer protein are combined in a conformation that displays the minimum possible energy. Amino acid 15 (selected because it belongs to the pair with the highest binding probability, but similar results are produced with amino acid 18) is fixed to the centre of the lattice. The minimum energy conformation of the folding nucleus is E_c = -2.15. The result, generated by solving problem S for the amino acids belonging to the elementary structures only, is presented in Figure 4.4.

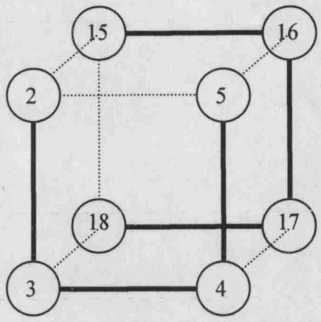

Figure 4.4: Folding core for 27mer sequence.

All the contacts of the amino acids of the nucleus are maintained by incorporating pairs **2–15**, **3–18**, **4–17** and **5–16** in set IJ_{fixed} (which already included pairs **2–5** and **15–18** from step 1) for the next step of the procedure.

4.5.1.3. Step 3: Native structure

In a cubic lattice, a large number of the available positions are symmetrical, increasing in this way the search space for the native conformation of a protein. To overcome part of this problem, one amino acid needs to be fixed at a vertex of the lattice. Figure 4.5 shows the four different vertices (3 black and one grey) we need to test as possible positions for the fixed amino acid; all other vertices in the lattice are symmetrical to these.

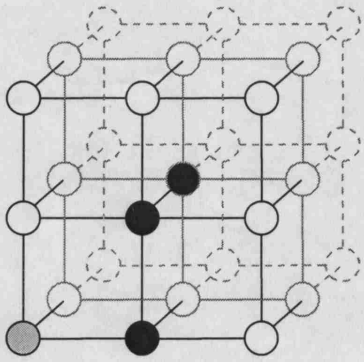

Figure 4.5: Eliminating symmetry in a 3×3×3 cubic lattice. All white vertices are symmetrical to the 4 highlighted ones. The grey vertex only has 3 neighbouring positions available, and is therefore unsuitable for fixing amino acid 15.

We select to fix the amino acid with the most contacts identified in the nucleus, because this tactic even further reduces the search space of the problem. The amino acids in the nucleus with the most identified contacts are 2, 5, 15 and 18 (two contacts each). From these, 15 has the contacts with the lowest sum of contact energies and is therefore selected. Amino acid 15 has two identified contacts in the nucleus, and also has two more adjacent amino acids (14 and 16). For this reason, it cannot be placed at the grey vertex of Figure 4.5, leaving only three candidate positions for amino acid 15.

After solving problem S for all three positions, the structure with the minimum possible energy (E_c = -8.8) is identified. Solution statistics are presented in Table 4.4. The 27mer lattice-designed protein (see Table 4.3) folds into the native conformation presented in Figure 4.6. The grey vertices in the figure specify where the folding nucleus lies. The predicted structure is identical to the one presented in the literature for this 27mer protein (Abkevich *et al.*, 1994; Sali *et al.*, 1994a and 1994b).

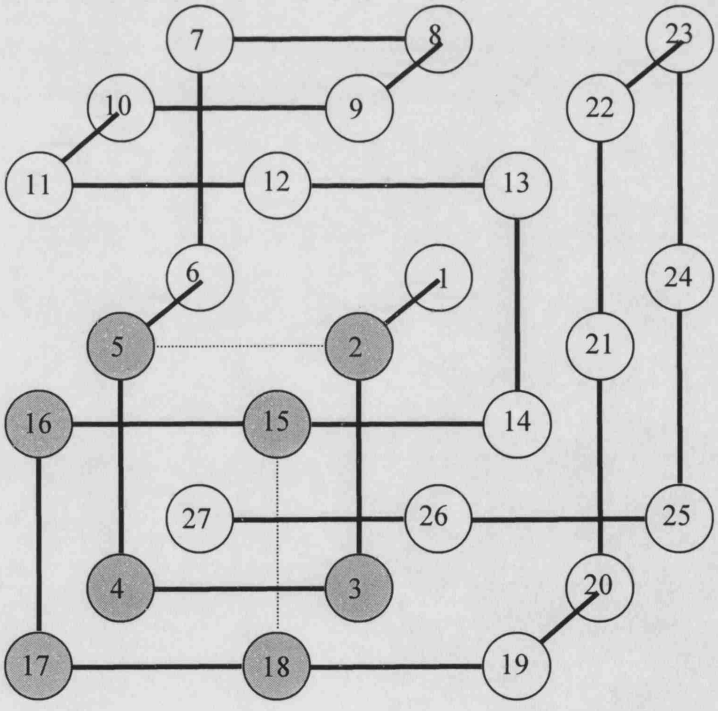

Figure 4.6: Native folding conformation for 27mer lattice-designed sequence. The folding core is highlighted in grey.

4.5.2. Example 2

4.5.2.1. Step 1: Elementary structures

Next, we apply the three-step strategy to a larger protein with 36 amino acids. Figure 4.7 displays the results of P_{ij} calculations using equation (4.20), associated with each couple of monomers to predict the formation of possible elementary structures. The following bonds have a value of P_{ij} much larger than all the other values and thus qualify as good candidates for elementary structures: **3–6**, **11–14** and **27–30**. The elementary structures are selected with the condition that these monomers form the contacts of Figure 4.8 and pairs **3–6**, **11–14** and **27–30** are added to set IJ_{fixed}.

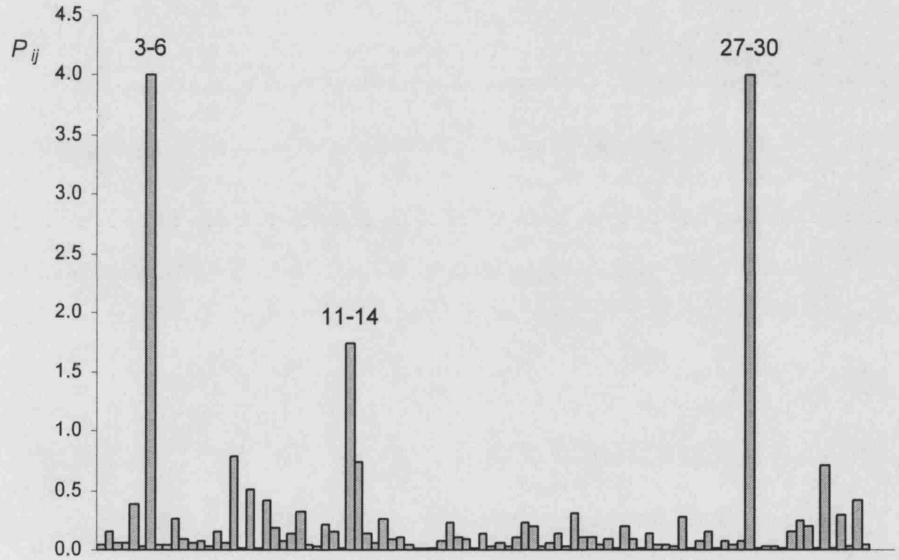

Figure 4.7: Binding probability for the 36mer. Pairs that demonstrate high binding probability form between amino acids 3-6, 11-14 and 27-30.

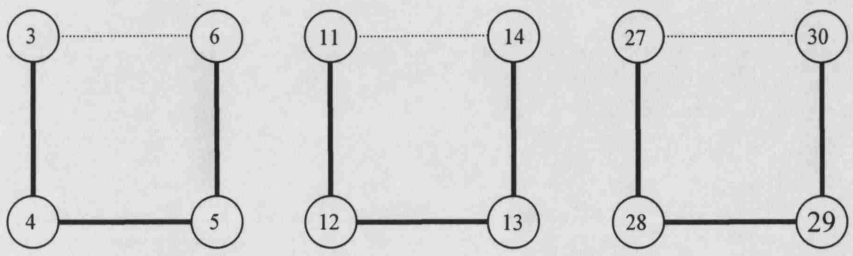

Figure 4.8: Elementary structures for 36mer sequence.

4.5.2.2. Step 2: Folding nucleus

The elementary structures of Figure 4.8 are combined in a conformation that displays the minimum possible energy to form the core of the folding procedure. Amino acid 3 is arbitrarily selected as the one that will be fixed to a position as close to the centre of the lattice as possible. Similar results can be produced with the selection of amino acids 6, 27 or 30, as all of these have the same binding probability. The minimum energy conformation of the folding nucleus for the 36mer protein is E_c = -7.81. Figure 4.9 presents the suggested core from the solution of problem S. The result is generated by only considering the identified elementary structures. Pairs **3–6**, **3–30**, **4–29**, **5–12**, **5–28**, **6–11**, **6–27**, **11–14**, **13–28**, **14–27** and **27–30** are included in set IJ_{fixed} for step 3.

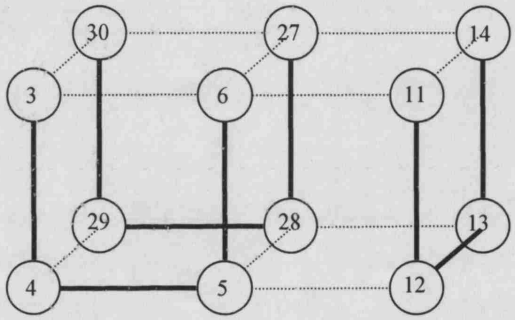

Figure 4.9: Folding core for 36mer sequence.

4.5.2.3. Step 3: Native structure

The amino acids of the nucleus with the most identified contacts are 6 and 27. From these, amino acid 6 has the contacts with the lowest sum of contact energies and is therefore selected as the initially fixed amino acid for the third step of the strategy. In a 4×3×3 cubic lattice, there are six different positions (3 black and 3 grey) we need to examine for fixing an amino acid; all other vertices in the lattice are symmetrical to these. Because amino acid 6 already has three contacts from the nucleus, and also has two more adjacent amino acids (5 and 7) it is only possible to place it at the three black vertices of Figure 4.10.

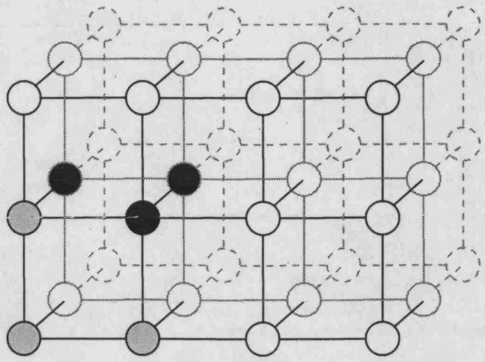

Figure 4.10: Eliminating symmetry in a 4×3×3 cubic lattice. All white vertices are symmetrical to the 6 highlighted ones. The grey vertices all have less than 5 neighbouring positions available, and are therefore unsuitable for fixing amino acid 6.

The structure with the minimum possible energy (E_c = -15.74) is identified after solving problem S for all three available positions of amino acid 6. The optimal solution is presented in Figure 4.11.

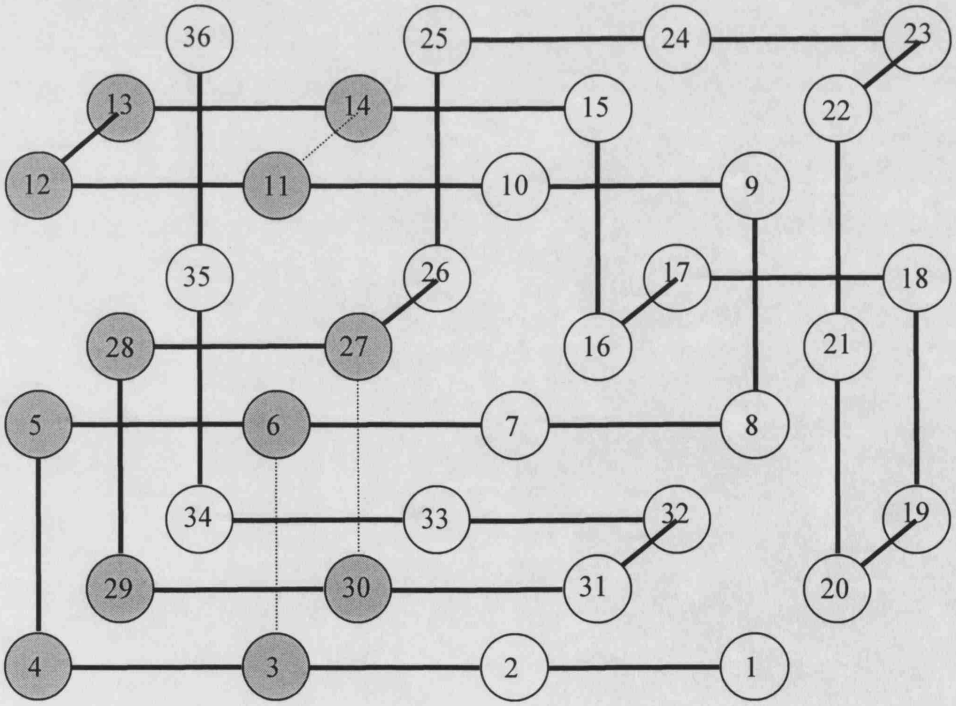

Figure 4.11: Native folding conformation for 36mer lattice-designed sequence. The folding core is highlighted in grey.

The grey vertices in Figure 4.11 specify the position of the folding nucleus. The predicted structure is identical to the one presented in the literature for the 36mer protein (Abkevich *et al.*, 1994; Klimov and Thirumalai, 1996; Broglia and Tiana, 2001a).

Table 4.4 presents computational statistics for the optimisation of the two presented examples. Model sizes in terms of discrete and continuous variables and constraints are given. Also shown are CPU times and objective values. Notice that the optimal solution for the 36mer sequence has an objective value of -16.12, which is different from the minimum energy of the native conformation reported above (E_c = -15.74). This is because the MILP model does not consider pairs with positive contact energies in order to reduce the size of the problem. In the final predicted structure of the protein, one such pair exists ($CE_{15,24}$ = 0.38) and a correction of the minimum energy is required, which however does not influence the accuracy of the predicted structure. Also, for the last computation, the solution after 10,000 seconds is presented. The model does in fact identify an optimal solution if it is allowed to run for longer (E_c = -13.52 after 91,700 seconds), but the energy is still higher than the first attempt for the 36mer protein (amino acid 6 fixed at k_6).

Table 4.4: Summary of computational statistics.

protein	step	fixed aa	lattice position	discrete/continuous variables	constraints	nodes	CPU time (s)	obj. value (*RT* units)
27mer	2	15	k_{14}	148/109	144	419	4.3	-2.15
27mer	3	15	k_2	324/682	626	-	0.2	infeasible
27mer	3	15	k_5	341/682	626	8325	1801.1	**-8.80**[a]
27mer	3	15	k_{14}	374/682	626	-	0.1	infeasible
36mer	2	6	k_6	291/187	229	11318	191.9	-8.20
36mer	3	6	k_6	596/1201	1076	4558	1423.7	**-16.12**[a]
36mer	3	6	k_{17}	564/1201	1076	10622	2471.9	-12.10
36mer	3	6	k_{18}	648/1201	1076	15510	10000	-12.43[b]

[a] Minimum energy

[b] Terminated after 10,000 seconds

4.6. Conclusions

Protein folding is an area of biology that concentrates a lot of interest, but the size of the problem makes predictions extremely difficult. Despite the large number of research groups that work on the subject, no overwhelmingly successful prediction method has been created as of yet. Lattice models offer a simplification of the complexity of the problem.

Here, an optimisation-based framework for positioning amino acids in unique positions of a three-dimensional cubical lattice has been presented. The framework utilises only the knowledge of the amino acid sequence and contact energies among amino acids, and can easily be extended to consider larger or different kinds of 2D or 3D lattices, and a different set of interactions (*e.g.* the HP model instead of contact energies between all 20 amino acids). Nevertheless, the application of a 20-letter amino acid alphabet considerably reduces problems with solution degeneracy, which is a common problem of other lattice-based prediction methods (especially the HP model). The overall problem was formulated as an MILP model and a three-step solution procedure has been proposed to identify small elementary structures, then a folding core and finally, the 3D structure of a small protein, in order to further simplify the task at hand.

An advantage of the presented methodology is that it utilises optimisation techniques and can predict optimal energy conformations for each step of the strategy. The approach was successfully applied to the two examples of section 4.5 to demonstrate its applicability. It was shown that lattice-designed proteins with sizes of up to 36 monomers can be solved with the proposed solution procedure. In both examples, the predicted conformation was identical to the native structure described in literature.

PART III

CHROMATOGRAPHIC PURIFICATION OF PROTEINS USING PEPTIDE TAGS

Chapter 5

MINLP models for the synthesis of optimal peptide tags and downstream protein processing

The development of systematic methods for the synthesis of downstream protein processing operations has seen growing interest in recent years, as purification is often the most complex and costly stage in biochemical production plants. The objective of the work presented here is to develop a mathematical model based on mixed integer optimisation techniques, which integrates the selection of optimal peptide purification tags into an established framework for the synthesis of protein purification processes. Peptide tags are comparatively short sequences of amino acids fused onto the protein product, capable of reducing the required purification steps. The methodology is illustrated through its application on two example protein mixtures involving up to 13 contaminants and a set of 11 candidate chromatographic steps. The results are indicative of the benefits resulting from the appropriate use of peptide tags in

purification processes and provide a guideline for both optimal tag design and downstream process synthesis.

5.1. Protein purification

Recent advances in biotechnology have given immense impetus to the introduction of biopharmaceutical and biotechnological products, which usually require special techniques and equipment for their production. After fermentation or extraction, a multi-step protocol has to be followed in order to achieve the specified product purity. Downstream processing of proteins is typically a major component of the manufacturing and investment costs in a biochemical production plant (Datar, 1986). The quality of the product is predominantly determined at the purification level, which is therefore regarded as the most significant production stage.

Of all separation methods during the downstream process, chromatographic operations are of major interest in the production of high-value biomolecules. A specified purity level of the target protein product is usually achieved by applying several chromatographic steps. In each of these steps, the protein mixture is split into two streams, one containing the product and one that is discarded. Such flowsheets are usually optimised on a unit per unit basis, thus creating the need for a more systematic synthesis and design procedure for purification steps, which considers the entire process instead of each unit individually.

A number of systematic bioprocess synthesis and design methods have been reported in the literature. Petrides (1994) developed a synthesis procedure, which uses expert knowledge to select unit operations in order to synthesise economically favourable processes. Lienqueo *et al.* (1996) also presented a knowledge based expert system for the selection of separation and purification steps for protein mixtures. The technique was then validated experimentally by its application with model protein mixtures (Lienqueo *et al.*, 1999).

Recently, methodologies based on optimisation techniques have been presented. An implicit enumeration algorithm developed for chemical processes (Fraga, 1998) was extended in order to synthesise optimal bioprocesses in a system-wide sense (Steffens *et al.*, 2000a). The technique incorporated heuristics based on physical property information in an implicit enumeration algorithm to solve the synthesis problem, thus reducing the search space. Evaluation of the technique was performed with two case studies.

An MILP framework using established criteria for modelling chromatographic techniques was presented (Vásquez-Alvarez *et al.*, 2001), in which mathematical models for each technique rely on physicochemical data on the protein mixture that contains the desired product, and provide information on its potential purification. The latter formulation was further improved by exploiting the advantages of convex hull representations (Vásquez-Alvarez and Pinto, 2001; Vásquez-Alvarez and Pinto, 2004) and by also considering the incorporation of product losses and the calculation of the amount of product recovered (Vásquez-Alvarez and Pinto, 2003).

The above methodologies can improve the production of a biotechnological product by optimising the purification sequence on the basis of physicochemical data for the product and the contaminant proteins. However, they do not consider whether any benefit is conferred by modifying these physicochemical properties of the product to enhance the separation, and thereby reduce the number of required downstream purification steps.

5.1.1. Purification tags

Such physicochemical properties modifications can be accomplished through the use of purification peptide tags. Purification tags are short sequences of amino acids that can be fused genetically onto the product protein in order to modify its physicochemical properties, in a way that will eventually enhance its separation from other contaminant proteins. It has recently been demonstrated that considerable improvements in yields and costs of downstream purification processes can be achieved with the use of such tags (Steffens *et al.*, 2000b).

The advantages of purifications tags are well-recognised; nevertheless it is difficult to select the right tag for a specific product protein. Each tag utilises a specific structural protein property to facilitate purification: affinity, charge, attraction to metal chelates, solubility or hydrophobicity. The performance of any particular fusion will depend on the physical properties of the product and contaminant proteins. For example, if the bulk of contaminant proteins have a negative charge it may be beneficial to fuse a series of positively charged amino acids onto the product protein and use cation exchange to purify the mixture.

The size of the purification tag is an important issue to consider. Purification tags range from full enzymes, such as *β-galactosidase*, which can be fused onto protein products and can be usually purified using a specific affinity interaction, to very short amino acid sequences (*e.g.* poly-his tag), for which a particular physicochemical property is exploited to accomplish separation.

The latter case of small peptide tags presents numerous advantages (Terpe, 2003); for example the need of only minor genetic modifications to the protein product, as these are small molecules. They are assumed to have a minimal effect on tertiary structure and biological activity and may not require cleavage for many applications due to their small size. The most commonly used small peptide tags are the polyarginine-tag or arg-tag (5 amino acids), for purification with cation exchange (Sassenfeld, 1984); the polyhistidine-tag or his-tag (6 amino acids), for immobilised metal affinity chromatography, a widely used tagging technique (Hochuli *et al.*, 1987); the FLAG-tag, a small (8 amino acids) hydrophilic tag (Hopp *et al.*, 1988); the Strep-tag (8 amino acids) (Schmidt and Skerra, 1993); the c-myc-tag (11 amino acids), a tag commonly used as a detection system, but rarely applied for purification purposes (Evan *et al.*, 1985); and the S-tag (15 amino acids) (Karpeisky *et al.*, 1994).

5.1.2. Tag design and synthesis of downstream processing

Selecting a purification tag that is optimal in a generic sense is a challenging task. Although there is a relative abundance of previous research in the use of recombinant technology to improve separation characteristics of protein products (Terpe, 2003;

Uhlen and Moks, 1990; Sassenfeld, 1990; Nygren *et al.*, 1994), this has mainly focused on specific tags, which have advantages in certain situations, but are not necessarily optimal. One study examines the development of a framework for selecting peptide tags in protein purification (Steffens *et al.*, 2000b), however predictions are based on a single physicochemical property (charge). The same is true with a recent study that only examines the behaviour of hydrophobic peptide tags (Fexby and Bulow, 2004). The need arises for a systematic methodology, which selects the most advantageous peptide tag and the appropriate steps to achieve the required purity, while taking into account a multitude of protein product physicochemical properties.

The aim of this chapter is to develop an MINLP framework that considers simultaneously the design of optimal peptide tags for each particular protein product and the synthesis of downstream protein processing. The systematic framework presented herein exploits the advantages of integer optimisation, considers the manipulation of two protein properties and can be expanded to more than two physicochemical features given that these are available. Physicochemical property data are used to specify the amino acid composition of the most advantageous and shortest tag, and concurrently select operations among a set of candidate chromatographic techniques that must achieve a specified purity level, while optimising a suitable performance criterion (*e.g.* minimisation of purification steps).

Next, the problem of optimal peptide tag design and synthesis of downstream processing is defined and a mathematical programming formulation is presented. The applicability of the resulting MINLP models is demonstrated through two illustrative examples.

5.2. Problem statement

Overall, the problem of simultaneous optimal tag design and synthesis of downstream protein processing can be stated as follows:

Given:

- a mixture of proteins (p: 1,...,P) with known physicochemical properties;
- a set of available chromatographic techniques (i: 1,...,I) each performing a separation by exploiting a specific physicochemical property (charge or hydrophobicity);
- the properties of the twenty amino acids (k: 1,...,20); and
- a specification for the desired product (dp) in terms of a minimum purity level.

Determine:

- the amino acid composition of the shortest and most advantageous peptide tag;
- the physicochemical properties of the tagged protein (desired product + tag); and
- the flowsheet of the high-resolution purification process.

So as to optimise a suitable performance criterion.

To solve the problem, a few assumptions need to be made. Physicochemical properties of the tagged protein are assumed to be calculated by adding the properties of the tag to the ones of the original protein. The amino acids that comprise the fused tags are assumed to have a fully exposed surface. The possibility of the tag burying itself into the protein is avoided by imposing an upper bound to the number of hydrophobic residues that may be included in the tag. An upper bound is imposed on the number of amino acids that can be present in the peptide tag, so as to avoid interference with the tertiary structure of the protein. At the same time, the formation of a secondary structure (*e.g.* an alpha-helix or a beta-sheet) from the tag itself should also be avoided; therefore the number of amino acids in the tag should not be larger than 6 or 7 (Creighton, 1993). The overall molecular weight of the protein

product is assumed to remain constant after the addition of the tag, as the combined molecular weight of a few amino acids is negligible compared to the one of the protein product. The methodologies used herein for the prediction of the physicochemical properties of the fused protein are theoretical estimations, nevertheless they are considered to be sufficiently adequate indications of the alteration of the property in question.

For process synthesis, it is assumed that the protein product is separated completely without any product loss; *i.e.* no product is left over in the discarded stream after each chromatographic step. Protein-protein interactions in chromatographic steps are assumed to be negligible. Finally, it is usually necessary to introduce membrane steps for buffer exchange and/or protein concentration between chromatographic steps, which could lead to some loss of protein product; however, for the needs of this study these losses are considered insignificant. Additional assumptions for process synthesis, such as the formulation of the models being based solely on physicochemical data and the approximation of the chromatographic peaks by isosceles triangles, can be found in Vásquez-Alvarez *et al.* (2001), Vásquez-Alvarez and Pinto (2001), Vásquez-Alvarez and Pinto (2003) or Vásquez-Alvarez and Pinto (2004).

5.3. Mathematical formulation

The proposed MINLP representations are based on a previously developed MILP formulation (Vásquez-Alvarez and Pinto, 2001) for the synthesis of purification bioprocesses. The optimisation framework selects a tag that modifies the properties of the protein product in the most beneficial way and concurrently minimises the number of chromatographic steps in the purification process. Next, the notation of the model is provided and the mathematical models are described in detail.

5.3.1. Nomenclature

The indices, sets and parameters associated with the problem are listed below:

Indices

dp	desired protein product
i	chromatographic techniques
k	amino acids
p	proteins in the mixture

Sets

AE	anion exchange chromatography
CE	cation exchange chromatography
IE	ion exchange chromatography
AA	acidic amino acid group
BA	basic amino acid group
HA	hydrophobic amino acid group

Parameters

$\hat{H}_{dp}$	initial product hydrophobicity
h_k	hydrophobicity of amino acid k
K_k	ionisation constants
KD_{ip}	retention time of protein p in chromatographic technique i
M	large positive number
m_{0p}	initial mass of protein p
MW_{dp}	molecular weight of product
N	maximum number of amino acids in tag (~ 6 or 7)
$\hat{Q}_{i,dp}$	initial product charge for chromatographic technique i
$\hat{S}_{dp}$	initial total surface area of product
s_k	total exposed area for amino acid k
SP	specified product purity
ε	small number

The formulation is based on the following key variables:

Integer variables

n_k number of occurrences of amino acid k in the tag

Binary variables

$x_{i,dp}$ 1 if product charge is greater than zero; 0 otherwise

w_i 1 if chromatographic technique i is selected; 0 otherwise

Continuous variables

$Q_{i,dp}$ product charge for chromatographic technique i

Positive continuous variables

CF_{ip} concentration factor of protein p after chromatographic technique i

DF_{ip} deviation factor of protein p after chromatographic technique i

H_{dp} product hydrophobicity

h_k hydrophobicity of amino acid k

$KD_{i,dp}$ product retention time in chromatographic technique i

m_{ip} mass of protein p after chromatographic technique i

r_k relative surface area for amino acid k

5.3.2. Model constraints

5.3.2.1. Peptide tag size constraints

An upper bound is imposed on the number of amino acids in each tag.

$$\sum_k n_k \leq N \tag{5.1}$$

As already discussed, smaller peptide tags have several practical advantages, including minimal effect on the protein structure, easier separation upon cleavage and, in many cases, no need for cleavage at all (Terpe, 2003). Therefore, in this study the bound is usually set to six, which will also help to eliminate the possibility of formation of a helix or a beta-hairpin from the tag.

A constraint imposed on the amino acid composition of the peptide tag is that *at most* half of the fused amino acids are permitted to have a hydrophobic nature.

$$\sum_{k \in HA} n_k \leq 0.5 \cdot \sum_{k} n_k \tag{5.2}$$

This ensures that the tag will not bury itself within the attached protein or form undesirable structures. It is difficult to specifically define the maximum fraction of hydrophobic amino acids in the peptide tag; nevertheless hydrophobic amino acids should be balanced by polar residues in the composition of the tag.

5.3.2.2. Physicochemical property constraints

The net charge ($Q_{i,dp}$) of the tagged protein is predicted based on the formula suggested by Mosher et al. (1993):

$$Q_p = \sum_{k \in BA} \frac{b_k}{\frac{K_k}{[H^+]_i} + 1} - \sum_{k \in AA} \frac{a_k}{\frac{[H^+]_i}{K_k} + 1} + z_L \tag{5.3}$$

where a_k and b_k are the number of acidic and basic amino acids respectively, K_k is the ionisation constant, $[H^+]$ is the concentration of hydrogen cations, and z_L is the total charge of ligands bound to the protein (typically metal ions).

According to equation (5.3), the net charge of a protein is approximated by considering the contribution of amino acids belonging to the basic group, minus the contribution of amino acids belonging to the acidic group. For this study, in order to estimate the charge $Q_{i,dp}$ of the tagged protein *dp* in ion exchange operation *i*, the charge contributions of any basic amino acids that exist in the tag are added to the initial charge of the desired product ($\hat{Q}_{i,dp}$) and, respectively, the charge contributions of any acidic amino acids are subtracted:

$$Q_{i,dp} = \hat{Q}_{i,dp} + \sum_{k \in BA} \frac{n_k}{\frac{K_k}{[H^+]_i} + 1} - \sum_{k \in AA} \frac{n_k}{\frac{[H^+]_i}{K_k} + 1} \qquad \forall i \in IE \tag{5.4}$$

Values for the ionisation constants K_k (Mosher *et al.*, 1993; Devereux *et al.*, 1984) are presented in Table 5.1. When proteins with complex structures are encountered, the interactions between various groups may be strong and some charged groups may not be exposed on the protein surface. For these reasons, the method described above may not always be accurate for large molecules, but is still useful as an indication of how much and in what way the addition of a peptide tag will modify the net charge of the desired product.

Table 5.1: Ionisation constants for the two amino acid groups (from Mosher et al., 1993)

Residue	*pK*
Basic group	
Arg	12.50
His	6.50
Lys	10.79
Acidic group	
Asp	3.91
Cys	8.30
Glu	4.25
Tyr	10.95

Many methods that predict the hydropathic character of different regions of a protein's amino acid chain based on well-established molecular thermodynamic theories exist (Hopp and Woods, 1981; Kyte and Doolittle, 1982); however, interaction parameters must be determined experimentally for each polymer system and protein. Generally, significant experimental work is required before a prediction method can be developed. The protein's hydrophobicity (H_p) is estimated here using a method developed by Lienqueo *et al.* (2002). An updated version of the method has also been published (Lienqueo *et al.*, 2003).

$$H_p = \sum_{aa} h_{aa} \cdot r_{aa} \tag{5.5}$$

$$r_{aa} = \frac{s_{aa}}{\sum_{aa'} s_{aa'}} \qquad \forall\, aa \qquad (5.6)$$

where h_{aa} is the value of the hydrophobicity assigned to each amino acid *aa* (*aa* = 1,...,20), r_{aa} is the relative surface area exposed for each amino acid *aa*, and s_{aa} is the total exposed area of amino acid *aa*. The denominator in equation (5.6) represents the total surface of the protein.

According to equations (5.5) and (5.6), the calculation of hydrophobicity is performed by considering the 3-dimensional structure of a protein molecule and the relative contribution of each amino acid on the surface of the protein to its properties (Berggren *et al.*, 2002). There are more than 40 hydrophobicity scales for amino acids in the literature (Lienqueo *et al.*, 2002), but for this study the normalised values of the scale proposed by Miyazawa and Jernigan (1985) are used (presented in Table 5.2).

The contribution of the original protein molecule to the hydrophobicity of the tagged product is considered to remain constant, therefore only the contributions of the amino acids in the tag need to be estimated and then added to the initial hydrophobicity of the protein product, $\hat{H}_{dp}$.

$$H_{dp} = \hat{H}_{dp} + \sum_{k} h_k \cdot r_k \qquad (5.7)$$

The relative surface area, r_k, for each kind of amino acid *k* in the tag is given by:

$$r_k = \frac{s_k \cdot n_k}{\hat{S}_{dp} + \sum_{k'} s_{k'} \cdot n_{k'}} \qquad \forall\, k \qquad (5.8)$$

where the total surface of the tagged protein is estimated from the exposed surface of the protein product, $\hat{S}_{dp}$ (considered to remain unchanged), plus the exposed surface of the amino acids in the tag, $\sum_{k'} s_{k'} \cdot n_{k'}$. These amino acids are assumed to have a fully exposed surface. In cases where this is not true, selecting to place the peptide

tag on the other terminus (*i.e.* the N-terminus instead of the C-terminus of the protein product *dp*, or *vice versa*) can solve this problem (Terpe, 2003). Values for the surface areas of fully exposed amino acids (Chothia, 1975) are presented in Table 5.2.

Table 5.2: Normalised hydrophobicity and exposed surface area for the 20 amino acids.

Residue	h_k [a]	s_k [b]
Phe [c]	1.000	210
Met [c]	0.987	185
Ile [c]	0.967	175
Leu [c]	0.908	170
Cys [c]	0.819	135
Trp [c]	0.775	255
Val [c]	0.770	155
Tyr	0.484	230
Ala	0.391	115
His	0.354	195
Thr	0.253	140
Gly	0.252	75
Arg	0.202	225
Ser	0.188	115
Gln	0.151	180
Pro	0.151	145
Asn	0.125	160
Glu	0.115	190
Asp	0.105	150
Lys	0.000	200

[a] Lienqueo *et al.*, 2002

[b] Chothia, 1975

[c] Hydrophobic group

5.3.2.3. Dimensionless retention time constraints

For the modelling of various chromatographic techniques, established criteria based on the retention time and on the width of the chromatographic peak were used. The necessary parameters have been determined experimentally using pure proteins, in order to predict chromatographic behaviour (Lienqueo *et al.*, 2002). Retention times are defined as a function of a physicochemical property and can be calculated using developed correlations, which relate retention times of each particular chromatographic technique to the appropriate physicochemical property (*e.g.* anion/cation exchange chromatography (*AE/CE*) and charge ($Q_{i,dp}$); hydrophobic interaction (*HI*) and hydrophobicity (H_{dp}), see Figure 5.1). The graphs of Figure 5.1 were produced from extensive experimentation (Lienqueo, 1999; Lienqueo *et al.*, 2002) and are expressed in mathematical relationships (5.9), (5.10) and (5.16).

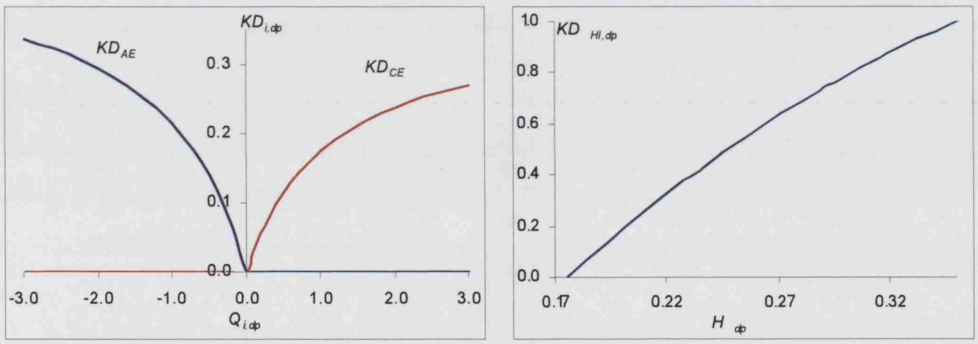

Figure 5.1: Correlation between retention times (KD) and appropriate protein property.

In ion exchange chromatography, proteins adsorb either to exchangers that bind negatively charged molecules (anion exchange) or to exchangers that bind positively charged molecules (cation exchange). Retention times are a function of charge density (Q_{ip}/MW_p), determined by electrophoretic titration curves (Watanabe *et al.*, 1994). The dimensionless retention times (KD_{ip}) are estimated based on mathematical expressions and on property data for the protein product and the contaminants (Lienqueo, 1999). The proposed correlations for ion exchange chromatography were obtained using bind-and-elution conditions; the elution was obtained with an increasing NaCl gradient between 0.0-2.0 M NaCl.

For anion exchange:

$$If\ Q_{ip} \geq 0, \qquad KD_{ip} = 0$$

$$If\ Q_{ip} < 0, \qquad KD_{ip} = \frac{8826 \cdot \left| Q_{ip} / MW_p \right|}{1 + 18845 \cdot \left| Q_{ip} / MW_p \right|} \qquad \forall i \in AE \qquad (5.9)$$

For cation exchange:

$$If\ Q_{ip} \leq 0, \qquad KD_{ip} = 0$$

$$If\ Q_{ip} > 0, \qquad KD_{ip} = \frac{7424 \cdot \left| Q_{ip} / MW_p \right|}{1 + 20231 \cdot \left| Q_{ip} / MW_p \right|} \qquad \forall i \in CE \qquad (5.10)$$

It should be noted that expressions (5.9) and (5.10) only need to be modelled where they refer to *dp*; dimensionless retention times for the contaminant proteins (KD_{ip}) remain constant and are used as parameters in the model. The modelling is performed using constraints (5.11) to (5.15).

$$Q^{+}_{i,dp} - Q^{-}_{i,dp} = Q_{i,dp} \qquad \forall i \in IE \qquad (5.11)$$

$$Q^{+}_{i,dp} \leq M \cdot x_{i,dp} \qquad \forall i \in IE \qquad (5.12)$$

$$Q^{-}_{i,dp} \leq M \cdot (1 - x_{i,dp}) \qquad \forall i \in IE \qquad (5.13)$$

where M is an appropriate upper bound. Binary variables $x_{i,dp}$ express whether the charge of the protein is positive or negative. The *absolute* value of $Q_{i,dp}$ is assigned to either $Q^{+}_{i,dp}$ or $Q^{-}_{i,dp}$ with constraint (5.11), because either $Q^{+}_{i,dp}$ or $Q^{-}_{i,dp}$ always has to be equal to zero due to constraints (5.12) and (5.13). For anion exchange chromatography:

$$KD_{i,dp} = \frac{8826 \cdot \left(Q^{-}_{i,dp} / MW_{dp} \right)}{1 + 18845 \cdot \left(-Q_{i,dp} / MW_{dp} \right)} \qquad \forall i \in AE \qquad (5.14)$$

If the protein charge is negative, it has to follow that $x_{i,dp} = 0$, and the retention time is given by constraint (5.14); otherwise, when the protein charge is positive, the retention time for anion exchange is zero, as $x_{i,dp}$ is forced to one and $Q^{-}_{i,dp} = 0$. Similarly, in the case of cation exchange chromatography:

$$KD_{i,dp} = \frac{7424 \cdot \left(Q^{+}_{i,dp} / MW_{dp}\right)}{1 + 20231 \cdot \left(Q^{+}_{i,dp} / MW_{dp}\right)} \qquad \forall i \in CE \qquad (5.15)$$

For a positive protein charge, $x_{i,dp} = 1$ and the retention time is given by constraint (5.15); otherwise, the retention time for cation exchange is zero.

Hydrophobic interaction chromatography utilises the hydrophobic character of the proteins to separate the mixture according to their relative hydrophobicity. Most hydrophobic amino acids are located near the core of the protein structure and away from the surface, but there usually are hydrophobic residues on the protein surface as well. A formula developed by Lienqueo *et al.* (2002) by evaluating a series of experimental and computational data is used here to estimate the dimensionless retention times for hydrophobic interaction ($KD_{HI,dp}$). Elution for hydrophobic interaction chromatography was obtained with a decreasing ammonium sulphate gradient between 2.0-0.0 M ammonium sulphate. Retention times are a function of hydrophobicity; the function in this case (on phenyl sepharose) is a quadratic equation.

$$KD_{HI,dp} = -12.14 \cdot H^{2}_{dp} + 12.07 \cdot H_{dp} - 1.74 \qquad (5.16)$$

As with ion exchange chromatography, dimensionless retention times for hydrophobic interaction in the case of the *contaminant proteins* ($KD_{HI,p}$) remain constant.

5.3.2.4. Concentration factor constraints

The concentration factor, CF_{ip}, represents the ratio between the mass of contaminant p after and before chromatographic step i. It can be calculated through a set of equations that describe the chromatographic peaks of the desired product and one of

the contaminants (Lienqueo *et al.*, 1996). The chromatographic peaks are approximated by two triangles (chromatograms), one referring to the product and the other to the contaminant protein (Figure 5.2).

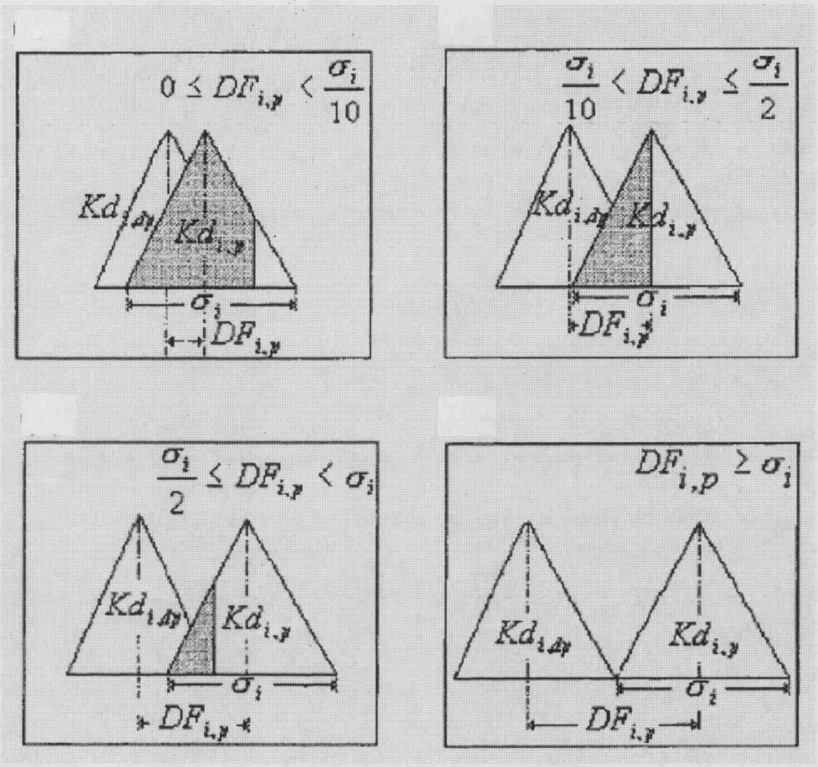

Figure 5.2: Representation of chromatographic peaks. The first triangle represents the protein product and the second contaminant p (from Vásquez-Alvarez et al., 2001).

In Figure 5.2, the left peak refers to the product and the one on the right to the contaminant protein. Assuming that the peaks have constant form, the area of the figure formed by the intersection of the two triangles (shaded areas) represents the amount of contaminant *p* that remains in the mixture after applying chromatographic technique *i* (Lienqueo *et al.*, 1996). Note that 100% of the product protein is assumed to be recovered.

The chromatograms are approximated with equations (5.19) presented below. In this set of equations, σ_i is the peak width parameter, which depends only on the type of chromatographic operations and is calculated by averaging over several proteins. For ion exchange, the value for the peak width is $\sigma_i = 0.15$; for hydrophobic interaction σ_i

= 0.22. Parameter Δ is a percentage of error, for which an experimentally determined value of 0.02 is used. Deviation factors, DF_{ip}, indicate the distance between the desired product's chromatographic peak and the chromatographic peak of one of the contaminants. Deviation factors express the driving force of the separation process; they are defined as the difference between the retention times of the product ($KD_{i,dp}$) and the contaminant in question (KD_{ip}) for each particular chromatographic step.

$$DF_{ip} = \left| KD_{i,dp} - KD_{ip} \right| \qquad \forall\, i, p \neq dp \qquad (5.17)$$

For the estimation of the deviation factor in this work the non-differential relationship (5.17) is modelled with the following constraint:

$$DF_{ip} = \left[\left(KD_{i,dp} - KD_{ip} \right)^2 + \varepsilon^2 \right]^{1/2} \qquad \forall\, i, p \neq dp \qquad (5.18)$$

The relationships (Lienqueo *et al.*, 1996) that describe the concentration factor, CF_{ip}, for protein p in chromatographic technique i are:

$$CF_{ip} = 1 \qquad if\ 0 \leq DF_{ip} < \frac{\sigma_i}{10} \qquad (5.19a)$$

$$CF_{ip} = (1+\Delta) \cdot \left(\frac{\sigma_i^2 - 2 \cdot DF_{ip}^2}{\sigma_i^2} \right) \qquad if\ \frac{\sigma_i}{10} \leq DF_{ip} < \frac{\sigma_i}{2} \qquad (5.19b)$$

$$CF_{ip} = 2 \cdot (1+\Delta) \cdot \frac{(\sigma_i - DF_{ip})^2}{\sigma_i^2} \qquad if\ \frac{\sigma_i}{2} \leq DF_{ip} < \sigma_i \qquad (5.19c)$$

$$CF_{ip} = \Delta \qquad if\ DF_{ip} \geq \sigma_i \qquad (5.19d)$$

Here, concentration factors for the various chromatographic steps are modelled with sigmoid functions. Because the peak width parameter, σ_i, is the same for both anion and cation exchange, there is only need for one equation for all ion exchange steps, but a separate one is required for hydrophobic interaction. Figure 5.3 demonstrates the sigmoid functions used, it can be observed that the sigmoid approximation (red dots) provide a satisfactory accurate approximation for CF_{ip}, as calculated from equations (5.19).

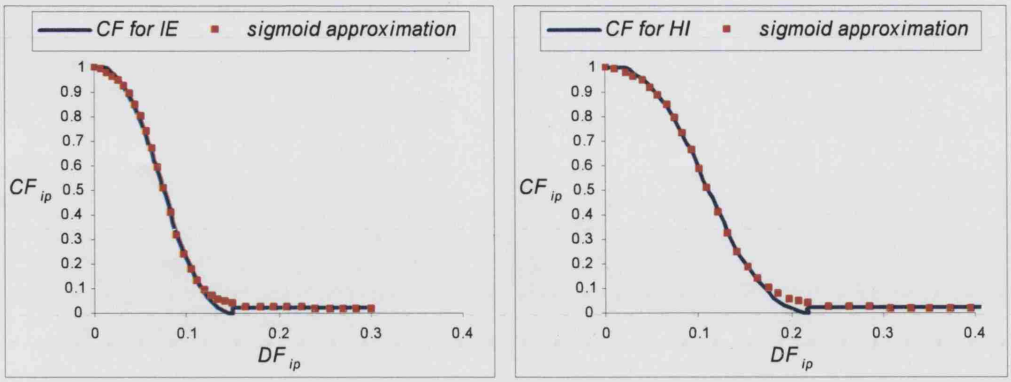

Figure 5.3: Sigmoid approximations of concentration factors (CF) for ion exchange (IE) and hydrophobic interaction (HI).

$$CF_{ip} = \frac{3.722}{3.727 + 0.579 \cdot e^{(54.410 \cdot DF_{ip} - 2.176)}} + 0.019 \qquad \forall\, i \in IE, p \neq dp \qquad (5.20)$$

$$CF_{HI,p} = \frac{3.937}{3.933 + 0.105 \cdot e^{(36.005 \cdot DF_{HI,p} - 0.299)}} + 0.018 \qquad \forall\, p \neq dp \qquad (5.21)$$

In the case of the desired product (*dp*), the concentration factor $CF_{i,dp}$ is always equal to one, since the assumption of no product loss has been made. Therefore:

$$CF_{i,dp} = 1 \qquad \forall\, i \qquad (5.22)$$

These concentration factors are introduced in the process synthesis constraints (section 5.3.2.5) that address the synthesis problem.

5.3.2.5. Process synthesis constraints

The convex hull representation applied for contaminant separation is presented in Vásquez-Alvarez and Pinto (2001); the formulation for the selection of appropriate chromatographic steps and the indication of the remaining amount of protein after each step *i* is also applied here. The contaminant constraints in consecutive chromatographic steps can be represented by the following disjunction, which relates the mass values of each protein in subsequent steps:

$$\begin{bmatrix} w_i \\ m_{ip} = CF_{ip} \cdot m_{i-1,p} \end{bmatrix} \vee \begin{bmatrix} \neg w_i \\ m_{ip} = m_{i-1,p} \end{bmatrix} \qquad \forall i, p \qquad (5.23)$$

Disjunction (5.23) generates the following convex hull constraints on mixed integer representation:

$$m_{1p} = CF_{1p} \cdot w_1 \cdot m_{0p} + m_{0p} \cdot (1 - w_1) \qquad \forall p \qquad (5.24)$$

$$m_{ip} = CF_{ip} \cdot m^1_{i-1,p} + m^2_{i-1,p} \qquad \forall p, i = 2,...,I \qquad (5.25)$$

$$m_{i-1,p} = m^1_{i-1,p} + m^2_{i-1,p} \qquad \forall p, i = 2,...,I \qquad (5.26)$$

$$0 \le m^1_{i-1,p} \le m_{0p} \cdot w_i \qquad \forall p, i = 2,...,I \qquad (5.27)$$

$$0 \le m^2_{i-1,p} \le m_{0p} \cdot (1 - w_i) \qquad \forall p, i = 2,...,I \qquad (5.28)$$

The mass of protein that is left after the first chromatographic step is given by constraint (5.24). In the following steps, constraints (5.25) to (5.28) hold. When chromatographic step i is selected (*i.e.* $w_i = 1$) the mass of contaminants is reduced (*i.e.* $m_{ip} = CF_{ip} \cdot m_{i-1,p}$), whereas when step i is not selected the mass remains constant (*i.e.* $m_{ip} = m_{i-1,p}$).

The mass of the desired protein product must meet the purity demand after the last chromatographic step I.

$$m_{I,dp} \ge SP \cdot \sum_p m_{I,p} \qquad (5.29)$$

5.3.3. Solution Approach

Typical performance criteria to be optimised in the problem defined above are the number of purification steps and/or the size of the tag. The overall problem is non-convex; non-linearities arise in constraint (5.8), for the estimation of the relative surface area of amino acids in the tag; constraints (5.14) – (5.16), for the evaluation of retention times from the values of the protein product's properties; constraint

(5.18), for the calculation of the deviation factor; sigmoid constraints (5.20) and (5.21), for the computation of the concentration factor; and constraints (5.24) and (5.25), for estimating the mass of proteins after each chromatographic step. The presented models constitute MINLP formulations. A two-stage solution procedure is proposed, in order to identify the shortest amino acid sequence that can produce the optimal flowsheet for the purification process.

5.3.3.1. Stage 1

Designing flowsheets with fewer purification steps can significantly reduce costs (Atkinson and Mavituna, 1991). The overall objective is to minimise the total number of selected chromatographic steps in the purification process, subject to the constraints described above (problem P1). A tag is also selected in this stage, but the selection is further improved at stage 2.

Problem P1

$$minimise \sum_i w_i \tag{5.30}$$

subject to:

peptide tag size constraints (5.1) and (5.2);

physicochemical property constraints (5.4), (5.7) and (5.8);

dimensionless retention time constraints (5.11), (5.12), (5.13), (5.14), (5.15) and (5.16);

concentration factor constraints (5.18), (5.20),(5.21) and (5.22);

process synthesis constraints (5.24), (5.25), (5.26),(5.27), (5.28) and (5.29);

$$x_{i,dp}, w_i \in \{0,1\} \qquad \forall i \in IE \tag{5.31}$$

$$n_k \in Z^+ \qquad \forall k \tag{5.32}$$

$$CF_{ip}, DF_{ip}, H_{dp}, KD_{i,dp}, m_{ip}, Q^{+}_{i,dp}, Q^{-}_{i,dp}, r_k \geq 0 \qquad \forall i,p,k \qquad (5.33)$$

5.3.3.2. Stage 2

The objective here (problem P2) is to minimise the number of amino acids in the tag, subject to the same constraints as before, plus an additional constraint that fixes the number of chromatographic steps as identified in stage 1. Then, the smallest tag that can produce the optimal flowsheet is finally determined.

Problem P2

$$minimise \sum_k n_k \qquad (5.34)$$

subject to:

constraints (5.1), (5.2), (5.4), (5.7), (5.8), (5.11) - (5.16), (5.18), (5.20) - (5.29), (5.31) - (5.33), and

$$\sum_i w_i \leq i^* \qquad (5.35)$$

where i^* is the number of steps identified in stage 1.

5.5. Computational results

The proposed formulation is applied to two example mixtures. Solutions were obtained with the network-enabled, problem-solving environment NEOS Server 4.0 (http://www-neos.mcs.anl.gov/; Gropp and Moré, 1997; Czyzyk *et al.*, 1998; Dolan, 2001) using the SBB solver for the solution of the MINLP models and CONOPT3 as the NLP solver.

5.5.1. Example 1

The first example is based on experimental data (Lienqueo *et al.*, 2002; Lienqueo, 1999) for a mixture of four proteins: thaumatin (*dp*), conalbumin (p_1), chymotripsinogen A (p_2) and ovalbumin (p_3). The physicochemical properties of the mixture are presented in Table 5.3. The purity level requirement for the desired product (*dp*) is 98%. Overall, there are 11 candidate chromatographic techniques: anion exchange chromatography (AE) at pH 4, AE at pH 5, AE at pH 6, AE at pH 7, AE at pH 8, cation exchange chromatography (CE) at pH 4, CE at pH 5, CE at pH 6, CE at pH 7, CE at pH 8 and hydrophobic interaction (HI).

Table 5.3: Physicochemical properties of protein mixture in first example.

	m_{0p}	MW_p	H_p	S	$Q_{ip} \times 10^{-17}$ (C/molecule)				
proteins	(mg/mL)	(Da)			pH 4	pH 5	pH 6	pH 7	pH 8
dp	2	22200	0.27	9573.15	1.60	1.57	1.64	1.55	0.75
p_1	2	77000	0.23	29287.60	0.93	0.33	-0.12	-0.34	-0.50
p_2	2	23600	0.31	10910.80	2.15	1.46	1.17	0.780.38	
p_3	2	43800	0.28	15880.90	1.16	-0.63	-1.36	-1.82	-1.95

In order to acquire a point of reference, the example was first solved without any tag fused to the protein product, *i.e.* using the formulation of problem P1, with an upper bound of zero imposed on the number of amino acids in the tag (*i.e.* $N = 0$). The resulting mathematical model involves 317 constraints, 41 discrete variables and 275 continuous variables and was solved in 4.14 seconds. The optimal solution is presented in Figure 5.4. The model was able to identify a solution that achieves a purity of 98.1% for the desired product, for which four chromatographic steps are needed: cation exchange chromatography at pH 6, cation exchange chromatography at pH 7, cation exchange chromatography at pH 8 and hydrophobic interaction. Note that the product mass remains constant, but its purity increases after each step: From 25.0% in the original mixture, to 51.4% after the first chromatographic step (CE pH 6); 64.6% after CE pH 7; 71.9% after CE pH 8; and finally 98.1% (above the required purity level) after the final chromatographic operation (HI).

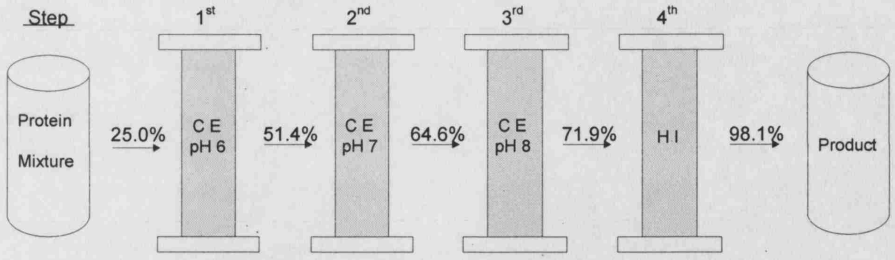

Figure 5.4: Optimal flowsheet for purification of protein mixture in example 1 without tag.

An improved solution was suggested when using a peptide tag. The minimum number of steps is identified by solving problem P1 with an upper bound of 6 amino acids per tag (stage 1). The model was solved in 8.49 seconds. Only three separation steps are needed: cation exchange chromatography at pH 7, cation exchange chromatography at pH 8 and hydrophobic interaction. Next, the number of purification steps was fixed ($i^* = 3$) and the model was solved again using the formulation of problem P2 (stage 2). The CPU time was 20.57 seconds. The solution is a tag of 2 lysine residues; the purity demand was achieved (98.0%) and the process included the same three purification techniques (CE pH 7, CE pH 8, HI) as in the solution of problem P1. The results are presented in Figure 5.5 and in Table 5.4.

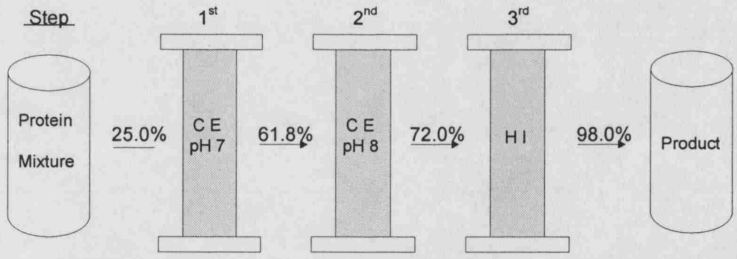

Figure 5.5: Optimal flowsheet for purification of protein mixture in example 1 with a tag of 2 lysines.

The selection of a tag exclusively with lysine amino acids suggests that the optimal solution to the problem at hand is to increase the charge of the desired protein. The use of other basic amino acids would have a stronger effect on the charge than lysine, but their presence would also increase hydrophobicity, which remains unchanged

with lysine (h_k = 0, Table 5.2). The observation implies that an increase in hydrophobicity is not beneficial. This was tested by forcing the inclusion of hydrophobic amino acids in the tag (*e.g.* phenylalanines); the experiments showed that a purity of 98% (as required here) is not achievable when hydrophobicity is increased even by a small amount.

Table 5.4: Values of retention times, deviation factors and concentration factors as estimated for the solution of example 1 (98% purity, 3 steps, 2-lys tag).

	dp			*p1*			*p2*			*p3*		
step[a]	*KD*	*DF*	*CF*	*KD*[b]	*DF*	*CF*	*KD*[b]	*DF*	*CF*	*KD*[b]	*DF*	*CF*
AE4	0.000	0.000	1.000	*0.000*	0.000	1.000	*0.000*	0.000	1.000	*0.000*	0.000	1.000
AE5	0.000	0.000	1.000	*0.000*	0.000	1.000	*0.000*	0.000	1.000	*0.100*	0.100	0.216
AE6	0.000	0.000	1.000	*0.013*	0.013	0.983	*0.000*	0.000	1.000	*0.173*	0.173	0.024
AE7	0.000	0.000	1.000	*0.036*	0.036	0.907	*0.000*	0.000	1.000	*0.206*	0.206	0.020
AE8	0.000	0.000	1.000	*0.051*	0.051	0.797	*0.000*	0.000	1.000	*0.214*	0.214	0.019
CE4	0.219	0.000	1.000	*0.072*	0.147	0.037	*0.238*	0.019	0.972	*0.128*	0.091	0.300
CE5	0.218	0.000	1.000	*0.029*	0.189	0.021	*0.204*	0.014	0.982	*0.000*	0.218	0.019
CE6	0.222	0.000	1.000	*0.000*	0.222	0.019	*0.184*	0.038	0.898	*0.000*	0.222	0.019
CE7	0.217	0.000	1.000	*0.000*	0.217	0.019	*0.147*	0.070	0.580	*0.000*	0.217	0.019
CE8	0.153	0.000	1.000	*0.000*	0.153	0.033	*0.090*	0.063	0.670	*0.000*	0.153	0.033
HI	0.629	0.000	1.000	*0.413*	0.216	0.039	*0.832*	0.203	0.051	*0.701*	0.072	0.809

[a] AE4: anion exchange at pH 4; etc.

[b] The values of *KD* for p_1, p_2 *and* p_3 (denoted in italics) are calculated before the computational experiments and are used as parameters in the model

Table 5.4 presents the dimensionless retention times, deviation factors and concentration factors for each protein in each chromatographic step for the solution of example 1 with the formulation of problem P2 (as shown in Figure 5.5). All values

are calculated from the mathematical model, as they depend on the physicochemical properties of the protein product, which are liable to change. It should be noted that contaminant retention times remain constant and are therefore used as parameters for the mathematical model.

5.5.2. Example 2

For the purpose of further testing the models, a larger and more challenging example was created. The physicochemical properties of this second mixture of 13 proteins are presented in Table 5.5.

Table 5.5: Physicochemical properties of protein mixture in second example.

	m_{0p}	MW_p	H_p	$Q_{ip} \times 10^{-17}$ (C/molecule)				
proteins	(mg/mL)	(Da)		pH 4	pH 5	pH 6	pH 7	pH 8
dp	2	77000	0.28	2.04	1.06	-0.37	-0.81	-1.13
p_1	2	22200	0.27	1.60	1.57	1.56	1.55	0.75
p_2	2	23600	0.31	2.15	1.46	1.17	0.78	0.38
p_3	2	13500	0.23	1.83	0.65	0.26	-0.20	-0.33
p_4	2	43800	0.28	1.16	-0.63	-1.36	-1.82	-1.95
p_5	2	15900	0.27	2.89	2.81	2.80	2.64	2.07
p_6	2	14400	0.32	-0.46	-0.47	-0.63	-1.21	-1.25
p_7	2	17500	0.21	0.45	-0.62	-0.79	-1.26	-1.70
p_8	2	50000	0.27	-0.12	-0.32	-0.76	-0.91	-1.04
p_9	2	12100	0.18	1.46	0.62	-1.02	-1.33	-1.52
p_{10}	2	25500	0.30	1.01	-0.63	-1.27	-1.59	-1.76
p_{11}	2	26000	0.28	2.96	1.26	0.92	0.54	0.01
p_{12}	2	19900	0.25	0.93	0.33	-0.12	-0.34	-0.50

The properties for the proteins of example 2 were determined using a random number generator to produce values within certain parameters, for example the charge of each protein was allowed to range between -3×10^{-17} and 3×10^{-17} Coulomb per molecule. A similar procedure was followed for the random assignment

of hydrophobicities and molecular weights. The exposed surface of the protein product was considered to be proportional to the molecular weight (*i.e.* the larger the molecular weight, the larger the exposed surface of the molecule), and was set to $\hat{S}_{dp} = 29287.6$ (see constraint (5.8)). The purity level requirement for protein *dp* was set to 95%. The same 11 candidate chromatographic steps as for example 1 are available: anion exchange chromatography at pH 4, AE pH 5, AE pH 6, AE pH 7, AE pH 8, cation exchange chromatography at pH 4, CE pH 5, CE pH 6, CE pH 7, CE pH 8 and hydrophobic interaction.

Example 2 was first solved without tags (*i.e.* $N = 0$). The formulation of problem P1 was applied; the resulting mathematical model involves 884 constraints, 41 discrete variables and 752 continuous variables. The optimal solution was identified in 71.78 seconds and is presented in Figure 5.6. The product is recovered with 95.3% purity and six chromatographic steps are needed: anion exchange chromatography at pH 6, AE pH 7, AE pH 8, cation exchange chromatography at pH 4, CE pH 5, and hydrophobic interaction.

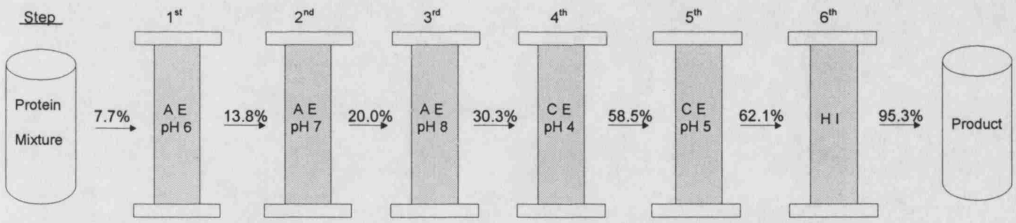

Figure 5.6: Optimal flowsheet for purification of protein mixture in example 2 without tag.

Using model P1, an improved solution with only 4 chromatographic steps (AE pH 6, AE pH 8, CE pH 4, HI) is suggested. The results were produced with an upper bound of 6 amino acids per tag (*i.e.* $N = 6$). In order to test whether this solution can be achieved with a smaller tag, model P2 was applied to identify solutions with the smallest possible number of amino acids in the tag. Using the minimum number of steps for 95% purity (*i.e.* $i^* = 4$), a minimum number of amino acids in the tag was specified. The results are presented in Figure 5.7. The suggested tag consists of 1 phenylalanine, 1 methionine and 2 tyrosine amino acids; the process has 4

chromatographic steps (AE pH 7, AE pH 8, CE pH 4, HI) and a purity of 95.4%. Note that this solution is different from the one produced with model P1; a different set of chromatographic steps was selected for the purification process.

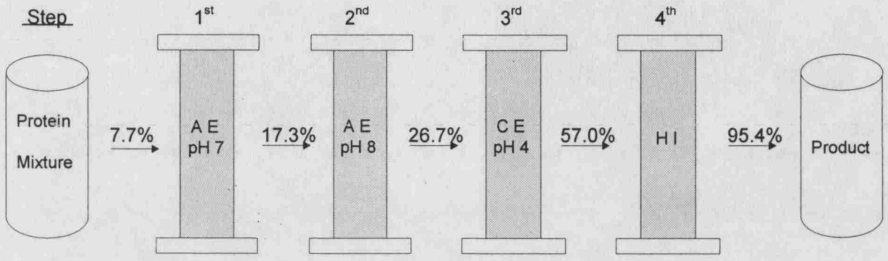

Figure 5.7: Optimal flowsheet for purification of protein mixture in example 2 with a minimised tag of 1 phenylalanine, 1 methionine and 2 tyrosine amino acids (purity requirement: 95%).

Next, a higher level of purity for the protein product was tested. Example 2 was solved again with a higher demand of 98% for product purity, which leads to an infeasible solution when the model is solved without tags. With an upper bound of 6 amino acids per tag, an improved solution with a purity of 98.4% and 5 chromatographic steps (AE pH 6, AE pH 8, CE pH 4, CE pH 5, HI) was suggested. The problem was solved again using the formulation of model P2 and an upper bound of 5 for the number of chromatographic steps (*i.e.* $i^* = 5$). The results are presented in Figure 5.8. The selected tag consists of 1 phenylalanine, 1 tryptophan and 2 tyrosines. The process has 5 chromatographic steps (AE pH 7, AE pH 8, CE pH 4, CE pH 5, HI) and a purity of 98.1%.

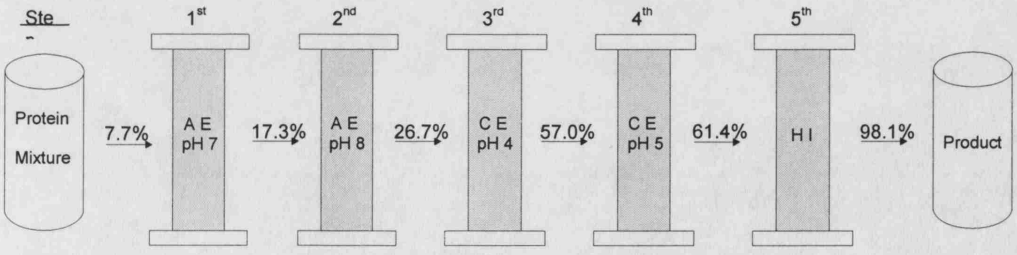

Figure 5.8: Optimal flowsheet for purification of protein mixture in example 2 with a tag of 1 phenylalanine, 1 tryptophan and 2 tyrosine molecules (increased purity requirement: 98%).

Table 5.6 presents computational statistics for the optimisation of the two presented examples with models P1 and P2. Model sizes in terms of discrete and continuous variables and constraints are given. Also shown are CPU times, nodes enumerated and objective values.

Table 5.6: Summary of computational statistics.

example	model	max aas	max steps	discrete/continuous variables	constraints	nodes	CPU time (s)	objective value
1	P1	≤ 0	-	41/275	317	114	4.14	4 [a]
1	P1	≤ 6	-	41/275	317	192	8.49	3 [a]
1	P2	≤ 6	≤ 3	41/275	318	636	20.57	2 [b]
2	P1	≤ 0	-	41/752	884	363	71.78	6 [a]
2	P1	≤ 6	-	41/752	884	346	68.84	4 [a]
2	P2	≤ 6	≤ 4	41/752	885	1777	175.61	4 [b]
2 [c]	P1	≤ 0	-	41/752	884	340	53.62	infeasible [a]
2 [c]	P1	≤ 6	-	41/752	884	514	94.50	5 [a]
2 [c]	P2	≤ 6	≤ 5	41/752	885	2881	320.67	4 [b]

[a] Number of chromatographic steps

[b] Number of amino acids in the tag

[c] 98% purity demand

5.6. Conclusions

An optimisation framework for the simultaneous selection of optimal peptide tags and the synthesis of chromatographic steps for the purification of protein mixtures in downstream processing has been presented. The framework utilises the advantages of integer optimisation and mathematical programming techniques, incorporates recent developments in the synthesis and optimisation of downstream purification processes and can be extended to consider application to larger examples, use of additional chromatographic techniques, or manipulation of other physicochemical properties.

The overall problem has been formulated as an MINLP model and a two-stage solution procedure has been proposed.

Two examples of protein mixtures were tested to demonstrate the efficiency of the optimisation based methodology. In both examples, small peptide tags were used (from two to four residues), and only specific physicochemical properties were modified, especially hydrophobicity and charge, without significant conformational changes or bio-activity. These smaller changes have allowed an important decrease in the number of purification steps. In the first example, only the modification of charge benefited the purification, while hydrophobicity was the most influential property in the second example. Results were indicative of the benefits of the application of optimisation-based techniques in the use of purification tags in biotechnological production plants, and have provided a useful guideline for both downstream process synthesis and optimal tag design. However, it will be interesting to validate the generated hypotheses by evaluating experimentally the chromatographic behaviour of the proteins together with the peptide tags.

Testing the mathematical framework with larger examples and investigating alternative solution strategies is a possible extension to this work. Consideration of additional types of chromatographic steps would also be very interesting, provided that the appropriate correlations become available. For example, ion exchange chromatography with pH gradient could be considered, which would potentially provide the same degree of purification but using a smaller number of steps. Finally, the modelling of the purification can be extended to incorporate a number of issues, including sequencing of the purification steps, product loss, protein-protein interactions, use of membrane steps between chromatographic steps and/or application of alternative objective functions.

In the next chapter, the MINLP framework presented here and applied for the simultaneous selection of optimal peptide tags and the synthesis of chromatographic steps for the purification of protein mixtures in downstream processing is linearised, in order to develop an MILP formulation.

Chapter 6

An MILP model for optimal peptide tag design and synthesis of downstream processing

As seen in chapter 5, downstream protein processing in biochemical production plants is typically among the most difficult and complex stages and the source of a large portion of the manufacturing and investment costs in a biochemical production plant. Early systematic methods for the synthesis of downstream protein processing made use of expert knowledge systems for selecting operations (Lienqueo *et al.*, 1996). Vasquez-Alvarez and Pinto (2004) presented an MILP framework, in which mathematical models for each chromatographic technique rely on physicochemical data on the protein mixture that contains the desired product, and provide information on its potential purification.

A two-step MINLP framework for the optimal design of case-specific peptide tags that alter the properties of a particular protein product in the most beneficial way, and the concurrent synthesis of downstream protein processing was presented in the

previous chapter. It was demonstrated that considerable improvement of downstream protein purification processes can be achieved with the use of comparatively short sequences of amino acids (purification tags), genetically fused on the protein product, which modify the physicochemical properties of the protein in a way that enhances the separation, thus simplifying the purification flowsheet.

The above framework is non-convex; non-linearities arise in the estimation of the relative surface area of amino acids in the tag; in the evaluation of retention times from the values of the protein product's properties; in the calculation of the deviation factor; in the computation of the concentration factor; and in estimating the mass of proteins after each chromatographic step. The resulting MINLP formulation may provide modelling flexibility, but in some cases feasible solutions cannot be identified. In order to improve solution quality, non-convexities are avoided by reformulating the framework as an MILP model through piecewise linear approximations of the non-linear functions. The new MILP model, derived from the previous MINLP, utilises physicochemical property data to specify the amino acid composition of the shortest and most advantageous peptide tag configuration, and concurrently select operations among a set of candidate chromatographic techniques in order to achieve a specified purity level. The applicability of the MILP framework is demonstrated by an example that relies on experimental data.

6.1. Problem statement

The problem in hand is identical to the one in chapter 5, but the models used for its solution differ (MILP formulation instead of MINLP). Overall, the problem of simultaneous optimal tag design and synthesis of downstream protein processing can be stated as follows:

Given:

- a mixture of proteins with known physicochemical properties;

- a set of available chromatographic techniques;
- the properties of the twenty amino acids; and
- a minimum purity level for the product protein.

Determine:

- the amino acid composition of the shortest and most advantageous tag;
- the physicochemical properties of the tagged protein; and
- the flowsheet of the purification process.

So as to minimise the number of chromatographic steps chosen *and* the number of amino acids present in the selected tag. All the assumptions presented in section 5.2 hold; any new assumptions will be discussed as the new MILP formulation is presented in section 6.2.

6.2. Mathematical formulation

Most of the notation of this chapter is identical to the one presented in chapter 5 (section 5.3.1). Only the additional symbols required are provided next.

6.2.1. Nomenclature

The additional indices, sets and parameters associated with the problem are listed below:

Parameters

c	penalty for selection of amino acids
ε	small number

Binary variables

y_{ip}	1 if $(KD_{i,dp} - KD_{ip})$ is positive; 0 otherwise

Continuous variables

sl_i slack variables

Positive continuous variables

$\overline{CF}_{ip}$ CF_{ip} if chromatographic technique i is selected; 1 otherwise

DF_{ip}^{+}, DF_{ip}^{-} auxiliary variables for absolute value of DF_{ip}

Negative continuous variables

ξ_{ip} $\ln CF_{ip}$ if chromatographic technique i is selected; 0 otherwise (*i.e.* $\xi_{ip} \equiv (\ln CF_{ip}) \cdot w_i$)

6.2.2. An alternative MINLP model

First, an alternative version of the MINLP formulation is presented, in order to facilitate the application of the piecewise linear approximations.

6.2.2.1. Problem P1

The MINLP model presented in section 5.3.7 (Problem P1) is summarised below.

$$minimise \sum_i w_i \tag{6.1}$$

subject to:

Peptide tag size constraints

$$\sum_k n_k \le N \tag{6.2}$$

$$\sum_{k \in HA} n_k \le 0.5 \cdot \sum_k n_k \tag{6.3}$$

Physicochemical property constraints

$$Q_{i,dp} = \hat{Q}_{i,dp} + \sum_{k \in BA} \frac{n_k}{\frac{K_k}{[H^+]_i} + 1} - \sum_{k \in AA} \frac{n_k}{\frac{[H^+]_i}{K_k} + 1} \qquad \forall i \in IE \tag{6.4}$$

$$H_{dp} = \hat{H}_{dp} + \sum_k h_k \cdot r_k \tag{6.5}$$

$$r_k = \frac{s_k \cdot n_k}{\hat{S}_{dp} + \sum_{k'} s_{k'} \cdot n_{k'}} \qquad \forall k \qquad (6.6)$$

Dimensionless retention time constraints

$$Q^+_{i,dp} - Q^-_{i,dp} = Q_{i,dp} \qquad \forall i \in IE \qquad (6.7)$$

$$Q^+_{i,dp} \le M \cdot x_{i,dp} \qquad \forall i \in IE \qquad (6.8)$$

$$Q^-_{i,dp} \le M \cdot (1 - x_{i,dp}) \qquad \forall i \in IE \qquad (6.9)$$

$$KD_{i,dp} = \frac{8826 \cdot \left(Q^-_{i,dp} / MW_{dp}\right)}{1 + 18845 \cdot \left(-Q_{i,dp} / MW_{dp}\right)} \qquad \forall i \in AE \qquad (6.10)$$

$$KD_{i,dp} = \frac{7424 \cdot \left(Q^+_{i,dp} / MW_{dp}\right)}{1 + 20231 \cdot \left(Q_{i,dp} / MW_{dp}\right)} \qquad \forall i \in CE \qquad (6.11)$$

$$KD_{HI,dp} = -12.14 \cdot H^2_{dp} + 12.07 \cdot H_{dp} - 1.74 \qquad (6.12)$$

Concentration factor constraints

$$DF_{ip} = \left[\left(KD_{i,dp} - KD_{ip}\right)^2 + \varepsilon^2\right]^{1/2} \qquad \forall i, p \ne dp \qquad (6.13)$$

$$CF_{ip} = \frac{3.722}{3.727 + 0.579 \cdot e^{\left(54.410 \cdot DF_{ip} - 2.176\right)}} + 0.019 \qquad \forall i \in IE, p \ne dp \qquad (6.14)$$

$$CF_{HI,p} = \frac{3.937}{3.933 + 0.105 \cdot e^{\left(36.005 \cdot DF_{HI,p} - 0.299\right)}} + 0.018 \qquad \forall p \ne dp \qquad (6.15)$$

$$CF_{i,dp} = 1 \qquad \forall i \qquad (6.16)$$

Process synthesis constraints

$$m_{1p} = CF_{1p} \cdot w_1 \cdot m_{0p} + m_{0p} \cdot (1 - w_1) \qquad \forall p \qquad (6.17)$$

$$m_{ip} = CF_{ip} \cdot m^1_{i-1,p} + m^2_{i-1,p} \qquad \forall p, i = 2,...,I \qquad (6.18)$$

$$m_{i-1,p} = m^1_{i-1,p} + m^2_{i-1,p} \qquad \forall p, i = 2,...,I \qquad (6.19)$$

$$0 \le m^1_{i-1,p} \le m_{0p} \cdot w_i \qquad \forall p, i = 2,...,I \qquad (6.20)$$

$$0 \le m^2_{i-1,p} \le m_{0p} \cdot (1 - w_i) \qquad \forall p, i = 2,...,I \qquad (6.21)$$

$$m_{I,dp} \ge SP \cdot \sum_p m_{I,p} \qquad (6.22)$$

Additional constraints

$$x_{i,dp}, w_i \in \{0,1\} \qquad \forall i \in IE \qquad (6.23)$$

$$n_k \in Z^+ \qquad \forall k \qquad (6.24)$$

$$CF_{ip}, DF_{ip}, H_{dp}, KD_{i,dp}, m_{ip}, Q^+_{i,dp}, Q^-_{i,dp}, r_k \geq 0 \qquad \forall i, p, k \qquad (6.25)$$

6.2.2.2. Process synthesis constraints revisited

Process synthesis constraints (6.17) - (6.22) are used to calculate mass m_{ip} of protein p after each chromatographic step i. Constraint (6.18) in particular, presents a serious difficulty in the effort of linearising the model, because of the multiplication of CF_{ip} and $m^1_{i-1,p}$, which are both positive continuous variables. But in practice, because the various chromatographic steps are treated as "black-boxes", there is no direct need to calculate all intermediate masses after each chromatographic step as constraints (6.17) - (6.22) do; the only one that we are interested in is the final mass of protein p after the last chromatographic technique ($m_{I,p}$). A transformation of the process synthesis constraints can be applied to alleviate the problem.

Mass $m_{I,dp}$ of protein product dp after the last chromatographic step I must meet a specified purity level, SP. Since it is assumed that the separation is performed without product loss (see section 5.2), the final product mass $m_{I,dp}$ is constant and equal to the initial mass $m_{0,dp}$. From constraint (6.22), we have:

$$m_{I,dp} \geq SP \cdot \sum_p m_{I,p} \Rightarrow (1-SP) \cdot m_{0,dp} \geq SP \cdot \sum_{p \neq dp} m_{I,p} \qquad (6.26)$$

The mass, $m_{I,p}$, of each contaminant protein p remaining after the final chromatographic technique I can be calculated from the initial contaminant mass $m_{0,p}$ of each contaminant by:

$$m_{I,p} = m_{0,p} \cdot \prod_{i=1}^{I} \overline{CF}_{ip} \qquad \forall p \neq dp \qquad (6.27)$$

where
$$\begin{aligned} \overline{CF}_{ip} &= CF_{ip}, \quad \text{if} \quad w_i = 1 \\ \overline{CF}_{ip} &= 1, \quad \text{if} \quad w_i = 0 \end{aligned} \qquad \forall i, p \neq dp$$

Variables $\overline{CF}_{ip}$ can be expressed as an exponential function of concentration factors CF_{ip} and decision variables w_i:

$$\overline{CF}_{ip} = e^{(\ln CF_{ip})\cdot w_i} \text{ or } \overline{CF}_{ip} = e^{\xi_{ip}} \qquad \forall i, p \neq dp \qquad (6.28)$$

where $\xi_{ip} \equiv (\ln CF_{ip}) \cdot w_i$. Therefore, using expressions (6.27) and (6.28), purity constraint (6.26) can now be expressed as:

$$(1 - SP) \cdot m_{0,dp} \geq SP \cdot \sum_{p \neq dp} m_{0,p} \cdot e^{\sum_i \xi_{ip}} \qquad (6.29)$$

$$\xi_{ip} = (\ln CF_{ip}) \cdot w_i \qquad \forall i, p \neq dp \qquad (6.30)$$

6.2.2.3. New MINLP formulations

Constraints (6.29) and (6.30) substitute constraints (6.17) - (6.22) in the new MINLP models that are now formulated. Problem P3 is the equivalent of Problem P1 from chapter 5, and Problem P4 is the equivalent of Problem P2. The two problems constitute a two-stage solution procedure to identify the shortest amino acid sequence that can produce the optimal flowsheet for a purification process.

Problem P3

$$\textit{minimise} \sum_i w_i \qquad (6.31)$$

subject to:

constraints (6.2)-(6.16),(6.23)-(6.25),(6 .29) and (6.30)

Problem P4

$$\textit{minimise} \sum_k n_k \qquad (6.32)$$

subject to:

constraints (6.2)-(6.16),(6.23)-(6.25),(6 .29), (6.30) and

$$\sum_i w_i \leq i^* \tag{6.33}$$

where i^* is the identified minimum required number of chromatographic techniques.

6.2.2.4. Solution of problems P3 and P4

Problems P3 and P4 were solved with NEOS Server 4.0 (http://www-neos.mcs.anl.gov/; Gropp and Moré, 1997; Czyzyk *et al.*, 1998; Dolan, 2001) using the SBB solver and CONOPT3 as the NLP solver for the solution of Examples 1 and 2 from chapter 5. The results produced were identical to the ones from the previous chapter (problems P1 and P2). All results are presented in Table 6.1, which should be read in comparison with Table 5.6.

Table 6.1: Summary of computational statistics for Problems P3 and P4.

example	model	max aas	max steps	discrete/continuous variables	constraints	CPU time (s)	objective value
1	P3	≤ 0	-	41/162	175	10.87	4 [a]
1	P3	≤ 6	-	41/162	175	16.21	3 [a]
1	P4	≤ 6	≤ 3	41/162	176	20.37	2 [b]
2	P3	≤ 0	-	41/459	472	34.44	6 [a]
2	P3	≤ 6	-	41/459	472	39.72	4 [a]
2	P4	≤ 6	≤ 4	41/459	473	30.42	4 [b]

[a] Number of chromatographic steps

[b] Number of amino acids in the tag

The problem size was reduced considerably, and in many cases, in particular for the larger example, there was a decrease in CPU time as well. This improvement may not be enough to justify the transformation, but this was not the motivation behind the new formulations; the goal as explained in section 6.2.2.2 was to overcome a major difficulty in the linearisation of the models, before we apply the piecewise linear approximations to the non-convex functions.

6.2.3. An MILP approach

Next, the proposed MILP representation, designed for the synthesis of purification bioprocesses so as to consider the optimal design of purification tags, is described in detail. The proposed representation extends the formulation of Problem P1 from section 5.3.7.

6.2.3.1. Peptide tag size constraints

Constraints (6.2) and (6.3) are linear and can be used here unchanged. A maximum number of 6 is imposed on the number of amino acids that can be present in the peptide tag, so as to avoid structural interference. At the same time, hydrophobic amino acids should be balanced by polar residues so that the tag is soluble and does not bury itself within the protein.

6.2.3.2. Physicochemical property constraints

The tagged protein's net charge (Q_{dp}) is predicted based on the methodology suggested by Mosher *et al.* (1993). In order to avoid the calculation of the charge of the tagged protein product for all chromatographic steps irrespective of whether they are selected in the final flowsheet or not (as is the case in the formulation of Problems P3 and P4 above), a slack variable is introduced to constraint (6.4). The slack variable is forced to zero when the corresponding chromatographic step i is selected ($w_i = 1$), allowing the proper estimation of the property. When step i is not selected, constraint (6.34) is relaxed.

$$Q_{i,dp} = \hat{Q}_{i,dp} + \sum_{k \in BA} \frac{n_k}{\frac{K_k}{[H^+]_i} + 1} - \sum_{k \in AA} \frac{n_k}{\frac{[H^+]_i}{K_k} + 1} + sl_i \qquad \forall i \in IE \qquad (6.34)$$

$$-M \cdot (1 - w_i) \leq sl_i \leq M \cdot (1 - w_i) \qquad \forall i \in IE \qquad (6.35)$$

where M is an appropriate large positive number. Values for the ionisation constants K_k (Mosher *et al.*, 1993) are presented in Table 5.1.

The tagged protein's hydrophobicity (H_{dp}) is estimated using the work by Lienqueo *et al.* (2002). Values for the normalised hydrophobicity and the surface areas of fully exposed amino acids (Lienqueo *et al.*, 2002) are presented in Table 5.2. Constraint (6.5) remains unchanged. Additionally, an extra assumption is made for the product surface of the tagged product protein. It is the same assumption as the one made in chapter 5 about the molecular weight, *i.e.* that the total surface does not change significantly from the addition of a very small number of amino acids to the chain of the protein, *i.e.* $\hat{S}_{dp} + \sum_{k'} s_{k'} \cdot n_{k'} \approx \hat{S}_{dp}$. Constraint (6.6) changes accordingly:

$$r_k = \frac{s_k \cdot n_k}{\hat{S}_{dp} + \sum_{k'} s_{k'} \cdot n_{k'}} \Rightarrow r_k = \frac{s_k \cdot n_k}{\hat{S}_{dp}} \qquad \forall k \qquad (6.36)$$

6.2.3.3. Dimensionless retention time constraints

Dimensionless retention times (KD_{ip}) are defined as a function of a physicochemical property of the product protein, $P_{i,dp}$ (either net charge, $Q_{i,dp}$, or hydrophobicity, H_{dp}). For ion exchange chromatography, retention times for the tagged protein product are estimated based on approximations of the chromatograms by isosceles triangles and on physicochemical property data for the product and contaminants (Vasquez-Alvarez *et al.*, 2001). The methodology presented by Lienqueo *et al.* (2002) is used to estimate the dimensionless retention times for hydrophobic interaction ($KD_{HI,p}$). Both relationships between retention time and physicochemical property are non-linear; so a piecewise linear approximation is used for their linearisation, as described by constraints (6.37)-(6.40):

$$KD_{i,dp} = \sum_j \alpha_{ij} \cdot \lambda_{ij} \qquad \forall i \qquad (6.37)$$

$$P_{i,dp} = \sum_j \beta_{ij} \cdot \lambda_{ij} \qquad \forall i \qquad (6.38)$$

$$\sum_j \lambda_{ij} = w_i \qquad \forall i \qquad (6.39)$$

$$\lambda_{ij} \in \text{SOS2} \qquad \forall i, j \qquad (6.40)$$

where j is a special set used for the piecewise linear approximation of $KD_{i,dp}$, a_{ij} and β_{ij} are appropriate parameters that help the calculation of $KD_{i,dp}$ for anion exchange chromatography, cation exchange chromatography and hydrophobic interaction, and λ_{ij} is a Special Order Set of type 2 variable (SOS2 variable) for the piecewise linear approximation of $KD_{i,dp}$. At most two variables within a SOS2 set can have non-zero values and they have to be adjacent.

Notice that in constraint (6.39), the summation is equal to decision variable w_i, instead of being equal to 1. This is to prevent the calculation of retention times whenever chromatographic step i is not selected (*i.e.* $w_i = 0 \Rightarrow KD_{i,dp} = 0$).

For anion and cation exchange chromatography, the piecewise linear approximations are presented in Figure 6.1. For hydrophobic interaction, the piecewise linear approximation is described by Figure 6.2.

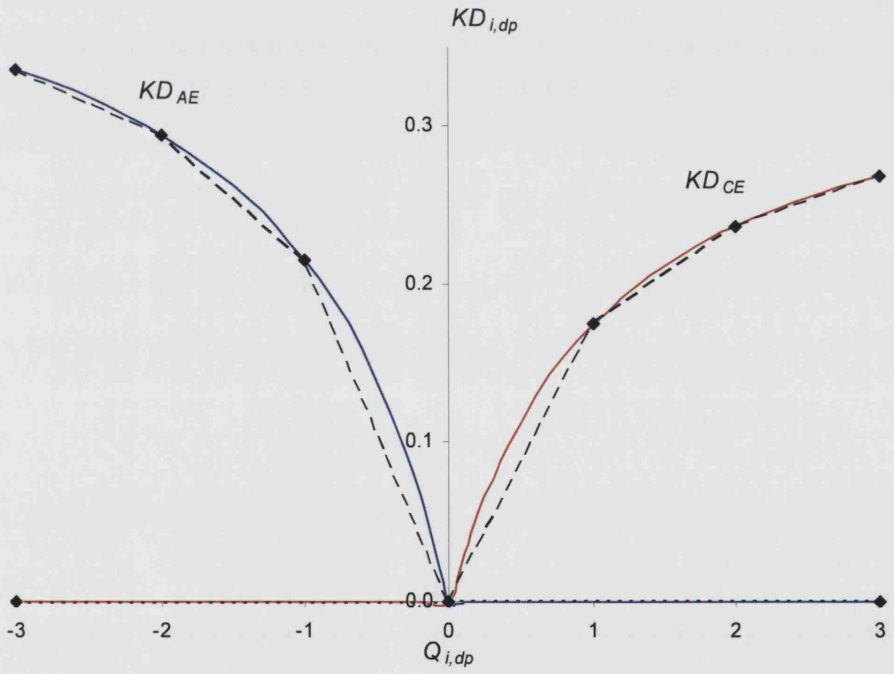

Figure 6.1: Piecewise linear approximations of retention times for ion exchange chromatography (AE: anion exchange; CE: cation exchange).

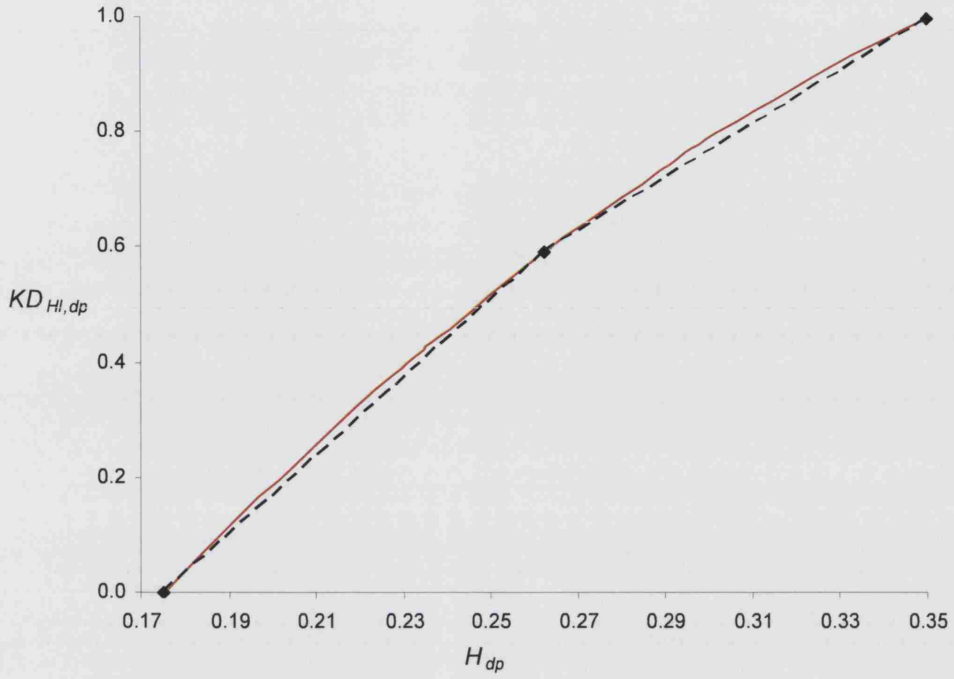

Figure 6.2: Piecewise linear approximation of retention time for hydrophobic interaction (HI).

6.2.3.4. Deviation factor constraints

Deviation factors, DF_{ip}, indicate the distance between the protein product's chromatographic peak and a contaminant's chromatographic peak. In particular, they are defined as the difference between the dimensionless retention times of the product and each contaminant p for each particular chromatographic step i. A new approach to the calculation of DF_{ip} is needed, to avoid the non-linearities of constraint (6.13):

$$DF_{ip}^{+} - DF_{ip}^{-} = KD_{i,dp} - KD_{ip} \qquad \forall i, p \neq dp \qquad (6.41)$$

$$DF_{ip}^{+} \leq M_2 \cdot y_{ip} \qquad \forall i, p \neq dp \qquad (6.42)$$

$$DF_{ip}^{-} \leq M_2 \cdot (1 - y_{ip}) \qquad \forall i, p \neq dp \qquad (6.43)$$

$$DF_{ip} = DF_{ip}^{+} + DF_{ip}^{-} - KD_{ip} \cdot (1 - w_i) \qquad \forall i, p \neq dp \qquad (6.44)$$

$$y_{ip} \leq w_i \qquad \forall i, p \neq dp \qquad (6.45)$$

where M_2 is an appropriate large positive number. Binary variables y_{ip} express whether the difference $KD_{i,dp} - KD_{ip}$ is positive or negative. The *absolute* value of the difference is assigned to either DF_{ip}^{+} or DF_{ip}^{-} with constraint (6.41), because either DF_{ip}^{+} or DF_{ip}^{-} always has to be equal to zero due to constraints (6.42) and (6.43). When $w_i = 1$, DF_{ip} is assigned the correct value (either DF_{ip}^{+} or DF_{ip}^{-}) with equation (6.44). When $w_i = 0$, constraint (6.45) forces y_{ip} (and therefore DF_{ip}^{+}) to zero and, because of constraint (6.41), $DF_{ip}^{-} = KD_{ip}$ ($KD_{i,dp}$ is equal to zero when $w_i = 0$). Therefore, from equation (6.44), we have $DF_{ip} = KD_{ip} - KD_{ip} = 0$.

6.2.3.5. Concentration factor constraints

Deviation factors are used to calculate concentration factors, CF_{ip}, which represent the ratio of the mass of contaminant p after chromatographic step i to the mass of contaminant p before step i. The relationship between deviation factors and concentration factors is also non-linear (see section 5.3.2.4), so another piecewise linear approximation is needed, as presented in Figure 6.3 and described by constraints (6.46)-(6.49). Deviation factors (DF_{ip}) are correlated directly with the logarithm of concentration factors CF_{ip} ($\ln CF_{ip}$), which in constraint (6.47) has been substituted with new variable ξ_{ip} in order to maintain the linearity of the model.

$$DF_{ip} = \sum_l \gamma_{ipl} \cdot \mu_{ipl} \qquad \forall i, p \neq dp \qquad (6.46)$$

$$\xi_{ip} = \sum_l \delta_{ipl} \cdot \mu_{ipl} \qquad \forall i, p \neq dp \qquad (6.47)$$

$$\sum_l \mu_{ipl} = w_i \qquad \forall i, p \neq dp \qquad (6.48)$$

$$\mu_{ipl} \in \text{SOS2} \qquad \forall l, i, p \neq dp \qquad (6.49)$$

where l is a special set used for the piecewise linear approximation of DF_{ip}, γ_{ipl} and δ_{ipl} are appropriate parameters that help in the calculation of DF_{ip}, and μ_{ipl} is a SOS2 variable for the piecewise linear approximation of DF_{ip}. Notice that when $DF_{ip} = 0$, it follows that $\ln CF_{ip} = 0$ and therefore $CF_{ip} = 1$, and that is what we want to accomplish when chromatographic step i is not selected.

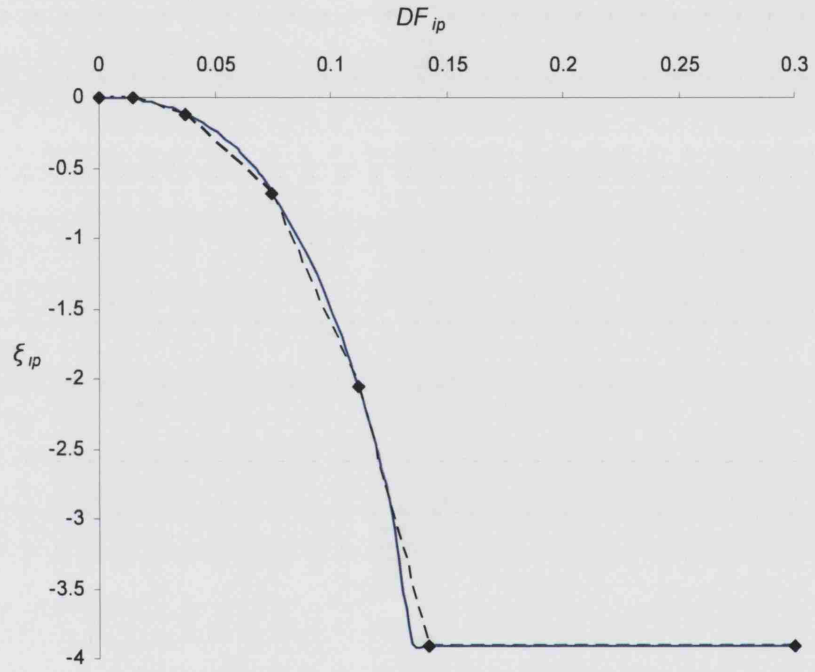

Figure 6.3: Piecewise linear approximation for concentration factors CF_{ip}.

6.2.3.6. Process synthesis constraints

Constraint (6.29) is used in place of the old purity constraint. Expression (6.29) incorporates the non-linear term $e^{\sum_i \xi_{ip}}$, which also needs to be linearised with a piecewise linear approximation (Figure 6.4). To accomplish this, factor $e^{\sum_i \xi_{ip}}$ is substituted with a new variable, $\overline{\mathrm{E}\xi}_p$ ($\overline{\mathrm{E}\xi}_p \equiv e^{\sum_i \xi_{ip}}$).

$$\sum_i \xi_{ip} = \sum_m \zeta_m \cdot \nu_{pm} \qquad \forall p \neq dp \qquad (6.50)$$

$$\overline{\mathrm{E}\xi}_p = \sum_m e^{\zeta_m} \cdot \nu_{pm} \qquad \forall p \neq dp \qquad (6.51)$$

$$\sum_m \nu_{pm} = 1 \qquad \forall p \neq dp \qquad (6.52)$$

$$\nu_{pm} \in SOS2 \qquad \forall m, p \neq dp \qquad (6.53)$$

where m is a special set used for the piecewise linear approximation of $\overline{\mathrm{E}\xi}_p$, ζ_m is an appropriate parameter for the calculation of $\overline{\mathrm{E}\xi}_p$, and ν_{pm} is a SOS2 variable for the piecewise linear approximation of $\overline{\mathrm{E}\xi}_p$.

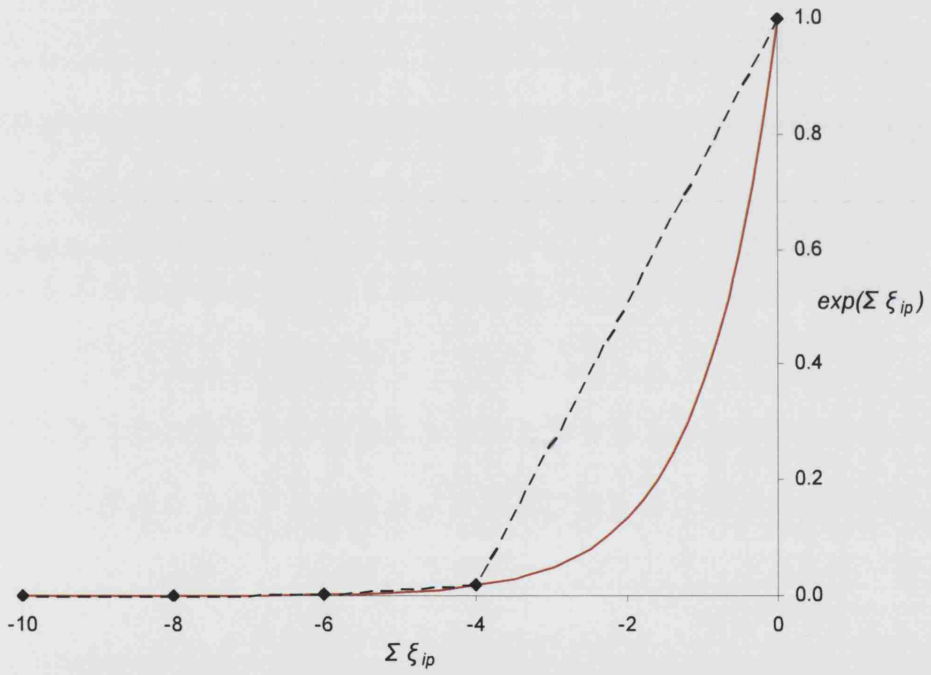

Figure 6.4: Piecewise linear approximation for exp(ξ_{ip}).

The approximation of Figure 6.4 begins with a very crude linear piece, because there are no observed occurrences of $\sum_i \ln CF_{ip}$ with a value higher than -4. After that point the approximation becomes much more detailed, because a very high accuracy is required in order to correctly approximate the values of $e^{\sum_i \xi_{ip}}$.

Finally, equation (6.29) is re-written using new variable $\overline{\mathrm{E}\xi}_p$ in place of the non-linear term $e^{\sum_i \xi_{ip}}$.

$$(1-SP)\cdot m_{0,dp} \geq SP\cdot \sum_{p\neq dp}(m_{0,p}\cdot \overline{\mathrm{E}\xi}_p) \tag{6.54}$$

6.2.3.7. Objective function

The overall problem is formulated as an MILP model in order to identify the chromatographic techniques and the shortest amino acid sequence that can produce the optimal flowsheet of the purification process. The objective is to minimise the total number of selected chromatographic steps i in the purification process and, using the penalty parameter c, to force the model to select the minimum number of amino acids n_k in the tag.

Problem P5

$$\textit{minimise} \sum_i w_i + c\cdot \sum_k n_k \tag{6.55}$$

subject to:

peptide tag size constraints (6.2) and (6.3);

physicochemical property constraints (6.5), (6.34)-(6.36);

dimensionless retention time constraints (6.37)-(6.40);

deviation factor constraints (6.41)-(6.45);

concentration factor constraints (6.46)-(6.49);

process synthesis constraints (6.50)-(6.54); and

$$y_{ip}, w_i \in \{0,1\} \qquad \forall i, p\neq dp \tag{6.56}$$

$$n_k \in \mathrm{Z}^+ \qquad \forall k \tag{6.57}$$

$$DF_{ip}, DF_{ip}^{+}, DF_{ip}^{-}, \overline{E\xi}_{p}, H_{dp}, KD_{i,dp}, r_k \geq 0 \qquad \forall i,p,k \qquad (6.58)$$

$$\xi_{ip} \leq 0 \qquad \forall i,p \qquad (6.59)$$

To reduce the combinatorial nature of the problem, Problem P5 is solved using a 2-step procedure. First, the MILP model is solved without the use of a peptide tag for the purification of protein *dp* (*i.e.* $N = 0$ in constraint (6.2)), and a set of chromatographic techniques is obtained. Then, the MILP is solved again with a tag fused to the product protein; but this time the candidate chromatographic techniques *i* are chosen only among those selected in the first step of the solution procedure.

6.3. Computational results

Solutions were obtained with GAMS (Brooke *et al.*, 1998), using the CPLEX 6.5 solver. All computational experiments were performed on an IBM RS6000 workstation. The methodology was tested with a four-protein mixture: thaumatin (*dp*), conalbumin (*p1*), chymotripsinogen A (*p2*) and ovalbumin (*p3*). The physicochemical properties of the mixture are presented in Table 6.2.

Table 6.2: Physicochemical properties of protein mixture.

Protein	$m_{0,p}$	MW_p	H_p	$Q_{ip} \times 10^{-17}$ (*C/molecule*)				
	(mg/mL)	(Da)		pH 4.0	pH 5.0	pH6.0	pH7.0	pH8.0
dp	2	22200	0.27	1.60	1.57	1.64	1.55	0.75
p1	2	77000	0.23	0.93	0.33	-0.12	-0.34	-0.50
p2	2	23600	0.31	2.15	1.46	1.17	0.78	0.38
p3	2	43800	0.28	1.16	-0.63	-1.36	-1.82	-1.95

The purity level required for the desired product (*dp*) is 98%. There are 11 available chromatographic steps: anion exchange chromatography (AE) at pH 4, pH 5, pH 6, pH 7, pH 8, cation exchange chromatography (CE) at pH 4, pH 5, pH 6, pH 7, pH 8 and hydrophobic interaction (HI). From these, CE pH 6, CE pH 7, CE pH 8 and HI

are needed for the purification without the use of a peptide tag fused to protein *dp*, which achieves a product purity of 98.1%. The solution is significantly improved with a tag of 3 lysine residues; a purity of 98.1% can be achieved with only three separation steps: CE pH 7, CE pH 8 and HI. The resulting mathematical model involves 151 constraints, 36 discrete variables and 211 continuous variables and was solved in 5.92 seconds. The results are illustrated in Figure 6.5.

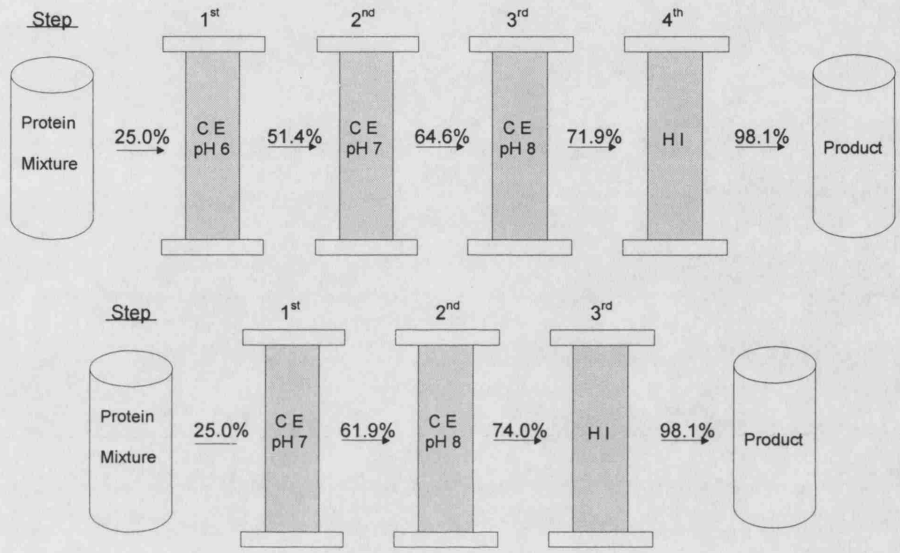

Figure 6.5: Optimal result for protein mixture with no tag and with a tag of 3 lysines.

The MILP solution is almost identical to the one provided by the MINLP model presented in chapter 5, which selected the same 3 chromatographic steps and also a tag with lysines only. The selection of a peptide tag that only contains lysine amino acids implies that the increase of the product charge benefits the purification and that a hydrophobicity increase would be detrimental. Even though there are amino acids with a stronger effect on charge than lysine, they would increase hydrophobicity as well, which remains unchanged when lysine is used.

The disparity in the results produced from the MILP and MINLP models (3-lysine-tag and 2-lysine-tag respectively) can be considered of no serious consequence and attributed to small differences in the approximations of the two models. In the MINLP solution, 2 lysine amino acids are just enough to get the purity of the protein

product to exactly 98% (the specified purity); in the MILP solution, 2 lysines fall just short of that (97.95%), so a third one is necessary to "push" the protein's charge enough in order for the required purity is achieved. That is also the reason why the MILP solution produces a slightly higher level of purity (98.1%) than the MINLP solution. The salient conclusions of the two studies though remain the same: significant improvements in the purification of proteins can be achieved with the use of very short peptide tags, the flowsheet of the example process was reduced by one chromatographic step, and the purification was shown to benefit by the modification of the charge of the product protein only, as any alteration of hydrophobicity was proven detrimental to the result.

6.4. Conclusions

An optimisation framework for the simultaneous selection of optimal peptide tags and the synthesis of chromatographic steps for the purification of protein mixtures in downstream protein processing has been presented. The framework was formulated as an MILP mathematical model, developed from a previous MINLP model, presented in chapter 5, through piecewise linear approximations of non-linear functions. The methodology was validated through its application to an example protein mixture involving 3 contaminants and a set of 11 candidate chromatographic steps. Results were indicative of the benefits of peptide tags in purification processes.

The results and conclusions produced from the application of the MILP formulation were the same as the ones of the previous chapter. In that respect, this study did not offer any new insight, but helped to confirm the lessons learned from the application of the MINLP formulation. Nevertheless, the main benefits of this work are the mathematical improvements that were integrated in the model. Three significant enhancements were incorporated: i) the transformation of the process synthesis constraints with a logarithmic function, which helped avoid the unnecessary estimation of the mass of contaminant proteins in all intermediate stages of the purification; ii) the prevention of the estimation of all relevant variables when a

chromatographic technique is not selected for the purification; and iii) the linearisation of the non-linear functions of the model with piecewise linear approximations.

Chapter 7

Conclusions and future directions

The aim of this thesis was *to facilitate biological studies by applying mathematical programming techniques to problems of biochemical nature*. Towards that goal, a number of mathematical models and solution algorithms have been developed in order to assist biologists in the analysis and quantification of some distinctive problems that biology is faced with. The key contributions of the thesis are summarised in the next section, while section 7.2 suggests promising new directions for future research work.

7.1. Contributions of the thesis

7.1.1. Analysis of biological networks

Part I of the thesis was concerned with the study of biological networks. A single-source, shortest path algorithm formulated as an LP model was presented, capable of calculating the shortest path from one node in a network to every other. The algorithm is characterised by its simplicity and deals efficiently with network

circularity. A literature survey was conducted in order to familiarise the reader with the current status of research and highlight recent academic contributions in the area. We focused on two specific biochemical networks: the metabolic network of *E. coli* and the p53 apoptotic control protein network.

In order to contribute to the search for evolutionary relationships among the *E. coli* enzymes, the correlations between pathway distance and genome distance and also between pathway distance and enzyme function for the SMM pathways of *E. coli* were investigated. A demonstrable relationship between pathway distance and genome distance was observed, with genes nearby in the genome far more likely to encode nearby enzymes in a metabolic pathway. No clear trend was observed when examining the relationship between pathway distance and conservation of function. This lack of obvious correlation, along with evidence from other metabolic studies concerning sequence and structural similarity of SMM enzymes, supports the theory of a patchwork model of pathway evolution: enzymes were almost randomly recruited on a need-only basis within the metabolic network of an organism.

The second biochemical network that has been investigated was the p53 cell cycle and apoptosis control network. The diameter of the network, which was used as a proxy of network navigability, was calculated using the LP shortest path algorithm. Two modes of attack (one random and one directed) were inflicted on the network, to study the response of its diameter. The p53 network was observed to be inherently robust to random knockouts of its proteins, which equates to resilience against mutational perturbation. This robustness was found to be a result of the structure of the network itself; however, the reliance on highly-connected nodes also makes the network vulnerable to the loss of its hubs. Tumour inducing viruses exploit this very weakness in order to disrupt the p53 network by targeting specific proteins. This study has identified the same proteins as the network hubs, which explains why tumour inducing viruses are so effective.

The applicability of the presented LP technique was illustrated by its application to the two aforementioned case studies. The LP algorithm is a fast and effective method of analysing certain properties of biochemical networks and was proven to be a valuable analysis tool for such complex systems.

7.1.2. Protein structure prediction

The prediction of the way proteins fold was the object of the second part of the thesis. An extensive literature survey was performed in order to give the reader an idea of the research in an area of biology that accumulates a huge amount of interest. An optimisation framework for positioning amino acids in unique positions of a three dimensional cubical lattice was then developed. The framework utilises only the knowledge of the amino acid sequence and the contact energies among amino acids.

The overall problem was formulated as an MILP model and a three-step solution procedure was proposed in order to simplify the extremely complex task of protein structure prediction. The three steps of the proposed methodology were: first, identify small elementary structures, which are created from a small number of residues close-by on the amino acid chain and are remarkably stable; then, using the developed MILP model, we optimally position the elementary structures relatively to each other to form a folding core; and finally, the 3D structure of the small protein is predicted by optimally placing the rest of the residues of the amino acid chain around the folding nucleus, again by applying the developed MILP model. The applicability of the optimisation-based framework was successfully demonstrated with two illustrative examples: a lattice-designed protein with 27 residues and a second, larger protein with 36 residues in its amino acid chain. Therefore, the proposed strategy can efficiently identify the native conformation of proteins with sizes of at least up to 36 monomers.

7.1.3. Chromatographic purification of proteins using peptide tags

Part III of the thesis was concerned with the purification of proteins during downstream processing in biochemical plants, and the optimisation of the flowsheet of the process through manipulation of the properties of the product protein with the use of peptide tags. Initially, current approaches in the area of downstream biotechnology were highlighted. Based on a recently developed MILP framework (Vásquez-Alvarez and Pinto, 2001) for the synthesis of purification bioprocesses, we then presented a systematic MINLP approach, which selects a tag that modifies the

properties of the protein product in the most beneficial way and concurrently minimises the number of chromatographic steps in the purification process. The framework employed a two-stage procedure that can be extended to consider application to larger examples, use of additional chromatographic techniques, or manipulation of other physicochemical properties. Two illustrative examples were solved so as to validate the efficiency of the proposed methodology. Small peptide tags were selected, which allowed an important decrease in the number of purification steps required for the purification. Results were indicative of the benefits of the application of optimisation-based techniques for the use of purification tags in biotechnological production plants, and have provided a useful guideline for both downstream process synthesis and optimal tag design.

An MILP mathematical model for the simultaneous selection of optimal peptide tags and the synthesis of chromatographic steps for the purification of protein mixtures in downstream protein processing was also developed. The framework was formulated based on the previously presented MINLP model and the main benefits of this work were the mathematical improvements that were integrated in the MILP model. The process synthesis constraints were transformed with a logarithmic function, which considerably reduced the size of the problem by avoiding the estimation of the mass of contaminant proteins in intermediate stages of the purification. Appropriate constraints also helped avoid the estimation of all relevant variables when a chromatographic step is not selected for the purification. Finally, the non-linear functions of the MINLP model were linearised with piecewise linear approximations. The methodology was validated through its application to an example protein mixture involving 3 contaminants and a set of 11 candidate chromatographic steps.

7.2. Recommendations for future work

A number of promising future research directions related to the application of mathematical programming to biochemical systems are presented in this section. The aim is to provide the reader with some future insight in the area as well as highlight a

number of emerging research issues that could benefit from the developed mathematical modelling frameworks presented in the thesis. Next, we consider those future research issues in detail.

7.2.1. Biological networks

Chapters 2 and 3 present the application of an LP shortest path algorithm to the analysis of two biological networks. The algorithm is thus proven to be a valuable analytical tool for the examination of the structure of biological networks and can be easily applied in the future to the study of other networks.

7.2.1.1. Network directionality

The biological networks studied in this thesis were both considered to be undirectional to simplify the process, but also because consideration of direction was not applicable in the case studies: for the study of *E. coli* metabolism we were measuring proximity, whereas for the p53 network only 5% of the interactions were directed. Nevertheless, the LP algorithm is fully capable of dealing with network directionality: it would be very interesting to apply the algorithm to the analysis of systems such as gene regulatory networks, where directionality plays a crucial part.

7.2.1.2. Metabolic pathways

In the case of the study of metabolic networks, conclusions regarding the evolution of metabolism were drawn. A stronger case would require the study of sequence, structural similarity and homology of SMM enzymes, similar to the work of other researchers in the field (Tsoka and Ouzounis, 2001; Teichmann *et al.*, 2002; Rison *et al.*, 2002). Such investigations go beyond the scope of this thesis, but would provide fertile ground for future research. The LP algorithm can also be applied to the analysis of metabolic networks of organisms other than *E. coli*, provided of course that reliable metabolic data are available.

7.2.1.3. p53 network

In the case of the p53 cell cycle and apoptotic control network, the application of the LP algorithm has provided insight into the mode of attack utilised by tumour-inducing viruses upon the apoptotic control system. We have used a connectionist model of the network with a concentrated focus on signalling networks and a specific and well-understood function. As more data regarding the p53 network become available in the future, it will be possible to extend the network model in order to attach directions and strength values to the connections, to make accurate predictions about the importance of individual nodes and edges. This will allow comparative analyses of how and why the variable dynamic network components operate under different evolutionary strains and cell type conditions. The LP framework presented here represents the first step in this exciting process.

7.2.2. Protein folding

Chapter 4 presents a mathematical programming framework for the prediction of protein structure. The developed MILP formulation is flexible and can easily be extended to consider larger examples, different kinds of lattices, and different sets of interactions (*i.e.* different energy functions from the 20-letter one used here).

7.2.2.1. Face-centred-cubic lattice

A promising direction is the application of the MILP model to different lattice configurations. The methodology can be adjusted to work with other kinds of lattices, such as the face-centred-cubic (FCC) lattice, which present certain advantages over the simple cubic lattice. The FCC lattice in particular can model real protein conformations with good quality (Park and Levitt, 1995). It was shown (Bagci *et al.*, 2002) that the FCC lattice closely approximates the positions of amino acids in the folded conformation of real proteins, making the FCC very suitable for modelling proteins.

7.2.2.2. De novo protein design

The application of our MILP model to the engineering of proteins with specific properties can be considered. In the so-called "inverse protein folding problem" or *de novo* protein design (Drexler, 1981; Pabo, 1983), the goal is to determine an amino acid sequence that folds into a given three-dimensional structure. As structure is typically equated to function, the design of a protein that would fold to a pre-specified conformation would mean that we would be able to design a protein with the exact desired attributes.

7.2.3. Purification tags in downstream protein processing

Chapters 5 and 6 present mathematical programming frameworks for the concurrent design of optimal peptide tags and the synthesis of downstream protein processing in biochemical production plants. Testing the mathematical framework with larger examples and investigating additional examples of protein purification is a possible extension to this work. Also, it would be very interesting to be able to validate the generated hypotheses by evaluating experimentally the chromatographic behaviour of the product proteins with the fused peptide tags.

7.2.3.1. Additional chromatographic techniques

Another possible research direction would be the consideration of additional types of chromatographic steps, such as gel filtration, ion exchange chromatography with pH gradient, or affinity chromatography, which would potentially provide the same degree of purification but using a smaller number of steps. The main constraint for the extension of the methodology in order to consider other chromatographic techniques is the availability of the appropriate correlations and the lack of reliable prediction methods for the physicochemical properties of the proteins.

7.2.3.2. Model enhancements

The modelling of the purification can be extended to incorporate a number of further issues, including sequencing of the purification steps (Vásquez-Alvarez and Pinto, 2001), product loss (Vásquez-Alvarez and Pinto, 2003) and/or application of

alternative performance criteria. In particular, the application of a financial objective function could benefit the optimisation, by improving the accuracy of the results. From a mathematical programming point of view, the aforementioned modification to the proposed model would also help reduce degeneracy of the solutions.

The possibility of product loss due to protein-protein interactions in any of the chromatographic steps and due to membrane steps for buffer exchange and/or protein concentration could also be considered, for example protein-protein interactions could cause the formation of aggregates or product-impurity binding, both of which would significantly reduce the purification achieved by the chromatography. A percentage of loss of protein product because of these effects can be introduced or more sophisticated representations of the chromatographic techniques that would account for the product loss could be developed.

7.3. Epilogue

In this thesis, we have presented some examples of mathematical programming applications to problems concerning biochemical systems. These paradigms illustrate the importance of integer optimisation as a solution tool for many problems related to systems biology and bioinformatics.

The implementation of process systems methodologies to biological problems is not always a straightforward procedure. Apart from the possible complexity of the mathematical formulations, detailed biological knowledge of the intricacies of the problem is also required. Furthermore, it is often the case that the available data are incomplete, or inconsistent, or plainly erroneous. Even though in recent years there has been an explosive increase in the availability of biological information, thanks to the development of new technologies, one should not forget that this information is still frequently amended and modified: metabolic reactions are often corrected or removed from/included in metabolic networks, new amino acid sequences and protein structures are continuously being added to databases, new genes are being

discovered, and an ever-expanding number of complete DNAs of organisms are being transcribed.

All the above obstacles demonstrate the need for further and continuous research in the application of mathematical programming to computational and systems biology. There remain many open questions about the problems discussed here, which could be an excellent starting point for future studies. Nevertheless, their number pales in comparison with the vast number of existing biological problems which could benefit greatly from efficient optimisation models and mathematical programming techniques. There is a world of opportunity out there that will surely remain a big challenge for researchers of the area for years to come.

Bibliography

Abkevich, V.I., Gutin, A.M. and Shakhnovich, E.I. (1994) Free energy landscape for protein folding kinetics: Intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052-6062.

Abkevich, V.I., Gutin, A.M. and Shakhnovich, E.I. (1995) Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460-471.

Albert, R., Jeong, H. and Barabási, A.L. (1999) The diameter of the World Wide Web. *Nature*, **401**, 130-131.

Albert, R., Jeong, H. and Barabási, A.L. (2000) Error and attack tolerance of complex networks. *Nature*,**406** , 378-382.

Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N. and Barabasi, A.L. (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**, 839-843.

Aloy, P., Stark, A., Hadley, C. and Russell, R.B. (2003) Predictions without templates: New folds, secondary structure, and contacts in CASP5. *Proteins* **53**, 436-456.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

Alves R., Chaleil, R.A., and Sternberg, M.J.E. (2002) Evolution of enzymes in metabolism: A network perspective. *J. Mol. Biol.* **310**, 311-325.

An, Y. and Friesner, R.A. (2002) A novel fold recognition method using composite predicted secondary structures. *Proteins* **48**, 352-366.

Androulakis, I.P. (2005) Selecting maximally informative genes. *Comput. Chem. Eng.* **29**, 535-546.

Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science* **181**, 223-230.

Antonov, A.V., Tetko, I.V., Mader, M.T., Budczies, J. and Mewes, H.W. (2004) Optimization models for cancer classification: Extracting gene interaction information from microarray expression data. *Bioinformatics* **20**, 644-U145.

Arita, M. (2000) Metabolic reconstruction using shortest paths. *Simulat. Pract. and Theory* **8**, 109-125.

Arita, M. (2004) The metabolic world of *Escherichia coli* is not small. *Proc. Natl. Acad. Sci. USA* **101**, 1543-1547.

Atkinson, B. and Mavituna, F. (1991) *Biochemical engineering and biotechnology handbook*, Macmillan Press, New York, USA.

Backofen, R. (1998) Using constraint programming for lattice protein folding. In *Proc. Pacific Symposium on Biocomputing*, pp. 389-400.

Backofen, R., Will S. and Bornberg-Bauer E. (1999) Application of constraint programming techniques for structure prediction of lattice proteins with extended alphabets. *Bioinformatics* **15**, 234-242.

Backofen, R. and Will S. (2003) A constraint-based approach to structure prediction for simplified protein models that outperforms other existing methods. *Lect. Notes Comput Sc.* **2916**, 49-71.

Bagci, Z., Jernigan, R.L. and Bahar, I. (2002) Residue packing in proteins: Uniform distribution on a coarse-grained scale. *J. Chem. Phys.* **116**, 2269-2276.

Banks, L., Pim, D. and Thomas, M. (2003) Viruses and the 26S proteasome: Hacking into destruction. *Trends Biochem. Sci.* **28**, 452-459.

Barabási, A.L. and Oltvai, Z.N. (2004) Network biology: Understanding the cell's functional organisation. *Nat. Rev. Genet.*, **5**, 101-113.

Bellman, R.E. (1958) On a routing problem. *Quart. Appl. Math.* **16**, 87-90.

Berger, B. and Leighton, T. (1998) Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. In *Proc. of the RECOMB '98*, 30-39.

Berggren, K., Wolf, A., Asenjo, J.A., Andrews, B.A. and Tjerneld, F. (2002) The surface exposed amino acid residues of monomeric proteins determine the partitioning in aqueous two-phase systems. *BBA-Protein Struct. M.* **1596**, 253-268.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1474.

Bornberg-Bauer, E. (1997) Chain growth algorithms for HP-type lattice models. In *Proc. 1st Annual International Conference on Computational Molecular Biology*, ACP Press, pp. 47-55.

Bowie, J., Luthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170.

Bradley, P., Chivian, D., Meiler, J., Misura, K.M.S., Rohl, C.A., Schief, W.R., Wedemeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C.E.M. and Baker, D. (2003) Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. *Proteins* **53**, 457-468.

Broglia, R.A., Tiana, G., Pasquali, S., Roman, H.E. and Vigezzi, E. (1998) Folding and aggregation of designed protein chains. *Proc. Natl. Acad. Sci. USA* **95**, 12930-12933.

Broglia, R.A. and Tiana, G. (2001a) Reading the three-dimensional structure of lattice model-designed proteins from their amino acid sequence. *Proteins* **45**, 421–427.

Broglia, R.A. and Tiana, G. (2001b) Hierarchy of events in the folding of model proteins. *J. Chem. Phys.***114** , 7267-7273.

Broglia, R.A., Tiana, G. and Provasi, D. (2004) Simple models of protein folding and of non-conventional drug design. *J. Phys.* **16**, R111-R144.

Brooke, A., Kendrick, D., Meeraus, A. and Raman, R. (1998) *GAMS: A user's guide*. GAMS Development Corporation, Washington, USA.

Brower, R.C., Vasmatzis, G., Silverman, M. and Delisi, C. (1993) Exhaustive conformational search and simulated annealing for models of lattice peptides. *Biopolymers* **33**, 329-334.

Burgard, A.P. and Maranas, C.D. (2001) Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.* **74**, 364-375.

Burgard, A.P. and Maranas, C.D. (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng.* **82**, 670-677.

Burgard, A.P., Nikolaev, E.V., Schilling, C.H. and Maranas, C.D. (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* **14**, 301-312.

Burgard, A.P., Pharkya, P. and Maranas, C.D. (2003) OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647-657.

Burgert, H.G., Ruzsics, Z., Obermeier, S., Hilgendorf, A., Windheim, M. and Elsing, A. (2002) Subversion of host defence mechanisms by adenoviruses. *Curr. Top. Microbiol. Immunol.* **269**, 273-318.

Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* **32**, 339-77.

Chang, Y. and Sahinidis, N.V. (2005) Optimization of metabolic pathways under stability considerations. *Comput. Chem. Eng.* , **29**, 467-479.

Chothia, C. (1975) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1-14.

Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein C. (2001) *Introduction to algorithms*. The MIT Press, Cambridge, MA, USA.

Creighton, T.E. (1993) *Proteins: Structures and molecular properties*. W.H. Freeman, New York, USA.

Crippen, G.M. (1991) Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry* **30**, 4232-4237.

Cui, Y., Wong, W.H., Bornberg-Bauer, E. and Chan, H.S. (2002) Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci. USA* **99**, 809-814.

Czaplewski, C., Liwo, A., Pillardy, J., Oldziej, S. and Scheraga, H.A. (2004) Improved conformational space annealing method to treat beta-structure with the UNRES force-field and to enhance scalability of parallel implementation. *Polymer* **45**, 677-686.

Czyzyk, J., Mesnier, M.P. and Moré J.J. (1998) The NEOS server. *IEEE Comput. Sci. Eng.***5** , 68-75.

Dasika, M.S., Gupta, A. and Maranas, C.D. (2004) A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks. *Pac. Symp. Biocomput.* **9**, 474-485.

Datar, R. (1986) Economics of primary separation steps in relation to fermentation and genetic engineering. *Process Biochem.* **21**, 19-26.

Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12**, 387-395.

Dill, K.A., Bromberg, K., Yue, K.M., Yee, D.P., Thomas, P.D. and Chan, H.S. (1995) Principles of protein folding – A perspective of simple exact models. *Protein Science* **4**, 561-602.

Dill, K.A. and Chan, H.S. (1997) From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10-19.

Dill, K.A., Fiebig, K.M. and Chan, H.S. (1993) Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci. USA* **90**, 1942-1946.

Dinner, A.R., Sali, A. and Karplus, M. (1996) The folding mechanism of larger model proteins: Role of native structure. *Proc. Natl. Acad. Sci. USA* **93**, 8356-8361.

Dolan, E. (2001) The NEOS server 4.0 administrative guide. In *Technical Memorandum ANL/MCS-TM-250*, Mathematics and Computer Science Division, Argonne National Laboratory, USA.

Drexler, K.E. (1981) Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. USA* **78**, 5275-5278.

Du, R., Pande, V.S., Grosberg, A.Y., Tanaka, T. and Shakhnovich, E.S. (1998) On the transition coordinate for protein folding. *J. Chem. Phys.* **108**, 334-350.

Edwards, J.S., Ibarra, R.U. and Palsson, B.O. (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotecnol.* **19**, 125-130.

Edwards, J.S. and Palsson, B.O. (2000a) The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA* **97**, 5528-5533.

Edwards, J.S. and Palsson, B.O. (2000b) Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnol. Prog.* **16**, 927-939.

Enzyme Nomenclature (1992) *Recommendations of the nomenclature committee of the international union of biochemistry (NC-IUB)*. Academic Press, San Diego, USA.

Evan, G.I., Lewis, G.K., Ramsay, G. and Bishop, J.M. (1985) Isolation of monoclonal antibodies specific for human c-myc proto-oncogene product. *Mol. Cell. Biol.* **5**, 3610-3616.

Fell, D.A. and Wagner, A. (2000) The small world of metabolism. *Nature Biotechnol.* **18**, 1121-1122.

Fexby, S. and Bulow, L. (2004) Hydrophobic peptide tags as tools in bioseparation. *Trends Biotechnol.* **22**, 511-516.

Fong, S.S. and Palsson, B.O. (2004) Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.* **36**, 1056-1058.

Forster, J., Famili, I., Fu, P., Palsson, B.O. and Nielsen, J. (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**, 244-253.

Fraga, E.S. (1998) The generation and use of partial solutions in process synthesis. *Chem. Eng. Res. Des.* **76**, 45-54.

Gerrard, J.A., Sparrow, A.D. and Wells, J.A. (2001) Metabolic databases – What next? *Trends. Biochem. Sci.* **26**, 137-140.

Ghosh, S., Zhu, T., Grossmann, I.E., Ataai, M.M. and Domach, M.M. (2005) Closing the loop between feasible flux scenario identification for construct evaluation and resolution of realized fluxes via NMR. *Comput. Chem. Eng.* **29**, 459-466.

Govindarajan, S. and Goldstein, R.A. (1997) Evolution of model proteins on a foldability landscape. *Proteins* **29**, 461-466.

Greenberg, H.J., Hart, W.E. and Lancia, G. (2004) Opportunities for combinatorial optimization in computational biology. *INFORMS J. Comput.* **16**, 211-231.

Gropp, W. and Moré, J. (1997) Optimization environments and the NEOS server. In *Approximation Theory and Optimization*, Buhmann, M.D. and Iserles, A. (eds.), Cambridge University Press, Cambridge, UK, pp. 167-182.

Gupta, A., Varner, J.D. and Maranas, C.D. (2005) Large-scale inference of the transcriptional regulation of *Bacillus subtilis. Comput. Chem. Eng.* **29**, 565-576.

Hart, W.E. and Istrail, S.C. (1996) Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal.*J. Comput. Biol.* **3**, 53-96.

Hatzimanikatis, V. and Bailey, J.E. (1997) Effects of spatiotemporal variations on metabolic control: Approcimate analysis usig (log)linear kinetic models. *Biotechnol. Bioeng.* **54**, 91-104.

Hatzimanikatis, V., Emmerling, M., Sauer, U. and and Bailey, J.E. (1998) Application of mathematical tools for metabolic design of microbial ethanol production. *Biotechnol. Bioeng.* **58**, 154-161.

Hatzimanikatis, V., Floudas, C.A. and Bailey, J.E. (1996) Optimization of regulatory architectures in metabolic reaction networks. *Biotechnol. Bioeng.* **52**, 485-500.

Hatzimanikatis, V., Lee, K.H. and Bailey J.E. (1999) A mathematical description of regulation of the G1-S transition of the mammalian cell cycle. *Biotechnol. Bioeng.* **65**, 631-637.

Haupt, S., Berger, M., Goldberg, Z. and Haupt, Y. (2003) Apoptosis - The p53 network, *J. Cell Sci.* **116**, 4077-4085.

Hochuli, E., Döbeli, H. and Schacher, A. (1987) New metal chelate adsorbent selective for proteins and peptides containing neighboring histidine-residues. *J. Chromatogr.* **411**, 177-184.

Hopp, T.P., Pricket, K.S., Price, V.L., Libby, R.T., March, C.J., Cerretti, D.P., Urdal, D.L. and Conlon, P.J. (1988) A short polypeptide marker sequence useful for recombinant protein identification and purification. *Bio-Technol.* **6**, 1204-1210.

Hopp, T.P. and Woods, K.R. (1981) Prediction of protein antigenic determinants from amino-acid sequences. *Proc. Natl. Acad. Sci. Biol.* **78**, 3824-3828.

Horowitz, N. H. (1945) On the Evolution of Biochemical Syntheses. *Proc. Natl. Acad. Sci. USA* **31**, 153-157.

Ingraham, J.L., Maaloe, O. and Neidhardt, F.C. (1983) *Growth of the bacterial cell.* Sinauer Asocciates Inc., Sunderland, MA, USA.

Irwin, M.S. and Kaelin, W.G. (2001) p53 family update: p73 and p63 develop their own identities. *Cell Growth Differ.* **12**, 337-349.

Jensen, R.A. (1976) Enzyme recruitment in the evolution of new function. *Annu. Rev. Microbiol.* **30**, 409-425.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabási, A.L. (2000) The large-scale organization of metabolic networks. *Nature* **407**, 651-654.

Jeong, H., Mason, S.P., Barabási, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41-42.

Jones, D.T. (1999a) Protein secondary structure prediction based on position specific scoring matrices. *J. Mol. Biol.* **292**, 195–202.

Jones, D.T. (1999b) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797-815.

Jones, D.T. and Guffin, L.J. (2003) Assembling novel protein folds from super-secondary structural fragments. *Proteins* **53**, 480–485.

Kanehisa M., Goto S., Kawashima S. and Nakaya A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42-46.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro, S. (2002) The EcoCyc database. *Nucleic Acids Res.* **30**, 56-58.

Karpeisky, M.Y., Senchenko, V.N., Dianova, M.V. and Kanevsky, V.Y. (1994) Formation and properties of S-protein complex with S-peptide-containing fusion protein. *FEBS Lett.* **339**, 209-212.

Karplus, K., Barret, C., Cline, M., Diekhans, M., Grate, L. and Hughey, R. (1999) Predicting protein structure using only sequence information. *Proteins* **S3**, 121-125.

Kim, D., Xu, D., Guo, J., Ellrott, K. and Xu, Y. (2003) PROSPECT II: Protein structure prediction program for genome-scale applications. *Protein Eng.* **16**, 641-650.

Kingsford, C.L., Chazelle, B. and Singh, M. (2005) Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* **21**, 1028-1036.

Kirkpatrick, S., Gelatt C.D. and Vecchi, M.P. (1983) Optimization by simulated annealing. *Science* **220**, 671-680.

Klepeis, J.L. and Floudas, C.A. (2002) Ab initio prediction of helical segments in polypeptides. *J. Comput. Chem.* **23**, 245-266.

Klepeis, J.L. and Floudas, C.A. (2003a) ASTRO-FOLD: A combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.* **85**, 2119-2146.

Klepeis, J.L. and Floudas, C.A. (2003b) Prediction of beta-sheet topology and disulfide bridges in polypeptides. *J. Comput. Chem.* **24**, 191-208.

Klepeis, J.L., Pieja, M.T. and Floudas, C.A. (2003a) A new class of hybrid global optimization algorithms for peptide structure prediction: Integrated hybrids. *Comput. Phys. Commun.* **151**, 121-140.

Klepeis, J.L., Pieja, M.T. and Floudas, C.A. (2003b) Hybrid global optimization algorithms for protein structure prediction: Alternating hybrids. *Biophys. J.* **84**, 869-882.

Klimov, D. and Thirumalai, D. (1996) Criterion that determines the foldability of proteins. *Phys. Rev. Lett.* **76**, 4070-4073.

Kohn, K.W. (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell*, **10**, 2703-2734.

Kolesov, G., Mewes, H.W. and Frishman, D. (2001) SNAPing up functionally related genes based on context information: A colinearity-free approach. *J. Mol. Biol.* **311**, 639-656.

Kopp, J. and Schwede, T. (2004) Automated protein structure homology modeling: A progress report. *Pharmacogenomics* **5**, 405-416.

Krippahl, L. and Barahona, P. (1999) Applying constraint programming to protein structure determination. *Lect. Notes Comput Sc.* **1713**, 289-302.

Kussell, E., Shimada, J. and Shakhnovich, E.I. (2003) Side-chain dynamics and protein folding. *Proteins* **52**, 303-321.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105-132.

Lau, K.F. and Dill, K.A. (1989) A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules* **22**, 3986-3997.

Lawler, E.L. (1976) *Combinatorial optimization: Networks and matroids.* Holt, Rinehart and Winston, New York, USA.

Lee, J., Kim, S.Y., Joo, K., Kim, I. and Lee, J. (2004) Prediction of protein tertiary strucure using PROFESY, a novel method based on fragment assembly and conformational space annealing. *Proteins* **56**, 704-714.

Lee, S., Phalakornkule, C., Domach, M.M. and Grossmann, I.E. (2000) Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput. Chem. Eng.* **24**, 711-716.

Lee, J., Ripoll, D.R., Czaplewski, C., Pillardy, J., Wedemeyer, W.J. and Scheraga, H.A. (2001) Optimization of parameters in macromolecular potential energy functions by conformational space annealing. *J. Phys. Chem. B* **105**, 7291-7298.

Levine, A.J. (1992) *Viruses*. Scientific American Library, New York, USA.

Levinthal, C. (1969) How to fold graciously. In *Mossbauer Spectroscopy in Biological Systems.* Debrunner, P., Tsibris, J.C.M. and Münck, E. (eds.), University of Illinois Press, Urbana, USA, pp. 22-24.

Li, C., Henry, C.S., Jankowski, M.D., Ionita, J.A., Hatzimanikatis, V. and Broadbelt, L.J. (2004) Computational discovery of biochemical routes to specialty chemicals. *Chem. Eng. Sci.* **59**, 5051-5060.

Lienqueo, M.E. (1999) *Development of an Expert System for the Rational Selection of Protein Purification Processes: Optimisation of Sequence Selection Criteria.* PhD Thesis (in Spanish), University of Chile, Santiago, Chile.

Lienqueo, M.E., Leser, E.W. and Asenjo, J.A. (1996) An expert system for the selection and synthesis of multistep protein separation processes. *Comput. Chem. Eng.* **20S**, S189-S194.

Lienqueo, M.E., Mahn, A. and Asenjo, J.A. (2002) Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography. *J. Chromatogr. A* **978**, 71-79.

Lienqueo, M.E., Mahn, A., Vasquez, L. and Asenjo, J.A. (2003) Methodology for predicting the separation of proteins by hydrophobic interaction chromatography and its application to a cell extract. *J. Chromatogr. A* **1009**, 189-196.

Lienqueo, M.E., Salgado, J.C. and Asenjo, J.A. (1999) An expert system for selection of protein purification processes: Experimental validation. *J. Chem. Technnol. Biot.* **74**, 293-299.

Light, S. and Kraulis, P. (2004) Network analysis of metabolic enzyme evolution in *Escherichia coli*. *BMC Bioinformatics* **5**, art. no. 15.

Lin, X.X., Floudas, C.A., Wang, Y. and Broach, J.R. (2003) Theoretical and computational studies of the glucose signaling pathways in yeast using global gene expression data. *Biotechnol. Bioeng.* **84**, 864-886.

Liwo, A., Arlukowicz, P., Czaplewski, C., Oldziej, S., Pillardy, J. and Scheraga, H.A. (2002) A method for optimizing potential-energy functions by hieracrchical design of the potential-energy landscape: Application to the UNRES force field. *Proc. Natl. Acad. Sci. USA* **99**, 1937-1942.

MacDonald, D., Joseph, S., Hunter, D.L., Moseley, L.L., Jan, N. and Guttmann, A.J. (2000) Self-avoiding walks on the simple cubic lattice. *J. Phys. A* **33**, 5973-5983.

Madigan, M., Martinko, J. and Parker, J. (1997) *Brock biology of microorganisms.* Prentice-Hill, Upper Saddle River, NJ, USA.

Mahadevan, R. and Palsson, B.O. (2005) Properties of metabolic networks: Structure versus function. *Biophys. J.* **88**, L7-L9.

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86.

Margulis, L. and Schwartz, K. (1998) *Five kingdoms: An illustrated guide to the phyla of life on earth.* W.H. Freeman, New York, USA.

Martin, A.C., Orengo, C.A., Hutchinson, E.G., Jones, S., Karmirantzou, M., Laskowski, R.A., Mitchell, J.B., Taroni, C. and Thornton, J.M. (1998) Protein folds and functions. *Struct. Fold. Des.* **6**, 875-884.

Mavrovouniotis, M.L., Stephanopoulos, G. and Stephanopoulos G. (1990) Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.* **36**, 1119-1132.

Michal, G (1998) *Biochemical pathways: An atlas of biochemistry and molecular biology*. John Wiley & Sons, London, UK.

Millgram, S. (1967) The small world problem. *Psychol. Today* **2**, 60-67.

Miyazawa, S. and Jernigan, R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules* **18**, 534-552.

Moore, G.L., Maranas, C.D., Gutshall, K.R. and Brenchley, J.E. (2000) Modeling and optimization of DNA recombination. *Comput. Chem. Eng.* **24**, 693-699.

Moore, G.L. and Maranas, C.D. (2002a) eCodonOpt: A systematic computational framework for optimizing codon usage in directed evolution experiments. *Nucleic Acids Res.* **30**, 2407-2416.

Moore, G.L. and Maranas, C.D. (2002b) Predicting out-of-sequence reassembly in DNA shuffling. *J. Theor. Biol.* **219**, 9-17.

Mosher, R.A., Gebauer, P. and Thormann, W. (1993) Computer-simulation and experimental validation of the electrophoretic behavior of proteins: III. Use of titration data predicted by the protein's amino acid composition. *J. Chromatogr.* **638**, 155-164.

Nikolaev, E.V., Burgard, A.P. and Maranas, C.D. (2005) Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophys. J.* **88**, 37-49.

Notredame, C. (2002) Recent progress in multiple sequence alignment: A survey. *Pharmacogenomics* **3**, 131-144.

Nygren, P.A., Stahl, S. and Uhlen, M. (1994) Engineering proteins to facilitate bioprocessing. *Trends Biotechnol.* **12**, 184-188.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896-2901.

Pabo, C. (1983) Molecular technology - Designing proteins and peptides. *Nature* **301**, 200.

Palu, A.D., Dovier, A. and Fogolari, F. (2004) Constraint logic programming approach to protein structure prediction. *BMC Bioinformatics* **5**, art. no. 186.

Papin, J.A., Hunter, T., Palsson, B.O. and Subramaniam, S. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell. Bio.* **6**, 99-111.

Park, B.H. and Levitt, M. (1995) The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **249**, 493-507.

Petkov, S.B. and Maranas, C.D. (1997) Quantitative assessment of uncertainty in the optimization of metabolic pathways. *Biotechnol. Bioeng.* **56**, 145-161.

Petrides, D.P. (1994) BioPro designer – An advanced computing environment for modeling and design of integrated biochemical processes. *Comput. Chem. Eng.* **18S**, S621-S625.

Pillardy, J., Czaplewski, C., Liwo, A., Wedemeyer, W.J., Lee, J., Ripoll, D.R., Arlukowicz, P., Oldziej, S., Arnautova, E.A. and Scheraga, H.A. (2001) Development

of physics-based energy functions that predict medium resolution structure for proteins of α, β and α/β structural classes. *J. Phys. Chem. B* **105**, 7299-7311.

Phalakornkule, C., Lee, S., Zhu, T., Koepsel, R., Ataai, M.M., Grossmann, I.E. and Domach, M.M. (2001) A MILP-based flux alternative generation and NMR experimental design strategy for metabolic engineering. *Metab. Eng.* **3**, 124-137.

Pharkya, P., Burgard, A.P. and Maranas, C.D. (2003) Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol. Bioeng.* **84**, 887-899.

Pramanik, J. and Keasling, J.D. (1997) Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependant biomass composition and mechanistic energy requirements. *Biotechnol. Bioeng.* **56**, 398-421.

Przybylski, D. and Rost, B. (2004) Improving fold recognition without folds. *J. Mol. Biol.* **341**, 255–269.

Ramakrishna, R., Edwards, J.S., McCulloch, A. and Palsson, B.O. (2001) Flux-balance analysis of mitochondrial energy metabolism: Consequences of systemic stoichiometric constraints. *Am. J. Physiol.-Reg. I.* **280**, R695-R704.

Reed, J.L. and Palsson, B.O. (2004) Genome-scale in silico models of *E-coli* have multiple equivalent phenotypic states: Assessment of correlated reaction subsets that comprise network states. *Genome Res.* **14**, 1797-1805.

Regan, L., Bogle, I.D.L. and Dunnill, P. (1993) Simulation and optimization of metabolic pathways. *Comput. Chem. Eng.* **17**, 627-637.

Rieger, T.R., Morimoto, R.I. and Hatzimanikatis, V. (2005) Mathematical modeling of the eukaryotic heat-shock response: Dynamics of the *hsp70* promoter. *Biophys. J.* **88**, 1646-1658.

Riley, M (1998) Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res.* **26**, 54.

Rison, S.C.G. (2002) *Of proteins and Pathways – Investigating protein functional classifications and the small molecule metabolism of Escherichia coli.* PhD Thesis, UCL, London, UK.

Rison, S.C.G., Teichmann, S.A. and Thornton, J.M. (2002) Homology, pathway distance and chromosomal localisation of Small Molecule Metabolism enzymes in *Escherichia coli. J. Mol. Biol.* **318**, 911-932.

Rison, S.C.G. and Thornton, J.M. (2002) Pathway evolution, structurally speaking. *Curr. Opin. Struct. Biol.* **12**, 374-382.

Robles, A.I., Linke, S.P. and Harris, C.C. (2002) The p53 network in lung carcinogenesis. *Oncogene* **21**, 6898-6907.

Rohl, C.A., Strauss, C.E.M., Chivian, D. and Baker, D. (2004) Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins* **55**, 656-677.

Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides, J. (2000) Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA* **97**, 6652-6657.

Sali, A., Shakhnovich, E.I. and Karplus M. (1994a) Kinetics of protein folding – A lattice model study of the requirements for folding to the native-state. *J. Mol. Biol.* **235**, 1614-1636.

Sali, A., Shakhnovich, E.I. and Karplus, M. (1994b) How does a protein fold? *Nature* **369**, 248-251.

Sargent, R.W.H. (1977) The decomposition of systems of procedures and algebraic equations. In *Proc. Biennial Conference on Numerical Mathematics*, Watson, G.A. (ed.), Spinger Verlag, Dundee.

Sassenfeld, H.M. (1990) Engineering proteins for purification. *Trends Biotechnol.* **8**, 88-93.

Sassenfeld, H.M. and Brewer S.J. (1984) A polypeptide fusion designed for the purification of recombinant proteins. *Bio-Technol.* **2**, 76-81.

Savageau, M.A. (1969) Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J. Theor. Biol.* **25**, 365-369.

Savageau, M.A. (1969) Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. *J. Theor. Biol.* **25**, 370-379.

Savageau, M.A. (1970) Biochemical systems analysis. III. Dynamic solutions using a power-law approximation. *J. Theor. Biol.* **26**, 215-226.

Schilling, C.H., Covert, M.W., Famili, I. Church, G.M., Edwards, J.S. and Palsson, B.O. (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582-4593.

Schilling, C.H., Edwards, J.S. and Palsson, B.O. (1999) Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnol. Prog.* **15**, 288-295.

Schilling, C.H. and Palsson, B.O. (2000) Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. *J. Theor. Biol.* **203**, 249-283.

Schmidt, T.G.M. and Skerra, A. (1993) The random peptide library-assisted engineering of a C-terminal affinity peptide, useful for the detection and purification of a functional Ig Fv fragment. *Protein Eng.* **6**, 109-122.

Schmidt, S., Sunyaev, S., Bork, P. and Dandekar, T. (2003) Metabolites: A helping hand for pathway evolution? *Trends Biochem. Sci.* **28**, 336-341.

Segre, D., Vitkup, D., Church, G.M. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**, 15112-15117.

Shakhnovich, E.I. and Gutin, A.M. (1990) Kinetics of protein folding. *J. Mol. Biol.* **235**, 1614-1636.

Shakhnovich, E.I. and Gutin, A.M. (1993) Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA* **90**, 7195-7199.

Skolnick, J., Kihara, D. and Zhang, Y. (2004) Development and large scale benchmark testing of the PROSPECTOR 3 threading algorithm. *Proteins* **56**, 502-518.

Steffens, M.A., Fraga, E.S. and Bogle, I.D.L. (2000a) Synthesis of bioprocesses using physical properties data. *Biotechnol. Bioeng.* **68**, 218-230.

Steffens, M.A., Fraga, E.S. and Bogle, I.D.L. (2000b) Synthesis of purification tags for optimal downstream processing. *Comput. Chem. Eng.* **24**, 717-720.

Strogatz, S.H. (2001) Exploring complex networks. *Nature* **410**, 268-276.

Tamames, J., Casari, G., Ouzounis, C.A. and Valencia, A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66-73.

Teichmann, S.A., Rison, S.C.G., Thornton, J.M., Riley, M., Gough, J. and Chotia, C. (2001) The evolution and structural anatomy of the small molecule metabolic pathways of *Escherichia coli*. *J. Mol. Biol.* **311**, 693-708.

Terpe, K. (2003) Overview of tag protein fusions: From molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biot.* **60**, 523-533.

Thomas, R., Mehrotra, S., Papoutsakis, E.T. and Hatzimanikatis, V. (2004) A model-based optimization framework for the inference on gene regulatory networks from DNA array data. *Bioinformatics* **20**, 3221-3235.

Tiana, G. and Broglia, R.A. (2001) Statistical analysis of native contact formation in the folding of designed model proteins. *J. Chem. Phys.* **114**, 2503-2510.

Tiana, G., Shakhnovich, B.E., Dokholyan, N.V. and Shakhnovich, E.I. (2004) Imprint of evolution on protein structures. *Proc. Natl. Acad. Sci. USA* **101**, 2846-2851.

Todd, A.E., Orengo, C.A. and Thornton, J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113-1143.

Tramontano, A. and Morea, V. (2003) Assessment of homology-based predictions in CASP5. *Proteins* **53**, 352-368.

Tsoka, S. and Ouzounis, C.A. (2001) Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*. *Genome Res.* **11**, 1503-1510.

Ueda, Y., Taketomi, H. and Go, N. (1975) Studies on protein folding, unfolding, and fluctuation by computer simulation. 1. Effect of Specific amino-acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Prot. Res.* 7, 445-459.

Uhlen, M. and Moks, T. (1990) Gene fusions for purpose of expression – An introduction. *Method. Enzymol.* **185**, 129-143.

Unger, R. and Moult, J. (1996) Local interactions dominate folding in a simple protein model. *J. Mol. Biol.* **235**, 1614-1636.

Varma, A. and Palsson, B.O. (1993) Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. *J. Theor. Biol.* **165**, 503-522.

Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J.P., Oltvai, Z.N. and Barabási, A.L. (2004) The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc. Natl. Acad. Sci. USA* **101**, 17940-17945.

Vásquez-Alvarez, E., Lienqueo, M.E. and Pinto, J.M. (2001) Optimal synthesis of protein purification processes. *Biotechnol. Progr.* **17**, 685-696.

Vásquez-Alvarez, E. and Pinto, J.M. (2001) MILP models for the synthesis of protein purification processes. In *Proc. ESCAPE-11*, pp. 579-584.

Vásquez-Alvarez, E. and Pinto, J.M. (2003) A mixed integer linear programming model for the optimal synthesis of protein purification processes with product loss. *Chem. Biochem. Eng. Q.* **17**, 77-84.

Vásquez-Alvarez, E. and Pinto, J.M. (2004) Efficient MILP formulations for the optimal synthesis of chromatographic protein purification processes. *J. Biotechnol.* **110**, 295-311.

Voet, D. and Voet, J.G. (1995) *Biochemistry.* John Wiley & Sons, New York, USA.

Vogelstein, B., Lane, D. and Levine, A.J. (2000) Surfing the p53 network. *Nature* **408**, 307-310.

Voit, E.O. (1992) Optimization in integrated biochemical systems. *Biotechnol. Bioeng.* **40**, 572-582.

Wagner, M., Meller, J. and Elber R. (2004) Large-scale linear programming techniques for the design of proteins folding potentials. *Math. Program.* **101**, 301-318.

Watanabe, E., Tsoka, S. and Asenjo, J.A. (1994) Selection of chromatographic protein purification operations based on physicochemical properties. *Annals N.Y. Acad. Sci.* **721**, 348-364.

Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature* **393**, 440-442.

Williams, H.P. (1999) *Model building in mathematical programming*. John Wiley, New York, USA.

Wolkenhauer, O. (2002) Mathematical modelling in the post-genome era: Understanding genome expression and regulation - A system theoretic approach. *Bio Systems* **65**, 1-18.

Woolston, P.W. (1994) *A physicochemical database for an expert system for the selection of recombinant protein purification processes*. PhD Thesis, University of Reading, Reading, UK.

Xia, Y., Huang, E.S., Levitt, M. and Samudrala, R. (2000) Ab initio construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* **300**, 171-185.

Xu, Y. and Xu, D. (2000) Protein threading using PROSPECT: Design and evolution. *Proteins* **40**, 343-354.

Xu, J.B., Li, M., Kim, D. and Xu, Y. (2003) RAPTOR: Optimal protein threading by linear programming. *J. Bioinf. Comp. Biol.* **1**, 95-117.

Xu, J.B., Li, M. and Xu, Y. (2004) Protein threading by linear programming: Theoretical analysis and computational results. *J. comb. Optim.* **8**, 403-418.

Yue, K. and Dill K.A. (1995) Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci. USA* **92**, 146-150.

Zhang, Y., Kolinski, A. and Skolnick, J. (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. *Biophys. J.* **85**, 1145–1164.

Refereed research articles from this thesis

Journal articles

Dartnell, L., Simeonidis, E., Hubank, M., Tsoka, S., Bogle, I.D.L. and Papageorgiou, L.G. (2005) Robustness of the p53 network and biological hackers. *FEBS Lett.* **579**, 3037-3042.

Simeonidis, E., Lienqueo, M.E., Tsoka, S., Pinto, J.M. and Papageorgiou, L.G. (2005) MINLP models for the synthesis of optimal peptide tags and downstream protein processing. *Biotechnol. Prog.* **21**, 875-884.

Simeonidis, E., Rison, S.C.G., Thornton, J.M., Bogle, I.D.L. and Papageorgiou L.G. (2003) Analysis of biochemical networks using a pathway distance metric through linear programming. *Metab. Eng.* **5**, 211-219.

Conference articles

Simeonidis, E., Pinto, J.M. and Papageorgiou, L.G. (2005) An MILP model for optimal design of purification tags and synthesis of downstream processing. In *Proc. ESCAPE-15*, L. Puigjaner and A. Espuña (eds.), Barcelona, Spain, pp.1537-1542.

Simeonidis, E., Dartnell, L., Bogle, I.D.L. and Papageorgiou, L.G. (2005) Analysis of biochemical networks using linear programming, In *Proc. 7th World Congress of Chemical Engineering*, Glasgow, Scotland, UK, in press.

Simeonidis, E., Pinto, J.M. and Papageorgiou, L.G. (2004) Optimal peptide tag design and synthesis of downstream protein processing. In *Proc. ESCAPE-14*, A. Barbosa-Povoa and H. Matos (eds.), Lisbon, Portugal, pp.289-294.