INVERSE PROTEIN FOLDING, HIERARCHICAL OPTIMISATION AND TIE KNOTS

Thomas M. A. Fink

ST. JOHN'S COLLEGE
UNIVERSITY OF CAMBRIDGE

A DISSERTATION
SUBMITTED FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY AT THE
UNIVERSITY OF CAMBRIDGE
1998

TO MY MOTHER

Contents

List of Figures

Preface

I HAVE BEEN FORTUNATE to have learnt physics from excellent physicists. My introduction to research began under the instruction of Thomas Tombrello as a first year undergraduate at Caltech. His influence on my manner of physical insight has been significant. Subsequent collaboration with Brad Werner was especially rewarding, and his guidance has since continued. I am, of course, most indebted to my research supervisor, Robin Ball, who has proved to be as effective a supervisor as he is a theoretical physicist. If supervisor and student may be said to be the analogue of master and apprentice, I hope I have picked up something of the style and command with which Robin practices the craft.

Physicists function best within a community, and I have been no less fortunate in working among the Theory of Condensed Matter group at the Cavendish Laboratory. I acknowledge, in particular, Ahmed Al-Falou, Robert Farr, Mark Gibbs, Mehul Khimasia, David Khmelnitskii, Peter King, Hernan Larralde, Yong Mao, Chris Pickard and Christian Waite, all of whom have made a mark on this dissertation.

I have at Cambridge valued most of all my friends, from whom I have received — notwithstanding the subject of this dissertation — great insight outside physics. I acknowledge here Keith and Judy Bates, Fr. Philip Dixon, Neil Hallinan, Thomas and Eva Harte, Luke and Elizabeth Howard, Laszlo Koczy, Christian Waite and Fr. Allan White.

Additionally, from America I mention Shreyas Vasanawala, Charles Waite and Christian Waite and Stephen Bowen, Michael Coleman, Tae Kim, James Reiner and Nathan Steinke.

Particular acknowledgement is made of my family, all of whom, despite my time away, have remained close.

I gratefully recognise financial support from a Benefactors' Scholarship from St. John's College and a National Science Foundation Graduate Research Fellowship.

Finally, I acknowledge Sarah Aitken, who knows no physics but remains my greatest source of insight.

Effort has been made to keep each chapter of this dissertation largely self-contained, repeating, where necessary, definitions and arguments made elsewhere. This is meant to be a convenience to the reader and reflects the intention for publication with which portions of this dissertation were written. Reference to work published or intended for publication is made to the chapter which describes it. The list of publications in the first chapter relates chapters to manuscripts.

Except where otherwise stated, this dissertation is the result of my own work and contains nothing which is the outcome of work done in collaboration. This dissertation has not been submitted in whole or in part for any degree or diploma at this or any other university.

Thomas M. A. Fink
St. John's College
May 1998

Nomenclature

Inverse Protein Folding

A	number of amino acid species ($A = 20$ for biological proteins)
A_c	critical number of species necessary for protein design
b	number of non-backbone bonds in lattice protein
B	number of non-backbone bonds in compact lattice protein
C	lattice protein contact map
C_{tot}	linear combination of p lattice protein contact maps
E	conformational energy
f_A	distribution of first-passage time of sequence A
f_A^n	distribution of mean of n first-passage times of sequence A
F	conformational free energy
g	number of monomers in a blob; width of funnel
$g_{\max}$	maximum width of funnel
H_{tot}	Hamiltonian by which sequence is trained
I	information
k_B	Boltzmann's constant
n	number of first-passage time measurements; $\frac{1}{\sqrt{n}}$ is effective temperature
N	length of protein chain (in units of amino acid residues)
p	number of conformations to which sequence is trained
$p_{\max}$	capacity of sequence
P	probability
$P_{A\to B}$	probability that sequence B replaces sequence A
S	protein sequence *or* conformational entropy
t	time

t_{MC}	time measured in units of Monte Carlo steps
t_A	first-passage time of sequence A
t_A^n	mean of n first-passage times of sequence A
T	temperature or effective temperature
U	$A \times A$ pair potential matrix (amino acid interactions)
$\tilde{U}$	$N \times N$ extended pair potential matrix
w	weight associated with a conformation
z	coordination number of protein lattice
Z	relative conformational energy
α	shift of pair potential mean
β	$\frac{1}{k_B T}$
Γ	protein conformation
Γ_0	ground state protein conformation
κ	compact conformational freedom per amino acid
$\tilde{\kappa}$	conformational freedom per amino acid
μ_A	mean first-passage time of sequence A
σ	standard deviation of amino acid interaction strengths
σ_{tot_i}	standard deviation of Hamiltonian of amino acid i
σ_A	standard deviation of first-passage time of sequence A
σ_A^n	standard deviation of mean of n first-passage times of seq. A

Hierarchical Optimisation

HOP	hierarchical optimisation problem
TSP	traveling salesman problem
PTSP	probabilistic traveling salesman problem
a	action (decision)
a_i	ith TSP city
$A(\Delta h^n)$	probability of accepting downhill Δh^n transition
A	set of TSP or PTSP cities
b_n	probability of a spanning cluster from level n downwards
B	active subset of PTSP cities
B_n	expected optimal number of costs paid at level n
c_n	average cost associated with a node purchased at level n

C	cost function
d	distance of TSP tour t or expected distance of PTSP tour u
D	distance matrix for TSP or PTSP cities
D_A	design A
f_x	distribution of cost x
f_A	distribution of cost H_A associated with design D_A
$f_{\Delta H}$	distribution of cost difference $H_B - H_A$
$f_{\Delta H}^n$	distribution of mean of n samples of ΔH
h_A	realisation of random variable H_A
Δh	realisation of random variable ΔH
Δh^n	realisation of random variable ΔH^n
H_A	distributed cost associated with design A
ΔH	distributed cost difference $H_B - H_A$
ΔH^n	distributed mean of n samples of ΔH
n	hierarchical optimisation stage number *or* number of samples in estimating mean of ΔH
N	total number of hierarchical optimisation stages *or* number of cities in TSP or PTSP
p_i	TSP or PTSP city probabilities
p_n	probability of purchasing a node in economic model
P	probability
q_n	average cost of purchased node $\times$ probability of purchasing it
$P_{A\to B}$	probability that the transition from A to B is accepted
$Q_{A\to B}$	probability that the transition from A to B is considered
R	reward function (negative of cost function)
s	state (realisation of a random variable)
t	TSP or *a priori* PTSP tour
t_{MC}	time measured in Monte Carlo steps
T	connection matrix for TSP or *a priori* PTSP tour T
T_{e}	effective temperature
u	pruned PTSP tour
U	connection matrix for pruned PTSP tour u

V_n { value assoc. with all future descendent actions from a stage n state (Ch. 8) *or* / $-$value assoc. with all future descendent nodes from a level n node (Ch. 10)

x random cost associated with a node

z number of daughter nodes descending from parent node

β_{e} effective inverse temperature $\frac{1}{kT_{\mathrm{e}}}$

β_{s} effective inverse temperature assoc. with stochastic annealing

λ mean of random costs associated with nodes

μ_A mean of cost H

$\Delta\mu$ mean of ΔH and ΔH^n

ν_A number of an ensemble of systems in state A

σ_A standard deviation of H_A

σ_{Δ_H} standard deviation of ΔH

$\sigma^n_{\Delta_H}$ standard deviation of ΔH^n

Tie Knots

b knot balance

$F_{\hat{\mathbf{r}}}$ number of walks beginning with $\hat{\mathbf{l}}$ and ending with $\hat{\mathbf{r}}$, *etc.*

h half-winding number (number of moves)

$\{h, \gamma\}$ knot class

$K(h)$ number of knots as a function of h

$K(h, \gamma)$ number of knots in a class

n number of random walk steps

r, c, l random walk axes

$\hat{\mathbf{r}}, \hat{\mathbf{c}}, \hat{\mathbf{l}}$ random walk steps

R, C, L knot regions

$R_\odot, R_\otimes, C_\odot, C_\otimes, L_\odot, L_\otimes$ knot moves

s knot symmetry

U_n probability of occupation after n steps

γ number of center moves or center steps

$\odot, \otimes$ knot directions

$\curlyvee$ triagonal basis about which tie is wound

INVERSE PROTEIN FOLDING, HIERARCHICAL OPTIMISATION AND TIE KNOTS

Chapter 1

Introduction

> Герой же моей повести, которо го я люблю всеми силами души, которого старался воспроизвести во всей красоте его и который всегда был, есть и будет прекрасен – правда.
>
> LEO TOLSTOY
> *Sevastopol in May*

This dissertation is composed of three parts: inverse protein folding, hierarchical optimisation and tie knots. Chapters 2 – 7 describe inverse folding, the design of proteins which quickly and stably fold to specified target conformations. In Chapters 8 – 10 we introduce hierarchical optimisation, a generalisation of conventional optimisation in which the solution must be determined stage-wise in light of successive information learnt. We provide a mathematical model of necktie knots in Chapter 11, with the express intention of recovering the traditional, and predicting new, aesthetic tie knots.

Here we summarise each of the chapters and outline their interrelationships within each part. Portions of this dissertation have been published or are intended for publication; the final section relates chapters to manuscripts.

1.1 Inverse Protein Folding

We review in Chapter 2 the development of protein folding and design, which has occurred almost entirely during the last four decades. Central to recent theoretical advances are the conformational energy landscape and lattice models of proteins. We present the statistical mechanical interpretation of protein folding afforded by the energy

landscape picture and consider its topographic structure. Landscapes associated with the kinetic and free energy barriers impeding stable, fast folding are identified and serve to indicate ideal protein landscapes.

In Chapter 3 we first examine lattice models of proteins, which contain the fundamental attributes of proteins in the absence of biological periphery. We describe lattice folding dynamics and introduce an analytic representation of lattice proteins, both of which are used extensively throughout this dissertation. We then review thermodynamically oriented sequence selection, an important method of protein design for lattice proteins. Sequences made to be thermodynamically stable in a desired target conformation are observed in simulation to fold more quickly to the target as well.

In Chapter 4 we characterise protein folding by thermodynamic stability and kinetic accessibility of the native state, which are the benchmarks of a sequence determined by successful design. The relationship between folding speed and stability is captured by the accessibility-stability phase space. These qualities, observed through folding simulation to be correlated in Chapter 3, are shown to be in conflict near the extremes of either. It is demonstrated, in particular, that thermodynamically oriented sequence selection does not favour optimal accessibility.

In Chapter 5 we interpret the ability of a protein to organise itself from any of the set of unfolded conformations into a unique folded structure as pattern recognition. Accordingly, proteins designed to recognise multiple independent conformations correspond to conventional associative memories. We propose a multiple target training scheme and calculate the maximum number of structures a sequence can be trained to simultaneously recognise. Driving the protein toward capacity incurs a loss of stability of the target structures such that at saturation recall becomes fragmentary.

Continuing from Chapter 5, we consider in Chapter 6 the weighted recognition of correlated targets with the intent of introducing a funnel of low energy structures sloping toward a single desired conformation. For our target patterns we choose the conformations sampled in unfolding, which we propose designate the folding paths of least kinetic

constraint. We derive an analytic bound on the basin of attraction such that the desired conformation is a free energy minimum; along with the capacity result from Chapter 5, this suggests a limit to the extent to which we can manipulate the conformational energy landscape.

We present in Chapter 7 a method of kinetically oriented sequence selection whereby sequences are optimised explicitly with respect to folding time. Since protein folding is a statistical event, ordering sequences according to folding ability requires knowledge of the distribution of folding times, which we estimate. By analogy with simulated annealing, we use the uncertainty in folding time measurements in a controlled way to avoid trapping in locally, but not globally, optimal sequences.

1.2 Hierarchical Optimisation

In Chapter 8 we generalise conventional optimisation to include problems whose solutions must be determined stage-wise in the light of information learnt at each level. Key to hierarchical optimisation is the balance of local optimality against minimising expected cost over the set of possible futures. Problem complexity may result from one or more individually difficult decisions or emerge from the concatenation of many elementary decisions. We conclude the chapter by deriving the general optimality equation, which we apply to model problems in Chapters 9 and 10.

We introduce in Chapter 9 stochastic annealing, the analogue of simulated annealing for two-stage hierarchical optimisation problems. Outlined for kinetically oriented protein design in Chapter 7, the algorithm may be applied to any optimisation problem in which the cost function is a distributed (random) variable. We show that repeated application of the transition probability closely approximates a thermal distribution of solutions. The algorithm is used to solve the probabilistic traveling salesman problem, a central problem of probabilistic optimisation.

We consider in Chapter 10 a model hierarchical optimisation problem consisting of many elementary decisions which must be made in

anticipation of future information learnt. The problem may be interpreted as a model of economic growth, the decision to buy representing investment in future return in the form of negative costs. The optimal decision policy, which may be framed as a novel form of percolation through a decision tree (Bethe lattice), exhibits a phase transition and finite size scaling.

1.3 Tie Knots

Necktie knots, the subject of Chapter 11, are inherently topological structures; what makes them tractable is the particular manner in which they are constructed. This observation motivates a map between tie knots and persistent random walks on a triangular lattice. The topological structure embedded in a tie knot may be determined by appropriately manipulating its projection; we derive the corresponding grammar for tie knot sequences. We classify knots according to their size and shape, measured by the half-winding number and the number of center moves, and provide an expression for the number of knots in a class. Aesthetic knots are characterised by the conditions of symmetry and balance. Of the 85 knots which may be tied with a conventional tie, we recover the four traditional knots (Four-in-Hand, Half-Windsor, Windsor and Pratt) and introduce six new aesthetic ones.

1.4 Schematic Organisation

The organisation of inverse protein folding is schematically outlined in Figure 1.1; hierarchical optimisation and tie knots are included in Figure 1.2.

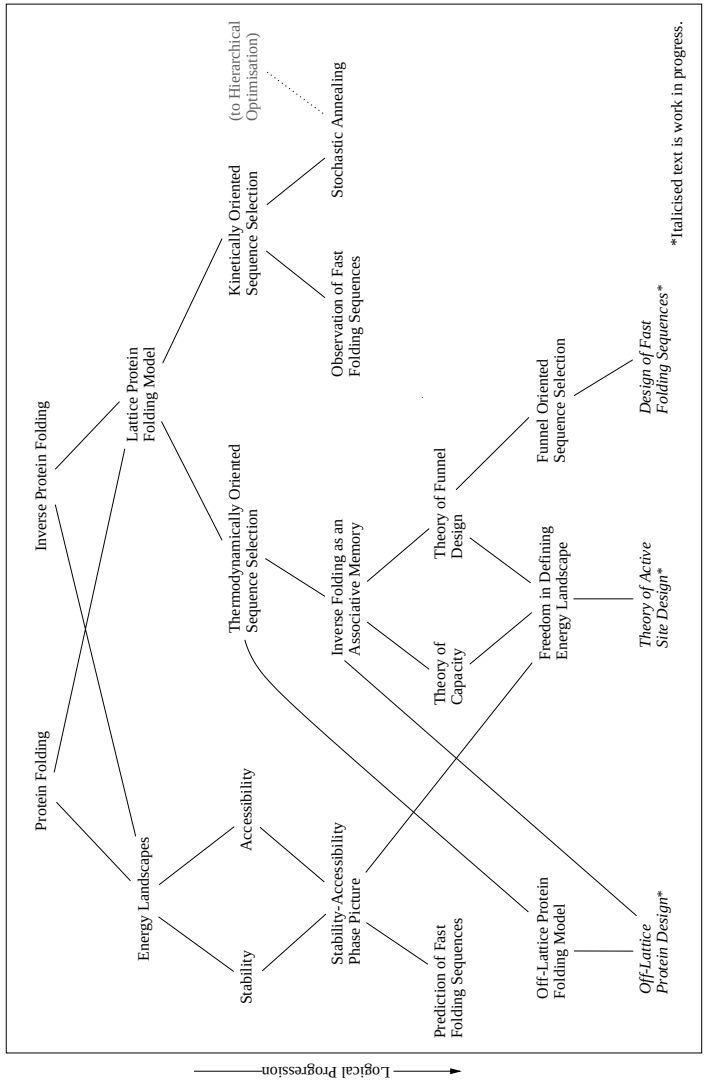

Figure 1.1: Schematic outline of inverse protein folding.

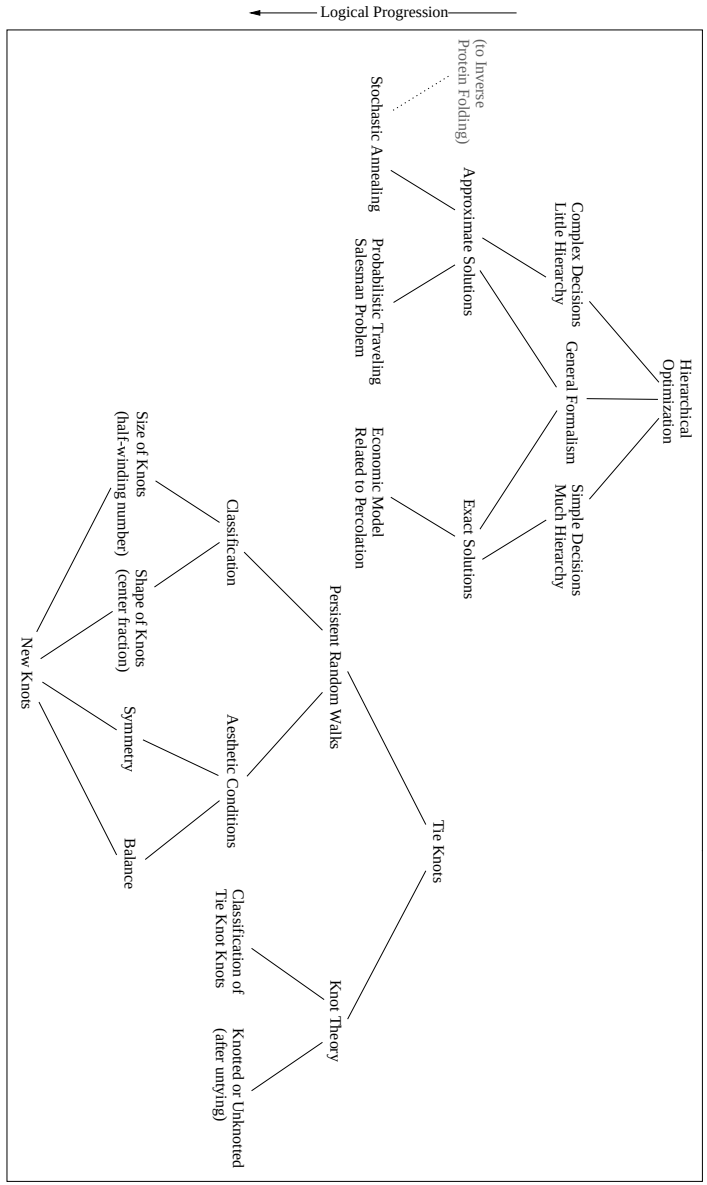

Figure 1.2: Schematic outline of hierarchical optimisation and tie knots.

1.5 Publications

Parts of this dissertation have been published or are intended for publication as follows.

Ch. 4 Thomas M. Fink and Robin C. Ball, "Robustness and Efficiency in Inverse Protein Folding," *Physica D* **107**, 199 (1997).

Ch. 5, 6 Thomas M. Fink and Robin C. Ball, "Inverse Protein Folding as an Associative Memory," submitted to *Physical Review Letters* (1997).

Ch. 7 Thomas M. Fink and Robin C. Ball, "Kinetically Oriented Sequence Selection," for *Physical Review Letters* (1998).

Ch. 8 Thomas M. Fink and Robin C. Ball, "Hierarchical Optimization Problems," in preparation (1997).

Ch. 9 Robin C. Ball and Thomas M. Fink, "Stochastic Annealing," for *Physical Review Letters* (1998).

Ch. 10 Thomas M. Fink and Robin C. Ball, "Exactly Solvable Hierarchical Optimization Problem Related to Percolation," *Physical Review Letters* **76**, 2827 (1996).

Ch. 11 Thomas M. Fink and Yong Mao, "Tie Knots and Random Walks," *Nature*, in press (1998).

Ch. 11 Thomas M. Fink and Yong Mao, "A Mathematical Theory of Tie Knots," submitted to *J. Phys. A* (1998).

Chapter 2

Protein Folding, Inverse Protein Folding and Energy Landscapes

> But in what shape they choose,
> Dilated or condensed...
> Can execute their aery purposes.
>
> JOHN MILTON
> *Paradise Lost*

Current understanding of protein folding was borne out of reconciling experimental evidence that proteins fold reversibly with the well accepted notion that they fold quickly. Only recently have similar advances occurred in inverse protein folding. Understanding of both is facilitated by the notion of the conformational energy landscape and the statistical mechanical interpretation of folding on it. Typical landscapes are rugged and exhibit topographic features which impede stable, fast folding exhibited by biological proteins; both real and successfully designed proteins require special energy landscapes.

2.1 Protein Folding

The foundations of protein folding began in the early 1960s when Anfinsen *et al.* [3] showed that proteins can fold reversibly. Under thermodynamic control, they observed the denaturation (unfolding) of a compact protein into a random coil of amino acids and the spontaneous assembly back to its original configuration (see, *e.g.*, Figure 2.1). Two conclusions could be drawn: 1) proteins organise themselves without assistant machinery into one of a myriad of possible

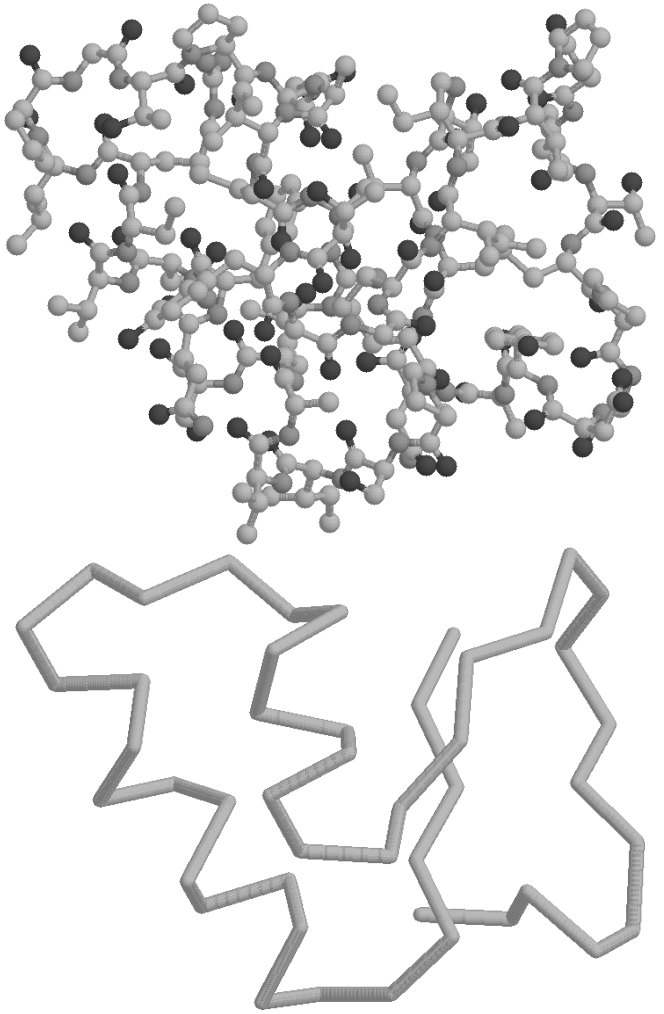

Figure 2.1: Top: Atomic representation of the protein crambin (46 amino acid residues). The linearly connected nature of the structure is not apparent. Bottom: The same protein with only the α-carbon backbone visible.

conformations; 2) the native conformation[1] of the heteropolymer is thermodynamically stable and, accordingly, the global minimum of the free energy landscape.

Anfinsen's revelation was at odds with the common view at the time, that proteins fold along a well-defined reaction pathway. Proteins were, after all, the product of a chemical reaction and should be expected to react accordingly. Pathway models dictate that the unfolded conformation must sequentially traverse a series of intermediate configurations before finally arriving at the folded conformation, in which intermediate $\mathcal{I}_k$ is in chemical equilibrium with intermediates $\mathcal{I}_{k-1}$ and $\mathcal{I}_{k+1}$. Schematically, this is written as

$$\mathcal{U} \rightleftharpoons \mathcal{I}_1 \rightleftharpoons \ldots \rightleftharpoons \mathcal{I}_n \rightleftharpoons \mathcal{F}, \tag{2.1}$$

where $\mathcal{U}$ is the unfolded (denatured) state and $\mathcal{F}$ the fully folded state.

The classical view of pathways meant that proteins travel quickly downhill toward the local (and presumably global) minimum corresponding to the folded state along a set itinerary. The observation of thermodynamic reversibility implied that proteins seek out the global minimum along a path directed as much by thermal fluctuations as by the local gradient. These two views became known as kinetic and thermodynamic control.

The path-dependence of kinetic control and path-independence of thermodynamic control are clearly incompatible. The essential impediment to accepting the thermodynamic view is the exponential size of the conformational landscape which the protein must explore. It would seem that a proportionally long search time would be necessary for it to find its ground state structure. If each additional amino acid can take on, say, two orientations with respect to the polypeptide chain, then the number of conformations available to a 100 amino

[1]The ground state of a protein is the minimum energy conformation in which the protein spends the greatest time at equilibrium. The native state is the most occupied conformation on the time scale of the functional life of the protein. If this time is less than that necessary for the protein to reach equilibrium from its denatured state, the native and ground states may differ. However, it is believed in Nature that these two conformations are generally equivalent. Unless otherwise stated, we use the two terms interchangeably.

acid protein is 2^{100}. Assuming (conservatively) the protein explores one conformation every picosecond [5], the time necessary to find a particular conformation would take 10^{18} s., comparable to the age of the universe. But proteins fold in times on the order of milliseconds, not years (for a biological overview of protein folding, see [6]). How can a protein navigate a vast landscape without a set path yet still find its target quickly? This apparent contradiction, posed by Cyrus Levinthal [15] in the late 1960s and since coined the 'Levinthal paradox,' began the extensive search for folding pathways via folding kinetics experiments. Only recently has a new understanding of protein folding based on the statistical mechanical interpretation of folding on an energy landscape come to view.

The paradox rests on the assumption that the unfolded state $\mathcal{U}$ in (2.1) from which the reaction begins is unique. But the denatured state is not a single conformation — it is all conformations apart from the folded state. Since the unfolded conformation is really a distribution over the entire conformation space, an ensemble of folding proteins requires an ensemble of independent folding pathways. These pathways will converge and intertwine and eventually coalesce as they approach the native conformation, all along traveling further downhill.

Of course, this picture has more to do with the thermodynamic exploration of an energy landscape funnel than with a well defined pathway. We are led to reject the view that proteins travel along a single deterministic pathway and instead consider the new statistical view of proteins scattered about the energy landscape making their way toward the funnel. Proteins navigate the landscape in ways that bring them downhill, all the while being buffeted by Brownian motion, occasionally knocking them uphill as well.

2.2 Inverse Protein Folding

The self-organisation of a denatured sequence to its native conformation, discussed in the previous section, has received significant recent attention [5, 6, 8]. Here we ask how the requisite sequence emerges from the functional need for the conformation via evolution (or any other method of sequence selection). The answer to this question is

the objective of inverse protein folding.

The natural end of inverse protein folding is the prediction of stable, fast folding amino acid sequences which fold *in situ* to biologically useful target conformations. Proteins are densely packed macromolecules which function primarily by topographic surface recognition; accordingly, proteins should be designed to match the surface topography of their intended targets.

In Chapters 3 – 7 we attempt to do just this. Along the way we answer questions such as: 1) *Is the inverse problem well posed?* We want to design proteins to fold to arbitrary (compact) conformations. Are protein targets limited in practise to those conformations to which some sequence can fold? 2) *Are biological proteins special?* In the previous section we saw that biological proteins fold to stable native conformations over short times scales. Is this typical of all proteins or does stable, fast folding occur for exceptional sequences only? 3) *How do we select a sequence to fold (quickly and stably) to a target?* Since the number of sequences grows exponentially, selecting for folding performance by exhaustion quickly becomes prohibitive. What properties of sequences correlate to folding ability?

Inverse protein folding inevitably tells us about the forward folding problem as well. To design sequences which successfully fold to their targets, we must understand what characterises good folding, on the one hand, and how such such sequences are selected, on the other. Folding ability may depend on our method of selection — and the size of the space we select from — in surprising ways.

The collapse of a linear chain of structural building blocks is an effective method of constructing complex macromolecules. Not surprisingly, ideas useful in protein design have applications outside their original context, such as the design of drugs and enzymes. Much of what we have to say applies to the engineering of useful heteropolymers in general.

2.3 Energy Landscapes

Protein folding and design are best understood within the framework of the energy landscape, the energy of a protein as a function of the

individual atomic coordinates (Figures 2.3 – 2.6). The landscape offers many levels of description; what we see depends on the resolution at which it is examined. Our aim is to survey the landscape on a scale such that protein folding dynamics may be resolved but surface variations peripheral to folding remain coarse grained.

To this end, we divide the degrees of freedom of the protein into two sets. The first contains the dihedral bond angles $\phi_1, \psi_1, \phi_2, \psi_2, \ldots, \phi_N, \psi_N$ (Figure 2.2) along the peptide backbone. Into the second set we place everything else: hydrogen bonding, torsion angle energies, rotations of individual side chains, *etc.* We label the first set conformational freedom and the second set internal freedom.

The internal degrees of freedom allow typical free energy changes on the order of k_BT, comparable to the thermal energy of individual atoms of the system. Since these fluctuations are peripheral to folding dynamics, we average over the free energy of the internal freedom available to a particular conformation. Accordingly, the vertical axis of the energy landscape represents the conformational energy of a unique backbone configuration plus the mean internal free energy of that configuration. We refer to this quantity simply as the energy E.

The many lateral axes of the energy landscape represent the $2N$ conformational degrees of freedom. Each configuration is represented by a point on the conformation space such that similar conformations are nearby. Since each additional link in the protein chain brings with it two additional degrees of freedom, the number of available conformations grows exponentially with chain length. For moderate N, the space of possible conformations is very large indeed.

The macroscopic properties of an ensemble of proteins at equilibrium are governed by thermodynamics. Equilibrium implies a distribution of conformations such that the probability of occupation at finite temperature is proportional to the Boltzmann factor, $\exp(\frac{-E}{k_BT})$, where T is the temperature and k_B is Boltzmann's constant. Accordingly, an ensemble of identical sequences will distribute itself such that, while some proteins are to be found on all parts of the landscape, greater fractions lie at lower elevations, in proportion to the Boltzmann factor. We say that a configuration is thermodynamically stable if at any one time a macroscopically significant fraction of the

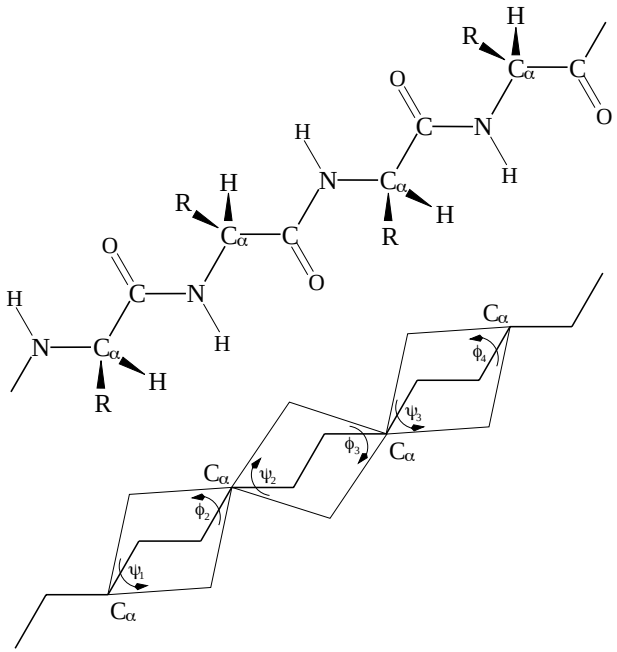

Figure 2.2: Top: Chemical diagram for a four amino acid fragment of protein. Bottom: Corresponding structural degrees of freedom along the backbone. Each α-carbon is doubly bonded along the backbone, about each of which protein segments may rotate freely.

ensemble is found there.

The return to equilibrium of an ensemble of proteins (protein folding) is described by kinetics, which describes the passage through phase space from isolated to densely populated regions. In the language of the energy landscape, kinetic evolution depends on the local topographic features which must be traversed and surmounted. Rough landscapes make direct navigation toward the ground state difficult: proteins are likely to get lost along the way and must travel nearly as much uphill as downhill, both of which effect slow folding. Ground state conformations readily visited by a non-equilibrium ensemble of

proteins are said to be kinetically accessible.

2.4 Smooth vs. Rugged Landscapes

The roughness of an energy landscape may be quantified by the presence of structural hierarchy: within a closed contour of constant elevation, there exist several closed contours of lower elevation, within each of which are more contours of lower elevation, *etc.* Landscapes characterised by a hierarchy of sub-valleys within valleys are said to be rugged; trivially hierarchical landscapes, in which each closed contour contains not multiple but a single closed contour of lower elevation, are called smooth.

Deterministic (downhill) dynamics on a rugged landscape yields a myriad of local minima nearby in energy but conformationally distant and not related by any symmetry. An ensemble of systems on this landscape shows little preference for the ground state. Similar difficulties can face stochastic systems (say, in thermal equilibrium at finite temperature) operating on a rugged landscape: below the glass transition temperature, mobility slows to unrealistic time scales and the system becomes effectively non-ergodic. This corresponds to deterministic exploration on a still rugged free energy landscape.

The conformational energy landscape of a protein is heavily constrained by the self-avoiding and non-crossable nature of the protein. To pass to a conformationally near but topologically distant conformation, the protein must swell and recollapse, overcoming a large energy barrier. Within these topological boundaries, the folding of a random protein sequence also exhibits frustration, the inability of chain segments to cooperatively align. Together, these suggest that the conformational landscape of a typical sequence is indeed rugged, which is generally (*e.g.*, [4]) thought to be the case.

2.5 Folding Funnels and Free Energy Traps

Topographic features of the conformational energy landscape determine the proficiency with which proteins fold to their native conformation. Low energy minima away from the ground state or entropically

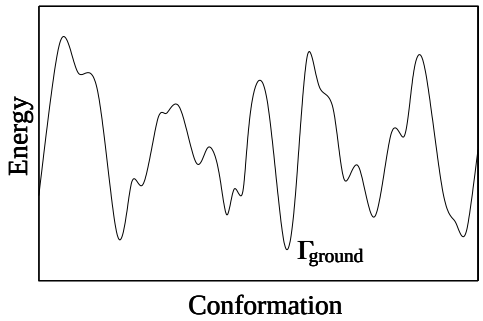

Figure 2.3: A heteropolymer conformational energy landscape. The well above the ground state is narrow and the global minimum is only marginally deeper than distant local minima. The ground state is neither easily accessible nor thermodynamically stable.

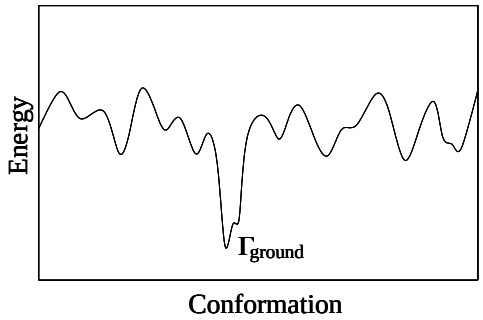

Figure 2.4: A rugged landscape with a single deep minimum provides thermodynamic stability but takes no measure to insure the ground state is kinetically accessible.

favourable plateaus act as free energy traps by prohibiting occupation of the ground state minimum. Valleys and mountains acting collectively in place of a broad funnel slow folding kinetics by trapping in local minima before imposing slow uphill climbs.

The energy landscape[2] of a random heteropolymer, shown in Figure 2.3, contains all of the ingredients detrimental to stable, fast folding. A hierarchy of valleys makes the ground state conformation only marginally lower in energy than quasi-degenerate local minima, which act as energetic traps by housing misfolded proteins. The absence of a steep gradient sloping toward the target slows navigation; like on a flat energy surface, proteins must meander through the landscape without guidance.

The more hospitable energy landscape shown in Figure 2.4 contains a single deep minimum above the ground state conformation separated by a large gap from the set of non-native structures. While this makes the target thermodynamically stable, it takes no measure to ensure that it is accessible kinetically. A significant fraction of an ensemble of proteins will, at equilibrium, reside in the target configuration, but the ensemble will take a very long time to equilibrate.

Stable folding not only requires a large energy gap — conformational entropy may reduce stability as well. This is observed, for example, on the champagne glass landscape [8] in Figure 2.5. After traveling down the initially steep basin toward the ground state, proteins meander aimlessly along the shallow ridge before discovering the well above the target. At any time in equilibrium, most of the chains are found along the shallow plateau, where there are many more accessible conformations at a disproportionately small cost in energy.

An ideal folding landscape, drawn in Figure 2.6, is characterised by a deep, broad funnel centered about the ground state conformation. As proteins fall to lower energies, the conformational freedom is reduced such that between a near-native configuration 2 and a less

[2]Our energy landscape schematics may be misleading. Real protein landscapes are $2N$-dimensional and contain many features not found in one dimension, such as saddles, moats, *cul-de-sacs* and other structures not expressible in low dimensions. Especially important is the significantly greater number of ways of getting from one point on the landscape to another. Accordingly, landscape schematics shown here should be thought of as slices through more realistic high-dimensional landscapes.

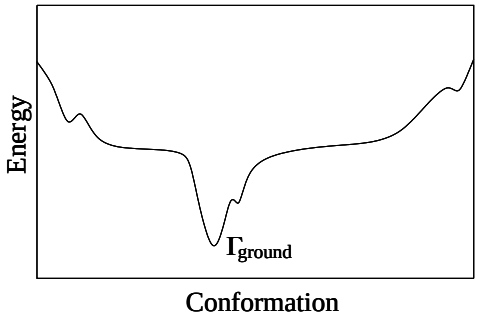

Figure 2.5: The champagne glass energy landscape illustrates how conformational entropy can lead to a free energy barrier to stable folding. A significant fraction of proteins, at any given moment, aimlessly wanders about the shallow plateau.

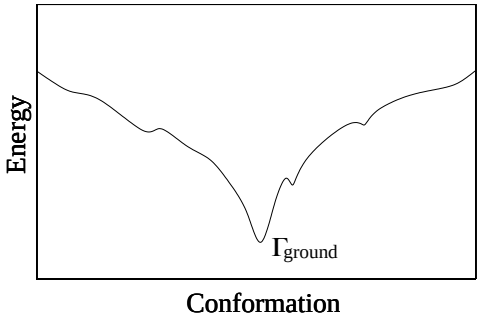

Figure 2.6: An ideal funnel landscape. The ground state conformation may be reached from all denatured conformations quickly and without free energy barriers.

near configuration 1,

$$\Delta G = \Delta E - T\Delta S < 0, \tag{2.2}$$

where $\Delta G = G_2 - G_1$, *etc.* This simply says that the funnel is everywhere sufficiently steep such that the loss in entropy is overcome by the decrease in energy.

Chapter 3

Lattice Models and Thermodynamic Sequence Selection

> The temperature of Heaven can be rather accurately computed. Our authority is Isaiah 30:26, 'Moreover, the light of the Moon shall be as the light of the Sun and the light of the Sun shall be sevenfold, as the light of seven days.' Thus Heaven receives from the Moon as much radiation as we from the Sun, and in addition 49 times as much as the Earth from the Sun, 50 times in all. Using the Stefan-Boltzmann law for radiation, $(H/E)^4 = 50$, where E is the absolute temperature of the earth, gives H as 525 °C. The exact temperature of Hell cannot be computed.... [However] Revelation 21:8 says 'But the fearful, and unbelieving...shall have their part in the lake which burneth with fire and brimstone.' A lake of molten brimstone must be at or below [its] boiling [temperature], 444.6 °C. We have, then, that Heaven, at 525 °C, is hotter than Hell at, 445 °C.
>
> *Applied Optics*

Biologically realistic protein models have yet to prove to be effective in the study of protein folding and design. In this chapter we consider a lattice model of proteins, later used extensively in folding simulations and analytic models, and discuss its scope and limits. Here we outline lattice folding dynamics and define an analytic representation of lattice proteins. We then review a method of protein design based on selecting sequences that are thermodynamically stable in the target conformation. Model proteins fold from denatured states to their target conformations in much less time than random heteropolymers need to reach their ground states.

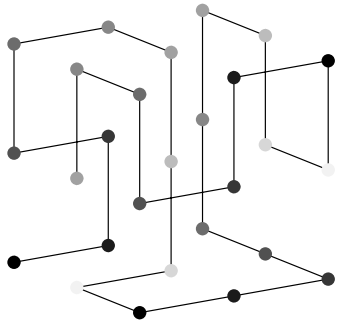

Figure 3.1: Proteins may be represented by self-avoiding walks on an infinite 3-dimensional cubic lattice. Shown here is a compact 27-mer.

3.1 Lattice Models

The simulation of protein dynamics, as well as mathematical models describing the statistical properties of proteins, are impeded by the complex atomic structure of the protein chain. Detailed computer modelling (*e.g.*, molecular dynamics) remains computationally infeasible and mathematical analysis is intractable in the absence of any symmetries. It appears, nevertheless, that much of the biological detail is peripheral and that the underlying physics driving the spontaneous organisation of proteins applies universally to linear heteropolymers. Physicists are led to consider simple exact models of proteins in which this fundamental behaviour might become apparent.

We represent proteins as self-avoiding random walks on an infinite 3-dimensional cubic lattice (Figure 3.1). Vertices indicate amino acids and edges represent the peptide chain. The continuous degrees of freedom provided by the dihedral angles are replaced by the z available discrete steps along the lattice axes, where z is the lattice coordination number. Intra-chain interactions occur according to a nearest-neighbour pair potential, in which amino acids are considered isotropic beads, or monomers.

While lattice models lack atomic detail, they contain the fundamental microscopic attributes of proteins: linear connectivity, chain flexibility, excluded volume and sequence dependent intra-chain interactions [8]. The most apparent limitations of lattice proteins are the artificial discrete degrees of freedom and the short range and isotropic nature of the interactions.

Direct comparison of the degrees of freedom of on- and off-lattice (biological) proteins is not insightful; while off-lattice bond rotations allow a continuous range of orientations, more relevant is the number of local minima through which these orientations pass. This quantity estimates the conformational freedom per monomer; it and its lattice counterpart κ (introduced in Chapter 4) provide a more equitable comparison. Nor do the different degrees of freedom significantly effect macroscopic conformations; the down-chain orientation correlation diminishes exponentially for both models.

For maximally or near-maximally compact conformations,[1] in which steric effects are the dominant interaction, nearest neighbour interactions are suitable approximations to the amino-acid residue potential. More open conformations, in which the number of amino acids composing the force field is much greater than z, of course rely on the tails of the biological potentials. However, such structures are much less thermodynamically stable than their compact counterparts and are not observed in the native states of biological proteins [7].

A set of anisotropic potentials not fixed in orientation with respect to the peptide chain provides similar means of reducing frustration as a larger alphabet of isotropic amino acid species. For example, a bead with different potentials on either hemisphere, when rotated by π, appears to its neighbours to be an altogether new species. In Nature, however, amino acids are rigidly connected to the backbone and the chain must reconfigure to make use of the additional variety, receiving little loss in frustration. The use of isotropic interactions to approximate anisotropic potentials therefore need not worry us.

As discussed in Chapter 2, the smallest scale features of the protein energy landscape correspond to conformational changes not along the

[1] When the context is clear, we use the terms protein, sequence, conformation, *etc.*, to refer to their respective biological or lattice counterparts.

protein polypeptide backbone. These fluctuations are typically of the order k_BT [5], the thermal energy of the atoms themselves. Lattice models avoid computation on this scale by neglecting atomic detail, capturing solely the more thermodynamically relevant conformational changes along the peptide chain. It is this reason especially that makes simulation of large lattice proteins computationally feasible.

Unlike their biological counterparts, lattice proteins are amenable to analytic analysis resulting from incorporated symmetries. These include the isotropy of the amino acid potentials and excluded volumes, the constant bond length and the discrete degrees of freedom. In practice, this means we need to integrate over fewer variables (in the case of simulation, less phase space must be explored). These simplifications allow the application of analytic techniques borrowed from other phenomena governed by complex energy landscapes, *e.g.*, spin glasses and neural networks.

3.2 Folding Dynamics

The folding of lattice proteins described above amounts to exploration of the ensemble of self-avoiding walk (SAW) configurations. The sampling must be ergodic (all conformations can be explored) and satisfy detailed balance (a transition and its reverse are equally probable), as well as define a sensible measure on the conformation space. This measure should reflect the physical view that proteins fold to conformationally similar structures in less time than conformationally distant configurations. It may be achieved by sampling, at any one time step, from among conformations derived from spatially localised perturbations only.

We fold proteins according to above by the repeated application of the move set containing end bends, corner flips and crankshaft motions (Figure 3.3); the dynamics are typified in Figure 3.2. This set respects linear connectivity and is applied such that the condition of excluded volume is maintained. Each move satisfies detailed balance and, apart from a vanishingly small number of pathological configurations (*e.g.*, a figure eight with both holes plugged by the chain ends), the move set is ergodic. We observe self-avoiding walk statistics at

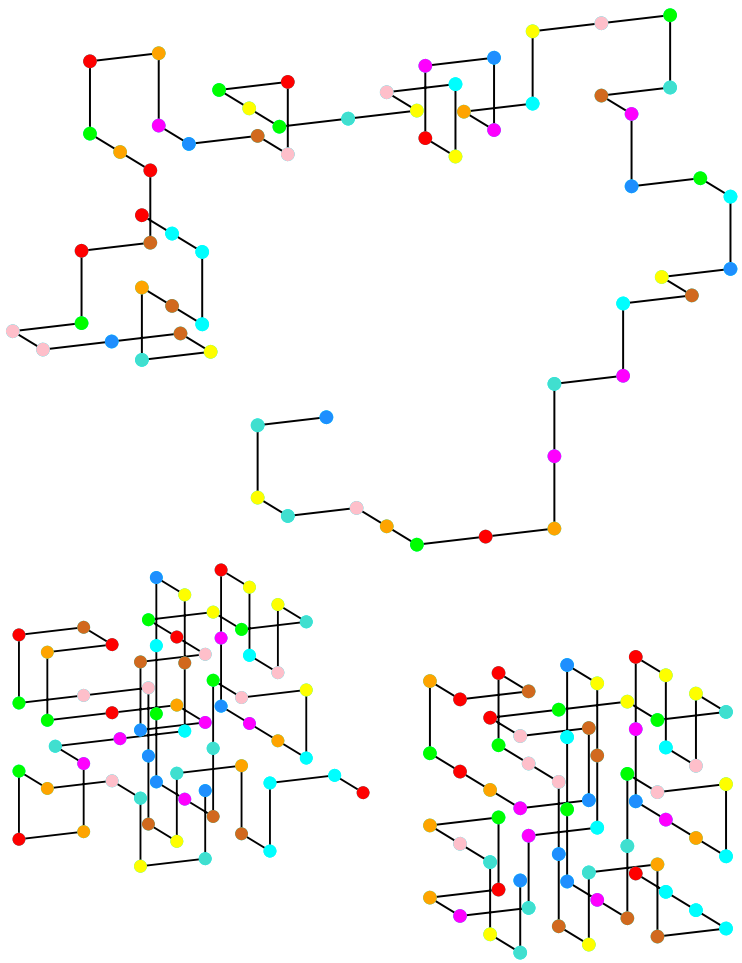

Figure 3.2: Spontaneous folding of a 64-mer trained to compact target conformation. Top: Denatured protein at infinite temperature. Folding begins at finite temperature. Left: Intermediate collapsed state. Right: Mirror image of desired target conformation. Since there are no chiral interactions, chirality of target is not preserved.

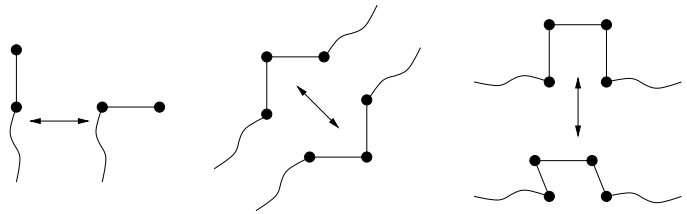

Figure 3.3: Move set for the simulation of a protein on a cubic lattice. From left: end bends, corner flips and crank-shaft motions.

infinite temperature.

At finite temperature, the protein has a preference for low energy configurations, and the sampling must be biased with respect to energy accordingly. This is achieved by the application of the Metropolis algorithm, in which moves are conditioned by an acceptance probability dependent on the resultant change in energy. Let P_{12} be the probability of a transition (selected, say, from the move set above) from conformation 1 to conformation 2; it must satisfy

$$P_{12} = \begin{cases} 1, & \Delta E < 0, \\ \exp(\frac{-\Delta E}{k_B T}), & \Delta E > 0, \end{cases} \tag{3.1}$$

where ΔE is the change in energy, T is the temperature and k_B is Boltzmann's (or some suitably redefined) constant. The repeated application of (3.1) gives rise to a Boltzmann distributed ensemble of conformations.

3.3 Analytic Representation

A protein conformation may be represented by its contact map, which, for an N-monomer sequence, consists of an $N \times N$ symmetric matrix C with $C_{ij} = 1$ if the ith and jth monomers are nearest neighbours and $C_{ij} = 0$ otherwise (Figure 3.4). The contact map is symmetric with zeroes along the diagonal (the *ith* monomer cannot be its own neighbour) and ones along the remaining tri-diagonals (due to backbone interactions). For a compact (maximally bonded) conformation,

the contact map is believed to be (though we know of no proof) a unique representation. More open conformations, however, may have degenerate maps; completely open conformations, for example, have no bonds apart from along the backbone.

Each monomer can take on one of A species, where A is the size of the amino acid alphabet. These species interact according to an $A \times A$ pair potential U (*e.g.*, Figure 3.5); the 20×20 potential derived in [16] from the distribution of contact energies in native proteins is typical. The interactions available to a particular protein sequence are conveniently expressed by the $N \times N$ extended pair potential $\tilde{U}$ (Figure 3.5), where $\tilde{U}_{ij}$ is the interaction energy of the ith and jth monomers according to their species S_i and S_j, *i.e.*, $\tilde{U}_{ij} = U_{S_i S_j}$. As indicated by the contact map, only bonds between monomers of opposite parity are topologically possible.

The energy of a protein sequence S embedded in conformation Γ may be compactly expressed

$$E(S, \Gamma) = \frac{1}{2} \sum_{ij=1}^{N} C_{\Gamma_{ij}} \tilde{U}_{S_{ij}}, \tag{3.2}$$

where C_Γ is the contact map corresponding to Γ and $\tilde{U}_S$ is the extended pair potential associated with S.

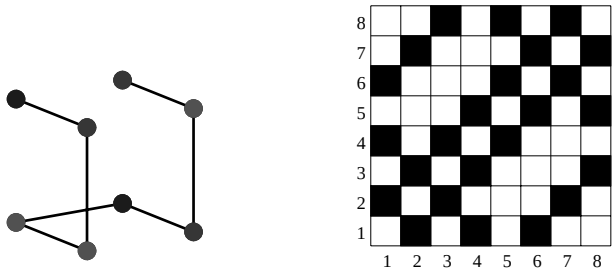

Figure 3.4: Left: An 8 monomer protein in a compact $2 \times 2 \times 2$ conformation Γ. Right: The corresponding 8×8 contact map C.

1 2 3 4 5 6 7 8
B−A−C−C−B−A−C−A

	A	B	C
C	-3	1	-2
B	2	3	1
A	-1	2	-3

	B	A	C	C	B	A	C	A
A	2		-3		2		-3	
C		-3		-2		-3		-3
A	2		-3		2		-3	
B		2		1		2		2
C	1		-2		1		-2	
C		-3		-2		-3		-3
A	2		-3		2		-3	
B		2		1		2		2

Figure 3.5: Left: An 8 monomer sequence S composed of 3 species, with the 3×3 pair potential U below. Right: The corresponding 8×8 extended pair potential $\tilde{U}$. Only bonds between monomers of opposite parity are topologically possible on a cubic lattice.

3.4 Shakhnovich Selection Scheme

As suggested in Chapter 2 and considered in detail in Chapter 4, successful protein design relies on two conditions: thermodynamic stability and kinetic accessibility of the target conformation. In the language of energy landscapes, the target configuration must be a global minimum sufficiently deep such that there exists a large gap between the native and set of non-native states and this target must lie at the bottom of a funnel of conformations sloping toward it.

The construction of a basin of attraction centred about the target conformation has proved to be a challenging task. As yet there exists no function of the target coordinates which (when extremised) yields a sequence whose energy landscape possesses this topography. We discuss two fundamentally different approaches to this problem in Chapters 6 and 7; in the meantime, we look to stability instead.

Ensuring that the target state is a deep global minimum in the space of conformations is relatively simple. Shakhnovich [19] introduced a straightforward selection algorithm in which the energy of a protein embedded in the target conformation is minimised over sequence space. He demonstrated, via simulation of lattice proteins, that a sequence trained[2] in this manner repeatedly folds back to its

[2]We refer to the selection of a sequence on its apparent ability to fold to a specified target conformation as training.

target conformation (Figure 3.2).

3.5 Minimisation of E

A sequence S is said to fold to a conformation Γ_{target} if (over all conformations) S minimises its energy. Let the folding function F map a sequence to the conformation to which it folds. The condition for folding may be compactly expressed

$$F(S) = \Gamma_{\text{target}} \quad \Leftrightarrow \quad E(S, \Gamma_{\text{target}}) = \min_{\Gamma}[E(\Gamma|S)]. \tag{3.3}$$

That S folds to Γ_{target} is, of course, no guarantee that S is thermodynamically stable in Γ_{target}; this depends on the folding temperature and the gap between Γ_{target} and the set of non-native conformations. It simply ensures that, after a long time, it will have spent more time there than in any other conformation.

Shakhnovich [19] observed in folding simulations that if S_{deep} is a deep local minimum in Γ_{target} with respect to sequence, then Γ_{target} is the *global* minimum of S_{deep} with respect to conformation. In the language of (3.3),

$$\operatorname*{deep}_{S}[E(S|\Gamma_{\text{target}})] = \min_{\Gamma}[E(\Gamma|S_{\text{deep}})], \tag{3.4}$$

and hence

$$E(S_{\text{deep}}, \Gamma_{\text{target}}) = \min_{\Gamma}[E(\Gamma|S_{\text{deep}})], \tag{3.5}$$

from which it follows that S_{deep} folds to Γ_{target}.

Equation (3.4) can be qualitatively understood as follows. Divide the space of possible compact structures into two sets: those which are conformationally near to the target and the much larger set of structures which are conformationally distant. Since the trained sequence is selected on the basis of reducing the energy of the target, the energies of the first set become increasingly deep as the conformations approach the target.

The trained sequence is not correlated with conformations in the second set; consequently, the statistics of the energy behave as though it were a random heteropolymer [4]. Under certain conditions of the interaction matrix [Ch. 4] and the amino acid alphabet size [Ch. 5], the fluctuations in energy away from the target conformation remain above the energy of the target. Accordingly, the target conformation coincides with the ground state conformation.

The Shakhnovich training scheme, apart from selecting sequences which fold to desired (compact) targets, is noteworthy for two reasons. Perhaps not surprisingly, trained sequences provide a significant gap between the energies of the native and set of non-native conformations. This means that at a temperature above the glass transition temperature T_c, the target conformation is thermodynamically stable. Below T_c, where any ground state can be made to be stable, the ruggedness of the energy landscape dominates the kinetics and exploration becomes extremely slow.

Thermodynamic stability without kinetic accessibility does not save us from the Levinthal paradox — golf-course landscapes are as difficult to traverse as heteropolymer landscapes. Remarkably, it appears that proteins trained to be thermodynamically stable in the target conformation are more kinetically accessible as well. The extent of the (presumably) attendant basin of attraction is unknown[3]. We provide a qualitative study of the correlation between well depth and funnel width in Chapter 4 and present a quantitative analysis in Chapter 6.

It is of practical importance that the training scheme require a deep, but not global, minimum in sequence space. Determining the global minimum sequence is a difficult optimisation problem; finding a *deep* minimum is more readily achieved, for example, by simulated annealing (the method used by Shakhnovich in [19] and outlined in [18]).

[3] Interestingly, the correlation of non-native conformations to the target configuration could be used to provide an estimate of the width of the corresponding funnel.

3.6 Minimisation of Z

Minimisation with respect to sequence of the energy of the target conformation only approximates the maximisation of its thermodynamic stability, on which the selection scheme described above rests. Unconstrained optimisation of the sequence, for example, yields a homopolymer or alternating bipolymer, depending on the minimum interaction energy of the pair potential. While this sequence has minimal energy, the gap between the native and set of non-native states has vanished. This difficulty is overcome, in practice, by imposing a constant composition of the amino acids in the sequence; optimisation occurs by varying the order.

The approximation may be made more precise by minimising the relative energy Z, proposed by Abkevich and coworkers in [1]. It estimates the difference between the target energy and the mean energy of the set of compact non-target structures, scaled with respect to the standard deviation of the interaction energies available to the sequence;

$$Z = \frac{E_{\text{target}} - E_{\text{av}}}{\sigma}. \tag{3.6}$$

The average energy of the compact conformations E_{av} is equivalent to Be_{av}, where e_{av} is the mean bond energy and B is the number of bonds (neglecting the backbone) in a compact conformation; for a cube of side n, $B = 2n^3 - 3n^2 + 1$. To calculate e_{av}, it is necessary to integrate over all compact conformations the mean bond energy of a single configuration. We may estimate e_{av} instead by taking the mean of the interaction energies available to the protein, that is, the mean of the extended pair potential $\tilde{U}$. Likewise, the standard deviation of the interaction energies σ may be approximated by the standard deviation of the elements of $\tilde{U}$.

Minimisation of the numerator selects for amino acids corresponding to interaction energies from the tails of the distribution — negative energies for bonds in the target conformation and positive energies for bonds favoured by distant configurations. Note that this is precisely the means by which σ is maximised (and sequence species diversity minimised). This is offset by minimising the denominator,

which draws bonds from the interior of the distribution, consequently increasing species diversity. The relative energy Z remains the standard Hamiltonian for thermodynamically oriented sequence selection.

Chapter 4
Stability and Accessibility

> By just exchange one for the other giv'n...
> There never was a better bargain driv'n.
>
> SIR PHILIP SIDNEY

Successful protein design is characterised by two criteria: thermodynamic stability, the probability of occupation of the target conformation, and kinetic accessibility, the ability of the protein to quickly fold to its target conformation. We observe a conflict between stability and accessibility and argue that modest reduction in the former allows significant increase in the latter. It follows that thermodynamically oriented sequence selection is not suitable for optimal protein design.

4.1 Introduction

Protein folding is concerned with determining the ground state conformation of a given sequence of amino acids. Inverse protein folding, or protein design, asks what sequence of amino acids possesses a given conformation as its ground state. The driving force of protein folding is the statistical mechanical evolution toward minimal conformational free energy, observed *in vitro* and verified by simulation. Given the need for a biologically useful target conformation, what mechanism drives protein design?

Unlike protein folding, protein design occurs in Nature on evolutionary time scales too long for our observation. Moreover, practical applications of protein design suggest that we do not necessarily wish to mimic nature in determining appropriate sequences. We begin, instead, by studying those qualities associated with successful folding to specific conformations.

4.2 Stability and Accessibility

The creation of a global minimum above the target conformation, as discussed in Chapter 2, is not sufficient for successful protein design. The protein must fold stably and efficiently to its target as well. We characterise (inverse) protein folding according to the extent to which these two conditions are satisfied.

Stability requires that the protein spend a significant fraction of its time in its biologically useful native conformation. This necessitates a pronounced energy gap between the native and set of non-native states such that the target conformation is occupied at the folding temperature. In principle, stability depends on the energy of the set of non-native states as much as the target (via the Boltzmann factor). In practice, the random energy model [4] allows us to approximate it with the (relative) energy of the native state.

Efficiency relates to the time necessary for a protein to fold, which corresponds to the first-passage time from a denatured state to the target conformation. But protein first-passage time is a broadly distributed random variable [Chapter 7]; accordingly, we measure efficiency by the mean first-passage time (MFPT).

Both criteria may be expressed in terms of the corresponding energy landscape. Stability coincides with a deep energy well above the target state and the absence of conformationally distant deep traps. Accessibility requires a landscape topography characterised by a folding funnel sloping toward the target, sufficiently steep to ensure that the loss of entropy is compensated by the increase in energy.

It has been suggested that the former condition implies the latter, *i.e.*, that thermodynamically oriented selection of sequences solves the problem of kinetic accessibility as well [19]. We claim that while selecting for a pronounced minimum of the native state energy makes protein design feasible, maximising stability does not provide optimal accessibility. Moreover, we argue that above some critical thermodynamic stability, these two conditions are in conflict. In particular, modest reduction in stability can provide significant increase in accessibility.

4.3 Details of Simulation

We study the design and folding of model proteins N monomers in length, each of A possible species, constrained to an infinite cubic lattice with nearest neighbour interactions. Proteins are designed by minimising the target state energy with respect to the (fixed composition) sequence; we observe their folding dynamics via Monte Carlo simulations from self-avoiding walk (SAW) conformations to their respective target structures.

A protein conformation Γ is designated by its contact map C, an $N \times N$ matrix where $C_{ij} = 1$ if monomers i and j are nearest neighbours not along the backbone and 0 otherwise. Since they cannot be avoided by the conformation, backbone interactions cannot influence the folding dynamics. Nearest neighbour monomers interact according to their species S_i and S_j by the $A \times A$ pair potential U_{S_i,S_j}, where S_i is the species of monomer i along the sequence S, *etc.* We used the 20×20 potential derived in [16] from the distribution of contact energies in native proteins. We represent the protein sequence S by the $N \times N$ extended pair potential $\tilde{U}$, where $\tilde{U}_{ij}$ is the interaction of monomers i and j such that $\tilde{U}_{ij} = U_{S_i,S_j}$.

The energy of the protein may thus be expressed

$$E(S,\Gamma) = \frac{1}{2}\sum_{ij=1}^{N} C_{\Gamma_{ij}}\tilde{U}_{S_{ij}}, \tag{4.1}$$

which is the folding Hamiltonian.

Given a desired target conformation Γ_{target}, we design a sequence S_{design} by annealing the energy with respect to the sequence variables while the conformation remains quenched at Γ_{target} [19]. Folding of S_{design} is simulated on a cubic lattice at constant temperature by the Metropolis application of the moveset containing end bends, corner flips and crank-shaft motions, where multiple occupation of lattice sites is forbidden. Such a move set is ergodic (apart from a vanishingly small number of pathological configurations) and generates SAW statistics at infinite temperature.

Starting from denatured (SAW) initial conformations, simulation continues until the energy of the target conformation $E(\Gamma_{\text{target}}|S_{\text{design}})$

is reached. The number of attempted moves required for this to occur is the first-passage time. We find that the energy of S_{design} embedded in Γ_{target} is achieved only by Γ_{target} for compact (maximally bonded) structures, in accordance with [19]; in other words, Γ_{target} is the global minimum conformation of S_{design}.

4.4 Shift of Pair Potential

For a given sequence, the probability that the ground state conformation Γ_0 is occupied is

$$P_{\Gamma_0} = \frac{e^{-\beta E_{\Gamma_0}}}{\sum_\Gamma e^{-\beta E_\Gamma}}, \tag{4.2}$$

where E_Γ is the energy of the sequence embedded in conformation Γ. Consider shifting the pair potential matrix U by a constant α,

$$U_{ij}(\alpha) := U_{ij} + \alpha, \tag{4.3}$$

where α may be positive or negative and U, taken from [16], has near zero mean,

$$\sum_{ij=1}^{N} U_{ij} = 0.018 \simeq 0, \tag{4.4}$$

where $U_{ij} \equiv U_{ij}(\alpha = 0)$. Note that varying α has no influence on sequence optimisation for a fixed conformation. For a compact native state, the probability of occupation of the ground state then appears as

$$P_{\Gamma_0}(\alpha) = \frac{e^{-\beta\alpha B}e^{-\beta E_{\Gamma_0}}}{\sum_\Gamma e^{-\beta\alpha b}e^{-\beta E_\Gamma}}, \tag{4.5}$$

where B is the number of bonds in the compact target structure Γ_0 and b is the number of bonds in conformation Γ. Since $b \leq B$ for all Γ, it follows that

$$\frac{\partial P_{\Gamma_0}}{\partial\alpha} = \beta(\langle b\rangle - B)P_{\Gamma_0} < 0, \tag{4.6}$$

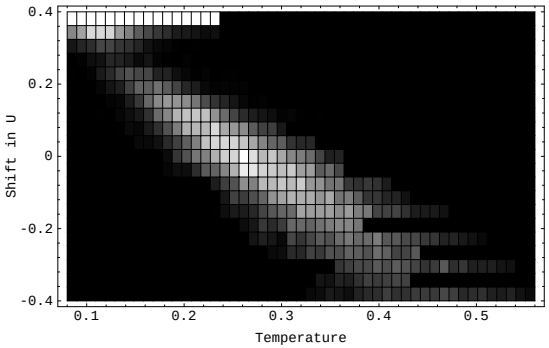

Figure 4.1: Inverse mean first-passage time to the target conformation for a 27 monomer sequence as a function of T and α. Each square represents the number of times the native state energy is reached ($N_{\rm ns}$) in 2×10^7 mcsteps. $N_{\rm ns}^{\rm max} = 164$ corresponds to white, 0 to black.

where $\langle b \rangle$ is the average number of bonds over the thermal ensemble of conformations at reciprocal temperature β.

From (4.6), the probability P_{Γ_0} that the native state Γ_0 is occupied is decreasing with α. Accordingly, negatively shifting the potential increases the energy gap between the native and the set of non-native states, thus increasing the thermodynamic stability of the desired conformation.

The extent to which this imposed increase in stability effects kinetic accessibility can be investigated via simulation. We define efficiency as the mean first-passage time $\langle t_{\rm fp} \rangle$ to a given target conformation, that is, the average number of Monte Carlo time steps necessary for a protein in a denatured state to fold to its target. We are interested in the MFPT as a function of the shift in the pair potential, α. However, the optimal folding temperature $T_{\rm opt}$ itself depends on α, so we consider instead $\langle t_{\rm fp}(\alpha, T) \rangle$.

As shown in 4.1, the mean first-passage rate exhibits a peak in the T, α plane near $\alpha = 0$. Making the mean interaction more negative promotes folding at higher temperatures, but decreases the attainable

folding efficiency. As $\alpha \to -\infty$, the protein effectively quenches to a suboptimal compact state; to explore conformational phase-space, it must pass over large energy barriers in order to unravel and collapse to other local minima. Thus, the gain in thermodynamic stability is made at the loss of folding efficiency. When $\alpha > 1$, the pair potential is purely repulsive and the protein is unable to fold; when $\alpha > 0.3$, the ground state ceases to be the target state.

4.5 Protein Folding is Many to One

The consistent success of the Shakhnovich training scheme relies on the existence of at least one sequence for each target conformation, *viz.*, that sequence which folds to the target. It follows that there must be at least as many sequences as compact conformations for this or any other satisfactory design procedure.

Improved protein design may be achieved by judiciously choosing a sequence, and hence an energy landscape, which more closely resembles a deep, broad funnel near the target conformation than does a thermodynamically oriented sequence. We thus require a sequence spectrum in which many sequences stably fold to a single target conformation. Assuming the ground state conformation of the large majority of sequences is compact, this implies the average number of sequences per compact conformation $\left\langle \frac{N_S}{N_{\Gamma_c}} \right\rangle$ must be much greater than 1.

For a random walk of N steps on a three dimensional lattice, the number of distinct conformations is approximately z^N, where z is the lattice coordination number. Less freedom is available to self-avoiding walks, to which our lattice proteins correspond; how much less depends on the protein's radius of gyration. As the protein approaches its compact native state, the conformational freedom diminishes rapidly. The number of configurations available to a protein can be estimated from Flory's theory of excluded volume for polymers [12]; for a compact lattice protein it has been approximated to grow as $\Omega_c(N) \sim \kappa^N$, $\kappa \sim 1.9$ [17].

The average folding degeneracy may be estimated to be

$$\left\langle \frac{N_{\rm S}}{N_{\Gamma_{\rm c}}} \right\rangle \simeq \frac{A^N}{\kappa^N} \gg 1, \tag{4.7}$$

where A^N is the number of sequences available to a protein made up of A amino acid species. Accordingly, the ability to design sequences which fold to specific targets requires $A > A_c \simeq 2$. This suggests that binary HP folding models, in which the number of species $A = 2$, are barely sufficient for the design of arbitrary compact proteins and incapable of producing stable, fast folding sequences.

This may be qualitatively argued as follows. Of the set of A^N possible sequences, a small fraction $(\frac{A}{\kappa})^N$ typically fold to a specific conformation. The large majority of these correspond to landscapes which have minima of comparable depth to the global (target) minimum; such a protein will consequently spend considerable time outside of the target conformation. Of those sequences which do fold stably, some will be more or less funnel oriented than others; the most kinetically accessible will not in general correspond to the most thermodynamically stable. The number of species A must be sufficiently large such that the first subset (those which fold to the target) contains at least one element. Further increase of A allows us to be more selective in our choice of sequence from the last subset (those which fold stably and quickly).

4.6 Accessibility-Stability Phase Space

Generally (*e.g.*, [19]) and in the present work, the Hamiltonians used to optimise protein sequence and structure are equivalent. That is, sequence design consists of minimising the energy (or relative energy) in an effort to maximise thermodynamic stability, with the tacit assumption that stable sequences (deep minima) fold quickly (are funnel shaped). While empirical observations suggest a correlation between the two, it does not imply that such sequences fold *most* efficiently to their target conformations. Roughly put, the deepest wells need not be the most funnel oriented.

This conjecture may be tested as follows. Consider the set of all sequences which fold to a given compact target conformation; such a

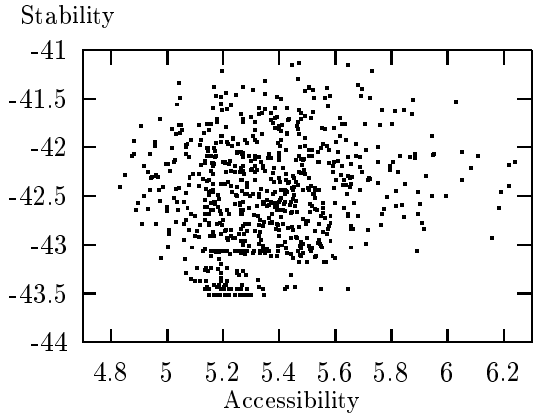

Figure 4.2: Ensemble of 27 monomer sequences independently trained to fold to a single compact target, plotted in accessibility-stability phase space. Accessibility is measured by the $\log_{10}$ of the mean first-passage time, stability by the energy of the sequence in the target configuration. Note that the most stable sequences do not correspond to the most quickly folding.

set is very large since in our case $A = 20 > A_c$. Plot each sequence, and hence landscape, in accessibility-stability phase space according to its mean first-passage time and target state energy. In practice, it is not feasible to examine all $(\frac{A}{\kappa})^N$ sequences which map to a specific compact conformation; the set must be suitably sampled.

We prepared an ensemble of 27-mers independently trained to fold to a single compact $3 \times 3 \times 3$ target conformation. The thermodynamic stability of each sequence is approximated by the energy of the sequence in the target conformation, the (presumably) ground state conformation of the sequence; the actual Boltzmann occupation probability depends on the energy gaps between the native and non-native states. The mean first-passage time is taken as the mean of $N_{\rm fold}$ folding times from a denatured to the target conformation, where $N_{\rm fold}$ is the number of times the sequence folds to its target in 2×10^7 Monte Carlo steps. Note that inversion of a rate gives better statis-

tics to those sequences which fold more efficiently, which are the ones in which we are particularly interested.

A landscape scatter plot is shown in Fig. 4.2. The bottom colinear points represent the sequentially degenerate ground state energy, due to rationality of the pair potential (of two significant figures) and certain sequence rearrangements. The spread of these mean first-passage times is significant, notwithstanding the accompanying uncertainties. A sequence annealed to minimal energy is effectively sampled from this range. Of greater interest is the large fraction of sequences with higher target energies and lower mean first-passage times. Approximately half, on average, of the sequences fold more efficiently than a typical ground state sequence, some significantly more so. Disregarding small variations about the target conformation due to entropic considerations, nearly all sequences shown spend a significant fraction of their time in the target conformation.

4.7 Conclusion

We have shown, both by shifting the mean of the pair potential and considering an ensemble of sequences trained to a single target, that thermodynamic stability and kinetic accessibility of the target conformation are in conflict. In particular, maximal stability does not correspond to maximal accessibility. A marginal reduction in the stability of the target conformation allows significant increase in folding efficiency.

While we have demonstrated that faster folding may be achieved, we have not addressed how the corresponding sequences should be selected.

We present in Chapter 6 a novel method of kinetically favoured sequence selection on the assumption that the widest possible funnel is that which least constrains the dynamics, which we propose is given by the conformations sampled during *unfolding* of the target conformation. Moreover, we provide arguments that lowering the energy of successive conformations, whether correlated, as in the case of a funnel, or independent, such as training to multiple targets, reduces the depth to which such conformations can be trained.

Alternatively, sequences may be annealed explicitly with respect to mean first-passage time; work to this end is reported in Chapter 7. Such a technique requires a statistical formalism for cost functions which are random variables (in the case of protein design, the folding time). This is the basis of stochastic annealing, presented in Chapter 9.

Chapter 5

Inverse Protein Folding as an Associative Memory

'It's a poor sort of memory that only works backwards,' the Queen remarked.

LEWIS CARROLL
Through the Looking Glass

We interpret a protein trained to fold to a target conformation as an associative memory. We generalise to the recognition of multiple conformations and provide capacity calculations based on energy fluctuations and information theory. Unlike the linear capacity of a Hopfield network, we find that the number of conformations which can be embedded in a protein sequence depends on the alphabet size as $\ln A$, independent of protein length.

5.1 Introduction

It is widely thought to be a design feature of real proteins that their native, biologically active state is both a deep global energy minimum and has a funnel of low energy configurations leading toward it [8]. The funnel guides the molecule to fold to its stable native conformation in a time much less than that required for it to explore all configurations, thus avoiding the so-called Levinthal paradox [15].

Inverse protein folding, or protein design, consists of designing a sequence of amino acids that quickly and stably folds to a desired target conformation. For simple lattice models, Shakhnovich and coworkers [1,19] have explored the folding of sequences designed to minimise a conformation's absolute and relative energies. This gives the molecule a deep ground state, but takes no measures to ensure that this state is readily accessible kinetically. We have previously provided evidence [Ch. 4] that such methods can be counterproductive toward achieving folding efficiency, *i.e.*, that stability and accessibility are at odds.

Here we show how Shakhnovich's approach can be generalised by analogy with neural network theory of associative memory to provide recognition of several configurations rather than a single target state. In doing this, we set the stage for addressing the problem of creating a broad funnel for a single conformation in Chapter 6.

5.2 Proteins as Associative Networks

A protein consists of a sequence S of N amino acids, or monomers, each of which can take on one of A possible species. We denote the species of the ith monomer of S by S_i, and monomers i and j interact according to the $N \times N$ extended pair potential $\tilde{U}$, where $\tilde{U}_{ij} = U_{S_i S_j}$ and U is the $A \times A$ pair potential.

Protein conformations may be represented by the contact matrix C, where $C_{ij} = 1$ if monomers i and j are nearest neighbours and 0 otherwise. For compact conformations, each interior monomer is surrounded by z neighbours, where z is the coordination number of the lattice. Accordingly, each row and column of C must contain z entries ($z - 2$ neglecting backbone connections). These entries will certainly be correlated, which we can quantify on the assumption of self-avoiding random walk statistics known to apply in polymer melts, but for the immediate purposes the correlations turn out to be of only secondary importance. Contact patterns are thought to be a unique representation of compact conformations (though this need not be the case for more open structures), and we approximate them as independent.

Protein folding may be considered pattern recognition in as much as the protein rapidly organises itself into the target pattern C upon entering the target basin of attraction (funnel). By analogy with pattern association, this idea may be generalised to the recognition of multiple patterns. This raises the question of how to train the se-

quence to recognise more than one conformation. The dilute representation in the contact pattern suggests that we may superimpose p patterns without saturation[1], providing us with a total pattern to which we train in the usual way [Ch. 3].

Interpreting the N monomers as neurons, the contact map may be viewed as a synapse map, with two neurons connected if they are nearest neighbours in space. The network of monomers is hence dilutely connected, with (for compact structures) asymptotically $\frac{z}{N}$ of the possible number of synapses present. The network corresponding to the total pattern is more heavily connected, averaging zp connections to each neuron.

The protein sequence is trained to its target pattern by keeping the synapse map (contact map C) quenched to the target while the synapse strengths (extended pair potential $\tilde{U}$) are annealed. Folding of a fixed sequence occurs at fixed synapse strengths through the evolution of the synapse map, from complete dilution in a denatured state to $\frac{z}{N}$ dilution in the target conformation.

5.3 Energy Function

The energy of a sequence in conformation C may be conveniently expressed

$$E = \frac{1}{2}\sum_{ij=1}^{N} C_{ij}\tilde{U}_{ij}. \tag{5.1}$$

For a sequence trained to have minimal energy in conformation Γ_μ, the energy appears as

$$E_\mu^{\min} = \min_{\tilde{U}}\left[\frac{1}{2}\sum_{ij=1}^{N} C_{\mu_{ij}}\tilde{U}_{ij}\right] = \frac{1}{2}\sum_{ij=1}^{N} C_{\mu_{ij}}\tilde{U}^*_{ij}, \tag{5.2}$$

where minimisation is over all $\tilde{U}$ corresponding to distinct sequences. The energy of a fixed sequence S_ν folded to its ground state conformation is

$$E_\nu^{\min} = \min_{C}\left[\frac{1}{2}\sum_{ij=1}^{N} C_{ij}\tilde{U}_{\nu_{ij}}\right] = \frac{1}{2}\sum_{ij=1}^{N} C^*_{ij}\tilde{U}_{\nu_{ij}}, \tag{5.3}$$

where $\tilde{U}^*$ minimises E_μ and C^* minimises E_ν. We refer to the typical value of E_ν for an untrained sequence as the copolymer energy $E_{\rm cp}$.

Throughout this chapter, the energy of a sequence realised in a particular conformation is indicated by E, while the Hamiltonian with which a sequence is trained (generally the linear combination of the energies realised in a number of conformations) is denoted by H.

5.4 Capacity from Energetics

We consider the capacity of the designed protein, that is, the number of conformations p that we can train the sequence to make simultaneously thermodynamically stable. For a protein to fold to a single target conformation, it is necessary that the energy of the trained sequence realised in that conformation, $E_\mu^{\min}$, be below the minimum fluctuations of the energy elsewhere, thereby making the target minimum global. Since the trained sequence is not correlated with distant conformations, energy fluctuations away from the target structure are statistically equivalent to those of a random copolymer sequence. We therefore require that the trained energy be less than the minimum energy of a random sequence, that is, $E_\mu^{\min} < E_{\rm cp}^{\min}$. Folding to a set of p conformations requires that the minimum energy of all of these lie below $E_{\rm cp}^{\min}$.

We now approximate the typical energy of a sequence optimally trained to a set of p target conformations and arranged in one of these configurations. The total contact map, to which we train by energy minimisation with respect to the sequence, is defined as a linear superposition of the p corresponding contact maps, that is

$$C_{{\rm tot}_{ij}} = \sum_{\mu=1}^{p} C_{\mu_{ij}}. \tag{5.4}$$

[1] This assumes that the number of patterns stored does not scale more quickly than N, which we find below is reasonable.

The minimum Hamiltonian associated with the total contact map may then be written

$$H_{\rm tot}^{\rm min} = \frac{1}{2}\sum_{ij=1}^{N} C_{\rm tot_{ij}}\tilde{U}_{ij}^{*} = \frac{1}{2}\sum_{ij=1}^{N}\sum_{\mu=1}^{p} C_{\mu_{ij}}\tilde{U}_{ij}^{*}; \tag{5.5}$$

here $\tilde{U}^{*}$ minimises $H_{\rm tot}$. It is simply the sum of the p individual conformational energies of the sequence implied by $\tilde{U}^{*}$. We re-express the right side of (5.5) as the sum over i of the total energy associated with monomer i, $H_{{\rm tot}_i}$, each minimised with respect to the choice of amino acid at monomer i, S_i,

$$H_{\rm tot}^{\rm min} = \sum_{i=1}^{N} \min_{S_i}[H_{{\rm tot}_i}]; \tag{5.6}$$

$H_{{\rm tot}_i}$ is obtained by summing over the connections to monomer i,

$$H_{{\rm tot}_i} = \frac{1}{2}\sum_{j=1}^{N}\sum_{\mu=1}^{p} C_{\mu_{ij}}\tilde{U}_{ij}. \tag{5.7}$$

Since C has z bonds connecting to monomer i, each $H_{{\rm tot}_i}$ is the sum of $\frac{zp}{2}$ random interaction energies freely chosen from the pair potential[2]. For simplicity of analysis, we approximate the distribution of $H_{{\rm tot}_i}$ by its central limit theorem form. Assuming a distribution of contact energies with zero mean (as is the case of that found in [16]) and standard deviation σ, this yields the probability density

$$f(H_{{\rm tot}_i}) \simeq \frac{1}{\sqrt{2\pi}\sigma_{{\rm tot}_i}}\exp(-\frac{H_{{\rm tot}_i}^2}{2\sigma_{{\rm tot}_i}^2}), \tag{5.8}$$

where $\sigma_{{\rm tot}_i}^2 = \frac{zp}{2}\sigma^2$. This estimation is valid out to $|H_{{\rm tot}_i}|$ of order $\frac{zp}{2}\sigma$.

The energy $H_{{\rm tot}_i}$ at each monomer is minimised with respect to the choice of amino acid by choosing the smallest of A samples from the gaussian $f(H_{{\rm tot}_i})$. Accordingly, the probability distribution of $H_{{\rm tot}_i}$

[2]Bringing the $\frac{1}{2}$ from (5.7) into the sum over bonds index accounts for frustration.

being the minimum of A samples of (5.8) appears as

$$f^{\rm min}(H_{{\rm tot}_i}) = Af(H_{{\rm tot}_i})(1-F(H_{{\rm tot}_i}))^{A-1}, \tag{5.9}$$

where $F(H_{{\rm tot}_i})$ is the usual cumulative distribution,

$$F(H_{{\rm tot}_i}) = \int_{-\infty}^{H_{{\rm tot}_i}} f(x){\rm d}x. \tag{5.10}$$

Maximising $f^{\rm min}$ with respect to $H_{{\rm tot}_i}$ yields the transcendental equation

$$H_{{\rm tot}_i}^{\rm min}(1-F(H_{{\rm tot}_i}^{\rm min})) = -\sigma_{{\rm tot}_i}^2(A-1)f(H_{{\rm tot}_i}^{\rm min}), \tag{5.11}$$

where $H_{{\rm tot}_i}^{\rm min}$ is the minimum of the $H_{{\rm tot}_i}$. For reasonably large A, $F(H_{{\rm tot}_i})$ is small and we estimate $H_{{\rm tot}_i}^{\rm min}$ as

$$H_{{\rm tot}_i}^{\rm min} \simeq -\sqrt{2}\sigma_{{\rm tot}_i}\sqrt{\ln A}, \tag{5.12}$$

from which it follows that[3]

$$H_{\rm tot}^{\rm min} \simeq -\sqrt{2}N\sigma_{{\rm tot}_i}\sqrt{\ln A}. \tag{5.13}$$

When the trained sequence is in one of the p target conformations, the energy of the sequence is, on average, given by

$$E_{\mu}^{\rm min} \simeq \frac{H_{\rm tot}^{\rm min}}{p} \simeq -\sqrt{\tfrac{z}{p}}N\sigma\sqrt{\ln A}. \tag{5.14}$$

The minimum copolymer energy $E_{\rm cp}$ may be estimated by similar arguments. Since the extended pair potential in the copolymer energy from (5.3) is untrained, we consider the usual product of it and the contact map as the sum of $\frac{zN}{2}$ random bonds. Accordingly, the energy is distributed as

$$f(E_{\rm cp}) \simeq \frac{1}{\sqrt{2\pi}\sigma_{\rm cp}}\exp(-\frac{E_{\rm cp}^2}{2\sigma_{\rm cp}^2}), \tag{5.15}$$

[3]This estimation is consistent with our use of the central limit theorem provided $\sqrt{2}\sigma_{{\rm tot}_i}\sqrt{\ln A} < \frac{zp}{2}\sigma$, that is, $\ln A < \frac{zp}{4}$.

where $\sigma_{\rm cp}^2 = \frac{zN}{2}\sigma^2$. Since the number of compact conformations of an N-mer follows $\Omega_c(N) \sim \kappa^N$ [17], the energy of the ground state is the minimum of κ^N samples of $f(E_{\rm cp})$ — again we wish to estimate the minimum of many samples of a gaussian. In accordance with (5.12),

$$E_{\rm cp}^{\rm min} \simeq -\sqrt{2}\sigma_{\rm cp}\sqrt{\ln(\kappa^N)} = -\sqrt{z}N\sigma\sqrt{\ln\kappa}, \tag{5.16}$$

where $\kappa \simeq 1.9$ [17] on a cubic lattice.

Comparing the minimum energy of the trained sequence (5.14) and the minimum copolymer energy (5.16) yields

$$p_{\rm max} \simeq \tfrac{\ln A}{\ln\kappa}. \tag{5.17}$$

5.5 Capacity from Information Theory

This result may also be achieved by information theoretic considerations. Consider the transmission of a message, which has been encoded as an N letter sequence and the $A \times A$ pair potential U. The message is decoded empirically by constructing the sequence (either *in vitro* or via computer simulation), allowing it to fold according to the pair potential and observing the p most occupied, and consequently lowest, target conformations.

The information retrieved by learning a single conformation may be determined as follows. Given κ^N possible compact conformations, the information contained in one conformation is equivalent to the number of nats[4] necessary to express a number between 1 and κ^N, *viz.*, $\ln(\kappa^N)$. Since the p target configurations are assumed to be independent, the total information captured scales linearly with p, namely $pN\ln\kappa$.

The information transmitted may be similarly determined. Since the number of sequences grows as A^N, the information associated with a sequence is $\ln(A^N)$. The information associated with the pair potential $I(U)$ is less easily quantified; among other things it depends on the precision of the interaction energies. For our purposes we only need know that it is finite and independent of N.

[4] A nat is the base e unit of information analogous to bits for base 2.

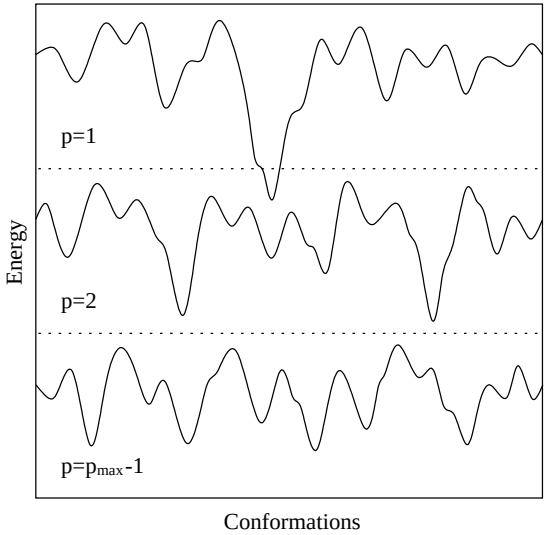

Figure 5.1: Energy landscapes of sequences trained to be thermodynamically stable in a single, multiple and $p_{\rm max}-1$ target conformations. As the number of targets increases, the depth to which the targets can be trained diminishes. At $p = p_{\rm max}$, the target wells are indistinguishable from nearby fluctuations.

The information retrieved must not be greater than the information transmitted, that is,

$$\ln(A^N) + I(U) \geq p\ln(\kappa^N). \tag{5.18}$$

For large N, $I(U)$ is negligible and the bound on p reduces to

$$p_{\rm max} \simeq \tfrac{\ln A}{\ln\kappa}, \tag{5.19}$$

which is identical to the result deduced from fluctuations in the energy landscape in (5.17). As p approaches $p_{\rm max}$, the typical well depth diminishes such that, at $p = p_{\rm max}$, the minima are lost among nearby fluctuations (Figure 5.1).

The agreement of these two results suggests that the training procedure is optimal, although only in the large N limit. In any event, our constant capacity result is not (as has been suggested) a shortcoming of the superposition rule.

5.6 Retrieving Memories from Proteins

The protein capacity may be understood in the context of Hopfield memories, whose capacities increase linearly with N [2]. Unlike proteins, Hopfield networks are globally connected — each of the N neurons is bonded to $N-1$ others. Since capacity is proportional to connectivity [2], the Hopfield capacity is linear. Locally connected networks, as is the case for proteins, accordingly possess constant capacity. Perhaps more surprising is that the answer is governed by the number of amino acids A rather than by the coordination number z.

For a uniform composition (*i.e.*, a homopolymer), zero conformations are encodable, as expected. Frequently studied binary models allow at most one configuration to be stored, while for a twenty amino acid set, $p_{\max}$ approaches 4.67 from above.

A Hopfield network recognises stored memories by associating different initial states of the system with the minima of the basins of attraction in which they lie. A biological protein has a single basin of attraction and all conformations lie in the basin or eventually make their way to it (though they may eventually leave it for a while as well). What happens to a protein designed to fold to multiple conformations? Since it begins folding from denatured states conformationally distant from any of the compact targets, it is difficult to control the first target minimum into which the protein falls; the protein meanders through the energy landscape, organising itself into that target to which it comes near. Moreover, since at equilibrium the probability of occupation of each state is proportional to its Boltzmann factor, in the long term all target states are visited in relation to the relative depths of their wells.

Thermodynamic target control relies on the variation of temperature to affect a change of the conformation occupied by the protein. Unfortunately, for a sequence trained, say, to $p=2$ conformations, at

equilibrium this allows at best (in the limit of infinite energy gaps) the constant occupation of Γ_1 at a lower temperature and an equally probable occupation of Γ_1 and Γ_2 at a higher temperature.

Interestingly, non-equilibrium target selection can be thermodynamically administered. Consider an ensemble of proteins on an energy landscape with a broad funnel and, far away, a narrower but deeper well. At a suitably high temperature the funnel is thermodynamically insignificant and the well becomes occupied. At very low temperatures the funnel is initially occupied by nearby proteins before equilibrium is reached.

Equilibrium target control can be implemented with the assistance of topographically distinct macromolecules which act as chaperones. This may be achieved by training to a set of conformations, to each of which folding is promoted by one of a set of chaperones. By introducing a single macromolecule species, folding is favoured toward the target state facilitated by that chaperone. This can be made more general by instead designing macromolecules which promote a pre-specified set of target conformations.

5.7 Conclusion

We have shown that a protein may be trained to recognise multiple conformations, analogous to an associative memory. We find that manipulation of the protein energy landscape by the introduction of independent minima is limited by $p_{\max} \simeq \frac{\ln A}{\ln k}$. As the number of independent minima approaches the capacity $p_{\max}$, the typical depth of the minima decreases until they are eventually lost in nearby fluctuations. Training to a single conformation requires an alphabet size $A > \kappa$; increasing A allows greater stability of the target structure.

Chapter 6

Funnel Design based on Unfolding Dynamics

Time shall unfold what plighted cunning hides.

WILLIAM SHAKESPEARE

The ability of a protein to recognise multiple independent target conformations was presented in Chapter 5. Here we generalise this idea to the recognition of correlated configurations, which may be applied to the problem of funnel design for a single conformation. The maximum basin of attraction, as parametrised in our model, depends on the alphabet size as $\ln A$.

6.1 Introduction

It is believed [Chapter 4] that a stable, fast folding protein requires a sequence whose conformational energy landscape contains both a deep global minimum above the native conformation and lies at the bottom of a basin of attraction sloping toward it. These conditions are known as thermodynamic stability and kinetic accessibility, respectively. While stability may be readily achieved by suppressing the energy of the sequence embedded in the target conformation, constructing a broad funnel leading toward the target has remained elusive.

The first satisfactory method of protein design, introduced by Shakhnovich in 1994 [19], relies on the correlation between stability and accessibility: stable sequences are found to fold more quickly as well. Minimising the energy or relative energy [1, 19] with respect to sequence (while the conformation remains quenched to the target) yields sequences whose conformational energy is a deep global minimum above the target and which fold much more rapidly than random heteropolymers of equal length. We have provided evidence, nonetheless, that the most stable sequences are not the fastest folding, and that a modest reduction in stability allows significant gain in efficiency.

We introduce in Chapter 7 a method of sequence design in which a sequence is optimised on the basis of folding time itself. Previous work to this end [13] was limited by the difficulty of accurately measuring the broadly distributed folding time. By exploiting this uncertainty to allow, in a controlled manner, uphill transitions as well as downhill moves, we were able to select sequences which fold in near-optimal time.

Kinetically oriented sequence selection, useful in exploring the limits of protein folding efficiency, is presently impractical as a means of designing long proteins. We provide in this chapter a method of design which relies on training to multiple targets discussed in Chapter 5. Unlike the independent conformations previously considered, here our patterns are correlated to a single target conformation.

Our approach to funnel design is to turn off all the monomer interactions (equivalent to an interacting system at infinite temperature) and to consider the dynamics by which a protein would then spontaneously unfold from the target state into a random ensemble. By the principle of detailed balance in equilibrium statistical mechanics, the ensemble of unfolding trajectories from the target state to random conformations is equivalent to the ensemble of folding trajectories from random configurations to the target — but of course the former ensemble is much more easily sampled. Therefore, observations of unfolding will tell us how the molecule would with least dynamical constraint fold.

We provide estimates of the unfolding contact map based on a simple blob model of unfolding. This leads to a definite proposal as to how different stages in the unfolding contact map should be weighted in training so as to create an optimal funnel.

6.2 Generalisation to Weighted Training

The capacity calculation of the previous section assumed equal weighting of the p target conformations. Here we generalise to weighted superposition: the total contact map is defined by summing over the individual maps with suitable weights,

$$C_{\mathrm{tot}_{ij}} = \sum_{\mu=1}^{p} w_\mu C_{\mu_{ij}}, \tag{6.1}$$

where w_μ is the weight associated with conformation $\Gamma\mu$. Let $\tilde{U}^*$ correspond to the sequence with minimum Hamiltonian in the total weighted contact map (6.1). Then

$$H_{\mathrm{tot}}^{\min} = \frac{1}{2}\sum_{ij=1}^{N} C_{\mathrm{tot}_{ij}}\tilde{U}^*_{ij} = \frac{1}{2}\sum_{ij=1}^{N}\sum_{\mu=1}^{p} w_\mu C_{\mu_{ij}}\tilde{U}^*_{ij}. \tag{6.2}$$

By analogy with calculations in the previous chapter, we reexpress (6.2) as a sum over H_{tot_i}, each minimised by the choice of S_i,

$$H_{\mathrm{tot}}^{\min} = \sum_{i=1}^{N} \min_{S_i}[H_{\mathrm{tot}_i}], \tag{6.3}$$

where H_{tot_i} is the sum over the connections to monomer i,

$$H_{\mathrm{tot}_i} = \frac{1}{2}\sum_{j=1}^{N}\sum_{\mu=1}^{p} w_\mu C_{\mu_{ij}}\tilde{U}_{ij}. \tag{6.4}$$

The local Hamiltonian H_{tot_i} is simply a weighted sum of the independent local conformational energies,

$$H_{\mathrm{tot}_i} = \sum_{\mu=1}^{p} w_\mu E_{\mu_i}. \tag{6.5}$$

Proceeding as before in Chapter 5, we find that H_{tot_i} is gaussianly distributed with standard deviation

$$\sigma^2_{\mathrm{tot}_i} = \frac{z\sigma^2}{2}\sum_{\mu=1}^{p} w_\mu^2. \tag{6.6}$$

We now consider the ith component of H_{tot} in (6.5) as a sum of two terms,

$$H_{\mathrm{tot}_i} = w_\mu E_{\mu_i} + \sum_{\nu=1,\nu\neq\mu}^{p} w_\nu E_{\nu_i} = H_{\mu_i} + H_{\mathrm{oth}_i}. \tag{6.7}$$

Since H_{μ_i} and H_{oth_i} are independently distributed with

$$\sigma^2_{\mu_i} = \frac{z\sigma^2}{2}w_\mu^2 \quad \text{and} \quad \sigma^2_{\mathrm{oth}_i} = \frac{z\sigma^2}{2}\sum_{\nu=1,\nu\neq\mu}^{p} w_\nu^2, \tag{6.8}$$

we may write their joint gaussian distribution as

$$f(H_{\mu_i}, H_{\mathrm{oth}_i}) \simeq \frac{1}{2\pi\sigma_{\mu_i}\sigma_{\mathrm{oth}_i}}\exp(-\frac{H^2_{\mu_i}}{2\sigma^2_{\mu_i}} - \frac{H^2_{\mathrm{oth}_i}}{2\sigma^2_{\mathrm{oth}_i}}). \tag{6.9}$$

The distribution of H_{μ_i} given $H_{\mu_i} + H_{\mathrm{oth}_i} = H^{\min}_{\mathrm{tot}_i}$ appears as

$$f(H_{\mu_i}|H^{\min}_{\mathrm{tot}_i}) \simeq c\exp\Big(-\frac{\sigma^2_{\mathrm{tot}_i}}{2\sigma^2_{\mu_i}\sigma^2_{\mathrm{oth}_i}}(H_{\mu_i} - \frac{\sigma^2_{\mu_i}}{\sigma^2_{\mathrm{tot}_i}}H^{\min}_{\mathrm{tot}_i})^2\Big), \tag{6.10}$$

where c is a normalising constant and $\sigma^2_{\mathrm{tot}_i} = \sigma^2_{\mu_i} + \sigma^2_{\mathrm{oth}_i}$. The value of H_{μ_i} which maximises (6.10) is defined as $H^{\min}_{\mu_i}$, that is,

$$H^{\min}_{\mu_i} = \frac{\sigma^2_{\mu_i}}{\sigma^2_{\mathrm{tot}_i}}H^{\min}_{\mathrm{tot}_i}, \tag{6.11}$$

which reduces to

$$H^{\min}_{\mu_i} = w_\mu E^{\min}_{\mu_i} = \frac{w_\mu^2}{\sum_{\mu=1}^{p} w_\mu^2}H^{\min}_{\mathrm{tot}_i}. \tag{6.12}$$

Recall [Chapter 5] that the minimum of A samples from a gaussian of zero mean and standard deviation σ_{tot_i} has typical value

$$H^{\min}_{\mathrm{tot}_i} \simeq -\sqrt{2}\sigma_{\mathrm{tot}_i}\sqrt{\ln A}. \tag{6.13}$$

Substituting (6.13) into (6.12) and summing over i yields

$$E^{\min}_{\mu} \simeq -\sqrt{z}N\sigma\sqrt{\ln A}\frac{w_\mu}{\left(\sum_{\mu=1}^{p} w_\mu^2\right)^{\frac{1}{2}}}, \tag{6.14}$$

which is the desired result. Under the conditions of equal weights, this reduces to the expression (5.14) obtained in the previous chapter.

6.3 Simple Blob Model of Unfolding

It is a well known trend in polymer physics that the larger scale features of molecular conformation have systematically longer relaxation times. For example, for non-interacting chains with simple kink-jump dynamics, a subsection of g monomer units has relaxation time $\tau(g)$ proportional to g^2. On this basis we assume that after time t, a spontaneously unfolding polymer will have equilibrated locally up to scale g, such that $\tau(g) = t$, but still reflect the folded conformation on larger scales.

This simple blob view of proteins (Figure 6.1), that time scales relate uniformly to length scales, is of course a particular view. The alternative, which we do not address here, would be to consider spatially localised nucleation events.

The folded protein, which we assume to be compact and associate with $g = 1$, consists of N single monomer blobs. The contact map $C(1)$ has z non-zero entries in each row and column, zN non-zero entries in total.

For the state unfolded up to length scale g, the protein may be thought of as a chain of $\frac{N}{g}$ blobs, folded to its coarse grained original conformation. Accordingly, the contact map $C(g)$ has $\frac{N}{g}$ intra-blob blocks along the diagonal and $\frac{zN}{g}$ inter-blob blocks corresponding to nearest neighbour blobs. Scaling theories for polymer configurations with excluded volume would imply that the average total number of contacts between two neighbouring blobs be of order unity. Averaging

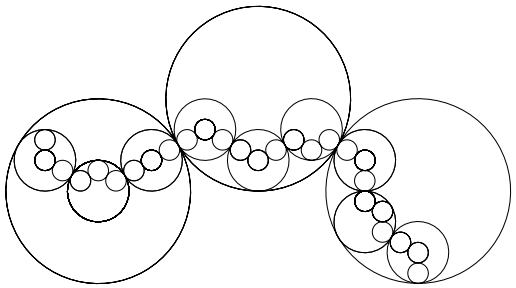

Figure 6.1: Two-dimensional representation of the blob model. The 27 monomer sequence is coarse grained to length scales $g = 1, g = 3$ and $g = 9$.

over an ensemble of conformations at constant g, this requires that each of the g^2 entries for each blob be of order $\frac{1}{g^2}$.

The total number of conformations (compact or otherwise) available to a protein grows as $\Omega(N) \sim \tilde{\kappa}^N$ (not to be confused with κ for compact structures only); this becomes $\tilde{\kappa}^{\frac{N}{g}}$ for a chain of $\frac{N}{g}$ blobs. Since the product of the internal and external conformational freedoms of a partially relaxed protein must equal $\tilde{\kappa}^N$, a sequence relaxed to length scale g can be estimated to take on $\tilde{\kappa}^{(N-\frac{N}{g})}$ configurations. It follows that the entropy gained in folding from a denatured configuration down to a conformation relaxed to length scale g may be expressed as

$$S(g) = -k_B\frac{N}{g}\ln\tilde{\kappa}. \tag{6.15}$$

6.4 Training to a Funnel

While an energy minimum significantly below the minimum copolymer energy ensures thermodynamic stability of the target conformation, rapid convergence necessitates a funnel of kinetic pathways sloping toward the target [Chapter 4]. The widest possible funnel is that which least constrains the dynamics, which we propose is given by the

conformations sampled in unfolding via the blob model. We thus consider combining the contact maps from different times (and values of g) of a noninteracting, spontaneously unfolding compact conformation with weights $w(g)$,

$$C_{\mathrm{tot}_{ij}} = \sum_{\ln g=1}^{\ln N} w(g) C_{ij}(g). \tag{6.16}$$

The minimum Hamiltonian associated with the total contact map then appears as

$$H^{\min}_{\mathrm{tot}} = \frac{1}{2}\sum_{ij=1}^{N} C_{\mathrm{tot}_{ij}} \tilde{U}^*_{ij} = \frac{1}{2}\sum_{ij=1}^{N}\sum_{\ln g=1}^{\ln N} w(g) C_{ij}(g) \tilde{U}^*_{ij}, \tag{6.17}$$

where again U^* corresponds to the sequence which minimises the total Hamiltonian. The total Hamiltonian associated with monomer i is the sum of the individual local Hamiltonians evaluated at different values of g,

$$H^{\min}_{\mathrm{tot}_i} = \sum_{\ln g=1}^{\ln N} H^{\min}_i(g), \tag{6.18}$$

where $H(g) = w(g)E(g)$. In accordance with our previous calculation, we estimate the variance in the choice of $H(g)$ available to a single monomer as

$$\sigma^2_{g_i} \simeq \frac{zg}{2}\Big(\frac{w(g)}{g^2}\Big)^2 \sigma^2, \tag{6.19}$$

where $\frac{zg}{2}$ is the number of contacts available to a given monomer equilibrated to scale g and $\frac{w(g)}{g^2}$ is the overall weighting for each one. The variance of the local energy per monomer integrated over all g is thus

$$\sigma^2_{\mathrm{tot}_i} \simeq \sum_{\ln g=1}^{\ln N} \sigma^2_{g_i} \simeq \frac{z\sigma^2}{2}\int_e^N \frac{dg}{g} g \frac{w^2(g)}{g^4}. \tag{6.20}$$

Again we require the relation between the Hamiltonian of the protein unfolded to length scale g and the total Hamiltonian. Proceeding as before, we find

$$H^{\min}_i(g) \simeq w(g) E^{\min}_i(g) \simeq \frac{\sigma^2_{g_i}}{\sigma^2_{\mathrm{tot}_i}} H^{\min}_{\mathrm{tot}_i}. \tag{6.21}$$

Substituting (6.13) and (6.19) into the above and summing over i, the minimum energy associated with matching the conformation at scale g can then be estimated as

$$E^{\min}(g) \simeq -\frac{z}{\sqrt{2}} N\sigma^2 \sqrt{\ln A}\, \frac{w(g)}{\sigma_{\mathrm{tot}_i} g^3}. \tag{6.22}$$

In order that the training reverse the unfolding dynamics, the required funnel must have sufficient slope, that is, $F(g) = E(g) - TS(g) < 0$. Equating the two expressions $T\times$ (6.15) and (6.22) gives

$$w(g) \simeq -\frac{\sqrt{2} k_B T \ln\tilde{\kappa}\ \sigma_{\mathrm{tot}_i}}{z\sigma^2\sqrt{\ln A}} g^2, \tag{6.23}$$

and thus $w(g) \propto g^2$. Unfortunately this form for w is inconsistent with a convergent (N-independent) evaluation of σ_{tot_i} in (6.20). Our assumption that the training energy could reverse the unfolding dynamics does not hold for all values of g.

We consequently introduce the cutoff scale $g_{\max}$, up to which our funnel extends. Substituting (6.23) into (6.20) and reducing the domain of integration yields

$$\sigma^2_{\mathrm{tot}_i} \simeq \frac{(k_B T)^2 \ln^2\tilde{\kappa}}{z\sigma^2 \ln A} \sigma^2_{\mathrm{tot}_i} \int_e^{g_{\max}} dg, \tag{6.24}$$

from which it follows that

$$g_{\max} \simeq \frac{z\sigma^2 \ln A}{(k_B T)^2 \ln^2\tilde{\kappa}}. \tag{6.25}$$

The width of our funnel, as parametrised by $g_{\max}$ above, increases strongly as folding temperature T decreases. At too low a temperature, however, the coil will collapse as a random copolymer into what

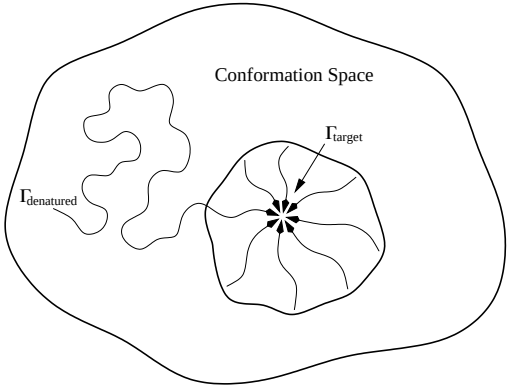

Figure 6.2: Folding in the presence of a funnel. The denatured protein wanders through conformation space until it matches the target structure coarse-grained to length scale $g_{\max}$, after which the funnel quickly guides the protein toward the target.

we presume to be a glassy state. The loss in entropy resulting from collapse will be equivalent to $-(6.15)$ evaluated at $g = 1$ (the collapsed copolymer will be fully folded). The modest decrease in energy afforded by the minimum copolymer energy can overcome this entropic loss only at low temperature $T_{\rm cp}$. Equating the minimum copolymer energy $E_{\rm cp}^{\rm min}$ in (5.16) from the previous chapter and $T_{\rm cp}$ times the loss in entropy $-(6.15)|_{g=1}$ leads to

$$k_B T_{\rm cp} \simeq \sigma \frac{\sqrt{z \ln \kappa}}{\ln \tilde{\kappa}}, \tag{6.26}$$

and hence at $T \simeq T_{\rm cp}$,

$$g_{\max} \simeq \frac{\ln A}{\ln \kappa} \simeq p_{\max}. \tag{6.27}$$

The cutoff $g_{\max}$ is the length scale of the structure below which the energy landscape corresponding to the trained sequence is characterised by a funnel. Above $g_{\max}$, the protein must organise itself into the desired (coarse grained) conformation without the help of kinetic

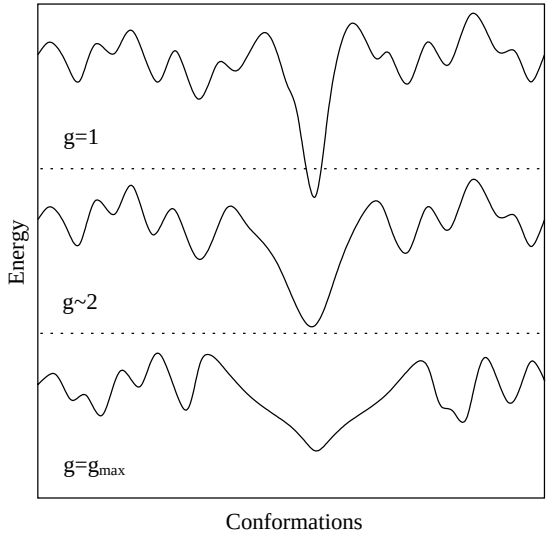

Figure 6.3: Energy landscapes of sequences trained to have increasingly broad funnels. Maximising stability (top) corresponds to a deep, narrow well. Training to an ensemble of unfolding targets (middle) provides a less deep, broader funnel. As the length scale g to which the funnel extends increases, the attainable depth of the target is reduced; at $g = g_{\max}$, the slope of the funnel is no longer sufficient to provide a free energy minimum (bottom).

guidance, that is, it must traverse an effective copolymer landscape (Figure 6.2). The effect of increasing the width of the funnel toward $g_{\max}$ is shown in Figure 6.3. As $g \to g_{\max}$, the slope of the funnel becomes sufficiently shallow such that, at $g = g_{\max}$, the decrease in energy no longer overcomes the loss of entropy.

Consider the protein as a sequence of $N/g_{\max}$ blobs, each of size $g_{\max}$. The benefit of the funnel is realised once the chain of blobs folds to its coarse grained target state. Assuming this to be the rate determining step, the time necessary for the protein to fold is reduced by the factor $\kappa^{-(1-1/g_{\max})N}$, which is significant even for small values of $g_{\max}$.

6.5 Conclusion

By considering the conformations sampled in unfolding, we have outlined a scheme of sequence selection which provides a basin of attraction around the target of order $g_{\max} = \frac{\ln A}{\ln \kappa}$. As the width of the funnel increases to include conformations matching the target conformation coarse grained to length scale $g_{\max}$, the free energy minimum above the target vanishes. Simulation to this end is ongoing and will be reported elsewhere.

Chapter 7

Kinetically Oriented Sequence Selection

> Auf Andere warte ich... auf Höhere, Stärkere, Sieghaftere... lachende Löwen müssen kommen
>
> FRIEDRICH NIETZSCHE
> *Thus Spoke Zarathustra*

We design proteins to rapidly fold to specified target structures by evolving sequences according to folding performance. In analogy with simulated annealing, we use the uncertainty of folding time estimates in a controlled way to efficiently traverse the sequential landscape. Kinetically oriented sequences are observed to fold to their targets significantly faster than sequences designed to maximise the thermodynamic stability of the target conformation.

7.1 Introduction

The ability to design protein sequences which fold to desired target structures has been achieved by optimising stability over sequence space in the target conformation [1, 19]. It was observed from simulation of simple lattice models of proteins that thermodynamically stable targets enjoy increased kinetic accessibility as well [20].

We have provided evidence in Chapter 4 that the correlation between stable and fast folding sequences is limited. In particular, it was demonstrated that maximising stability does not maximise accessibility. This can best be observed by considering a set of independently

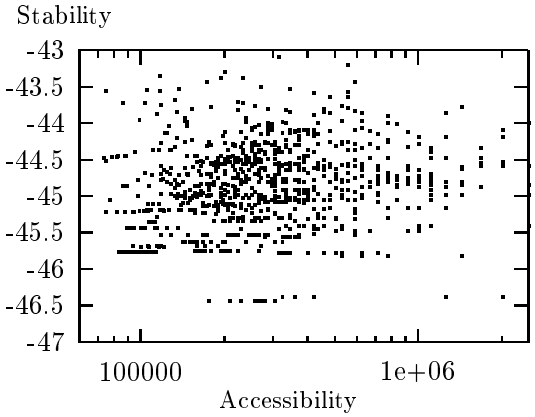

Figure 7.1: Ensemble of 27-mer sequences in a single target conformation independently selected from a low temperature ($T = 0.01$) bath, plotted in accessibility-stability phase space. Accessibility is estimated by the mean first-passage time to the target and stability is approximated by the relative energy of the target conformation. Note that the most stable sequences do not correspond to the most quickly folding.

annealed sequences which fold to a single target conformation plotted in accessibility-stability phase space (Figure 7.1). Sequences are plotted in accordance with the behaviour of their corresponding energy landscapes in the vicinity of the target conformation. Folding ability (and hence funnel width) is approximated by the mean first-passage time along the x-axis and relative well depth is estimated by the Z-score [Ch. 3] along the y-axis. While all of the sequences in Figure 7.1 are sufficiently stable (*i.e.*, a macroscopically large fraction of an ensemble of each occupies, at any one time, the target conformation), the most stable sequences do not correspond to the fastest folding.

It has been argued analytically [Chs. 5, 6] that the ability to manipulate the conformational energy landscape is limited by the finite amount of freedom in choosing the sequence. Allowing the energy funnel to become increasingly shallow (and the free energy well less deep)

enables us to make the target minimum increasingly broad. This gain in folding ability occurs at the expense of reduced stability, which eventually becomes prey to entropic traps [Ch. 6].

7.2 Kinetically Oriented Sequence Selection

The conflict between stability and accessibility, outlined above, suggests the selection of sequences on the basis of folding ability. The natural analogue of thermodynamically oriented sequence selection is kinetic sequence selection, in which folding time replaces relative energy as the favoured trait. The analogy is not perfect. Thermodynamic selection doesn't evaluate the actual stability of a protein in the target state; it optimises conformational stability by using sequential stability as a guide. Estimating accessibility by such means is not possible, since folding time (unlike energy) is a useful measure of conformational phenomena only.

Our method of kinetic sequence selection is simple. We begin with a sequence which folds (though not necessarily quickly) to a specified target conformation, which is the sequence occupied by the protein. A similar alternative sequence is derived by perturbing (*e.g.*, by point mutation) the occupied sequence. The alternative sequence is accepted as the occupied sequence on the basis of its folding performance.

Unlike its thermodynamic counterpart, the success of kinetically oriented sequence selection is impeded by its computational cost. This is manifested in two areas.

The essential difficulty is measuring folding time accurately. Since each new sequence is selected on the basis of its folding ability, uncertainties in estimates of folding time limit the degree to which the sequence may be optimised.

Also worrying is the risk that some amino acid mutations may make folding extremely inefficient. Since the cost of calculating the folding time grows in proportion to folding time itself, this may slow sequence optimisation to unrealistic time scales. This is of especial concern for short sequences, since even a single mutation may effect the target stability to an extent such that the target conformation

ceases to be the ground state conformation.

7.3 Measuring Folding Time

Once a protein assembles itself into its target state conformation, it remains there in proportion to the corresponding Boltzmann factor, $\exp(\frac{-E}{k_B T})$, which increases with the depth of the well. Furthermore, if the target well is of finite width near the bottom (as is generally the case), the protein will experience entropic fluctuations among conformations similar to the target. Accordingly, we measure folding efficiency by the first-passage time spent between leaving a (unique) denatured state and entering, however briefly, the target conformation. Since protein folding is path-independent [8], the first-passage time (FPT) t_A of an individual sequence S_A will deviate significantly from the mean first-passage time (MFPT) $\langle t_A \rangle$ of an ensemble of identical sequences S_A.

We can estimate the distribution of first-passage times $f_A(t_A)$ by constructing a histogram of FPTs of an ensemble of identical sequences. Figure 7.2 a shows the FPT distributions for two similar sequences trained to fold to the same target conformation. Notably, the long-time tail appears to decay geometrically, which suggests that occasionally proteins spend a long time folding. The geometric tail also implies that nearby distributions have considerable overlap, which makes the ordering of similar sequences according to their mean first-passage times difficult.

We consider the inference of the MFPT from multiple observations of the FPT. Let t_A^n be the mean of n independent foldings of sequence S_A to its target conformation;

$$t_A^n = \frac{1}{n} \sum_{i=1}^{n} t_A. \tag{7.1}$$

Accordingly, t_A^n is distributed by f_A^n about $\langle t_A \rangle$ with variance ${\sigma_A^n}^2 = \frac{1}{n}\sigma_A^2$. As $n \to \infty$, f_A^n approaches a δ-function centered about $\langle t_A \rangle$. The resulting decrease in distribution overlap allows more accurate sequence ordering (Figure 7.2 b).

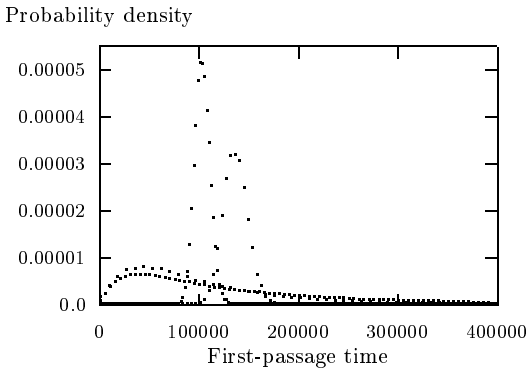

Figure 7.2: a) Distributions of first-passage times of two similar sequences (lower curves). Long tails cause significant overlap. b) Distributions of the mean of 100 first-passage times (upper curves). Note reduced overlap.

7.4 Simulated Annealing

Consider a problem for which there is a very large number of solutions, all of which may be ordered according to a cost function. How do we select the optimal or a near optimal solution in a realistic time scale, that is, in a time much less than that needed to examine all solutions? Simulated annealing is an efficient method of optimisation based on an elegant correspondence between complex optimisation problems and statistical mechanics.

The essence of simulated annealing is the controlled introduction of noise in updating the solution to avoid trapping in local, but not global, minima. The noise is reduced as the optimality of the solution improves such that, at the global minimum, determinism is recovered and the solution remains optimal.

Simulated annealing relies on a scheme of acceptance or rejection of a small change to the state of the system, known as the Metropolis algorithm, such that downhill moves occur with probability unity but uphill moves with probability proportional to the difference in Boltzmann weights. Consider the transition from state D_A to state D_B;

the transition probability $P_{A\to B} \equiv P_{D_A\to D_B}$ may be written

$$P_{A\to B} = \begin{cases} 1, & \Delta E < 0, \\ \exp(\frac{-\Delta E}{k_B T}), & \Delta E > 0, \end{cases} \tag{7.2}$$

where $\Delta E = E(D_B) - E(D_A)$ and k_B is the Boltzmann or another suitably chosen constant. Note that while these rules allow transitions both upward and downward in energy, they increasingly prefer the latter as the (*ad hoc*) effective temperature T is decreased.

Repeated application of (7.2) gives rise to a distribution of states which tends to be Boltzmann with respect to energy,

$$P(D) \propto \exp\left(\frac{-E(D)}{k_B T}\right). \tag{7.3}$$

This distribution has the property that, at sufficiently high temperature T, the probability of occupation of each state is uniform, while at $T = 0$, the ground state is occupied with probability unity. The transition from the readily attainable high temperature distribution to the elusive annealed distribution by the gradual decrease of T allows optimisation in a time short relative to complete enumeration. How slowly T must be reduced depends on details of the problem; in practice, only a deep local minimum is required and the annealing schedule may be chosen accordingly.

7.5 Probability of Moving Downhill

In analogy with simulated annealing, we wish to use the uncertainty in measuring protein first-passage time to our advantage in navigating the sequence FPT landscape. Unlike simulated annealing, in which the transition probabilities are chosen such that the state of the system is distributed according to a thermal (Boltzmann) distribution, the transition rules for competing sequences are imposed by the probabilistic nature of the estimate of mean first-passage time. We quantify the resulting transition probabilities below.[1]

[1] A more general treatment of optimisation of a system in which the cost function is a random variable is provided in Chapter 9. We summarise the case for proteins here.

Consider an occupied sequence S_A and an alternative sequence S_B. We assume S_A and S_B are sufficiently similar (differing, say, by a point mutation) such that we may approximate the distribution f_B by f_A, apart from a shift in mean;

$$f_B(t_B) \simeq f_A(t_B - \Delta\mu), \tag{7.4}$$

where $\Delta\mu = \mu_B - \mu_A$ and $\mu_A = \langle t_A \rangle$, *etc.*

Let f_A^n and f_B^n be the distributions of t_A^n and t_B^n defined in (7.1). Since t_A^n and t_B^n are each the sum of n random variables, it follows from the central limit theorem that, for $n \gg 1$, f_A^n and f_B^n are gaussian, which by (7.4) we assume to satisfy

$$f_B^n(t_B^n) \simeq f_A^n(t_B^n - \Delta\mu). \tag{7.5}$$

The transition from sequence S_A to sequence S_B occurs if our estimate of the MFPT of S_B is less than that of S_A, *i.e.*, if $t_B^n < t_A^n$. Then the probability $P_{A\to B} \equiv P(S_A \to S_B)$ equals $P(t_B^n < t_A^n)$, that is,

$$P_{A\to B} = \int_{-\infty}^{\infty} f_A^n(t_A^n) \int_{-\infty}^{t_A^n} f_B^n(t_B^n)\ \mathrm{d}t_B^n\ \mathrm{d}t_A^n. \tag{7.6}$$

Substituting (7.5) into f_B^n in (7.6) yields

$$\begin{aligned} P_{A\to B}(\Delta\mu) &= \int_{-\infty}^{\infty} f_A^n(t_A^n) \int_{-\infty}^{t_A^n - \Delta\mu} f_A^n(t_B^n)\ \mathrm{d}t_B^n\ \mathrm{d}t_A^n && (7.7) \\ &= -\int \int_{-\infty}^{\infty} f_A^n(t_A^n)\ f_A^n(t_A^n - \Delta\mu)\ \mathrm{d}t_A^n\ \mathrm{d}\Delta\mu && (7.8) \\ &= \frac{1}{2}\left(1 - \mathrm{erf}\frac{\Delta\mu}{2\sigma_A^n}\right), && (7.9) \end{aligned}$$

where the constant of integration in (7.9) is imposed by $P_{A\to B}(0) = \frac{1}{2}$. Since $\sigma_A^n = \frac{1}{\sqrt{n}}\sigma_A$,

$$P_{A\to B}(\Delta\mu, n) = \frac{1}{2}\left(1 - \mathrm{erf}\frac{\sqrt{n}\,\Delta\mu}{2\sigma_A}\right), \tag{7.10}$$

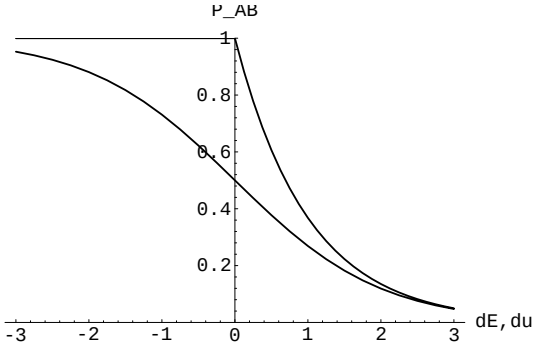

Figure 7.3: Comparison of simulated annealing (top curve) and sequence competition (bottom curve) transition probabilities, as a function of ΔE and $\Delta\mu$. Sequence competition denies some downhill moves but allows fewer uphill transitions.

which is the sequence competition transition probability (Figure 7.3).

Immediately we recognise $\frac{\Delta\mu}{2\sigma_A^n} = \frac{\sqrt{n}\Delta\mu}{2\sigma_A}$ to be the analogue of $\frac{\Delta E}{k_B T}$ from simulated annealing. We are led to analogously anneal our protein sequence by suitably increasing the number of samples of the FPT in our estimation of the MFPT. The correspondence between T and $\frac{1}{\sqrt{n}}$ suggests as a sensible annealing schedule $n(t_{\mathrm{MC}}) \propto t_{\mathrm{MC}}^2$, where t_{MC} is time measured in Monte Carlo time steps; this provides an effective temperature T decreasing as $\frac{1}{t_{\mathrm{MC}}}$.

7.6 Thermodynamic Guidance

Adverse sequence mutations may produce alternative sequences with very long mean first-passage times or, worse, sequences which no longer fold to their target conformations (*i.e.*, the target state ceases to be the ground state). This is a worrying prospect because, unlike most cost functions, calculating the folding time requires computation in proportion to the folding time itself. We can safeguard against this danger by using thermodynamic stability as a guide in admitting sequences whose FPTs are not excessive and, therefore, readily measurable.

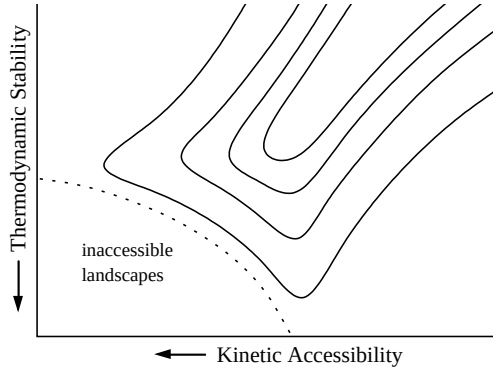

Figure 7.4: Contour plot schematic of accessibility-stability phase space. Stability and folding ability are correlated toward the top right but anticorrelated near the bottom left. The dotted line represents the limited ability to manipulate the conformational energy landscape. Along this curve, an increase in stability corresponds to a decrease in accessibility.

We first condition each mutation on the basis of the alternative sequence's Z-score by the application of a Metropolis rule at constant temperature. Accepted sequences are then subject to sequence competition.

The use of stability as an indication of folding ability relies on the correlation between the two. The forward correlation — that stable sequences are fast folding — was exploited in [19] in the folding of lattice proteins designed to be thermodynamically stable. Evidence for the reverse correlation — accessible sequences are stable — is provided in [13]. It should be remarked that the anticorrelation between folding stability and ability, discussed earlier in Chapter 4, applies only near the stability and accessibility extremes of the accessibility-stability phase space. This was borne out analytically in Chapters 5

and 6, in which maximising protein capacity and funnel width led to free energy traps away from the target state.

A contour plot of stability-accessibility phase space suggested by our research is drawn schematically in Figure 7.4. Only the stable, accessible tail of the distribution is shown. Stability and folding ability are correlated away from the tail (top right) but become anticorrelated near their extremes (bottom left). The dotted curve designates the limited freedom in manipulating the conformational energy landscape, along which increased stability results in decreased accessibility.

Ideally, the stability selection pressure should favour mutations such that sequences are driven along the correlated region of Figure 7.4, but not into the anticorrelated thermodynamic extreme. As long as the selection pressure is small, it should not trap sequences in this region.

7.7 Results

We have optimised 27 monomer sequences to fold to compact $3 \times 3 \times 3$ target conformations according to the method of sequence evolution described above. Initial sequences were determined by thermodynamic sequence selection outlined in Chapter 3, which relies on the minimisation over sequence of the target state energy. By annealing (the relative energy) down to various finite temperatures, we were able to exercise rough control over the folding ability of the initial sequences.

Evolution of fast and slowly folding sequences is shown in Figures 7.5 and 7.6 (both to the same structure). Each point represents a unique sequence, plotted according to mean first passage time (measured in Monte Carlo steps) and the number of accepted mutations. It is important to distinguish between the number of accepted mutations and the number of attempted mutations, which is significantly higher. As the folding performance improves, the likelihood of a random mutation causing further improvement diminishes. Moreover, the probability of accepting a less efficient sequence decreases with temperature.

The free parameters associated with our method of kinetic optimi-

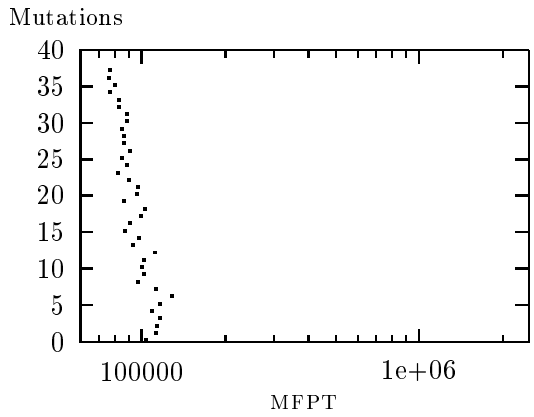

Figure 7.5: Evolution of an originally fast folding sequence. Each point represents a sequence plotted according to its mean first-passage time (x-axis) and number of accepted mutations (y-axis), where MFPT is measured in number of Monte Carlo steps. Initial temperature $T_i = \frac{1}{\sqrt{n}} = 0.10$, final temperature $T_f = 0.029$. Folding time has been reduced by approximately $\frac{1}{2}$.

sation are the annealing schedule, *i.e.*, the functional form of $n(t_{\mathrm{MC}})$, and the temperature T_{guide} at which mutations are conditioned on the resultant change in stability. Since the computational expense of assessing a sequence becomes prohibitive at lower temperatures, we wish to make as few low temperature mutations as possible, on the one hand, while giving disproportionate attention to low temperature exploration, on the other. We reduce the effective temperature $\frac{1}{\sqrt{n}}$ according to $n(t_{\mathrm{MC}}) \propto t_{\mathrm{MC}}^2$, where t_{MC} is the number of Monte Carlo time steps. The constant temperature T_{guide} was determined empirically to be 0.20.

Proteins were annealed from initial sequences of various optimality down to effective temperatures $T \simeq 0.025$, over up to 6000 (kinetically) attempted mutations. Despite the limited duration of these runs, it appears that some behaviour is typical of the evolution. Minimum

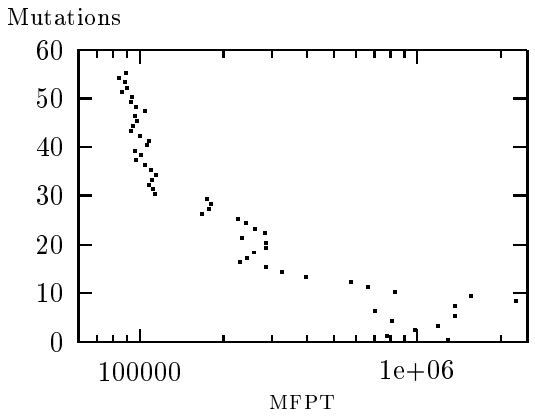

Figure 7.6: Evolution of an originally slowly folding sequence, with $T_i = 1.00$, $T_f = 0.023$. The original sequence folded poorly with MFPT $\simeq 2 * 10^6$ while the final sequence folded well with MFPT $= 8 * 10^4$, 1.4 orders of magnitude faster.

folding times were consistently found to be between 70,000 and 80,000 Monte Carlo time steps. Particularly noteworthy is the small number of accepted mutations, ~ 50, necessary to achieve these folding rates.

Gutin and coworkers [13] have also presented a method of designing fast folding 27-mers, in which mutated sequences are accepted according to a two step algorithm. The first test requires the mean of two FPTs of the alternative sequence to be less than the occupied sequence MFPT. If this is passed, the second test requires the mean of a further 10 FPTs of the alternative sequence to be 20% less than the occupied MFPT before the mutation may be accepted. Motivation for these particular acceptance criteria is not provided.

Unlike the selection method presented here, the Gutin algorithm is unable to quantify (and allows no control over) the degree of folding optimality. Sequences evolved according to the Gutin technique fold less quickly than those presented in Figures 7.5 and 7.6 above.

7.8 Sequential MFPT Landscape is Smooth

The evolution of what appear to be near optimal folding sequences in such short time scales is surprising. We propose this is due in part to the comparatively smooth surface of the sequential MFPT landscape.

The ability to traverse the conformational energy landscape is heavily constrained by the self-avoiding and non-crossable nature of the protein. To pass to a conformationally near but topologically distant conformation, the protein must swell and recollapse, overcoming a large energy barrier. These kinetic constraints, especially important in the context of compact proteins, makes conformational optimisation (folding) slow.

The sequential MFPT landscape, while significantly larger, does not contain kinetic fences crossing thermodynamically favourable paths. Unhindered access of nearby sequences allows rapid exploration of the landscape. For an entirely funnel-oriented MFPT landscape, in which each element of the sequence may be independently optimised, we may estimate the number of attempted and accepted monomer mutations necessary to achieve global optimality.

The number of attempted mutations needed to optimise a single monomer is $\frac{A}{2}$, while the number necessary to visit each of the N monomers once is $\frac{N \ln N}{2}$. Then the number of attempted mutations required to optimise the entire sequence may be bounded from above by

$$Q_{\text{att}} \simeq \frac{A\, N \ln N}{4}. \tag{7.11}$$

The number of accepted mutations is bounded by

$$Q_{\text{acc}} \simeq \frac{\ln \frac{A}{2} N \ln N}{2}. \tag{7.12}$$

For a 27 monomer protein composed of 20 amino acid species, $Q_{\text{att}} = 445$ and $Q_{\text{acc}} = 102$.

The perfect funnel outlined above is, of course, unrealistic. Nor do we believe our annealed sequences to be sequential global minima. Nevertheless, these estimates suggest that optimisation of the mean first-passage time can occur on time scales not dissimilar to those

observed in this Chapter.

7.9 Conclusion

We introduce a new method of sequence design based on the selection of sequences according to folding ability. Using the uncertainty of the first-passage time, which we find to be near-poisson distributed, in a controlled way allows us to avoid trapping in locally, but not globally, optimal sequences. Annealing down to equal effective temperatures $\propto \frac{1}{\sqrt{n}}$ allows comparison of mean first-passage times of sequences evolved to fold to different targets. Our kinetically oriented method of sequence selection yields sequences which fold in 70,000 – 80,000 Monte Carlo time steps, significantly faster than those trained according to thermodynamic design techniques.

Chapter 8

Hierarchical Optimisation Problems

> Suppose we solve all the problems it presents? We end up with more problems than we started with. That's the way problems propagate their species.
>
> N. SIMPSON

Hierarchical optimisation generalises conventional optimisation to include problems in which the solution must be determined stage-wise in the light of information learnt at each stage. We begin this chapter by considering conventional and probabilistic traveling salesman problems, which naturally leads us to hierarchical optimisation problems. We differentiate between decision difficulty and problem hierarchy, both of which give rise to global complexity. By introducing the concept of value, we derive the general optimality equation, applied to model problems in Chapters 9 and 10.

8.1 Traveling Salesman Problem

The prototypical complex optimisation problem is the traveling salesman problem (tsp). The task is to determine the minimal length tour through a set of cities such that each city is visited once and the path returns to its starting point (Figure 8.1 a,b). The tsp has been the subject of an enormous amount of literature and is the standard testing ground for methods of combinatoric[1] optimisation.

We formulate the traveling salesman problem as follows. Consider a set A of N cities on a Euclidian plane, labelled $a_1, a_2, \ldots, a_N$. Pairwise separations are given by the $N \times N$ distance matrix D, where

[1] By combinatoric we mean that the number of enumerable solutions scales exponentially or more rapidly with the size of the problem.

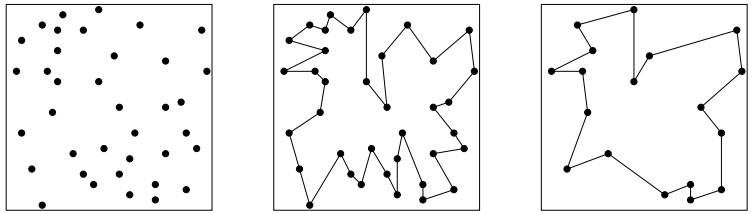

Figure 8.1: Conventional and probabilistic traveling salesman problems. The TSP seeks the minimal length closed tour through a set of cities (a,b). The PTSP seeks a tour which, when pruned to pass through a stochastically realised subset of cities, is of minimal length in the expected sense (a,b,c).

$D_{ij} \equiv D_{a_i a_j}$ is the distance between city i and city j. Cities are visited according to a tour t where $t \equiv t_1, t_2, \ldots, t_N$ is a permutation of the cities $a_1, a_2, \ldots, a_N$. We represent the tour by the $N \times N$ connection matrix T such that $T_{ij} = 1$ if cities a_i and a_j are neighbours along the tour and $T_{ij} = 0$ otherwise. The total distance traveled by tour t may be expressed

$$d(t) = \frac{1}{2} \sum_{ij=1}^{N} T_{ij}\, D_{ij}. \tag{8.1}$$

The objective of the TSP is the ground state tour t_0 which minimises d.

8.2 Probabilistic Traveling Salesman Problem

The introduction of probabilistic elements to combinatoric optimisations problems was made by Jaillet [27] in 1985 and soon after extended by Bertsimas [22]. Both considered, in particular, an extension of the TSP known as the probabilistic traveling salesman problem (PTSP), which has since become a paradigm problem of probabilistic combinatoric optimisation.

Unlike the TSP, the PTSP applies to a set of cities for which, on

any given instance, only a stochastically chosen subset need be visited. The objective of the PTSP is the construction of an *a priori* tour through all the cities such that, upon realising the active subset, the collapsed tour is of minimal length in the expected sense (by collapsed we mean that the reduced tour visit the subset of cities in the same order as the original).

We label the set of N cities $A \equiv a_1, a_2, \ldots, a_N$, with pair-wise separations D_{ij}. Each city is to be visited with probability $p_i = p$ such that, on any given instance of the problem, a subset $B \subseteq A$ cities must be visited. Consider some *a priori* tour t through A. The pruned tour u (represented by U) is a permutation of the cities in B such that they are visited in the same order as by t. The original tour t is chosen such that the pruned tour u, averaged over all possible realisations of the city probabilities, is of minimal expected length. The expected distance of the pruned tour may be written

$$\langle d(u|t) \rangle = \frac{1}{2} \sum_{B \subseteq A} P(B) \sum_{ij=1}^{N} U_{ij}(u|t, B)\, D_{ij}, \tag{8.2}$$

where the first summation runs over all subsets B of A and

$$P(B) = \prod_{i \subseteq B} p_i \prod_{i \subseteq A-B} (1 - p_i). \tag{8.3}$$

The objective of the PTSP is the tour t which minimises the expected value of $d(u)$.

The probabilistic TSP may be considered a conventional TSP in which the total distance of the *a priori* tour is a random variable parametrised by the city probabilities p_i. The task here is to determine the tour whose *mean* length is minimal. Optimisation of a complex problem in which the cost function is a random variable is considered in the following chapter. In the remainder of this chapter we generalise our example to multiple stage hierarchical optimisation problems.

8.3 Hierarchical Optimisation Problems

The probabilistic traveling salesman problem is one of a number of probabilistic combinatoric optimisation problems; others include the minimum spanning tree and vehicle routing problems [22, 23]. Characteristic of them all is optimisation in light of information learnt (the problem instance) followed by the realisation of probabilistic information and further optimisation. (The second, albeit elementary, stage of optimisation in the PTSP is pruning the tour to visit only the active subset.)

We consider the extension of probabilistic combinatoric optimisation problems to problems consisting of an arbitrary number of optimisation levels, each separated by the realisation of information. We describe them as follows:

1. they contain two or more levels of optimisation,
2. at each successive level, probabilistic information is realised before making a decision,
3. decision at any level can be made only after the preceding levels have been completed.

In light of the stage-wise manner in which they must be solved, we call such problems hierarchical optimisation problems[2] (HOPs).

Hierarchical optimisation is an appropriate framework for problems in which the solution must be determined (or imposed) over long time scales, during which the problem conditions fluctuate or experience feedback from past optimisation stages. Examples include investment strategy as a function of fluctuating market prices, the design of a multi-stage engineering project and models of economic growth. The decision must be made at each stage such that the emphasis on immediate reward is balanced by the maximisation of implicitly dependent expected future rewards. In a problem plagued by fluctuating conditions, this requires securing a wide range of future options in response to unknown future circumstances.

[2] The same phrase is used in [28] to describe probabilistic combinatoric optimisation.

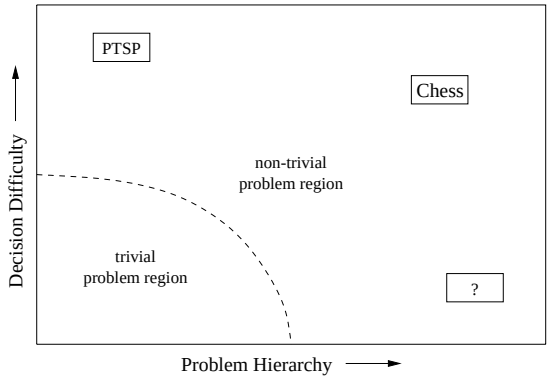

Figure 8.2: Hierarchical-combinatoric phase-space for hierarchical optimisation problems. Complexity increases away from the origin. The probabilistic traveling salesman problem has two decision levels, one of which is very difficult. We propose that complexity may also result from the concatenation of many elementary decisions, which is the subject of Chapter 10.

8.4 Combinatoric and Hierarchical Complexity

The complexity of hierarchical optimisation problems (often reflected in the structure of their solutions) may be generated by two independent properties: 1) the inclusion of decision levels for which there are a combinatorially large number of options and 2) the concatenation of many (inter-dependent) decision levels.

Probabilistic optimisation problems considered above are direct extensions of conventional combinatoric optimisation problems resulting from the inclusion of probabilistic elements in the problem instance. Choosing the optimal solution from a combinatoric number of possibilities is difficult in its own right; the addition of stochastic elements requires the solution of many instances of the problem. Such problems belong to a class of problems typically characterised by a small number of decision levels, one or more of which admits a combinatorially large number of solutions.

Alternatively, HOPs consisting of many (often straightforward) de-

cision levels may give rise to complex behaviour through the coupling of each decision to subsequent decisions implicitly dependent on it. Complexity in this case is an emergent property of the HOP itself.

We describe the first class of HOPs as combinatorially complex and the second as hierarchically complex. Figure 8.2 shows hierarchy-combination phase space, in which problem complexity grows with distance from the origin. Chess exhibits both combinatoric and hierarchical complexity and is considerably difficult to solve (*i.e.*, for which to construct an optimal decision policy). We propose that there exist many-level HOPs which, while consisting of individually elementary decisions, exhibit complex global behaviour; such an example is the subject of Chapter 10.

8.5 Definition of Hierarchical Optimisation

Hierarchical optimisation may be framed as a sequence of state-action stages. Each stage begins with the realisation of a random variable (state) and ends with the decision (action) chosen in light of this state. Figure 8.3 depicts the branching structure of possible state-action futures; in a typical HOP, different regions of the tree may be more or less ramified. States and actions of one stage can be learnt and taken only after completing the stage preceding it.

For reasons which will soon be evident, the stages are ordered in reverse chronological order; the first stage, corresponding to the top of the HOP tree, is labelled N and the last stage is labelled 1. Optimisation begins with realisation of the state s_N, after which action a_N is chosen from the set (or space) of possible actions. In the second stage of optimisation, state s_{N-1} is learnt and action a_{N-1} taken, and so on.

It is assumed that the options available to action a_n depend explicitly on the state s_n just learnt alone. The reward[3] associated with action a_n in light of state s_n may depend on s_n in addition to a_n and is written $R(a_n|s_n)$.

[3]We use the terms *reward* and *cost* interchangeably, with the understanding that one is the negative of the other.

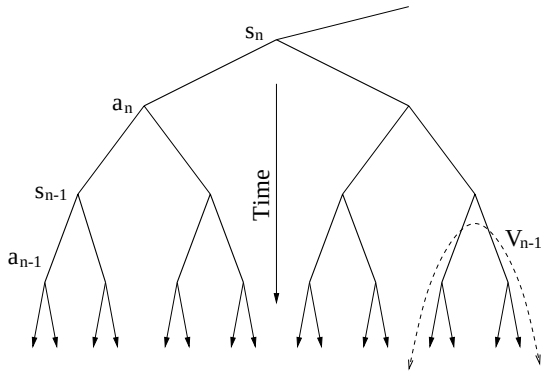

Figure 8.3: Hierarchical optimisation tree schematic. In each of the N stages, state s is realised before taking action a. Each path from top to bottom corresponds to a unique realisation of outcomes and decisions.

The transition probability from state s_n to state s_{n-1} may depend explicitly on both s_n and a_n; we write it as $P(s_{n-1}|s_n, a_n)$. Note that if the transition probabilities are all zero or unity (*i.e.*, the problem is no longer stochastic), then the hierarchical nature of the optimisation collapses. The value of all states is known *a priori* and, consequently, all decisions can be made up front.

For a conventional (single stage) optimisation problem, the optimal solution satisfies

$$a_1^{\text{opt}} = a_1 | \max_{a_1} \Big[R(a_1|s_1) \Big], \tag{8.4}$$

i.e., the optimal action a_1^{opt} is that action which, given the problem instance s_1, maximises the reward function. For the traveling salesman problem, a_1 is a tour connecting the set s_1 of city coordinates and R is the negative distance traveled.

Equation 8.4 may be generalised to describe hierarchical optimisation problems as well. For a two stage HOP,

$$a_2^{\text{opt}} = a_2|\max_{a_2}\Big[R(a_2|s_2) + \sum_{s_1} P(s_1|s_2,a_2)\,R(a_1^{\text{opt}}|s_1)\Big], \quad (8.5)$$

$$a_1^{\text{opt}} = a_1|\max_{a_1}\Big[R(a_1|s_1)\Big]. \quad (8.6)$$

In the context of the probabilistic traveling salesman problem, s_2 is the set of city coordinates through which we construct an *a priori* tour a_2. We then learn the stochastically chosen subset of cities s_1 which must be visited and choose a pruned tour a_1 accordingly. In this example a reward (equal to $-d$) is realised for the pruned tour but not the *a priori* tour.

8.6 General Optimality Equation

The extrapolation of (8.5), (8.6) to HOPs of higher order is, while straightforward, not very useful. We instead re-express the conditions of optimality in terms of the value V_n, the maximum expected reward incurred by the actions descending from state s_n. It represents the immediate reward associated with the optimal (in the global sense) action a_n in addition to the expected rewards associated with making optimal future actions. The value associated with state s_n may be expressed

$$V_n(s_n) = \max_{a_n}\Big[R(a_n|s_n) + \sum_{s_{n-1}} P(s_{n-1}|s_n,a_n)\ \max_{a_{n-1}}\Big[R(a_{n-1}|s_{n-1}) + \sum_{s_{n-2}} P(s_{n-2}|s_{n-1},a_{n-1})(\ldots)\Big]\Big]. \quad (8.7)$$

Immediately we recognise the expression maximised over a_{n-1} in (8.7) to be the value $V_{n-1}(s_{n-1})$. Accordingly, we may recursively express (8.7) as

$$V_n(s_n) = \max_{a_n}\Big[R(a_n|s_n) + \sum_{s_{n-1}} P(s_{n-1}|s_n,a_n)\,V_{n-1}(s_{n-1})\Big]. \quad (8.8)$$

At stage n, we choose as the optimal action that action which maximises the right side of (8.8). This depends implicitly on the future optimal actions a_{n-1}, a_{n-2}, *etc.* Accordingly, we must solve (8.8) from the bottom of the tree up, beginning with

$$V_1(s_1) = \max_{a_1}\Big[R(a_1|s_1)\Big], \quad (8.9)$$

since decision ends at stage 1. This is equivalent to saying that the boundary condition for (8.8) is $V_0 = 0$.

We refer to (8.8) as the general optimality equation. In the next chapter we consider two-stage hierarchical optimisation problems in which the first stage decision is sufficiently hard that we can only approximate the optimal solution given by (8.8). In Chapter 10 we apply the general optimality equation to an HOP consisting of many optimisation stages; while each action, taken independently, is elementary, the global solution exhibits non-trivial behaviour.

Chapter 9

Stochastic Annealing

> Of course Einstein's wrong. I just played craps with God yesterday.
>
> JOSEPH LONG

In this chapter we consider optimisation of a system for which the cost function is a random variable whose distribution depends on the state of the system. Competing states are ordered according to the mean of the cost, which we estimate by random sampling. By analogy with simulated annealing, we show that dynamically controlling the uncertainty associated with the cost estimate can improve the degree of optimisation obtained, whilst providing substantial gain in computational efficiency. We apply our technique to the probabilistic traveling salesman problem, a central problem in probabilistic optimisation.

9.1 Introduction

Hierarchical optimisation, introduced in Chapter 8, concerns the stagewise optimisation of a cost which is a function of a sequence of decisions separated by the realisation of probabilistic information. We consider in this chapter two-stage hierarchical optimisation problems (HOPS) for which there is a very large number of solutions (combinatoric complexity) and reward is provided in the final stage only.

The class of two-stage HOPS described above is an extension of conventional optimisation which includes problems in which the cost function is a random variable. Accordingly, the following results apply to any problem for which the cost function H is itself to be computed by random sampling over other variables. An example in finance might

be the value of a portfolio from sampled paths of future prices; in protein design, the time for the molecule to fold under sampled thermal fluctuations; in oil field development, the return from placement of wells under sampled reservoir geometry. In all cases it has to be assumed — or the sampling so designed — that unlikely to be sampled outcomes of the cost do not dominate considerations.

9.2 Simulated Annealing

The essence of simulated annealing, qualitatively outlined in Chapter 7, is that, to a controlled extent, unfavourable design (state) transitions should be accepted in order to reduce trapping in locally, but not globally, optimal designs. This is achieved by the controlled introduction of noise to otherwise deterministic downhill dynamics such that determinism is recovered as the system approaches optimality.

Consider an occupied design D_A and a similar alternative design D_B. For a (deterministic) cost function $H(D)$ to be minimised, design evolution is carried out such that downhill moves are accepted with probability unity but uphill moves with probability decreasing with the change in cost and with time via the control parameter T_e. This is the Metropolis algorithm [29], written

$$P_{A\to B} = \begin{cases} 1, & \Delta H < 0, \\ \exp(-\beta_e \Delta H), & \Delta H > 0, \end{cases} \tag{9.1}$$

where $\Delta H = H_B - H_A \equiv H(D_B) - H(D_A)$ and β_e is the reciprocal effective temperature $\frac{1}{kT_e}$, with k some constant.

Consider a very large ensemble of systems, in which ν_A is the number of systems in design D_A and ν_B is the number of systems in design D_B. We wish to determine the equilibrium distribution of designs imposed by the repeated application of (9.1).

Let us make transitions (moves) in all the systems of our ensemble, where $Q_{A\to B}$ is the probability that the move from D_A to D_B is considered and $P_{A\to B}$ the probability that, once considered, the move is accepted. Assume, without loss of generality, that $H_A > H_B$. Then the number of systems moving from design D_A to design D_B is

$$\nu_A \, Q_{A\to B} \, P_{A\to B} = \nu_A \, Q_{A\to B}, \tag{9.2}$$

and the number moving from D_B to D_A is

$$\nu_B \, Q_{B\to A} \, P_{B\to A} = \nu_B \, Q_{A\to B} \, e^{-\beta_e (H_A - H_B)}. \tag{9.3}$$

By assumption, $Q_{A\to B}$ is the same for all A and B, and we henceforth simply write Q. The net number of systems moving from D_A to D_B is

$$Q(\nu_A - \nu_B \, e^{-\beta_e (H_A - H_B)}). \tag{9.4}$$

It follows that net transfer from D_A to D_B will continue until (9.4) $= 0$, that is,

$$\frac{\nu_A}{\nu_B} = \frac{e^{-\beta_e H_A}}{e^{-\beta_e H_B}}, \tag{9.5}$$

or simply

$$\nu_C \propto e^{-\beta_e H_C} \quad \forall \quad D_C, \tag{9.6}$$

which is the condition for equilibrium. Equation (9.6) implies that, at any given time, the probability of a system being in design D_C is proportional to its Boltzmann factor.

The Metropolis transition rule in (9.1) may be replaced by the sigmoid function governing Glauber dynamics [26],

$$P_{A\to B}(\Delta H) = \frac{1}{1 + \exp(\beta_e \Delta H)}, \tag{9.7}$$

commonly used in finite temperature neural network models. Unlike the Metropolis rule, the Glauber rule (Figure 9.1) is a symmetrically shaped function and hence a more physical representation of stochasticity. This increase in realism, however, is at the expense of efficiency: the Glauber probability sometimes denies downhill moves and allows fewer uphill transitions. It may be readily verified that the Glauber rule also gives rise to the Boltzmann distribution (9.6).

The important feature of (9.1) and (9.7) is the ratio of acceptance

probability for a move and its reverse:

$$\frac{P_{A\to B}}{P_{B\to A}} = \frac{P_{A\to B}(\Delta H)}{P_{A\to B}(-\Delta H)} = e^{-\beta_e \Delta H}, \tag{9.8}$$

which ensures path independence of the transition probabilities between any two designs. If moves and their reverse are equally likely to be attempted, a detailed balance equilibrium is approached in which designs are sampled with probability proportional to $e^{-\beta_e H(D)}$. The system is thereby driven toward a global minimum of H if, with successive moves, the reciprocal temperature β_e is gradually increased sufficiently slowly (just how slowly being the difficult question of an annealing schedule).

9.3 Stochastic Annealing

Here we derive the transition rule analogous to (9.1) and (9.7) imposed by sampling a distributed cost function. Unlike in §7.5, here we consider the estimate of the difference in design costs rather than the difference of the estimates themselves.

Consider again an occupied design D_A and a similar alternative design D_B, whose cost functions H_A and H_B are now random variables, taking on values h_A and h_B. The costs are distributed according to f_A and f_B, with means $\mu_A = \langle H_A \rangle$ and $\mu_B = \langle H_B \rangle$ and standard deviations σ_A^2 and σ_B^2.

It may or may not be appropriate to use the same sampling for H_A and H_B. Accordingly, we do not assume the distributions to be identical (up to mean) but instead consider the random variable

$$\Delta H = H_B - H_A, \tag{9.9}$$

whose value $\Delta h = h_B - h_A$ is distributed according to $f_{\Delta H}$. We take the mean of ΔH to be $\Delta\mu = \mu_A - \mu_B$ and the variance to be $\sigma_{\Delta H}^2$; if σ_A and σ_B are approximately equal, then $\sigma_{\Delta H}^2 \simeq 2\sigma_A^2$.

We estimate the mean of ΔH by the statistic

$$\Delta H^n = \frac{1}{n} \sum_{i=1}^{n} \Delta h_i; \tag{9.10}$$

for large n, ΔH^n is gaussianly distributed as $f^n_{\Delta H}(\Delta h^n)$, in accordance with the central limit theorem, with mean $\Delta\mu$ and variance ${\sigma^n_{\Delta H}}^2 = \frac{1}{n}{\sigma_{\Delta H}}^2$.

Design evolution occurs such that the transition from the occupied design D_A to the alternative design D_B is accepted if $\Delta h^n < 0$.[1] Then the probability $P_{A\to B}$ may be expressed

$$P_{A\to B}(\Delta\mu) = \int_{-\infty}^{0} f^n_{\Delta H}(\Delta h^n)\,\mathrm{d}\Delta h. \tag{9.11}$$

Let the distribution $g^n_{\Delta H}$ be identical to $f^n_{\Delta H}$ but with zero mean, *i.e.*, $g^n_{\Delta H}(\Delta h^n) = f^n_{\Delta H}(\Delta h^n + \Delta\mu)$. Then (9.11) may be rewritten

$$P_{A\to B}(\Delta\mu) = \int_{-\infty}^{-\Delta\mu} g^n_{\Delta H}(\Delta h^n)\,\mathrm{d}\Delta h. \tag{9.12}$$

$$= \frac{1}{2}\left(1 - \mathrm{erf}\frac{\Delta\mu}{\sqrt{2}\,\sigma^n_{\Delta H}}\right). \tag{9.13}$$

Substituting $\sigma^n_{\Delta H} = \frac{1}{\sqrt{n}}\sigma_{\Delta H}$ into (9.13) yields

$$P_{A\to B}(\Delta\mu, n) = \frac{1}{2}\left(1 - \mathrm{erf}\frac{\sqrt{n}\Delta\mu}{\sqrt{2}\,\sigma_{\Delta H}}\right), \tag{9.14}$$

which is the stochastic annealing transition probability.

Evidently $\frac{\sqrt{n/2}}{\sigma_{\Delta H}}$ is the analogue (up to a constant) of β_e in (9.1). This suggests as an annealing schedule increasing the sample population n with time according to $n(t_\mathrm{MC}) \propto t^2_\mathrm{MC}$, where t_MC is time measured in Monte Carlo time steps; this corresponds to the effective temperature T_e in (9.1) decreasing as $\frac{1}{t_\mathrm{MC}}$.

[1] Because design D_A must previously have been tested, against others which it superceded, we could in principle seek to use also that earlier information about its cost in estimating ΔH^n. In the present discussion we will ignore this possibility, with the advantage that this makes acceptance decisions statistically independent of each other.

9.4 Comparison of Simulated and Stochastic Annealing

The transition probability curves for Metropolis, Glauber and stochastic annealing dynamics are shown in Figure 9.1. Stochastic annealing matches Glauber dynamics surprisingly well for $|\Delta\mu| < \frac{\beta}{2}$, which suggests it may be a good approximation to thermal selection.

The ratio of forward to backward acceptance for the stochastic annealing rule (9.14) is given by

$$\frac{P_{A\to B}}{P_{B\to A}} = \frac{\left(1 - \mathrm{erf}\frac{\sqrt{n}\Delta\mu}{\sqrt{2}\,\sigma_{\Delta H}}\right)}{\left(1 + \mathrm{erf}\frac{\sqrt{n}\Delta\mu}{\sqrt{2}\,\sigma_{\Delta H}}\right)}; \tag{9.15}$$

while this does not afford a thermal distribution, we can observe how well it approximates it. Expanding (9.8) and (9.15) to first order in

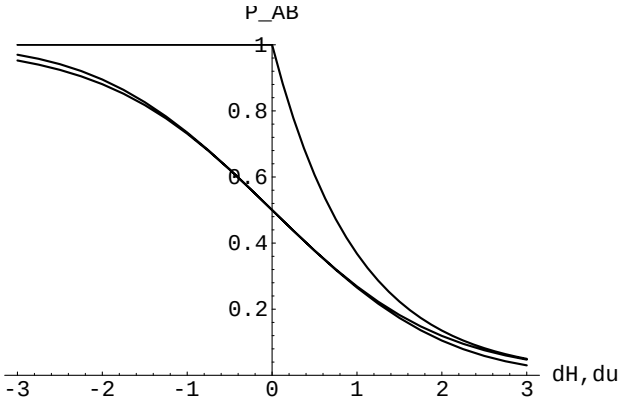

Figure 9.1: Transition probabilities for Metropolis, Glauber and stochastic annealing (right side, top to bottom). The Metropolis and Glauber curves are shown for $\beta_\mathrm{e} = 1$; the stochastic annealing probability is plotted accordingly with $\beta_\mathrm{s} = 1$ in (9.16). Note the similarity between (thermal) Glauber dynamics and stochastic annealing.

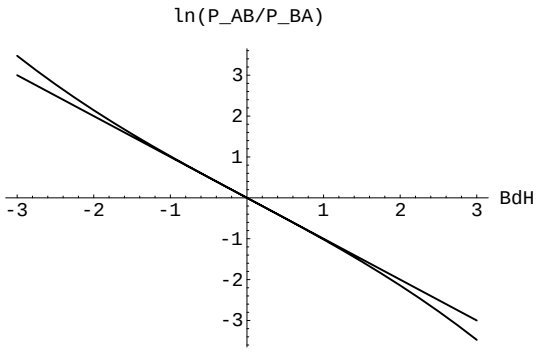

Figure 9.2: Logarithm of the ratio of forward to backward acceptance for simulated annealing (straight) and stochastic annealing (curved). Stochastic annealing satisfies thermal statistics well for cost changes within $\pm\frac{2}{\beta}$.

ΔH and $\Delta\mu$ and matching terms, we find that the inverse effective stochastic annealing temperature may be approximated by

$$\beta_{\mathrm{s}} \simeq \frac{\sqrt{8/\pi}}{\sigma^{n}_{\Delta H}} = \frac{\sqrt{8/\pi}}{\sigma_{\Delta H}}\sqrt{n}. \tag{9.16}$$

Setting $\beta_{\mathrm{e}} = 1$ in (9.8) and $\beta_{\mathrm{s}} = 1$ in (9.15) (*i.e.*, $\frac{\sqrt{n}}{\sigma_{\Delta H}} \simeq \sqrt{\frac{\pi}{8}}$), we compare the logarithms of each in Figure 9.2; despite having a fundamentally different functional form, the finite sample acceptance mirrors thermal acceptance rather well for cost changes within $\pm\frac{2}{\beta}$. Outside this range, the most significant errors are over-acceptance of highly favourable moves and under-acceptance of highly unfavourable transitions — something which from the standpoint of optimisation may not be too serious.

The analogy with thermal acceptance, embodied in equation (9.16), leads to a useful prediction of an equivalent thermal distribution if we can further approximate $\sigma_{\Delta H}$ as independent of the pair of designs D_A and D_B. Additional notekeeping, not discussed here, allows us to relax this condition if $\sigma_{\Delta H}$ changes sufficiently slowly such that

$\sigma_A \simeq \sigma_B$ for all designs D_A and their similar alternatives D_B. Note that this condition is in practice generally satisfied: in protein sequence optimisation (§7.5), sequences are compared with alternatives differing by a single monomer; in the probabilistic traveling salesman problem (§9.5), alternative tours differ by a single tour segment flip (reversal of a connected string of cities).

9.5 Probabilistic Traveling Salesman Problem

In this section we apply the method of stochastic annealing, presented in §9.3, to the probabilistic traveling salesman problem (PTSP), described in Chapter 8.

We outline the problem here for convenience. Consider a set A of N cities, each of which must be visited with probability p; on any given instance, an active subset $B \subseteq A$ must be visited. Given an *a priori* tour t through A, the corresponding pruned tour u through B visits the active cities in the same order. The objective is to determine the *a priori* tour t such that the tour u is minimal in the expected sense.

It should be remarked that the probabilistic nature of the PTSP induces behaviour distinctly different from that exhibited by the conventional traveling salesman problem (TSP). Notably, the optimal PTSP tour may intersect itself [22], a phenomenon easily shown to be absent from the optimal TSP tour. This suggests the practical question of how well the optimal TSP tour satisfies the PTSP. The optimal TSP and PTSP tours are guaranteed to coincide if the number of cities $N \leq 5$; larger tours correspond only under special circumstances, *e.g.*, when the cities lie at the vertices of a convex n-gon [22]. Jaillet [27] has provided examples in which substituting the optimal TSP tour into the corresponding PTSP gives arbitrarily large errors. It follows that, for practical applications (*e.g.*, mail routes) as well as theoretical interest, we must attempt to solve the PTSP directly.

We stochastically annealed PTSP tours defined over 100 cities uniformly distributed on the unit square. Simulations were run for $\sim 100,000$ time steps, with n increasing from 1 to ~ 4000. Figure 9.3 shows near optimal tours for various probabilities p. When $p = 1$, the

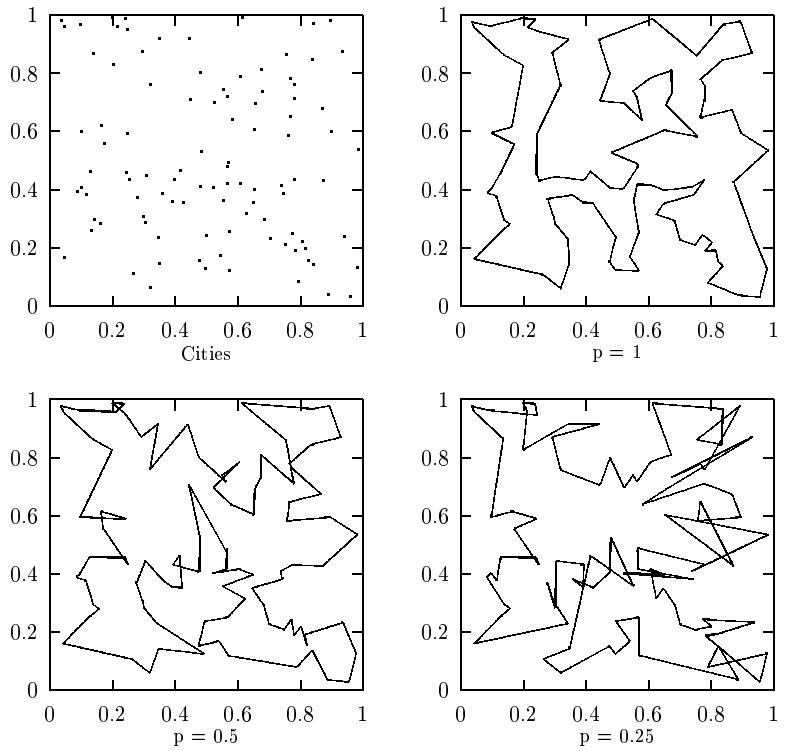

Figure 9.3: Near-optimal tours for the probabilistic traveling salesman problem, with $p = 1, \frac{1}{2}, \frac{1}{4}$. For $p = 1$, the problem is deterministic and reduces to the conventional TSP.

problem reduces to the conventional TSP.

Preliminary results (expected pruned tour length $d(u)$) are shown in Table 9.1. The 2-popt and 1-shift algorithms are both deterministic (downhill) and were applied to the exact mean tour length[2] by Bertsimas [22]. Stochastic annealing follows, where we have also included

[2]Jaillet derived a closed form expression for computing the exact expected length of an *a priori* tour over all realisations of the city probabilities [27]. This allows comparison of conventional optimisation techniques with probabilistic methods, such as the one introduced here.

p	2-popt $d(u)$	1-shift $d(u)$	SA $d(u)$	SA $d(t)$
1	8.9	8.5	8.5	8.5
$\frac{1}{2}$	6.6	6.2	6.3	9.2
$\frac{1}{4}$	4.7	4.5	4.7	11.9

Table 9.1: Optimised PTSP pruned tour lengths $d(u)$. Cities are uniformly distributed on the unit square. Results are shown for the deterministic algorithms 2-popt and 1-shift and stochastic annealing; the *a priori* tour length $d(t)$ is also included.

for comparison the *a priori* tour length $d(t)$.

Notwithstanding the advantage provided by exact measurement of the cost function, the solutions achieved by the deterministic techniques are nearly equivalent to those obtained by stochastic annealing. This is not truly surprising, since simulated annealing (which our technique approximates) is well known to provide at least as optimal solutions as downhill dynamics. However, for general randomly distributed cost functions, the mean cost cannot be exactly computed and neither deterministic methods nor conventional simulated annealing are applicable.

9.6 Conclusion

We have provided a general method of optimisation of a randomly distributed cost function. It effects a transition probability between two states satisfying $P_{A\to B}(\Delta\mu, n) = \frac{1}{2}(1-\mathrm{erf}\frac{\sqrt{n}\,\Delta\mu}{\sqrt{2}\sigma_{\Delta H}})$, where $\Delta\mu$ is the mean difference in costs and the effective temperature is proportional to $\frac{1}{\sqrt{n}}$. Comparison with simulated annealing suggests the technique provides a close approximation to thermal selection. Preliminary tests on the probabilistic traveling salesman problem are encouraging.

Chapter 10

Exactly Solvable Hierarchical Optimisation Problem

Le bien est l'ennemi du mieux.

FRANCOIS-MARIE VOLTAIRE

In contrast to the previous chapter, here we consider a sequence of elementary decisions which must be made in the light of successive information learnt. A key feature is that the decisions must balance the reduction of immediate cost against learning information and hence securing a wider range of future options — a conflict which motivates us to attach a *value* to information. We analytically derive an optimal decision policy; while each individual decision is elementary, the solution to the collective problem, which may be interpreted as a novel percolation model, exhibits a phase transition and finite size scaling.

10.1 Introduction

The introduction of probabilistic elements in combinatoric optimisation problems (COPs) began with Jaillet [27] in his study of the probabilistic traveling salesman problem (PTSP). The objective is to obtain a first stage solution which minimises the expected cost of a second stage tour. Bertsimas [22] extended this idea to other probabilistic COPs and suggested the *a priori* optimisation heuristic to solve them. These problems are characterised by a stochastically defined instance

(in the case of the PTSP, the city locations), which is learnt after initially optimising over all instances and allows further optimisation in light of the actual instance (see, *e.g.*, [23]).

The inherently *hierarchical* nature of probabilistic optimisation was identified in Chapter 8, where we introduced hierarchical optimisation generally. We propose in this chapter a hierarchical optimisation problem which consists of many elementary decisions but displays nontrivial global behaviour and derive an exact analytic solution.

10.2 Description of Problem

The *constrained* form of the problem, appropriate to the optimal development of a design, is as follows. Decision starts from a unique node at level $N+1$, from which the costs associated with z descendant nodes at level N are observed. It must be decided how many of these nodes to buy and hence pursue from level N; this process continues in like manner down to level 1. For each of the nodes bought at level n, the price of z descendant nodes at level $n-1$ are learnt. It is then decided which of all of these descendant nodes to buy before proceeding to the next level (Figure 10.1). The objective is to reach at least one node at level 1 with the minimum overall cost. This implies the constraint that at every level at least one node must be bought.

The problem can also be interpreted as one of economic growth, the decision to buy representing investment in future return in the form of negative costs, *i.e.*, profits. In this *economic* form of the problem it is not appropriate to deny the possibility of buying no nodes at some level, but of course the result corresponds to termination of the activity.[1]

It is a vital feature of our model that, in either form, the costs are only learnt one level at a time and previous decisions cannot be changed. The contrasting case, where all costs are known in advance, would correspond to the problem of directed branched polymers spanning a Cayley tree of sites. We will assume that the costs x

[1] It is important to distinguish between the tree of nodes, shown in Figure 10.1, from the corresponding tree of decisions (*cf.* Figure 8.3); in particular, the decision tree is much more highly branched, the decision at each level being the total number of nodes to purchase.

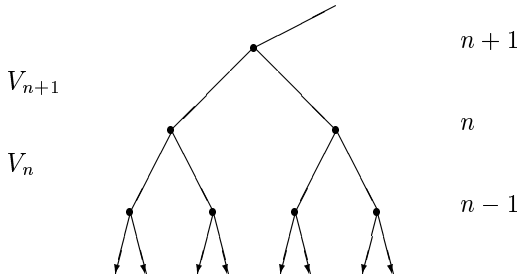

Figure 10.1: The decision problem may be summarised as a z-fold Cayley tree (where $z = 2$ in the example shown), N levels deep, with stochasticly chosen costs x associated with every node. The tree is traversed from top (level $N+1$) to bottom (level 1), such that costs on a given level are known (and may thus be purchased) only by paying the cost of the node from which they branch. The objective is to sequentially traverse the tree from top to bottom such that the total incurred cost is minimal.

are drawn independently from some *a priori* probability distribution such that they may be negative or positive. We concentrate mainly on the case where the distribution is uniform over the finite interval $x \in [\lambda - \frac{1}{2}, \lambda + \frac{1}{2}]$; we argue that the case of a general distribution qualitatively behaves similarly. Unless otherwise stated, we focus on the solution to the economic version of the problem.

10.3 Optimality Equation

We set out to obtain the decision policy which minimises total expected costs, henceforth called the optimal decision policy. We begin by defining V_n to be the expected minimum total cost incurred over $n-1$ levels, *i.e.*, stemming from a single (purchased) node at level n downwards. That is, $-V_n$ is the expected *value*[2] associated with the

[2]The definition of value presented here is not identical to the definition provided in Chapter 8 (this chapter was published earlier). We address the difference in definitions and its implication in Appendix 10.8.

subtree stemming from a node x_n, corresponding to the maximum *cost* that we are willing to pay for knowledge of, or access to, that subtree. Identifying $-V_n$ as both the value of a subtree and a bound on costs enables us to recursively define V_{n+1} in terms of the V_n associated with its descendant nodes.

The optimality equation for V_n may be expressed

$$V_{n+1} = \langle \min_a (\mathrm{C}(a(s)) + g(a(s))V_n) \rangle_s, \tag{10.1}$$

where $\mathrm{C}(a(s))$ is the cost incurred by choosing action a when in state s. The state s is the set of costs observed and the action a is the particular subset of those costs paid. The function $g(a(s))$ gives the number of costs in a.

The optimal policy $a_{opt}(s)$ is achieved by paying those costs x_n which are less than the maximum cost we are willing to pay, *i.e.*, satisfying $x_n < -V_n$. Summing over the states s, (10.1) appears as

$$V_{n+1} = \sum_s (P(s)\mathrm{C}(a_{opt}(s)) + P(s)g(a_{opt}(s))V_n), \tag{10.2}$$

where $P(s)$ is the probability that state s occurs. The expectation of the cost of a given action is equal to the average cost of a purchased node, c_n, times the expected number of nodes purchased. The optimality equation may then be written

$$V_{n+1} = \sum_s P(s)g(a_{opt}(s))(c_n + V_n). \tag{10.3}$$

The factor $P(s)g(a_{opt}(s))$ is the mean number of nodes purchased. With p_{ni} the probability of purchasing the i^{th} node, and noting that the p_{ni} are independent, we may alternatively express the mean number of purchased nodes as the sum of p_{ni} over the z descendant nodes, which yields

$$V_{n+1} = \sum_{i=1}^{z} p_{ni}(c_n + V_n) = z p_n (c_n + V_n). \tag{10.4}$$

Here, p_n is the probability that cost $x_n < -V_n$ and c_n is the mean of x_n given that $x_n < -V_n$, both of which are readily obtained from the cost distribution f_x;

$$p_n = \int_{-\infty}^{-V_n} f_x dx, \qquad c_n p_n = q_n = \int_{-\infty}^{-V_n} x f_x dx. \tag{10.5}$$

The optimality equation may then finally be expressed

$$V_{n+1} = z(p_n V_n + q_n). \tag{10.6}$$

We have thus derived the recursion relation which governs the optimal decision policy. It is important to note that while the decision process occurs sequentially going down the tree, the policy is defined recursively going up the tree. This means that the boundary condition is located at the bottom level; since there exists no descendant nodes at level 1, we clearly must have $V_1 = 0$.

The stability of (10.6) is implied by

$$\left|\frac{dV_{n+1}}{dV_n}\right| < 1, \tag{10.7}$$

which is satisfied for $zp_n < 1$, where zp_n, the number of descendant nodes times the probability than an unknown cost x_n is paid, is the branching rate. Thus the sequence V_n is convergent going up the tree if and only if the mean purchase of subtrees is decaying downwards.

The optimal decision problem may be alternatively expressed, on a given realisation of the costs x_n, as sequentially choosing the number of nodes to purchase at each level such that the total incurred cost is minimal. Let B_n be the expected number of costs paid at level n; clearly, $B_n \leq z^{N+1-n}$. The total cost incurred at level n, C_n, may then be expressed as

$$C_n = B_n c_n. \tag{10.8}$$

Since

$$B_n = z p_n B_{n+1} \quad \text{and} \quad B_{N+1} = 1, \tag{10.9}$$

we may express the total cost C as

$$C = \sum_{n=1}^{N} C_n = \sum_{n=1}^{N} q_n z^{N+1-n} \prod_{i=n+1}^{N} p_i. \tag{10.10}$$

10.4 Uniform Cost Distribution

The optimal decision policy is dependent on the probability density function, f_x, from which the costs x are chosen, the number of decisions to be made, N, and the degree of the decision tree, z. In this section we examine the behaviour associated with a uniform distribution, $f_x = 1, x \in [\lambda - \frac{1}{2}, \lambda + \frac{1}{2}]$, for which the optimal policy is most tractable. For convenience, we set the degree of the Cayley tree $z = 2$.

For $-V_n \leq \lambda + \frac{1}{2}$, we can express p_n and q_n from (10.5) as

$$p_n = -V_n - \lambda + \tfrac{1}{2}, \qquad q_n = \tfrac{1}{2}(V_n^2 - \lambda^2 + \lambda - \tfrac{1}{4}); \tag{10.11}$$

for $-V_n > \lambda + \frac{1}{2}$, the values of p_n and q_n simplify to 1 and λ, respectively. These physical bounds of p_n (as a probability) and q_n (as a mean) apply to all identities in the remainder of this section. Substitution of (10.11) into (10.6) yields the optimality equation in the form of a quadratic map,

$$V_{n+1} = -(V_n + \lambda - \tfrac{1}{2})^2, \qquad V_1 = 0. \tag{10.12}$$

Equation (10.12) may be expressed in terms of the more useful parameter p_n, the probability of paying an unknown cost x_n. Eliminating V_n from the left side of (10.11) and (10.12) yields

$$p_{n+1} = p_n^2 + \tfrac{1}{2} - \lambda, \qquad p_1 = \tfrac{1}{2} - \lambda, \tag{10.13}$$

the stable fixed point of which is given by

$$p_f = \tfrac{1}{2} - \sqrt{\lambda - \tfrac{1}{4}}, \qquad \lambda \geq \tfrac{1}{4}. \tag{10.14}$$

Without loss of generality, we restrict our attention to $\lambda \in [-\frac{1}{2}, \frac{1}{2}]$.

Observing that (10.12, 10.13) have stable fixed points for $\lambda \geq \frac{1}{4}$ and diverge otherwise, we henceforth refer to the separatrix $\lambda = \frac{1}{4}$ as the critical point λ_c.

First we consider the behaviour of the optimal economic policy. We then address the solution of the constrained form of the problem, which, it turns out, coincides with the economic solution for $\lambda < \lambda_c$.

From (10.9), the growth of B from level n to level $n+1$ vanishes for $zp_n \leq 1$. Substituting this condition into the left side of (10.11) implies $-V_n \leq \lambda$, which, it can be shown, is not satisfied for $\lambda \in [\frac{1}{4}, \frac{1}{2}]$. Thus, the branching rate obeys $zp_n < 1$ and the sequence $B_{N+1} \ldots B_1$ is decreasing for $\lambda \geq \lambda_c$, corresponding to a region of decreasing economic activity. Moreover, since the B_n scale geometrically with n by the factor zp_n, the tree of purchased nodes is finite in the limit of $N \to \infty$ for $\lambda \geq \lambda_c$.

To examine the behaviour of $p(n, \lambda)$, we may approximate the difference equation (10.13) by a differential equation,

$$\frac{dp}{dn} \simeq p^2 - p + \lambda - \frac{1}{2}, \tag{10.15}$$

provided p_n is slowly varying. For $\lambda \geq \lambda_c$, the solution to (10.15) appears as

$$p(n, \lambda) = \tfrac{1}{2} - \sqrt{\lambda - \lambda_c}\, \coth(\sqrt{\lambda - \lambda_c}\,(n + c)) \tag{10.16}$$

and approaches its fixed point (10.14) exponentially. For $\lambda \leq \lambda_c$, it is convenient to express the solution as

$$p(n, \lambda) = \tfrac{1}{2} - \sqrt{\lambda_c - \lambda}\, \cot(\sqrt{\lambda_c - \lambda}\,(n + c)), \tag{10.17}$$

where the constant of integration c is of order unity.

At the critical point $\lambda = \frac{1}{4}$, both of the above reduce to algebraic behaviour and p_n slowly approaches its fixed point as

$$p_n \simeq \frac{1}{2} - \frac{1}{n + c}. \tag{10.18}$$

When $\lambda = \frac{1}{2}$, there are no negative costs, and thus the obvious optimal policy is to buy zero nodes at all levels; indeed, this is the policy

suggested by (10.16).

With n_0 the level above which all probabilities $p_n = 1$, the optimal decision policy consists of an initial period of maximum growth down to level n_0, during which all nodes are purchased with probability unity. This corresponds to a phase of expansion in which all (positive and negative) costs are paid in the interest of securing future options. Below n_0, the p_n fall below 1, and there is a decrease in purchasing activity such that at the bottom level only negative costs are paid; this may be identified as a profit making regime. Note that since the number of iterations necessary for p_n to exceed 1 is independent of N (for $N > n_0$), the length of the second regime, n_0, is a function of λ only.

We begin our analysis of the constrained policy by examining how well the economic policy, which we have already solved, satisfies the added constraint of reaching the bottom level of the tree. Let b_n be the probability that the subtree of purchased nodes stemming from a single node on level n does not terminate before reaching level 1, *i.e.*, that at least one of the nodes on level 1 is purchased. We may explicitly calculate b_n by noting that it satisfies the recursion relation

$$b_{n+1} = -(p_n b_n)^2 + 2p_n b_n, \qquad b_1 = 1, \tag{10.19}$$

where p_n is the aforementioned probability of purchasing a node at level n. We are interested in the final term b_{N+1}, the probability that the tree of purchases extends from level $N+1$ to 1, as a function of λ. Clearly, this function depends on the value of N (Figure 10.2). As N approaches infinity, $b_{N+1}(\lambda)$ approaches a step function: below λ_c, the probability of purchasing at least one node at the bottom goes to unity, while above λ_c, the probability vanishes.

For finite N, the point at which the economic policy ceases to continue to the bottom, and thus at which the economic and constrained policies diverge, occurs not at λ_c but rather $\lambda_{\rm eff} < \lambda_c$. We are interested in characterising this divergence point, $\lambda_{\rm eff}$, as a function of N. This is most easily addressed in the framework of a percolation model, which we defer until the end.

For $\lambda > \lambda_c$, it is necessary to adapt our economic decision policy such that it does not exhibit terminating behaviour. Since the

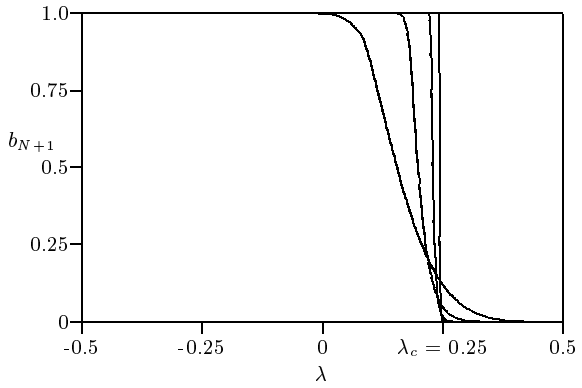

Figure 10.2: Critical transition and finite size effects resulting from the economic policy. Shown is the probability b_{N+1} of the tree of purchased nodes reaching the bottom level as a function of the mean cost λ. Curves are shown, from left to right, for decision tree lengths $N = 4, 8, 16, 32$. For large N, the function approaches a step function, discontinuous at the critical point λ_c.

optimal economic policy may be constrained to satisfy the additional constraint of reaching the bottom level of the tree simply by paying the minimum cost available when the economic policy dictates the purchase of none, the optimal constrained policy might be approximated as a concatenation of optimal economic decision policies, each with different initial conditions.

10.5 General Cost Distribution

For a general cost distribution $f_x = g_x(x - \lambda)$, we may write equation (10.6) as

$$V_{n+1} = z \int_{-\infty}^{-V_n} (V_n + x) g_x dx. \tag{10.20}$$

Change of variables $y = x - \lambda$ and $U_n = -V_n - \lambda$ and integration by parts yields

$$U_{n+1} = -z((y - U_n)G_y)\Big|_{-\infty}^{U_n} + z \int_{-\infty}^{U_n} G_y dy - \lambda, \tag{10.21}$$

where $G_y(y) = \int_{-\infty}^{y} g_y dy$ is the cumulative probability distribution. The first term vanishes for all distributions g_y which go to zero more rapidly than y^{-2} as $y \to -\infty$. Accordingly, the optimal policy for a general cost distribution is given by

$$U_{n+1} = z \int_{-\infty}^{U_n} \int_{-\infty}^{y} g_y dy\, dy - \lambda = zH(U_n) - \lambda. \tag{10.22}$$

We note that $H > 0$, $H(-\infty) = 0$, $H(y) \to y - c$ as $y \to \infty$, and H is concave upwards. There then exists a critical point λ_c such that $zH(U_n) - \lambda$ intersects the line $U_{n+1} = U_n$ zero times for $\lambda < \lambda_c$, once at $\lambda = \lambda_c$, and twice for $\lambda > \lambda_c$, as shown in Figure 10.3. To

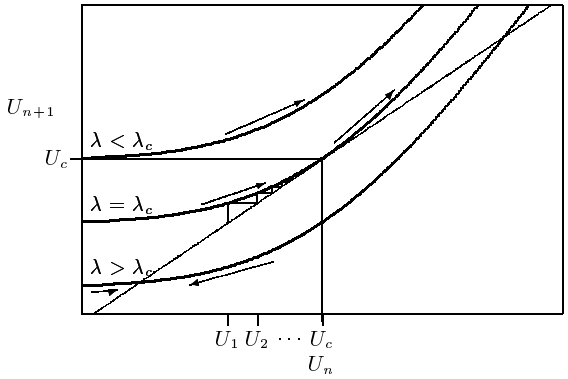

Figure 10.3: The function $zH(U_n) - \lambda$ for λ below, at, and above its critical value plotted in $(n+1, n)$ phase space. The arrows indicate the direction of convergence for increasing n. The sequence $U_n(\lambda)$, $U_1 = -\lambda$, may be obtained geometrically, as shown for $U_n(\lambda_c)$ slowly converging to U_c. For $\lambda < \lambda_c$, the sequence U_n increases to infinity; for $\lambda > \lambda_c$, the sequence U_n converges to $U_f < U_c$.

observe the near critical behaviour, we substitute $U_n = u_n + U_c$ in (10.22), where U_c is the fixed point U_f at λ_c, and expand H about U_c to second order to obtain

$$u_{n+1} + U_c \simeq zH(U_c) + u_n zH'(U_c) + \tfrac{1}{2}u_n^2 zH''(U_c) - \lambda. \quad (10.23)$$

Observing that $zH(U_c) = U_c + \lambda_c$ and $zH'(U_c) = 1$, which determine the critical parameters, and $H''(U_c) = zg(U_c)$, (10.23) reduces to

$$u_{n+1} \simeq u_n + \tfrac{1}{2}zg(U_c)u_n^2 + \lambda_c - \lambda, \quad (10.24)$$

which generalises the recursion relation (10.12) for a general cost distribution and reduces exactly to (10.12) for $z = 2$ and a uniform distribution.

It is now clear that the optimal policy for any well behaved distribution of costs behaves qualitatively similarly as the case for a uniform distribution, which we have already examined in detail.

10.6 Interpretation as a Percolation Model

The economic version of the problem may be interpreted as a percolation problem on a Bethe lattice of dimension z (Figure 10.4). Unlike a conventional percolation problem, the probabilities p are not uniform over the lattice sites, but rather satisfy the recursion relation (10.13). The analogy between our HOP and a percolation model provides mutual insight into the two seemingly disparate problems.

On a large but finite lattice, conventional percolation models exhibit a shift in the percolation threshold and finite size scaling of various quantities near p_c. We find similar behaviour characterised by a shift in λ_c and finite size scaling of properties nearby.

The quantity $b(N+1, \lambda)$ may be interpreted as the probability of finding a spanning cluster in our Cayley tree of linear dimension N at concentration $p(\lambda)$. The effective critical point λ_{eff} at which the cluster spans our finite tree, and consequently percolation occurs, satisfies $b(N+1, \lambda_{\text{eff}}) \simeq \frac{1}{2}$, that is, λ_{eff} is the value at which $b_{N+1}(\lambda)$ makes its sharp transition from 0 to 1.

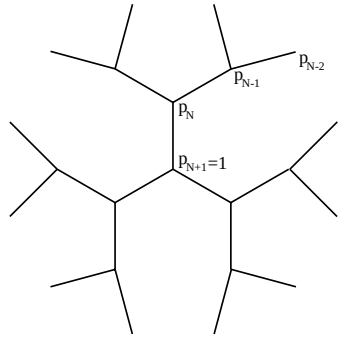

Figure 10.4: The economic version of the problem may be interpreted as percolation on a Bethe lattice with non-uniform probabilities.

Moreover, we find that

$$\prod_{n=1}^{N} zp_n(\lambda_{\text{eff}}) = B_1 \simeq 1. \quad (10.25)$$

and $p_N(\lambda)$ undergoes a sharp transition from $\frac{1}{2}$ to 1 at λ_{eff}. Imposing the latter condition on the analytic result (10.17) yields

$$\lambda_c - \lambda_{\text{eff}} \simeq (\frac{\pi}{N})^2 \quad (10.26)$$

independent of the precise value of p_N used. Equation (10.26) can be interpreted as $\lambda_c - \lambda_{\text{eff}} \propto N^{-1/\nu}$, where the conventional critical exponent $\nu = \frac{1}{2}$ agrees with ordinary percolation on a Bethe lattice [30].

It would be interesting to explore generalisations of the decision problem to different connectivities of the decisions, *i.e.*, corresponding to different percolation lattices. On a Euclidian lattice, in which the path to given decision is not unique, we believe our model would correspond to a form of invasion percolation in which the future consequences of each invasion step must be weighed in choosing which step to take.

10.7 Conclusion

We have shown that optimising a sequence of elementary decisions with limited information at each stage yields complex global behaviour with a percolation-like critical point. When the mean cost λ lies below a critical threshold λ_c, the optimal number of options to pursue grows exponentially before entering a steady profitable region. Above λ_c, we distinguished between the economic form of the model, in which the tree of options pursued tends to terminate, and the constrained version, in which at least one option must be pursued to the end and for which we can only offer an approximately optimal solution. We demonstrated that our solution is universal in the sense that its qualitative behaviour, which we examined in detail for a uniform cost distribution, does not depend on quantitative details of the model, the branching rate z and distribution of costs f_x.

Near the critical point, at which the solution to the two problem versions bifurcates, the economic solution exhibits behaviour analogous to a percolation model; we found a finite size shift in λ_c and finite size scaling in the probability b_{N+1} of the solution connecting to the bottom, corresponding to the probability of a spanning cluster in percolation. Our problem naturally maps to a novel percolation model on a Bethe lattice in which the probabilities p of occupation satisfy a recursion relation dependent on z and f_x. The percolation dynamics are characterised by λ; at λ critical the probability of occupation satisfies an inverse radius dependence.

10.8 Appendix A: Applying the General Optimality Equation

In our formulation of the economic problem in §10.2, we interpreted each stage of the problem as a set of simultaneous independent decisions. Hierarchical optimisation problems, however, allow (by definition) a single decision at each stage. Here we show that, formulated as an HOP proper, the economic optimality equation, already derived in §10.3, readily follows from the general optimality equation from Chapter 8. By doing so, we highlight conditions under which HOPs may be exactly solved.

We begin with the general optimality equation, which the reader may recall from (8.8) to be

$$W_{n+1}(s_{n+1}) = \max_{a_{n+1}} \Big[R(a_{n+1}|s_{n+1}) + \sum_{s_n} P(s_n|s_{n+1}, a_{n+1})\, W_n(s_n) \Big], \tag{10.27}$$

where we have replaced V with W to avoid confusion with the value $-V$ defined in this chapter. In the context of the economic problem, the action a_{n+1} which maximises the right side of (10.27) is the optimal number of nodes purchased, of expected value B_{n+1}; the state s_{n+1} is the set of nodes examined, of typical number $\frac{B_{n+1}}{p_{n+1}}$.

The impediment to solving (10.27) is summing over all possible realisations of the state s_n, on which $W(s_n)$ depends. Consider rewriting (10.27) in terms of the recursion parameter V_n, which satisfies

$$W_n = -B_n V_n; \tag{10.28}$$

$-V_n$ is the expected value attached to the subtree of nodes descending from (but not including) a single node at level n, the same V_n used previously in this chapter. Unlike W_n, V_n does not depend on the state s_n; it varies only with n. Since B_n is already an expectation over s_n, this allows us to forgo the summation over s_n.

Inserting (10.28) into (10.27) yields

$$-B_{n+1}V_{n+1} = -B_n c_n - B_n V_n, \tag{10.29}$$

where c_n is the mean of the cost x_n, given that $x_n < -V_n$ (10.8). Multiplying both sides of (10.29) by zp_n and noting that $B_n = zp_n B_{n+1}$ (10.9), where p_n is the probability of purchasing a node at level n, we may express (10.29) as

$$V_{n+1} = zp_n(c_n + V_n), \tag{10.30}$$

which is identical to the result derived by alternative means in (10.4).

Chapter 11

Tie Knots and Random Walks

A well-tied tie is the first serious step in life.

OSCAR WILDE

Here we introduce a mathematical model of tie knots and provide a map between tie knots and persistent random walks on a triangular lattice.[1] The topological structure of a knot may be characterised directly by the walk sequence. We classify knots according to their size and shape and quantify the number of knots in each class. The optimal knot in a class is selected by the proposed aesthetic conditions of symmetry and balance. Of the 85 knots which can be tied with a conventional tie, we recover the four knots in widespread use and introduce six new aesthetic ones. For large (though impractical) half-winding number, we present some asymptotic results.

11.1 Introduction

The simplest of conventional tie knots, the Four-in-Hand, has its origins in late nineteenth-century England: drivers tied their scarves round their necks lest they lose the reigns of their four-in-hand carriages. King Edward VIII, after abdicating in 1936, has been credited with introducing what is now known as the Windsor knot, whence its smaller derivative, the Half-Windsor, evolved [34]. More recently, in 1989, the Pratt knot was registered with the Neckwear Association of America, the first new knot to appear in 50 years.

Tie knots, evidently, do not come quickly. Rather than wait another half-century for the next knot to appear, we present in this chapter a formal approach. Our attention is limited to necktie knots, *viz.*, those that may be tied with a conventional necktie. An engaging account of tie history can be found in [34]; tie custom and sartorial gospel are considered in [31].

[1] This chapter is the result of work done in collaboration with Yong Mao [32,33]; the text included here was written by the author.

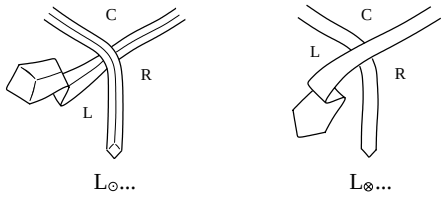

Figure 11.1: The two ways of beginning a knot. Both give rise to the triagonal basis $\curlyvee$ and divide the space into the three regions through which the active end can subsequently pass. For knots beginning with $L_\odot$, the tie must begin inside out.

11.2 Definition of Tie Knots

A tie knot is initiated by bringing the wide (active) end to the left and either over or under the narrow (passive) end, forming the triagonal basis $\curlyvee$ and dividing the space into right, centre and left (R, C, L) regions (Figure 11.1; this, and all other figures, are drawn in the frame of reference of a mirror image of the actual tie). Knots beginning to the right are identical upon reflection to their left-hand counterparts and are omitted from the discussion.

Once begun, a knot is continued by wrapping the active end around the triagonal basis; this process may be considered a sequence of half-turns from one region to another. The location and orientation of the active end are represented by one of the six states $R_\odot, R_\otimes, C_\odot, C_\otimes, L_\odot$ and $L_\otimes$, where R, C and L indicate the regions from which the active end emanates and $\odot$ and $\otimes$ denote the directions of the active end as viewed from in front, *viz.*, out of the page (shirt) and into the page (shirt), respectively.

The notational elements $R_\odot, R_\otimes, C_\odot$, *etc.*, initially introduced as states, may be considered moves in as much as each represents the

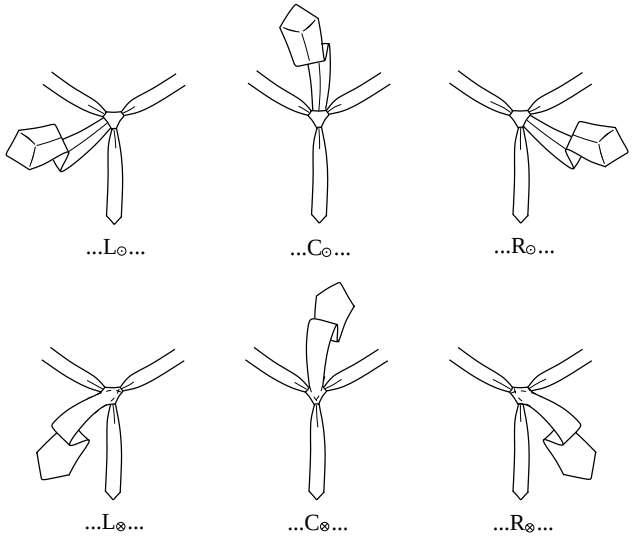

Figure 11.2: The six moves with which a tie knot is tied. The move $L_\odot$, for instance, indicates the move which places the active end into the left region and directed out of the page.

half-turn necessary to place the active end into the corresponding state (Figure 11.2). This makes the successive moves $R_\odot L_\odot$, for instance, impossible, and implies that $R_\otimes$ is the inverse of $R_\odot$. Accordingly, the move direction must oscillate between $\odot$ and $\otimes$ and no two consecutive move regions may be identical.

To complete a knot, the active end must be wrapped over the front, *i.e.*, either $R_\odot L_\otimes$ or $L_\odot R_\otimes$, then underneath to the centre, $C_\odot$, and finally through (denoted T but not considered a move) the front loop just made (Figure 11.3).

We can now formally define a tie knot as a sequence of moves chosen from the move set $\{R_\odot, R_\otimes, C_\odot, C_\otimes, L_\odot, L_\otimes\}$, initiated by $L_\otimes$ or $L_\odot$ and terminating with the subsequence $R_\odot L_\otimes C_\odot T$ or $L_\odot R_\otimes C_\odot T$. The sequence is constrained such that no two consecutive moves indicate the same region or direction. The complete sequence for the

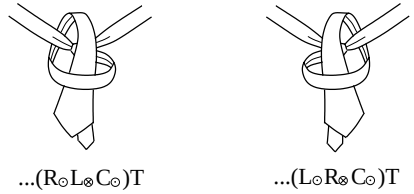

Figure 11.3: The two ways of terminating a knot. The active end is finally put through (denoted T) the front loop constructed by the last three moves.

Four-in-Hand, for example, is shown in Figure 11.4.

11.3 Tie Knots as Random Walks

We represent knot sequences as random walks on a triangular lattice. The axes r, c, l correspond to the three move regions R, C, L and the unit vectors $\hat{\mathbf{r}}, \hat{\mathbf{c}}, \hat{\mathbf{l}}$ represent the corresponding moves (Fig. 11.5); we omit the directional notation $\odot, \otimes$ and the terminal action T. Since all knot sequences end with $C_\odot$ and alternate between $\odot$ and $\otimes$, all knots of odd number of moves begin with $L_\odot$ and those of even number of moves begin with $L_\otimes$. Our simplified random walk notation is thus unique, and we only make use of the directional notation $\odot$ and $\otimes$ in the context of move sequences.

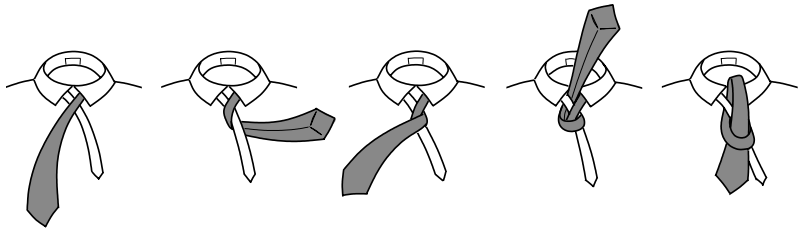

Figure 11.4: The Four-in-Hand, represented by the sequence $L_\otimes R_\odot L_\otimes C_\odot T$.

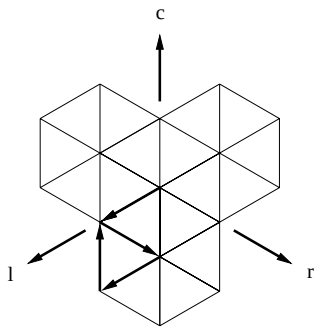

Figure 11.5: A tie knot may be represented by a persistent random walk on a triangular lattice, beginning with $\hat{\mathbf{l}}$ and ending with $\hat{\mathbf{l}}\,\hat{\mathbf{r}}\,\hat{\mathbf{c}}$ or $\hat{\mathbf{r}}\,\hat{\mathbf{l}}\,\hat{\mathbf{c}}$. Only steps along the positive r, c and l axes are permitted and no two consecutive steps may be the same. Shown here is the Four-in-Hand, indicated by the walk $\hat{\mathbf{l}}\,\hat{\mathbf{r}}\,\hat{\mathbf{l}}\,\hat{\mathbf{c}}$.

The three-fold symmetry of the move regions implies that only steps along the positive lattice axes are acceptable and, as in the case for moves, no consecutive steps can be identical; the latter condition makes our walk a second order Markov, or persistent, random walk. Nevertheless, every site on the lattice can be reached since, *e.g.*, $-\hat{\mathbf{c}} = \hat{\mathbf{r}} + \hat{\mathbf{l}}$ and $2\hat{\mathbf{c}} = \hat{\mathbf{c}} + \hat{\mathbf{l}} + \hat{\mathbf{c}} + \hat{\mathbf{r}} + \hat{\mathbf{c}}$.

The evolution equations for the persistent random walk are written

$$\begin{aligned} k_{n+1}(r,c,l) &= \tfrac{1}{2}p_n(r-1,c,l) + \tfrac{1}{2}q_n(r-1,c,l) \\ p_{n+1}(r,c,l) &= \tfrac{1}{2}k_n(r,c-1,l) + \tfrac{1}{2}q_n(r,c-1,l) \\ q_{n+1}(r,c,l) &= \tfrac{1}{2}k_n(r,c,l-1) + \tfrac{1}{2}p_n(r,c,l-1), \end{aligned} \tag{11.1}$$

where $k_n(r,c,l)$ is the conditional probability that the walker is at point (r,c,l) on the nth step, having just taken a step along the positive r-axis, p is conditioned on a step along the positive c-axis, *etc.* The unconditional probability of occupation of a site, U, may be written

$$U_n(r,c,l) = k_n(r,c,l) + p_n(r,c,l) + q_n(r,c,l). \tag{11.2}$$

11.4 Size of Knots

The size of a knot, and the primary parameter by which we classify it, is the number of moves in the knot sequence, denoted by the half-winding number h. The initial and terminal sequences dictate that the smallest knot be given by the sequence $L_\odot R_\otimes C_\odot T$, with $h = 3$. Practical (*viz.*, the finite length of the tie) as well as aesthetic considerations suggest an upper bound on knot size; we limit our exact results to half-winding number $h \leq 9$. The number of knots as a function of size, $K(h)$, corresponds to the number of walks of length h subject to the initial and terminal conditions.

We derive $K(h)$ by first considering all walks of length n beginning with $\hat{\mathbf{l}}$, our initial constraint. Let $F_{\hat{\mathbf{r}}}(n)$ be the number of walks beginning with $\hat{\mathbf{l}}$ and ending with $\hat{\mathbf{r}}$, $F_{\hat{\mathbf{c}}}(n)$ the number of walks beginning with $\hat{\mathbf{l}}$ and ending with $\hat{\mathbf{c}}$, *etc.* Accordingly, since at any given site the walker chooses between two steps,

$$F_{\hat{\mathbf{r}}}(n) + F_{\hat{\mathbf{c}}}(n) + F_{\hat{\mathbf{l}}}(n) = 2^{n-1}. \tag{11.3}$$

Because the only permitted terminal sequences are $\hat{\mathbf{r}}\,\hat{\mathbf{l}}\,\hat{\mathbf{c}}$ and $\hat{\mathbf{l}}\,\hat{\mathbf{r}}\,\hat{\mathbf{c}}$, we are interested in the number of walks of length $n = h-2$ ending with $\hat{\mathbf{r}}$ or $\hat{\mathbf{l}}$, after which the respective remaining two terminal steps may be concatenated.

We begin by considering $F_{\hat{\mathbf{l}}}(n)$. Now $\hat{\mathbf{l}}$ can only follow from $\hat{\mathbf{r}}$ and $\hat{\mathbf{c}}$ upon each additional step, that is,

$$F_{\hat{\mathbf{l}}}(n+2) = F_{\hat{\mathbf{r}}}(n+1) + F_{\hat{\mathbf{c}}}(n+1), \tag{11.4}$$

from which it follows that

$$F_{\hat{\mathbf{l}}}(n+2) = F_{\hat{\mathbf{r}}}(n) + F_{\hat{\mathbf{c}}}(n) + 2F_{\hat{\mathbf{l}}}(n). \tag{11.5}$$

Combining (11.3) and (11.5) gives rise to the recursion relation

$$F_{\hat{\mathbf{l}}}(n+2) = F_{\hat{\mathbf{l}}}(n) + 2^{n-1}. \tag{11.6}$$

With initial conditions $F_{\hat{\mathbf{l}}}(1) = 1$ and $F_{\hat{\mathbf{l}}}(2) = 0$, (11.6) is satisfied by

$$F_{\hat{\mathbf{l}}}(n) = \tfrac{2}{3}(2^{n-2} + (-1)^{n-1}). \tag{11.7}$$

The recursion relation for $F_{\hat{\mathbf{r}}}(n)$ is identical to (11.6), but with initial conditions $F_{\hat{\mathbf{r}}}(1) = 0$ and $F_{\hat{\mathbf{r}}}(2) = 1$. Accordingly,

$$F_{\hat{\mathbf{r}}}(n) = \tfrac{1}{3}(2^{n-1} + (-1)^{n-1}). \tag{11.8}$$

The number of knots of size h is equal to the number of walks of length $h-2$ beginning with $\hat{\mathbf{l}}$ and ending with $\hat{\mathbf{r}}$ or $\hat{\mathbf{l}}$, that is,

$$K(h) = F_{\hat{\mathbf{r}}}(h-2) + F_{\hat{\mathbf{l}}}(h-2) = \tfrac{1}{3}(2^{h-2} - (-1)^{h-2}), \tag{11.9}$$

where $K(1) = 0$, and the total number of knots is

$$\sum_{i=1}^{9} K(i) = 85. \tag{11.10}$$

11.5 Shape of Knots

The shape of a knot depends on the number of right, centre and left moves in the tie sequence. Since symmetry dictates an equal number of right and left moves (see below), knot shape is characterised by the number of centre moves γ. We use it to classify knots of equal size h; knots with identical h and γ belong to the same class. A large centre fraction $\frac{\gamma}{h}$ indicates a broad knot (*e.g.*, the Windsor), while a small centre fraction suggests a narrow one (*e.g.*, the Four-in-Hand).

For a knot of half-winding number h, the number of centre moves γ is an integer between 1 and $\frac{1}{2}(h-1)$. Accordingly, for large h, the range of the centre fraction $\frac{\gamma}{h}$ tends toward $[0, \frac{1}{2}]$. However, not all centre fractions allow for aesthetic knots; knots with $\frac{\gamma}{h} < \frac{1}{4}$ are too cylindrical and unbalanced (see below). We consequently limit our attention to centre fractions $[\frac{1}{4}, \frac{1}{2}]$, *i.e.*, $\gamma \in [\frac{1}{4}h, \frac{1}{2}(h-1)]$.

This, along with our size constraint, limits the knots classes of interest (canonical knot classes) to

$$\begin{aligned}\{\{h,\gamma\}\} = \;& \{3,1\},\ \{4,1\},\ \{5,2\},\ \{6,2\},\ \{7,2\} \\ & \{7,3\},\ \{8,2\},\ \{8,3\},\ \{9,3\},\ \{9,4\}\ .\end{aligned} \tag{11.11}$$

The number of knots in a class, $K(h,\gamma)$, corresponds to the number of walks of length h containing γ steps $\hat{\mathbf{c}}$, beginning with $\hat{\mathbf{l}}$ and ending with $\hat{\mathbf{r}}\,\hat{\mathbf{l}}\,\hat{\mathbf{c}}$ or $\hat{\mathbf{l}}\,\hat{\mathbf{r}}\,\hat{\mathbf{c}}$. The sequence of steps may be considered a coarser sequence of γ groups, each group composed of $\hat{\mathbf{r}}$s and $\hat{\mathbf{l}}$s and separated from other groups by a $\hat{\mathbf{c}}$ on the right; the Windsor knot, for example, contains three groups, $\hat{\mathbf{l}}\,\hat{\mathbf{c}}\,\hat{\mathbf{r}}\hat{\mathbf{l}}\,\hat{\mathbf{c}}\,\hat{\mathbf{r}}\hat{\mathbf{l}}\,\hat{\mathbf{c}}$, of lengths 1, 2, 2, respectively. We refer to a particular assignment of the centre steps as a centre structure.

Let n_1 be the number of groups of length 1 in a given sequence, n_2 the number of length 2, ..., $n_{h-2\gamma+1}$ the number of length $h-2\gamma+1$. These group numbers satisfy

$$n_1 + n_2 + \ldots + n_{h-2\gamma+1} = \gamma, \tag{11.12}$$

$$n_1 + 2n_2 + \ldots + (h-2\gamma+1)n_{h-2\gamma+1} = h-\gamma. \tag{11.13}$$

We desire the number of ordered non-negative integer solutions $n_1, n_2, \ldots, n_{h-2\gamma+1}$ to (11.12, 11.13), that is, the number of ordered ways of partitioning the integer $h-\gamma$ into γ positive integers. Call this function $P(h-\gamma,\gamma)$; it is given by

$$P(h-\gamma,\gamma) = \tbinom{h-\gamma-1}{\gamma-1}. \tag{11.14}$$

The number of centre structures is equivalent to $P(h-\gamma,\gamma)$ subject to the terminal condition, which requires that the final group cannot be of length one. The latter condition reduces the possible centre structures by $\binom{h-\gamma-2}{\gamma-2}$.

Since the steps within each group must alternate between $\hat{\mathbf{r}}$ and $\hat{\mathbf{l}}$, the steps of each group may be ordered in two ways, beginning with $\hat{\mathbf{r}}$ or beginning with $\hat{\mathbf{l}}$, except for the first, which by assumption begins with $\hat{\mathbf{l}}$. Accordingly, for a centre structure of γ groups, the number of walks is $2^{\gamma-1}$.

It follows that the number of knots in a class is

$$K(h,\gamma) = 2^{\gamma-1}\left(\tbinom{h-\gamma-1}{\gamma-1} - \tbinom{h-\gamma-2}{\gamma-2}\right) = 2^{\gamma-1}\tbinom{h-\gamma-2}{\gamma-1}. \qquad (11.15)$$

11.6 Symmetry

The symmetry of a knot, and our first aesthetic constraint, is defined as the number of moves to the right minus the number of moves to the left, *i.e.*,

$$s = \sum_{i=1}^{h} x_i, \qquad (11.16)$$

where $x_i = 1$ if the ith step is $\hat{\mathbf{r}}$, -1 if the ith step is $\hat{\mathbf{l}}$ and 0 otherwise. We limit our attention to those knots from each class which minimise s. For $h-\gamma$ even, the optimal symmetry $s = 0$; otherwise, optimal $s = \pm 1$.

The move composition, and hence the symmetry, of a knot sequence corresponds to the terminal coordinates of the analogous random walk. It is natural to inquire about the distribution of these coordinates; we take particular interest in the terminal coordinates of walks corresponding to knots in the class $\{h,\gamma\}$. These coordinates are gaussianly distributed about a point near the origin; the derivation of this distribution is provided in Appendix A.

11.7 Balance

Whereas the centre number γ and the symmetry s tell us the move composition of a knot, balance relates to the distribution of these moves; it corresponds to the extent to which the moves are well mixed. A well balanced knot is tightly bound and keeps its shape. We use it as our second aesthetic constraint.

Let σ_i represent the ith step of the walk. The winding direction $\omega_i(\sigma_i, \sigma_{i+1})$ is equal to 1 if the transition from σ_i to σ_{i+1} is, say, clockwise and -1 otherwise. (By clockwise we mean in the frame of reference of the mirror (*viz.*, $\hat{\mathbf{c}}\,\hat{\mathbf{r}}$, $\hat{\mathbf{r}}\,\hat{\mathbf{l}}$, $\hat{\mathbf{l}}\,\hat{\mathbf{c}}$), which is counter-clockwise in the frame of the shirt. Such distinctions, however, need not concern us.) The balance b may then be expressed

$$b = \tfrac{1}{2}\sum_{i=2}^{h-1} |\omega_i - \omega_{i-1}|. \qquad (11.17)$$

With σ and ω the analogue of angular position and velocity, respectively, the balance b may be considered the sum over the magnitude of the angular acceleration.

Of those knots which are optimally symmetric, we desire that knot which minimises b. Only knots with half-winding number $3i$ and $3i+2$ can have zero balance, where i is a positive integer; half-winding numbers $3i+1$ correspond to optimal balance 1.

11.8 Untying

A tie knot is most easily untied by pulling the passive end out through the knot. It may be readily observed that the resulting conformation, when pulled from both ends, yields either the straightened tie or a subsequent smaller knot [35]. More formally, when the passive end is removed and the two tie ends joined, the tie may either be knotted or unknotted, where any conformation that can be continuously deformed to a standard ring (the canonical unknot) is said to be un-

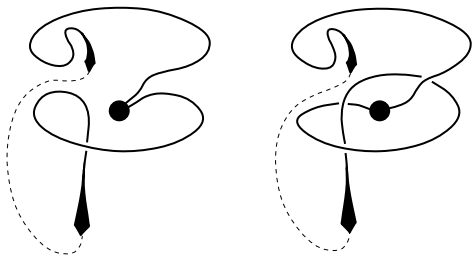

Figure 11.6: The left diagram, with terminal sequence $\ldots R_{\odot}L_{\otimes}C_{\odot}T$, is unknotted, while the right, with terminal sequence $\ldots L_{\odot}R_{\otimes}C_{\odot}T$, forms a trefoil knot.

knotted.

To determine the topological structure of such configurations, we first note that a knot tied up to but not including the terminal sequence corresponds, upon removing the passive end, to a string wound in a ball with the interior and exterior ends protruding. Since the ball can be undone by pulling the exterior end, all such conformations are unknotted. The terminal sequence, in particular the action T, is responsible for any remaining knot.

This can best be observed diagrammatically by projecting the knot onto the plane (Fig. 11.6). The solid spheres represent the non-terminal sequences (which cannot give rise to a knot), with the terminal sequences drawn explicitly. The dotted lines represent imaginary connections of the tie ends. The left diagram, with the terminal sequence $R_{\odot}L_{\otimes}C_{\odot}T$, can be continuously deformed to a loop and is hence unknotted. No amount of deformation of the right diagram, with terminal sequence $L_{\odot}R_{\otimes}C_{\odot}T$, will reduce the number of intersections below three. It is the simplest knotted diagram, a trefoil knot. The knotted status of all aesthetic tie knots is included in Table 11.1.

11.9 Topology

We began this chapter by considering tie knots as combinatoric constructs in light of the special manner in which they are constructed. Here we examine the topological structure of tie knots. As in the previous section, we imagine the tie ends to be connected, this time before removing the passive end.

Figure 11.7 shows the Windsor knot, for example, projected onto the plane. Let the projected diagram take precedence. By manipulating the diagram such that the corresponding knot is continuously deformed,[2] we see that the Windsor knot is topologically equivalent to a trefoil (for an excellent introduction to knot theory, see [36]). Other tie knots give rise to more complicated knots.

The topological complexity of any knot may be characterised by its crossing index, the minimum number of intersections allowed by its

[2]The diagrammatic manipulations associated with continuous deformation of a knot are called the Reidemeister moves; see [36].

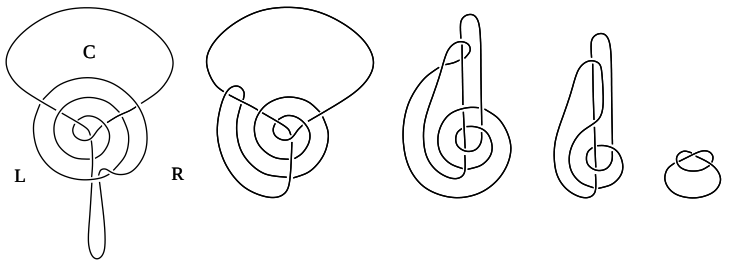

Figure 11.7: The Windsor knot (left) has the topological structure of a trefoil (right).

projection. The standard knot table is arranged by crossing index, for each of which there may be more than one knot. The number of knots per index appears to grow rapidly, but little is known about this number. Demonstrating equivalence between an arbitrary knot projection and its reduced (standard knot table) form by geometric manipulation is a tedious and often nontrivial task. We wish instead to determine the crossing index of tie knots by grammatically manipulating knot sequences.

The grammatical rules associated with diagrammatic reduction become apparent by considering a more tractable diagram projection, applied to the Windsor knot in Figure 11.8. The new projection may be derived from the projection used in Figure 11.7 by contracting the top two arms of the triagonal basis $\curlyvee$ and sliding the windings of the active end onto them. Preserving the old regions and directions, the new projection allows immediate recognition of the sequence which generates it.[3]

It may be verified by constructing appropriate projections that the rules of the grammar are

$$\ldots XRLCT, \ldots XLRCT \quad \rightarrow \quad \ldots XC, \tag{11.18}$$

[3]While all of the diagrams in Figure 11.8 are topologically equivalent to the first, only the first diagram, with terminal sequence intact, corresponds to a tie knot sequence.

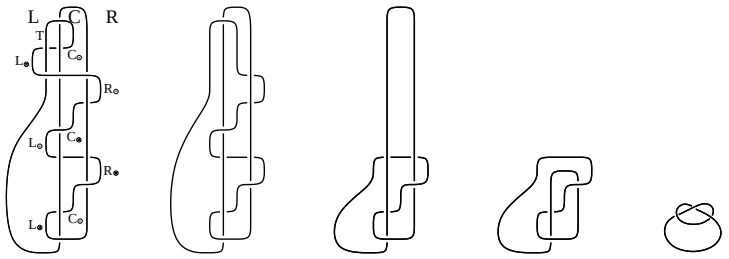

Figure 11.8: Alternative projection of the Windsor knot (left). Changes in sequence associated with reduction of intersections are now apparent.

$$\ldots XCC \quad \rightarrow \quad \ldots X, \tag{11.19}$$

$$\ldots Cf(R,L) \quad \rightarrow \quad \ldots C, \tag{11.20}$$

where X is any move region (R,C,L) and $f(R,L)$ is an alternating sequence of Rs and Ls (of any length). Knots with minimal sequence length of zero may be deformed to the unknot; all other knots have crossing index equal to the minimal sequence length plus one. The crossing index of all aesthetic knots is listed in Table 11.1.

11.10 Conclusion

The ten canonical knot classes $\{h,\gamma\}$ and the corresponding most aesthetic knots are listed in Table 11.1. The four named knots are the only ones, to our knowledge, to have received widespread attention, either published or through tradition (although we have recently learnt that the first entry, $L_\odot R_\otimes C_\odot T$, is extensively used by the communist youth organisation throughout China).

The first four columns describe the knot class $\{h,\gamma\}$, while the remainder relate to the corresponding most aesthetic knot. The centre fraction $\frac{\gamma}{h}$ provides a guide to knot shape, the higher fractions corresponding to broader knots; it, along with the size h, should be used in selecting a knot.

Certain readers will have observed the use of knots whose se-

h	γ	γ/h	$K(h,\gamma)$	s	b	Name	Sequence	KS	SKT
3	1	0.33	1	0	0		$L_\odot R_\otimes C_\odot T$	y	0_1
4	1	0.25	1	-1	1	Four-in-Hand	$L_\otimes R_\odot L_\otimes C_\odot T$	n	3_1
5	2	0.40	2	-1	0	Pratt Knot	$L_\odot C_\otimes R_\odot L_\otimes C_\odot T$	n	0_1
6	2	0.33	4	0	0	Half-Windsor	$L_\otimes R_\odot C_\otimes L_\odot R_\otimes C_\odot T$	y	0_1
7	2	0.29	6	-1	1		$L_\odot R_\otimes L_\odot C_\otimes R_\odot L_\otimes C_\odot T$	n	0_1
7	3	0.43	4	0	1		$L_\odot C_\otimes R_\odot C_\otimes L_\odot R_\otimes C_\odot T$	y	3_1
8	2	0.25	8	0	2		$L_\otimes R_\odot L_\otimes C_\odot R_\otimes L_\odot R_\otimes C_\odot T$	y	7_4
8	3	0.38	12	-1	0	Windsor	$L_\otimes C_\odot R_\otimes L_\odot C_\otimes R_\odot L_\otimes C_\odot T$	n	3_1
9	3	0.33	24	0	0		$L_\odot R_\otimes C_\odot L_\otimes R_\odot C_\otimes L_\odot R_\otimes C_\odot T$	y	4_1
9	4	0.44	8	-1	2		$L_\odot C_\otimes R_\odot C_\otimes L_\odot C_\otimes R_\odot L_\otimes C_\odot T$	n	5_2

Table 11.1: Aesthetic tie knots, characterised, from left, by half-winding number h, centre number γ, centre fraction $\frac{\gamma}{h}$, knots per class $K(h,\gamma)$, symmetry s, balance b, name, sequence, knotted status and standard knot table label. Unnamed knots are hereby introduced by the authors of [32, 33].

quences are equivalent to those shown in Table 1 apart from transpositions of $\hat{\mathbf{r}}, \hat{\mathbf{l}}$ groups, for instance, the use of $L_{\otimes}R_{\odot}C_{\otimes}R_{\odot}L_{\otimes}C_{\odot}T$ in place of the Half-Windsor; some will argue that this *is* the Half-Windsor. Such ambiguity follows from the variable width of conventional ties — the earliest ties were uniformly wide. Since the active end increases in width toward the end, a left move gives greater emphasis to the left than a preceding right move to the right. This makes some transpositions arguably favourable, namely the last $\hat{\mathbf{r}}, \hat{\mathbf{l}}$ group in the knots $\{5,2\}$, $\{6,2\}$, $\{7,2\}$, $\{8,3\}$, $\{9,3\}$ in Table 11.1. We make no attempt to distinguish between these knots and their counterparts; at last we call upon the sartorial discretion of the reader.

11.11 Appendix A: Distribution of End to End Distance of Walks in the Class $\{h, \gamma\}$

We begin by rewriting the evolution equations for the persistent random walk from (11.1) *given* that, for large h, the fraction of steps along the c-axis tends toward $\frac{\gamma}{h}$,

$$\begin{aligned} k_{n+1}(r,c,l \mid \tfrac{\gamma}{h}) &= \tfrac{\gamma}{h-\gamma} p_n(r-1,c,l \mid \tfrac{\gamma}{h}) + \tfrac{h-2\gamma}{h-\gamma} q_n(r-1,c,l \mid \tfrac{\gamma}{h}) \\ p_{n+1}(r,c,l \mid \tfrac{\gamma}{h}) &= \tfrac{1}{2} k_n(r,c-1,l \mid \tfrac{\gamma}{h}) + \tfrac{1}{2} q_n(r,c-1,l \mid \tfrac{\gamma}{h}) \qquad (11.21) \\ q_{n+1}(r,c,l \mid \tfrac{\gamma}{h}) &= \tfrac{h-2\gamma}{h-\gamma} k_n(r,c,l-1 \mid \tfrac{\gamma}{h}) + \tfrac{\gamma}{h-\gamma} p_n(r,c,l-1 \mid \tfrac{\gamma}{h}). \end{aligned}$$

Since we are only interested in the $\hat{\mathbf{r}}$ and $\hat{\mathbf{l}}$ step composition, we project the 2-dimensional walk on to the perpendicular to the c-axis, say, the x-axis, reducing the problem to a symmetric 1-dimensional persistent random walk of $m = h-\gamma$ steps of $\hat{\mathbf{r}}$ and $\hat{\mathbf{l}}$. In this simplified walk, a step to the left is followed with probability $\frac{\gamma}{2h-2\gamma}$ by another step to the left and $\frac{2h-3\gamma}{2h-2\gamma}$ by a step to the right; a step to the right is similarly biased toward the left.

With $x_i = 1$ if the ith step is $\hat{\mathbf{r}}$ and $x_i = -1$ if the ith step is $\hat{\mathbf{l}}$, the resulting evolution equations may be written

$$u_{n+1}(x) = \tfrac{\gamma}{2h-2\gamma} u_n(x-1) + \tfrac{2h-3\gamma}{2h-2\gamma} v_n(x-1) \qquad (11.22)$$

$$v_{n+1}(x) = \tfrac{2h-3\gamma}{2h-2\gamma} u_n(x+1) + \tfrac{\gamma}{2h-2\gamma} v_n(x+1),$$

where $u_n(x)$ is the conditional probability that the walker is at x on the nth step, having just taken a step along the positive x-axis, and v is conditioned on a step along the negative x-axis.

The terminal coordinate of the projected random walk is equivalent to the symmetry s, which is now written

$$s = \sum_{i=1}^{m} x_i. \qquad (11.23)$$

Since the projected walk is a finite-order Markov chain, the central limit theorem provides that the distribution of s approaches a gaussian for large m. Accordingly, we desire the projected walk's mean and variance.

The evolution equations (11.22) are symmetric about 0 apart from the initial step, which is $\hat{\mathbf{l}}$. Observing the possible paths taken in the first few steps of the unprojected walk, it is evident that the first moment $\langle s \rangle$ satisfies

$$\langle s \rangle = \tfrac{\gamma}{h-\gamma}(-1) + \tfrac{h-2\gamma}{h-\gamma}(\tfrac{\gamma}{h-\gamma}(0) + \tfrac{h-2\gamma}{h-\gamma}\langle s \rangle), \qquad (11.24)$$

and, accordingly, the mean μ_s is

$$\mu_s = \langle s \rangle = \tfrac{h-\gamma}{3\gamma-2h}. \qquad (11.25)$$

In what follows, we make use of the local correlation function, $\langle x_i x_{i+k} \rangle$. It may be observed that

$$\langle x_i x_{i+1} \rangle \equiv \langle x_i x_{i+1} \rangle_{i+1} = \tfrac{2\gamma-h}{h-\gamma} x_i x_i = \tfrac{2\gamma-h}{h-\gamma}, \qquad (11.26)$$

where $\langle ... x_{i+1} \rangle_{i+1}$ denotes the average over x_{i+1}. By considering the general average $\langle x_i x_{i+k} \rangle$ as successive averages over x_{i+k}, x_{i+k-1}, *etc.*, we have

$$\langle x_i x_{i+k} \rangle = \tfrac{2\gamma-h}{h-\gamma} \langle x_i x_{i+k-1} \rangle = \ldots = (\tfrac{2\gamma-h}{h-\gamma})^k. \qquad (11.27)$$

The second moment may be expressed in terms of the local correlation function as

$$\langle s^2 \rangle = \sum_{i,j=1}^{m} \langle x_i x_j \rangle. \tag{11.28}$$

Separating the sum into $i = j$ and $i \neq j$ terms, we have

$$\langle s^2 \rangle = m + 2 \sum_{j>i=1}^{m} \langle x_i x_j \rangle = m + 2 \sum_{i=1}^{m-1} \sum_{k=1}^{m-i} \langle x_i x_{i+k} \rangle. \tag{11.29}$$

Substituting in (11.27), it follows that

$$\langle s^2 \rangle \simeq \tfrac{\gamma}{2h-3\gamma} m = \tfrac{\gamma(h-\gamma)}{2h-3\gamma}. \tag{11.30}$$

Since the mean μ_s is always bounded by $[-1, 0]$, we approximate the variance as

$$\sigma_s^2 = \langle s^2 \rangle - \langle s \rangle^2 \simeq \tfrac{\gamma(h-\gamma)}{2h-3\gamma}. \tag{11.31}$$

Equations (11.25) and (11.31) specify the distribution of the terminal coordinate of walks in $\{h, \gamma\}$. For $\gamma = \frac{h}{3}$, the distribution of the terminal coordinates of the 2-dimensional persistent random walk (11.1) readily follows.

Chapter 12

Epilogue

> The works of the Lord are great, sought out of all them that have pleasure therein.
>
> *Psalms 111:2*

Much of the work presented in this dissertation is ongoing. Here we present open questions and possibilities for future consideration, some of which are already under investigation.

Stability of Kinetically Oriented Sequences

We have provided evidence (Chapters 5 - 7) that folding efficiency and thermodynamic stability are anticorrelated near the extremes of either. Notably, it is thought that biological proteins (presumably evolved to be fast folding) are only just stable in their native state conformations.

Our predictions may be verified through simulation by determining the stability of sequences directly trained to fold quickly to a target (Chapter 7); rather than estimating it by means of the energy or relative energy, we may measure stability via direct observation of time spent in the target structure. Biological experiments, such as the effect of stabilising point mutations on folding performance, could provide further insight.

Active Site Capacity

Proteins make use of local chemically sensitive active sites, which may be incorporated into existing design procedures. In accordance with our estimate of conformational capacity (Chapter 5) as a function of

the size of the amino acid alphabet, we may determine the active site capacity, *i.e.*, the maximum fraction of fixed monomers. The active site cannot be so large that the target conformation of the sequence ceases to be a deep stable global minimum.

Alternatively, we may ask how long a protein must be to successfully support an active site of fixed size (number of monomers). A more refined answer will depend on the nearby availability of free monomers (rather than the availability averaged over the entire protein) such that the fragment of protein housing the active site is conformationally stable.

Freedom in Defining Energy Landscape

Our ability to manipulate the conformational energy landscape, both by introducing independent minima and by constructing a broad funnel by way of correlated minima, was shown to be limited in Chapters 5 and 6. In both cases we are bound by the size of the amino acid alphabet as $\frac{\ln A}{\ln \kappa}$. The agreement of both limits is surprising, especially given the different methods of determining them; it suggests a second look at energy landscape constraints.

The information theoretic derivation of the protein capacity $p_{\max}$ [Chapter 5] is straightforward and rigorous. While efforts to apply it to the calculation of the funnel width $g_{\max}$ have yet to prove successful, we suspect information theory may be applicable generally to systems governed by energy landscapes which we wish to manipulate. The central impediment is quantifying the information contained in a specified class of landscape features, such as (in the case of $g_{\max}$) free energy funnels. Model problems may provide further insight.

Chain Length Scaling of Folding Time

Mean first-passage time for lattice proteins has been observed from simulation to scale with chain length as N^{λ}; for random heteropolymers, $\lambda \simeq 6$, while for sequences trained to be thermodynamically stable in the target structure, $\lambda \simeq 4$ [20]. How does folding time scale for optimally folding sequences?

We demonstrated in Chapter 7 that sequences evolved to have

minimal mean first-passage times fold significantly more quickly than thermodynamically oriented sequences. Moreover, the degree of optimality of a selected sequence is designated by the effective temperature, proportional to $\frac{1}{\sqrt{n}}$. By stochastically annealing proteins of different lengths to identical low temperatures, we can estimate the chain length dependence of folding time for kinetically oriented sequences; presumably, $\lambda_{\text{kinetic}} < 4$.

Off-Lattice Protein Folding and Design

Notwithstanding the theoretical and computational insight afforded by lattice models, the design of real proteins requires a realistic folding representation. An off-lattice protein folding simulation has recently been developed by Mehul Khimasia and Robin Ball [14] of the Theory of Condensed Matter group in the Cavendish Laboratory. It is a significant extension of the lattice protein model which takes into account the geometry imposed by the α-carbon bond angles (Figure 12.1). Interactions occur according to continuous isotropic residue-residue potentials rather than the previously assumed nearest neighbour interactions; this may readily be refined to include more realistic (*e.g.*, hydrogen bonding) potentials.

We have begun designing protein sequences intended to fold to specified off-lattice target structures, a task complicated, among other

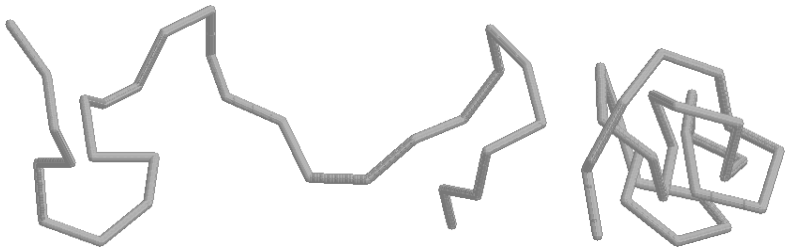

Figure 12.1: Off-lattice protein folding model allows rotation about the α-carbon bonds. Left: Unfolded 30 residue protein. Right: compact folded structure.

things, by the continuous space of available conformations. This space may, for any given sequence, be reduced to a finite set of distinct conformations corresponding to the many local conformational energy landscape minima. Because the set of one sequence will generally not correspond to the set of another, sequences must be trained to conformations which they can at best only approximately fold to. If the target and nearby minimum structural difference is significant, the design attempt may fail, creating instead an effective copolymer landscape (*i.e.*, one without a dominant global minimum).

Applications of Hierarchical Optimisation

The general optimality equation, derived in Chapter 8, can in many instances only be solved approximately, as is the case for probabilistic optimisation problems considered in Chapter 9. Many-stage optimisation problems not amenable to exact solutions require a search over an extensive tree of futures. The ultrametric property of decision policies on this tree requires novel forms of optimisation; simulated annealing on a tree, adapted from conventional simulated annealing, is one such possibility.

Chapters 8 and 10 (hierarchical optimisation problems) were presented at the 1998 Meeting on Statistical Finance in Rome, only the second conference to address the new field of econophysics[1]. Applying analytic (as well as computational) techniques borrowed from physics to reduced models of stocks and markets has obvious practical and surprising theoretical returns. It is hoped that parts of this dissertation will develop in that direction.

[1] Coined by H. Eugene Stanley.

Bibliography

> [Academic] reading, after a certain age, diverts the mind too much from its creative pursuits. Any man who reads too much and uses his own brain too little falls into lazy habits of thinking.
>
> ALBERT EINSTEIN

Inverse Protein Folding

[1] V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, "Impact of Local and Non-Local Interactions on Thermodynamics and Kinetics of Protein Folding," *J. Mol. Biol.* **252**, 460 (1995).

[2] D. J. Amit, *Modeling Brain Function* (Cambridge University Press, Cambridge, UK, 1989), Ch. 6, Ch. 7.

[3] C. B. Anfinsen *et al.*, "The Kinetics of Formation of Native Ribonuclease during Oxidation of the Reduced Polypeptide Chain," *Proc. Natl. Acad. Sci. USA* **47**, 1309 (1961).

[4] J. D. Bryngelson and P. G. Wolynes, "Spin Glasses and the Statistical Mechanics of Protein Folding," *Proc. Nat. Acad. Sci. USA* **84**, 7524 (1987).

[5] Joseph D. Bryngelson *et al.*, "Funnels, Pathways and the Energy Landscape of Protein Folding: A Synthesis," *Proteins: Structure, Function and Genetics*, **21**, 167 (1995).

[6] Thomas E. Creighton, *Protein Folding* (W. H. Freeman and Company, New York, 1992).

[7] Ken A. Dill *et al.*, "Principles of Protein Folding — A Perspective from Simple Exact Models," *Protein Sci.* **4**, 561 (1995).

[8] Ken A. Dill and Sun Chan, "From Levinthal to Pathways to Funnels," *Nature Struct. Biol.* **4**, 10 (1997).

[9] Thomas M. Fink and Robin C. Ball, "Robustness and Efficiency in Inverse Protein Folding," *Physica D* **107**, 199 (1997).

[10] Thomas M. Fink and Robin C. Ball, "Inverse Protein Folding as an Associative Memory," submitted to *Phys. Rev. Lett.* (1997).

[11] Thomas M. Fink and Robin C. Ball, "Kinetically Oriented Sequence Selection," for *Phys. Rev. Lett.* (1998).

[12] P. J. Flory, "The Configuration of Real Polymer Chains," *J. Chem. Phys.* **17**, 303 (1949).

[13] A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich, "Evolution-Like Selection of Fast-Folding Model Proteins," *Proc. Natl. Acad. Sci. USA* **92**, 1282 (1995).

[14] Mehul Khimasia and Robin C. Ball, "Off-Lattice Protein Folding," in preparation (1998).

[15] C. Levinthal, "Are there Pathways for Protein Folding?" *J. Chim. Phys.* **65**, 44 (1968).

[16] S. Miyazawa and R. Jernigan, "Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading" *J. Mol. Biol.* **256**, 623 (1996).

[17] W. J. C. Orr, *Trans. Faraday Soc.* **43**, 12 (1947).

[18] E. I. Shakhnovich and A. M. Gutin, "Engineering of Stable and Fast-Folding Sequences of Model Proteins," *Proc. Natl. Acad. Sci. USA* **90**, 7195 (1993).

[19] E. I. Shakhnovich, "Proteins with Selected Sequences Fold into Unique Native Conformation," *Phys. Rev. Lett.* **72**, 3907 (1994).

[20] E. I. Shakhnovich, "Chain Length Scaling of Protein Folding Time," *Phys. Rev. Lett.* **77**, 5433 (1996).

Hierarchical Optimisation

[21] Robin C. Ball and Thomas M. Fink, "Stochastic Annealing," for *Phys. Rev. Lett.* (1998).

[22] D. J. Bertsimas, *Probabilistic Combinatorial Optimisation Problems*, Ph.D. thesis, MIT (1988).

[23] D. J. Bertsimas, P. Jaillet and A. R. Odoni, "*A Priori* Optimization," *Opns. Res.* **38**, 1019 (1990).

[24] Thomas M. Fink and Robin C. Ball, "Exactly Solvable Hierarchical Optimization Problem Related to Percolation," *Phys. Rev. Lett.* **76**, 2827 (1996).

[25] Thomas M. Fink and Robin C. Ball, "Hierarchical Optimization Problems," in preparation (1997).

[26] R. J. Glauber, "Time-Dependent Statistics of the Ising Model," *J. Math. Phys.* **4**, 294 (1963).

[27] P. Jaillet, *Probabilistic Travelling Salesman Problems*, Ph.D. thesis, MIT (1985).

[28] M. Kubo and H. Kasugai, "Randomized Decision Strategy for the Hierarchical-Optimization Problems," *Opns. Res. Soc. Jap.* **33**, 335 (1990).

[29] N. Metropolis *et al.*, "Equation of State Calculations for Fast Computing Machines," *J. Chem. Phys.* **21**, 1087 (1953).

[30] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor and Francis, London, 1992), p. 60.

Tie Knots

[31] Hardy Amies, *The Englishman's Suit* (Quartet, London, 1994), Ch. 6.

[32] Thomas M. Fink and Yong Mao, "Tie Knots and Random Walks," *Nature*, in press (1998).

[33] Thomas M. Fink and Yong Mao, "A Mathematical Theory of Tie Knots," submitted to *J. Phys. A* (1998).

[34] Sarah Gibbings, *The Tie* (Studio Editions, London, 1990).

[35] Thomas P. Harte and Luke S. G. E. Howard, private communication.

[36] Charles Livingston, *Knot Theory* (Mathematical Association of America, Washington, 1993).

Colophon

The design of this book was inspired by the work of Jan Tschichold, prescribed in *The Form of the Book*. The book was typeset using LaTeX and the `farrfink.sty` style file. The text was set and illuminated in Computer Modern Roman 11pt and figures were drawn with `xfig`, `gnuplot` and *Mathematica*. The book was printed at 600 dpi on 100 gsm wove oyster paper and bound by Cambinders.