Geometrical and mechanical aspects of structure and flexibility in proteins

Hansen, Mikael Sonne; Røgen, Peter; Hansen, Vagn Lundsgaard

Publication date:
2008

Document Version
Early version, also known as pre-print

Citation (APA):
Hansen, M. S., Røgen, P., & Hansen, V. L. (2008). Geometrical and mechanical aspects of structure and flexibility in proteins.

Geometrical and mechanical aspects of structure and flexibility in proteins

Mikael Sonne Hansen

Ph.D. thesis

Department of Mathematics
Technical University of Denmark

Title of Thesis:
Geometrical and mechanical aspects of protein structure and flexibility

Author:
Mikael Sonne Hansen

Supervisors:
Peter Røgen & Vagn Lundsgaard Hansen

Address:
Department of Mathematics
Technical University of Denmark
Matematiktorvet, building 303S
DK-2800 Kgs. Lyngby

E-mail:
`M.S.Hansen@mat.dtu.dk`
`P.Roegen@mat.dtu.dk`
`V.L.Hansen@mat.dtu.dk`

Kgs. Lyngby, September 14, 2007

Mikael Sonne Hansen

Thesis submitted in partial fulfilment of the requirements for the
ph.d.-degree at the Technical University of Denmark

`This document was typeset in` TeX

Resumé (Danish)

Emnerne i denne Ph.D afhandling falder naturligt i tre dele:

I Kapitel 3 introducerer vi glatte kanalflader, givet ved indhyldningen af en 1-parameter familie af kugler. Kanalflader hvor kugleradius varier lineært udgør de fundamentale byggeblokke i de efterfølgende to kapitler.

I Kapitel 4 giver vi en fuldstændig løsning til problemet omkring bestemmelse af den korteste afstand mellem to keglesegmenter i rummet.

Et hovedresultat er formuleringen af den diskrete tube med ikke-uniform radius i Chapter 5, der bygger på resultater fra de to foregående to kapitler. Ved at maksimere skæringsvolumenet mellem tube og en proteinstruktur, får vi en proteintube der afspejler geometriske egenskaber ved proteinet. Efterfølgende undersøges nogle få aspekter af proteintuber.

I Kapitel 6 anvender vi 'normal mode' analyse (NMA), en mekanisk metode der antager et harmonisk energilandskab, til studiet af lav-energetisk proteindynamik. Vi præsenterer den første implementering af et enkelt-parameter potentiale, kvadratisk i dihedrale vinkler (DV), og sammenligner med resultater fra NMA i Kartesiske koordinater.

Først undersøger vi, om nogle få egenvektorer er tilstrækkeligt til at beskrive observerede forskelle mellem to former af det samme protein. Svaret afhænger af den givne bevægelse. I tilfældet *calmodulin* er DV vektorer signifikant bedre end Kartesiske. I to andre tilfælde er ingen af metoderne overbevisende.

Herefter kigger vi på de stereokemiske egenskaber for en proteinstruktur, der deformeres langs en egenvektorretning. Dette gør vi ved at bestemme indholdet af sekundærstruktur, samt den kovalente energi. I begge tilfælde udviser strukturer deformeret langs DV vektorer, betydeligt bedre egenskaber end deres Kartesiske ækvivalent.

Sammen med T. Novotný, J.N. Pedersen, T. Ambjörnsson og R. Metzler har vi benyttet statistisk mekaniske metoder til at studere dynamikken for to bobler i et stykke DNA. Spørgsmål omkring bobbelmøde kan formuleres ved en Fokker-Planck (FP) ligning for det førhen uløste problem omkring 'to ondsindede fodgængere i modsatrettede linære potentialer'.

Sandsynlighedsfordelingen for fodgængerpositioner, opnået ved FP metoden, bekræftes ved sammenligning med resultatet af en løsningen af masterligningen for det oprindelige bobbelproblem. Endelig bestemmes et sæt af biologisk realistiske parametre for hvilke FP tilnærmelsen holder, og vi foreslår en eksperimentel opstilling til studiet af bobbeldynamik.

Summary

This thesis falls naturally in three parts:

In Chapter 3 we introduce smooth canal surfaces, the envelopes of 1-parameter families of spheres. Canal surfaces with a linear variation of sphere radius constitute the basic building blocks in the following two chapters.

In Chapter 4 we provide an exact solution to the problem of finding the shortest distance between two cone segments in 3-space.

A main result is the formulation of a discrete self-avoiding tube of non-uniform radius in Chapter 5, which draws upon the results of the preceding two chapters. By maximizing the intersection volume between the tube and a protein structure, we obtain a protein tube that reflects geometric properties of the protein. We then proceed study a few aspects of protein tubes.

In Chapter 6 we use normal mode analysis (NMA), a mechanical method that assumes a harmonic energy landscape, to study low-energy dynamics of proteins. We report on the first implementation of a single-parameter potential in dihedral angles (DA) coordinates and compare with results from NMA in Cartesian coordinates.

First, we examine if a few normal modes can represent the observed differences between the open and closed conformations of a protein, and if DA present an improvement over Cartesian coordinates. This depends on the motion involved in the change. In the case of *calmodulin*, DA modes perform significantly better than Cartesian modes. In two other cases neither do well.

Second, we study the stereochemistry of a structure under deformations along eigenmodes. This is done by looking at secondary structure content and the bonded energy after deformation. On both accounts structures deformed along DA modes fare significantly better than the Cartesian equivalent.

Together with T. Novotný, J.N. Pedersen, T. Ambjörnsson, and R. Metzler we have investigated two-bubble breathing dynamics in a DNA construct, using statistical mechanical methods. The question of bubble coalescence is mapped to a Fokker-Planck (FP) equation for the previously unsolved problem of two vicious walkers in opposite linear potentials.

The probability distribution of walker positions from the FP approach, is validated by comparison with a solution of the master equation for the initial bubble problem. Finally, we determine a set of biologically reasonable parameters for which the FP approximation holds, and propose an experimental setup to study two-bubble coalescence dynamics.

Acknowledgements

I would like to thank my supervisors Peter Røgen and Vagn Lundsgaard Hansen for support during these last three years. Especially P. Røgen, who has donated endless hours and insights, has my sincere gratitude (had I made a dedication it would have been to him).

During my research stay abroad, I had the pleasure of visiting both Jason Cantarella (and his wife Tammy Coombs) at UGA, who took me in and treated me like family, and Patrice Koehl at UC Davis, who provided the initial idea and guidance-at-a-distance for a substantial part of this thesis.

Also thanks to Anders Nørgård and Jonas Nyvold Pedersen for useful comments on selected chapters, and Bernd Dammann for invaluable computer ”magic”. Finally, I cannot wait to go to Red Rocks!

Mikael Sonne Hansen
Kgs. Lyngby, 2007

Contents

Chapter 1

Introduction

The last couple of decades has seen an explosion in the amount and availability of experimental data on biological systems. The defining example is the complete sequencing of the human genome finished in 2003 [Jasny 01, Collins 03]. With the completely sequenced genomes of over hundred organisms[1], and an explosion in the number of sequences identified as putative genes, the hunting ground for genes has become enormous. This also goes for the number of gene products, namely proteins, ready for study.[2]

It is the structural properties of proteins that determine their function. To learn about function of proteins we must therefore study structure, which is fully encoded in the sequence [Anfinsen 73]. It is the hope, that new information from genome sequences can be combined with knowledge about protein structure, to increase our understanding of protein function.

The developments at the sequence level have been followed by an increase, though at a slower pace, in the amount of structural data on proteins. The structural data are made accessible to the public in the Protein Data Bank [Berman 03].[3] This large amount of structural data has called for automated and/or computationally inexpensive tools to supplement detailed atomistic simulations [Karplus 02] of individual proteins.

In the study of protein structure, geometrical considerations are relevant. The folding of a protein brings previously distant sites into close proximity. In fact, the formation of a loose *shape*, resembling the folded protein, is the rate-limiting step in the folding of small proteins [Lindorff-Larsen 04]. Furthermore, these loose shapes have implications for the organization of the protein structure universe [Koehl 02, Lindorff-Larsen 05]. This under-

[1] `www.genomenewsnetwork.org`

[2] Recently RNA has come into prominence, as a plethora of non-coding (and "protein-like") RNAs have been discovered. They are found the be involved in e.g. gene regulation [Eddy 01] and various types of cancer [Hall 05]. In the so-called 'RNA world' hypothesis of evolution, non-coding RNA is the missing link between the unorganized primordial soup and the complex biomolecular machinery observed today [Joyce 02].

[3] As of September 2007 the PDB comprises about 45,000 resolved protein structures.

standing has lead to an increasing popularity of geometrically based protein structure classification [Røgen 03a, Røgen 03b].

Geometrical considerations are also relevant at the single protein level, e.g. to study protein-solvent interactions [Eisenberg 86, Edelsbrunner 05]. Geometrical methods can also be used to detect cavities, which are known to act as hot spots for protein-ligand binding [Edelsbrunner 98]. Chapter 5, where we construct a shape and volume capturing tube, and subsequently use it to study aspects of protein structure, belongs in this line of work.

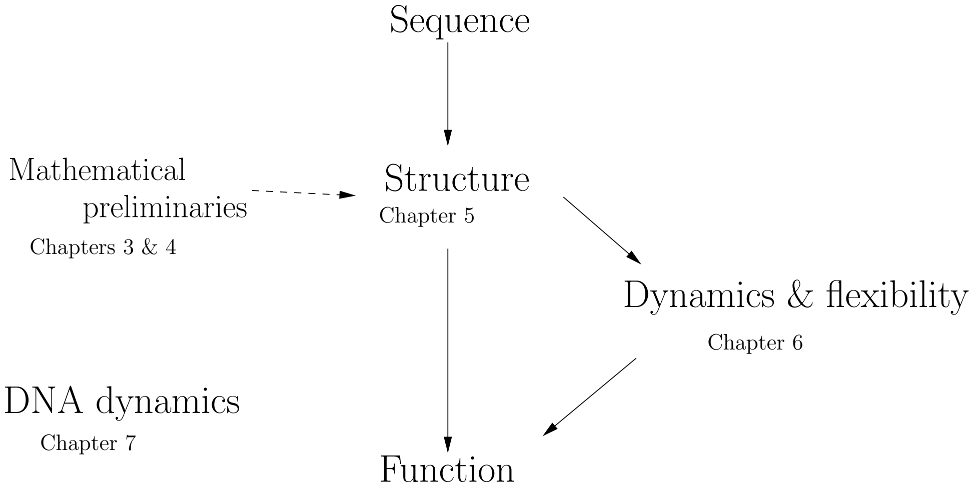


Figure 1.1: Thesis flow-chart

The working protein is not a static structure. There is abundant evidence that flexibility, and larger conformational changes, play a role in diverse functions such as enzyme catalysis [Benkovic 03], protein-ligand binding [Frauenfelder 91], and allosteric regulation [Ma 98]. It is easy to forget this dynamical aspect looking at the textbook drawings of proteins. More worryingly, many applications only use a single (static) representation of the protein structure to model a protein. This can lead to an unwanted bias in the output, as observed by [Fu 07] in the context of protein design. In some cases, the problems related to a single static structure can be alleviated by the use of a simple mechanical spring model to introduce flexibility [Fu 07]. In Chapter 6 we use a similar model to study flexibility and conformational change in a set of proteins.

The present thesis uses classical (differential) geometry and simple mechanical models to probe the role of shape, volume, and flexibility in the world of proteins. The flow-chart in Fig. 1.1 explains the relation between the individual chapters.

Bibliography

[Anfinsen 73] C. Anfinsen. *Principles that Govern the Folding of Protein Chains (Nobel lecture in chemistry 1972).* Science, vol. 181, no. 4096, pages 223–230, 1973.

[Benkovic 03] S.J. Benkovic & S. Hammes-Schiffer. *A Perspective on Enzyme Catalysis.* Science, vol. 301, pages 1196–1202, 2003.

[Berman 03] H.M. Berman, K. Henrick & H. Nakamura. *Announcing the worlwide Protein Data Bank.* Nature Structural Biology, vol. 10, no. 12, page 980, 2003.

[Collins 03] F.S. Collins, E.D. Green, A.E. Guttmacher & M.S. Guyer. *A Vision for the Future of Genomics Research.* Nature, vol. 247, pages 536–540, 2003.

[Eddy 01] S.R. Eddy. *Non-Coding RNA Genes and the Modern RNA World.* Nature Rev.: Genetics, vol. 2, pages 919–929, 2001.

[Edelsbrunner 98] H. Edelsbrunner, J. Liang & C. Woodward. *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.* Prot. Sci., vol. 7, pages 1884–1897, 1998.

[Edelsbrunner 05] H. Edelsbrunner & P. Koehl. *The Geometry of Biomolecular Solvation.* Discrete and Computational Geometry, vol. 52, pages 241–273, 2005.

[Eisenberg 86] D. Eisenberg & A.D. McLachlan. *Solvation energy in protein folding and binding.* Nature, vol. 319, pages 199–203, 1986.

[Frauenfelder 91] H. Frauenfelder, S.G. Sligar & P.G. Wolynes. *The Energy Landscapes and Motions of Proteins.* Science, vol. 254, no. 5038, pages 1598–1603, 1991.

[Fu 07] X. Fu, J.R. Apgar & A.E. Keating. *Modeling Backbone Flexibility to Achieve Sequence Diversity: The Design of Novel α-Helical Ligands for Bcl-x_L.* J. Mol. Biol., vol. 371, pages 1099–1117, 2007.

[Hall 05] P.A Hall & S.E.H. Russell. *New perspectives on neoplasia and the RNA world.* Hematol. Oncol., vol. 23, pages 49–53, 2005.

[Jasny 01] B.R. Jasny & D. Kennedy (editors). *The Human Genome.* Science, vol. 291, pages 1145–1434, 2001.

[Joyce 02] G.F. Joyce. *The antiquity of RNA-based evolution.* Nature, vol. 418, pages 214–221, 2002.

[Karplus 02] M. Karplus & J.A. McCammon. *Molecular dynamics simulations of biomolecules.* Nat. Struct. Bio., vol. 9, no. 9, pages 646–652, 2002.

[Koehl 02] P. Koehl & M. Levitt. *Protein topology and stability define defines the space of allowed sequences.* Proc. Nat. Acad. Sci. USA, vol. 99, pages 1280–1285, 2002.

[Lindorff-Larsen 04] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C.M. Dobson, F.M. Poulsen & M. Vendruscolo. *Transition states for protein folding have native topologies despite high structural variability.* Nat. Struct. Bio., vol. 11, pages 443–449, 2004.

[Lindorff-Larsen 05] K. Lindorff-Larsen, P. Røgen, E. Paci, M. Vendruscolo & C.M. Dobson. *Protein folding and the organization of the protein topology universe.* Trends Biochem. Sci., vol. 30, no. 1, pages 13–19, 2005.

[Ma 98] J. Ma & M. Karplus. *The allosteric mechanism of the chaperonin GroEL: a dynamic analysis.* Proc. Nat. Acad. Sci. USA, vol. 95, pages 8502–8507, 1998.

[Røgen 03a] P. Røgen & H. Bohr. *A new family of global protein shape descriptors.* Mathematical Biosciences, vol. 182, pages 167–181, 2003.

[Røgen 03b] P. Røgen & B. Fain. *Automatic classification of protein structure by using Gauss integrals.* Proc. Natl. Acad. Sci. USA, vol. 100, no. 1, pages 119–124, 2003.

Chapter 2

Selected elements of protein biology

This chapter provides a brief overview on protein structure with an emphasis on the properties most relevant for the coming chapters. We do not touch upon the vast subject of nucleic acids.

An excellent review of the principles behind the shapes of proteins is the introduction to [Koehl 06]. Most of the information in this chapter concerning protein structures can be found there.

All visualization of protein structures have been made with the amazing open-source molecular visualization system `PyMOL`, from DeLano Scientific [DeLano 02].

I Introduction

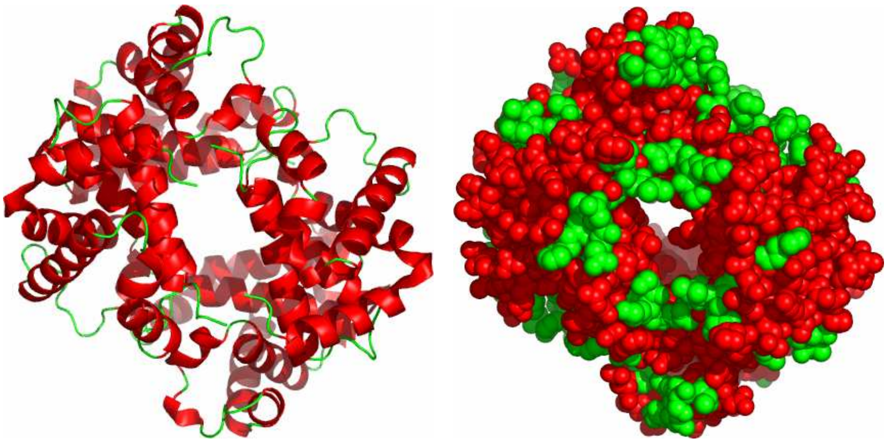
Proteins are involved in almost all cellular functions e.g. catalyzing reactions (enzymes), molecule transport (such as oxygen by hemoglobin), signal transmission (hormones), or channels between the interior and exterior of a cell. They can be divided into three main classes: (i) Membrane proteins are embedded in the lipid environment of a cell membrane. (ii) Fibrous proteins are long rod-like structures often providing structural scaffolding for softer components. (iii) Globular proteins are water soluble and fold into a unique compact 3-dimensional structure under physiological conditions. From hereon we concentrate on globular proteins.

To perform a given function most proteins fold into the a unique 3-dimensional structure known as the *native state*. Incorrect folding can have disastrous consequences as demonstrated by the existence of prions [Prusiner 98] and the proteinaceuos aggregates related to Alzheimer's disease [Bucciantini 02]. It is universally accepted that the native structure only depends on the underlying amino acid sequence - possibly with chaperone proteins guiding the kinetics under and after folding [Ellis 91, Frydman 01].

This is known as the *thermodynamic hypothesis* [Anfinsen 73], and to predict the native structure of a water soluble protein based on the sequence is the still unsolved *protein folding problem* [Honig 99, CASP6 05].

Experimental structure determination and the Protein Data Bank

The first resolved protein structures were myoglobin [Kendrew 58] and hemoglobin [Perutz 60] using X-ray crystallography (see Fig. 2.1). X-ray methods work by the scattering of X-rays off a protein crystal. Growing protein crystals is time-consuming, difficult, and sometimes even impossible [Weber 97]. Furthermore, the crystal environment can lead to artifacts such as the loss of oxygen-binding coorporativity in hemoglobin [Mozzarelli 91]. however, until the advent of nuclear magnetic resonance in the early 80s [Wüthrich 82] X-ray crystallography was the only source of structural information.



(a) Cartoon representation of hemoglobin. This emphasises the spatial organization of secondary structure elements.

(b) All-atom representation of hemoglobin.

Figure 2.1: Two representations of human hemoglobin (`1a3n`*, [Tame 98]) colored by secondary structure: α-helices (red) and random coil (yellow).*

Nuclear magnetic resonance (NMR) looks at spin-spin interactions between nuclei in different parts a protein tumbling in solution. The sample is thus free of the restrictive crystal environment, and much closer to the true *in vivo* environment of the protein. This means that NMR can provide direct information on flexible regions in protein structures. For larger molecules sensitivity and resolution becomes a problem but recently proteins up to 900kDa have been analyzed with NMR techniques [Fiaux 02] and is routinely performed for $\sim$ 25kDa proteins [Wider 00].

Resolved protein structures are made accessible in the Protein Data Bank (PDB) [Berman 03] which currently comprises around $\sim 45,000$ entries corresponding to $\sim 15,000$ different proteins.[1]

II Structural hierarchy in proteins

In our proposed geometrical and mechanical approach to the study of protein structure and flexibility, properties such as

- the number and distribution of atoms in amino acid sidechains,
- the covalent graph in sidechains, and
- the structure (shape) of the native protein,

each play an important role. To learn more about these aspects we give a short survey of the structural hierarchy in protein.

II.1 Amino acids and primary structure

Proteins are a family of heteropolymers build from a set of twenty naturally occurring amino acids. All the amino acids share a common backbone but each have a distinct sidechain that is responsible for the difference in physicochemical and stereochemical properties. Appendix A provides a table of the different sidechains classified according to their interaction with water. Two observations are worthy of notice: (i) There is a large variation in the number of atoms in, and hence the volume of, sidechains. From the single hydrogen atom in *glycine* to the 12 atoms in *argenine*. (ii) In the aromatic amino acids *phenylalanine*, *tryptophane*, *tyrosine*, and *histine* the rings are rigid. In *proline* the ring structure involves atoms from the backbone which makes this amino acid particularly rigid.

Primary structure

During protein synthesis amino acids are chemically linked and form a dipeptide of two amino acid *residues* connected by a rigid peptide bond. This is shown in Fig. 2.2. Continuing this process we obtain a 1-dimensional chain of residues known as the *primary structure* of the protein.

In the living cell proteins do not assemble by condensation but a more complicated process involving several proteins both in the initial transcription of DNA into RNA (by RNA polymerase) and in the final translation of the amino acids into a protein (by the ribosome) [Mathews 99]. This

[1] Two proteins are said to be identical if they exhibit higher than 70% sequence similarity (`www.rcsb.org/pdb/statistics/clusterStatistics.do`).

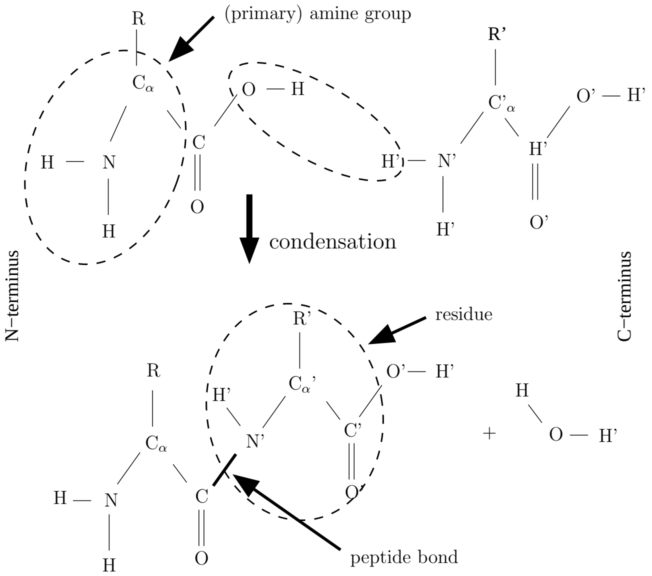


Figure 2.2: Formation of a dipeptide by a condensation process involving two amino acids. R, R' represent one of twenty sidechains. The sequence of atoms $\cdots N \cdot C_\alpha \cdot C \cdot N' \cdot C'_\alpha \cdot C' \cdots$ is referred to as the protein backbone. In the cell protein synthesis does not take place by simple condensation but a more complicated process involving several other proteins (see text).

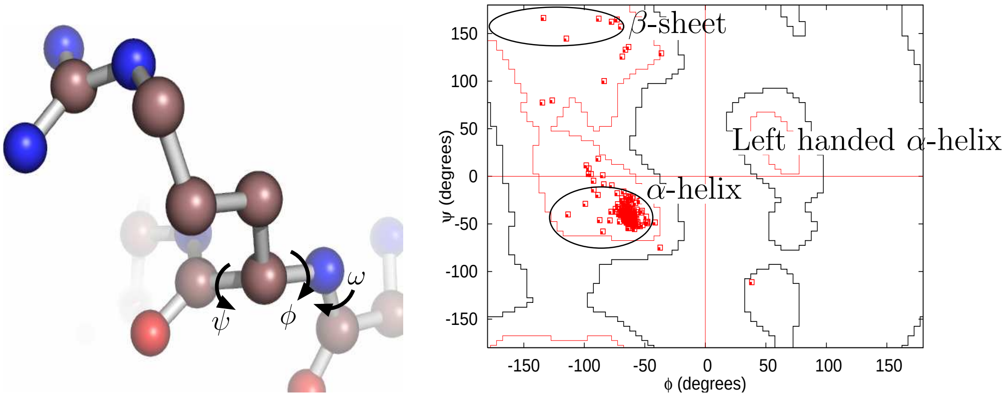
process always runs from the N-terminus to the C-terminus such that both the sequence and structure of a protein comes with an orientation.

Remark 2.1 *The directionality of protein assembly has caused some people to speculate about sequential folding, where the folding takes place during translation, and what implications this could have for the properties of the native structure [Laio 06].*

Remark 2.2 *The locus where the protein backbone joins with a sidechain, the C_α atom (see Fig. 2.2), plays a special role in the protein geometry [Lovell 03]. Often the C_α atoms are used to represent the position of the protein chain and we talk about the C_α-backbone*

II.2 Secondary structure: α-helices, β-sheets and all the rest

The presence of local structural motifs, or *secondary structure*, in proteins were predicted in [Pauling 51a, Pauling 51b] based on theoretical considerations of the protein geometry and hydrogen-bonding patterns. The criteria was to determine local structures that could accommodate all the different amino acids. It was found, that the only classes of such structures are the right-handed α-helices and (anti-parallel/parallel) β-sheets (see Fig. 2.4):



(a) The soft dihedral angles ϕ and ψ define the orientation of a sidechain relative to the backbone. ω is the angle associated to the rigid peptide bond.

(b) Ramachandran plot for 288 residues of hemoglobin (`1bbb`, [Silva 92]). Favored and allowed regions are delimited by stepped red and black lines respectively [Lovell 03].

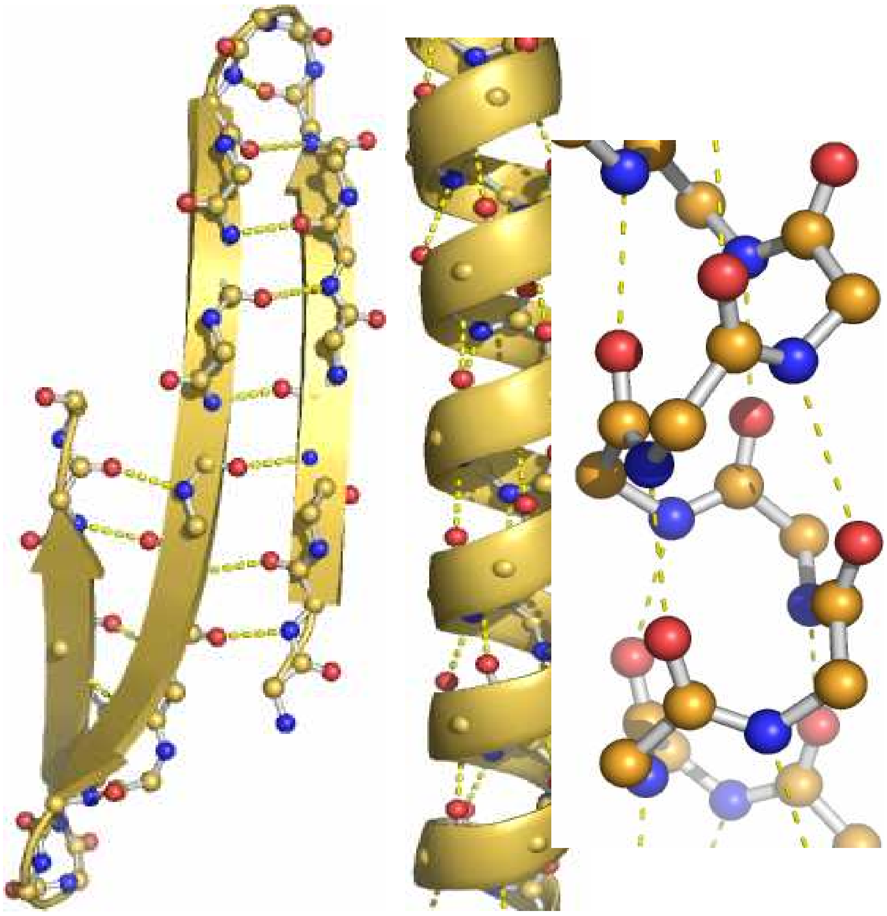
Figure 2.3: The dihedral angles ϕ and ψ of the protein backbone can only assume a very limited set of values due to steric hindrances. This information is conveyed by the Ramachandran plot.

- The stability of a right-handed α-helix is a competition between the energy gained from hydrogen-bonds between atoms in the backbone (aligned with the helix axis) and the decrease in conformational entropy due to steric hindrance of sidechains [Creamer 92, Srinivasan 99].
- When the entropic cost of helix formation rises, the chain can be driven towards the formation of β-sheets. Here steric clashes are less frequent while an ordered conformation, suitable for hydrogen bonding, is retained [Srinivasan 99].

On average proteins contain around 25% helices, 25% sheets, and 50% less regular structures such as loops, turns, and random coil [Brooks 88]. As we see, the two motifs are indeed the universal structures initally looked for [Pauling 51b].

Due to steric hindrances the soft dihedral angles in the protein backbone can only assume distinct values. This is the information contained in the Ramachandran plot [Ramachandran 63] which is shown for the hemoglobin structure in Fig. 2.3(b). The *allowed* and *favored* regions[2] for the angle pairs are almost identical for all residues. The exceptions are glycine, which

[2] Allowed and favored regions for angle values are defined by having 99.95% and 98% of the data points from a set of representative protein structures inside the contours [Lovell 03].



(a) β-sheet formed by hydrogen bonding between β-strands. Neighboring residues have bonds and sidechains that point in opposite directions.

(b) Regular (right-handed) α-helix. The hydrogen bonds are aligned with the helix axis and connect the oxygen of residue i with the polar hydrogen atom bound to the nitrogen of residue $i + 4$.

Figure 2.4: There are two main types of regular secondary structure in proteins both stabilized by hydrogen bonds (hatched lines). Atoms are colored by element: N (blue), C (green) and O (red). Sidechains and hydrogen atoms are omitted for simplicity.

has more freedom due to its single hydrogen atom, and proline, which has less freedom because of the cyclic structure formed with the backbone. In secondary structure elements the concentration of points in the Ramachandran plot is even more pronounced and together with the hydrogen bonding patterns this information can be used to assign secondary structure type to residues [Kabsch 83].

II.3 Tertiary structure and domains

The *tertiary structure* of a protein is the unique compact 3-dimensional structure adopted by an amino acid sequence (see Section I). When we talk about the *structure of a protein* this is often what we have in mind. A tertiary structure consists of one or several *domains* [Rose 79, Richardson 81]. Protein domains are difficult to define but often understood to be: 'a region of a structure that recurs in different contexts in different proteins and/or a compact, spatially distinct unit in protein structure' [Koehl 06].

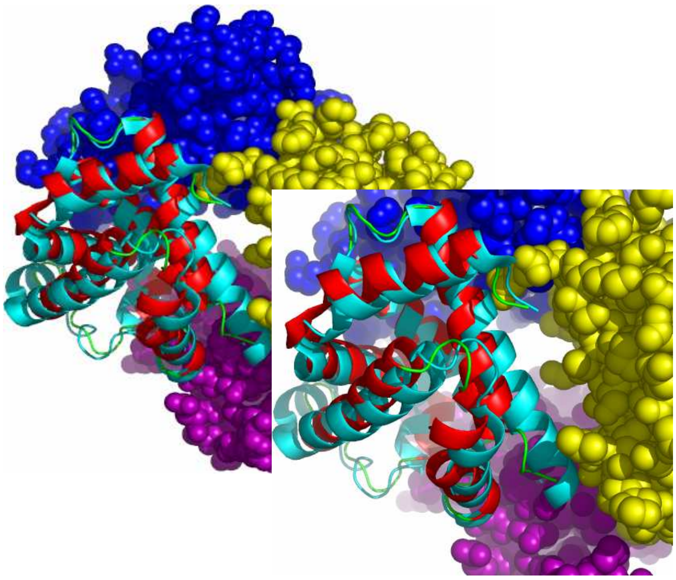


Figure 2.5: Myoglobin structure `2jho` *[Arcovito 07] (cyan) superimposed on one of four similar domains (blue,yellow and red+green) in hemoglobin.*

A concrete example is given by the four oxygen binding domains in hemoglobin. Fig. 2.5 shows an RMSD-minimizing superposition of a myoglobin structure on one of the hemoglobin domains. The structures are seen to be very similar, with an RMSD of 1.5Å. Hemoglobin and myoglobin are homologous proteins so the structural resemblance could have been detected by a sequence comparison. However, many proteins have similar structures but low sequence similarity [Rost 99] and in this case a direct structural comparison of domains is necessary.

The backbones in the hemoglobin domains are not connected and together they form the *quaternary structure*. Other proteins, e.g. calmodulin which we meet in Chapter 6, also has several domains but only a single connected backbone. In this case we do not talk about quaternary structure.

III Protein energetics

The thermodynamic hypothesis says, that the native state of a protein is the global minimum of the free energy

$$G = U - TS, \tag{2.1}$$

where the system includes solvent. Here U is the internal energy, T the temperature, and S the entropy. In going from the unfolded to the native state under physiological conditions, $\Delta G \sim (-5) - (-15)$kcal/mol [Pace 90]. Comparing this with the characteristic thermal energy, $k_B T \approx 0.62$kcal/mol, or the strength of a hydrogen bond, $1 - 5$kcal/mol [Rose 93], we see that the native state of a protein is a marginally stable structure in which opposing energetic factors balance out.

According to the thermodynamic hypothesis we can then fold a protein if we have a sufficiently accurate expression for the free energy G. Hybrid QM/MD methods are available [Liu 01] but solving the Schrödinger equation to obtain the protein wave-function is not possible for a system the size of a protein. Fortunately proteins are large enough ($\sim 1 - 20$nm) to behave as classical objects in most respects.

A typical classical semi-empirical potential used for molecular dynamics simulations is of the form

$$\begin{aligned} U = \quad & K_{\text{bond}} \sum_{\text{bond}} (l - l_0)^2 + K_{\text{ang}} \sum_{\text{angle}} (\phi - \phi_0)^2 + \\ & K_{\text{dih}} \sum_{\text{dihedral}} (1 + \cos(N\theta - \theta_0)) + \sum_{\text{atoms } i<j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{r_{ij}} \right). \end{aligned} \tag{2.2}$$

Here the first three terms represent bonded interactions: covalent bonds, valence angles, and dihedral angles. The last two terms are non-bonded interactions: a standard Lennard-Jones potential representing van der Waals forces and a Coulomb potential for electrostatic interactions (including the all important hydrogen bonds). The force constants $K_{\text{bond}}, K_{\text{ang}}, K_{\text{dih}}$, the equilibrium values l_0, ϕ_0, θ_0, the parameters A_{ij}, B_{ij}, and the atomic charges q, together define a force-field. Different force-fields are then constructed from experimental values on small organic molecules and *ab initio* quantum mechanical calculations [MacKerell Jr. 98, Kaminski 01, Ponder 03].

Finally, the entropic part of the free energy Eq. (2.1) is estimated by sampling the conformational space accessible to the system at temperature T.

Remark 2.3 *For a system of two identical atoms in vacuum the minimum of the Lennard-Jones potential*

$$U_{vdw}(r) = \frac{A}{r^{12}} - \frac{B}{r^6},$$

is used to define the van der Waals radius of an element. Values range from 1Å for hydrogen to 1.95Å for carbon. This is typically the radius used to represent an atom by a sphere in the all-atom representation of a protein (see Fig. 2.1(b)).

The effects of solvent: The hydrophobic effect

Proteins spontaneously fold into their native structure in a watery environment but are typically unfolded in the gas phase. This tells us that water plays a crucial role in the stability of the native state. The explicit inclusion of water in MD simulations cause an explosion in the computational costs, where most of the time is spent updating the configuration of the water molecules. This has lead to a widespread use of implicit solvent models, e.g. using continuum models for the electrostatic interactions (see [Simonson 03] and references therein) and various formulations involving the exposed surface area to account for the hydrophobic effect [Eisenberg 86, Koehl 94, Edelsbrunner 05].

The *hydrophobic effect* refers to the burial of non-polar (hydrophobic) residues in the core of the protein structure, which increases the fraction of polar (hydrophilic) residues on the surface. Part of the effect is entropic: Hydrophobic residues at the surface cause a frustration of the hydrogen bonding network formed by the water molecules. This then decreases the number of configurations accessible to the water [Lum 99]. The same effect is responsible for the organization of lipid-membranes and the formation of oil droplets in water. The hydrophobic effects plays a major role in the

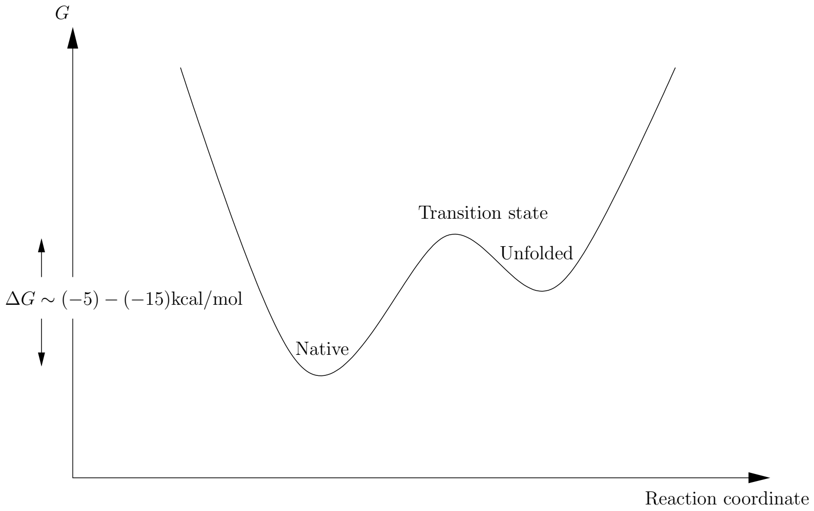


Figure 2.6: Schematic of the protein folding landscape. The rate-limiting step is a set of conformations know as the transition state (ensemble). A more complex landscape involving several rate-limiting could be envisioned.

formation of the hydrophobic core of a protein [Rose 93].

Folding heuristics

Any set of folding heuristics must explain the Levinthal paradox [Levinthal 69, Honig 99]. Namely, how does the protein find a unique native state within the huge conformational space *a priori* available to it?

In going from the unfolded to the native state a protein pass through one (or several) rate-limiting step(s) by a set of conformations known as the transition state (see Fig. 2.6). In the *nucleation-condensation* view of protein folding the transition state involves the formation of native contacts that provide a scaffolding, or folding nucleus, for the last part of the process [Daggett 03]. Here rate limiting step is due to a decrease in entropy by a restriction of the conformational space accessible to a protein [Baldwin 07], that still exhibits imperfect burial of non-polar residues in a hydrophobic core and only partial formation of secondary structures [Lindorff-Larsen 04].[3]

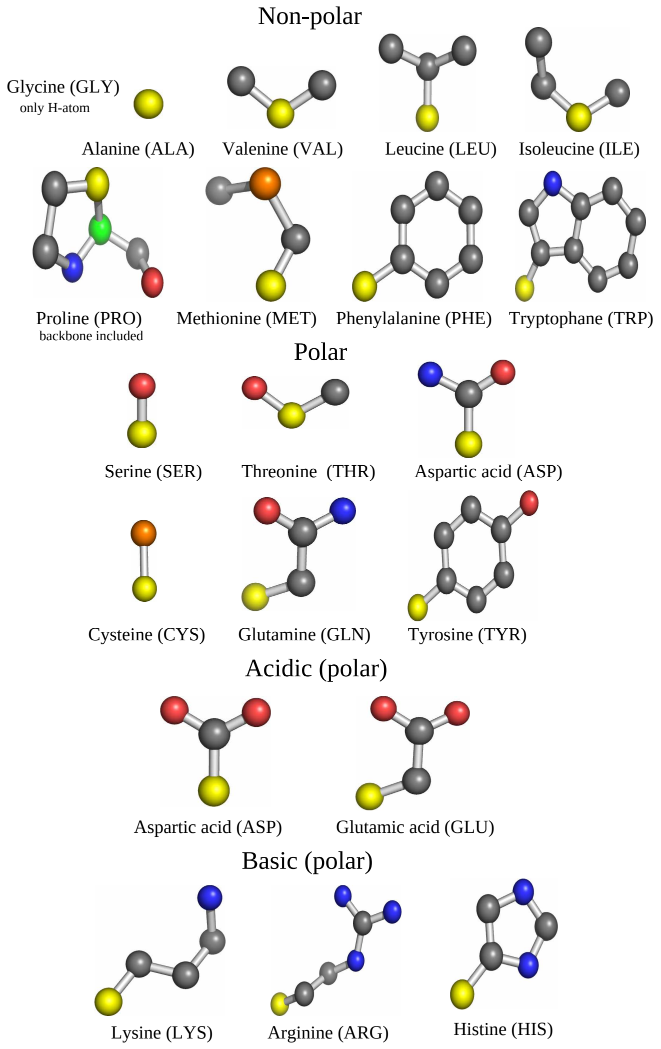
IV Conclusion

With small excerpts from the vast subject of structural protein biology we have hopefully awakened the interest of newcomers, without excessively boring the experts. Our aim has been to provide information and a general feel for the structural aspects most relevant for the following chapters. Energetic considerations only play a minor role later on, but was introduced here to clarify the interactions responsible for the native protein structure.

[3] Two other popular folding heuristics are/have been the framework model [Rose 79, Kim 82] and the hydrophobic collapse model (see [Daggett 03] and references therein).

A Amino acid sidechains

Here the sidechains of the twenty naturally occurring amino are presented and classified according to their interaction with water. It is favorable for polar water molecules to frequent the company of other polar (i.e. hydrophilic) molecules and shun non-polar (*hydrophobic*) molecules. Sidechains with a net charge are either *acidic* (negatively charged) or *basic* (positively charged).



The atoms are colored in the following way: C_α (green), C_β (yellow), other C atoms (gray), oxygen (red), nitrogen (blue), sulphur (orange). The two exceptional sidechains are *glycine*, with only a single hydrogen, and *proline*, where the sidechain form a cyclic structure involving the backbone (which has been included above). Hydrogen atoms are left for simplicity.

Bibliography

[Anfinsen 73] C. Anfinsen. *Principles that Govern the Folding of Protein Chains (Nobel lecture in chemistry 1972).* Science, vol. 181, no. 4096, pages 223–230, 1973.

[Arcovito 07] A. Arcovito, M. Benfatto, M. Cianci, S.S. Hasnain, K. Nienhaus, G.U. Nienhaus, C. Savino, R.W. Strange, B. Vallone & S.D. Longa. *X-Ray Structure Analysis of a Metalloprotein with Enhanced Active-Site Resolution Using in Situ X-Ray Absorption Near Edge Structure Spectroscopy.* Proc. Nat. Acad. Sci. USA, vol. 104, page 6211, 2007.

[Baldwin 07] R.L. Baldwin. *Energetics of Protein Folding.* J. Mol. Biol., vol. 371, pages 283–301, 2007.

[Berman 03] H.M. Berman, K. Henrick & H. Nakamura. *Announcing the worlwide Protein Data Bank.* Nature Structural Biology, vol. 10, no. 12, page 980, 2003.

[Brooks 88] B. Brooks, M. Karplus & M. Pettitt. *Proteins: A theoretical perspective of dynamics, structure and thermodynamics.* Adv. Chem. Phys., vol. 71, pages 1–259, 1988.

[Bucciantini 02] M. Bucciantini, E. Giannoni, F. Chiti, F. Baroni, L. Formigli, J. Zurdo, N. Taddei, G. Ramponi, C.M. Dobson & M. Stefani. *Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases.* Nature, vol. 416, pages 507–510, 2002.

[CASP6 05] CASP6. *Critical Assessment of Techniques for Protein Structure Prediction 6.* Proteins: Struct. Func. Bioinf., vol. 61, no. 7, pages 1–236, 2005.

[Creamer 92] T.P. Creamer & G.D. Rose. *Side-chain entropy opposes a-helix formation but rationalizes experimentally determined helix-forming propensities.* Proc. Nat. Acad. Sci. USA, vol. 89, pages 5937–5941, 1992.

[Daggett 03] V. Daggett & A.R. Fersht. *Is there a unifying mechanism for protein folding?* Trends Biochem. Sci., vol. 28, pages 18–25, 2003.

[DeLano 02] W.L. DeLano. The pymol molecular graphics system. DeLano Scientific, Palo Alto, 2002.

[Edelsbrunner 05] H. Edelsbrunner & P. Koehl. *The Geometry of Biomolecular Solvation.* Discrete and Computational Geometry, vol. 52, pages 241–273, 2005.

[Eisenberg 86] D. Eisenberg & A.D. McLachlan. *Solvation energy in protein folding and binding.* Nature, vol. 319, pages 199–203, 1986.

[Ellis 91] R.J. Ellis & S.M. van der Vies. *Molecular Chaperones.* Ann. Rev. Biochem., vol. 60, pages 321–347, 1991.

[Fiaux 02] J. Fiaux, E.B. Bertelsen, A.L. Horwich & K. Wüthrich. *NMR analysis of a 900K GroEL GroES complex.* Nature, vol. 418, pages 207–211, 2002.

[Frydman 01] F. Frydman. *Folding of Newly Translated Proteins in Vivo: The Role of Molecular Chaperones.* J. Mol. Biol., vol. 70, pages 603–647, 2001.

[Honig 99] B. Honig. *Protein Folding: From the Levinthal Paradox to Structure Prediction.* J. Mol. Biol., vol. 293, pages 283–293, 1999.

[Kabsch 83] W. Kabsch & C. Sander. *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.* Biopolymers, vol. 22, pages 2577–2637, 1983.

[Kaminski 01] G. Kaminski, R.A. Friesner J. Tirado-Rives & W.L. Jorgensen. *Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides.* J. Phys. Chem. B, vol. 105, pages 6474–6487, 2001.

[Kendrew 58] J.C. Kendrew, G. Bodo, H.M. Dintzis, R.G. Parrish, Wyckoff & D.C. Phillips. *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature, vol. 181, pages 662–666, 1958.

[Kim 82] P.S. Kim & R.L. Baldwin. *Intermediates in the Folding Reactions of Small Proteins.* Annu. Rev. Biochem., vol. 59, pages 631–660, 1982.

[Koehl 94] P. Koehl & M. Delarue. *Polar and Nonpolar Atomic Environments in the Protein Core: Implications for Folding and Binding.* Proteins: Struct., Func. and Gen., vol. 20, pages 264–278, 1994.

[Koehl 06] P. Koehl. Protein structure classication, volume 22 of *Reviews in Computational Chemistry*, pages 1–55. Wiley and Sons, 2006.

[Laio 06] A. Laio & C. Micheletti. *Are Structural Biases at Protein Termini a Signature of Vectorial Folding?* Proteins: Struct., Func. and Gen., vol. 62, pages 17–23, 2006.

[Levinthal 69] C. Levinthal. *How To Fold Graciously.* Univ. of Illinois Bulletin, vol. 41, pages 22–24, 1969. Mössbaun Spectroscopy in Biological Systems Proceedings.

[Lindorff-Larsen 04] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C.M. Dobson, F.M. Poulsen & M. Vendruscolo. *Transition states for protein folding have native topologies despite high structural variability.* Nat. Struct. Bio., vol. 11, pages 443–449, 2004.

[Liu 01] H. Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans & W. Yang. *Quantum Mechanics Simulation of Protein Dynamics on Long Timescale.* Proteins: Struct., Func. and Gen., vol. 489, pages 484–489, 2001.

[Lovell 03] S.C. Lovell, I.W. Davis, W.B. Arendall, III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson. *Structure Validation by C_α Geometry: ϕ, ψ, and C_β Deviation.* J. Mol. Biol., vol. 50, pages 437–450, 2003.

[Lum 99] K. Lum, D. Chandler & J.D. Weeks. *Hydrophobicity at Small and Large Length Scales.* J. Phys. Chem., vol. 103, pages 4570–4577, 1999.

[MacKerell Jr. 98] A. D. MacKerell Jr., D. Bashford, M. Bellott, R.L. Dunbrack Jr., J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher III, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin & M. Karplus. *All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins.* J. Phys. Chem. B, vol. 102, pages 3586–3616, 1998.

[Mathews 99] C.K. Mathews, K.E. van Holde & K.G. Ahern. Biochemistry. Prentice Hall, 3 edition, 1999.

[Mozzarelli 91] A. Mozzarelli, C. Rivetti, G. L. Rossi, E.R. Henry & W.A. Eaton. *Crystals of haemoglobin with the T quarternary structure bind oxygen noncooperatively with no Bohr effect.* Nature, vol. 351, pages 416–419, 1991.

[Pace 90] N.C. Pace. *Measuring and increasing protein stability.* Trends in Biotech., vol. 8, pages 93–98, 1990.

[Pauling 51a] L. Pauling & R.B Corey. *Configurations of polypeptide chains with favored orientations of the polypeptide around single bonds: Two pleated sheets.* Proc. Nat. Acad. Sci. USA, vol. 37, pages 729–740, 1951.

[Pauling 51b] L. Pauling, R.B Corey & H.R. Branson. *Two hydrogen-bonded helical configurations of the polypeptide chain.* Proc. Nat. Acad. Sci. USA, vol. 37, pages 205–211, 1951.

[Perutz 60] M.F. Perutz, M.G. Rossmann, A.F. Cullis, H. Muirhead, G. Will & A.C.T. North. *Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-. Resolution, Obtained by X-Ray Analysis.* Nature, vol. 185, pages 416–422, 1960.

[Ponder 03] J.W. Ponder & D.A. Case. *Force fields for protein simulations.* Adv. Prot. Chem., vol. 66, pages 27–85, 2003.

[Prusiner 98] S.B. Prusiner. *Prions (Nobel lecture in Medicine 1997).* Proc. Nat. Acad. Sci. USA, vol. 23, pages 13363–13383, 1998.

[Ramachandran 63] G.N. Ramachandran, C. Ramakrishnan & V. Sasisekharan. *Stereochemistry of polypeptide chain configurations.* J. Mol. Biol., vol. 7, pages 95–99, 1963.

[Richardson 81] J.S. Richardson. *The anatomy and taxonomy of protein structure.* Adv. Prot. Chem., vol. 34, pages 167–339, 1981.

[Rose 79] G.D. Rose. *Hierarchic organisation of domains in globular proteins.* J. Mol. Biol., vol. 234, pages 447–470, 1979.

[Rose 93] G.D. Rose & R. Wolfenden. *Hydrogen bonding, hydrophobicity, packing, and protein folding.* Annu. Rev. Biophys. Biomol. Struct., vol. 22, no. 3, pages 381–415, 1993.

[Rost 99] B. Rost. *Twilight zone of protein sequence alignments.* Prot. Eng., vol. 12, no. 2, pages 85–94, 1999.

[Silva 92] M.M. Silva, P.H. Rogers & A. Arnone. *A third quaternary structure of human hemoglobin A at 1.7-A resolution.* J. Biol. Chem., vol. 267, pages 17248–17256, 1992.

[Simonson 03] T. Simonson. *Electrostatics and dynamics of proteins.* Rep. Prog. Phys., vol. 66, pages 737–787, 2003.

[Srinivasan 99] R. Srinivasan & G.D. Rose. *A physical basis for protein secondary structure.* Proc. Nat. Acad. Sci. USA, vol. 96, no. 25, pages 14258–14263, 1999.

[Tame 98] J. Tame & B. Vallone. *Deoxy human hemoglobin*, 1998.

[Weber 97] P.C. Weber. Overview of protein crystallization methods, volume 276 of *Meth. Enzym.*, pages 13–22. Elsevier, 1997.

[Wider 00] G. Wider. *Structure Determination of Biological MAcromolecules in Solution Using Nuclear Magnetic Resonance Spectroscopy.* BioTechniques, vol. 29, pages 1278–1294, 2000.

[Wüthrich 82] K. Wüthrich, G. Wider, G. Wagner & W. Braun. *Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance.* J. Mol. Biol., vol. 155, pages 311–319, 1982.

Chapter 3

Canal surface with linear radial function

The unifying notion of the next three chapters is *canal surface*, the envelope of 1-parameter family of spheres with centers moving along a smooth curve, first introduced in [Monge 50].[1] Examples are of canal surfaces are: the cylinder, the torus, the dupin cyclides (of which an example is given in Fig. 3.1), and surfaces of revolution in general.

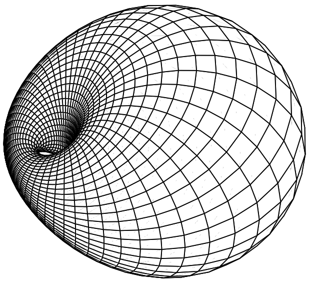


Figure 3.1: An example of Dupin cyclide: a Duping ring

In 4 dimensions canal surfaces have been used in the context of general relativity [Langevin 06] but the most obvious applications are in 3 dimensions, e.g. geometric design [Farouki 96] or distance computations [Tornero 91, Kim 03, Lee 07]. An application of the latter type is presented in Chapter 4 where we give an exact solution to the problem of finding the shortest distance between two spherically generated cone segments. This solution is then used to formulate the constraints on a self-avoiding tube which we use to represent the 3-dimensional structure of a protein. This is done in Chapter 5.

Our reason for including this short introduction to smooth canal surfaces

[1] 1850 that is.

in $\mathbb{R}^3$ is twofold: (i) The smooth theory is more developed than the discrete theory involving polygonal curves. (ii) The results will guide us towards an appropriate formulation of a discrete self-avoiding tube in Chapter 5.

We refer the reader to [do Carmo 76] for an introduction to classical differential geometry.

I Defining canal surfaces

An n-dimensional canal surface, $\mathcal{S}$, is the envelope a 1-parameter family of n-dimensional spheres of radius $R(s)$ with centers on a smooth curve $\boldsymbol{\gamma}(s)$ in $\mathbb{R}^{n+1}$. We restrict ourselves the case $n = 2$.

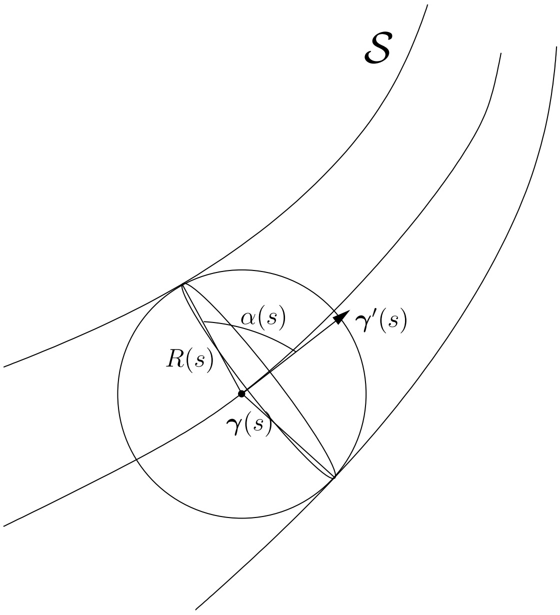


Figure 3.2: A canal surface can be defined as the envelope swept out by of a 1-parameter family of spheres moving along a regular curve $\boldsymbol{\gamma}$.

We consider a curve $\boldsymbol{\gamma}$ and a radial function $R(s)$. Then a variable point $\mathbf{x}$ on a canal surface is defined by the set of implicit equations

$$G(\mathbf{x}, s) = \|\mathbf{x} - \boldsymbol{\gamma}(s)\|^2 - R^2(s) = 0, \tag{3.1a}$$

$$\frac{\partial G}{\partial s}(\mathbf{x}, s) = -2\boldsymbol{\gamma}' \cdot (\mathbf{x} - \boldsymbol{\gamma}) - 2RR' = 0, \tag{3.1b}$$

which has a solution if and only if

$$\frac{|R'(s)|}{\|\boldsymbol{\gamma}'(s)\|} \leq 1. \tag{3.2}$$

Now let $\alpha(s)$ be the angle between the tangent $\boldsymbol{\gamma}'(s)$ and $\mathbf{x}(s)-\boldsymbol{\gamma}(s)$, called the *opening angle* at the point $\boldsymbol{\gamma}(s)$. Then Eq. (3.1b) can be rewritten as

$$\begin{aligned} \boldsymbol{\gamma}' \cdot (\mathbf{x}-\boldsymbol{\gamma}) &= \|\boldsymbol{\gamma}'\|\,\|\mathbf{x}-\boldsymbol{\gamma}\|\cos\alpha = -R\,R', \\ \cos\alpha(s) &= -\frac{R'(s)}{\|\boldsymbol{\gamma}'(s)\|}, \end{aligned} \tag{3.3}$$

where we have also made use of Eq. (3.1a). For each value of s this relation is satisfied by an S^1-family of points on the surface. For the sphere centered at $\boldsymbol{\gamma}(s)$ this circle is found at an angle $\alpha(s)$ to the generating curve. In other words, the envelope of the generating spheres is a 1-parameter family of circles as shown in Fig. 3.2.

Remark 3.1 *When $\boldsymbol{\gamma}$ is arc-length parametrized $\|\boldsymbol{\gamma}'(s)\| = 1$ and Eq. (3.2) reduces to $|R'(s)| \leq 1$.*

Remark 3.2 *When $R(s)$ is constant $\mathcal{S}$ is called a tubular surface. Examples are the helical canal surface and the torus.*

II Regular canal surfaces with linearly varying radial function

We now consider canal surfaces with a radial function varying linearly with arc-length, that is,

$$R(s) = as + b > 0, \quad a, b \in \mathbb{R}_+. \tag{3.4}$$

We call such canal surfaces *cones*. It is then quite natural to examine how much a cone can bend before it starts to self-intersect.

Theorem 3.1 *Let $\boldsymbol{\gamma} : I \to \mathbb{R}^3$ be a smooth arc-length parametrized regular curve with nowhere vanishing curvature κ. If $\boldsymbol{\gamma}$ is the generating curve of a regular cone $\mathcal{S}$ then it has a maximum curvature given by the supremum of*

$$\kappa(s) < \frac{\sqrt{1-(R')^2}}{R(s)}. \tag{3.5}$$

Proof: Since $\boldsymbol{\gamma}$ is regular with non-vanishing curvature there exists a global Frenet frame, $\{\mathbf{t}(s), \mathbf{n}(s), \mathbf{b}(s)\}$, which can be used to explicitly parametrize the cone, $\mathcal{S} : I \times S^1 \to \mathbb{R}^3$, as follows

$$\mathcal{S}(s,\phi) = \boldsymbol{\gamma}(s) + R(s)\left[\cos\alpha(s)\mathbf{t}(s) + \sin\alpha(s)\big(\cos\phi\,\mathbf{n}(s) + \sin\phi\,\mathbf{b}(s)\big)\right], \tag{3.6}$$

where $\cos\alpha(s) = -R'(s)$. The cone is then said to be a regular $\mathcal{C}^1$ surface if there exists a non-vanishing normal vector field for all pairs of values (s, ϕ). That is

$$\frac{\partial \mathcal{S}}{\partial s} \times \frac{\partial \mathcal{S}}{\partial \phi} \neq \mathbf{0}, \quad \text{or equivalently} \quad \|\frac{\partial \mathcal{S}}{\partial s} \times \frac{\partial \mathcal{S}}{\partial \phi}\|^2 \neq 0. \tag{3.7}$$

To determine when $\mathcal{S}$ regular we must see when Eq. (3.7) is satisfied. For cones we have $\frac{\mathrm{d}}{\mathrm{d}s}\cos\alpha(s) = -R''(s) = 0$ and repeated application of the Frenet equations[2] gives

$$\begin{aligned} \frac{\partial \mathcal{S}}{\partial s} &= (1 + R'\cos\alpha - R\kappa\cos\phi\ \sin\alpha)\mathbf{t} \\ &\qquad + (R'\cos\phi\ \sin\alpha - R\tau\sin\phi\ \sin\alpha + R\kappa\cos\alpha)\mathbf{n} \\ &\qquad + (R'\sin\phi\ \sin\alpha + R\tau\cos\phi\ \sin\alpha)\mathbf{b}, \end{aligned} \tag{3.9}$$

$$\frac{\partial \mathcal{S}}{\partial \phi} = -\sin\phi\ \sin\alpha\mathbf{n} + \cos\phi\ \sin\alpha\mathbf{b}, \tag{3.10}$$

where $\kappa = \kappa(s)$ and $\tau = \tau(s)$ is the curvature and the torsion of the generating curve γ respectively. After straightforward calculations using the orthogonality of the Frenet frame together with Eq. (3.3) we have

$$\|\frac{\partial \mathcal{S}}{\partial s} \times \frac{\partial \mathcal{S}}{\partial \phi}\|^2 = R^2(s)\sin^2\alpha(s)\left(\sqrt{1-(R')^2} - \kappa(s)R(s)\cos\phi\right), \tag{3.11}$$

which is nonzero for all values of $\phi \in [0, 2\pi]$, if and only if the curvature κ satisfies Eq. (3.5).

□

The above analysis is only valid when $|R'| < 1$. If $R' \to 1$ then $\kappa \to 0$ and a global Frenet frame is no longer guarantied to exist, the calculations break down, the generating curve tends to a straight line and the surface tends to a half-space. Eq. (3.5) is a special case of a more general result given in [Garcia 06].

In the next section we will see what type of generating curve lies behind the maximally bend cone.

Remark 3.3 *Torsion does not appear in Eq. (3.5) and so questions concerning the generating curve can be settled in the plane.*

[2] The Frenet equations are relations between the vectors in the Frenet frame $\{\mathbf{t}(s), \mathbf{n}(s), \mathbf{b}(s)\}$

$$\mathbf{t}' = \kappa\mathbf{n}, \quad \mathbf{n}' = -\kappa\mathbf{t} - \tau\mathbf{b}, \quad \mathbf{b}' = \tau\mathbf{n}, \tag{3.8}$$

where $\kappa(s) = \|\gamma''(s)\|$ is the curvature and the last equation implicitly defines the torsion $\tau(s)$ ([do Carmo 76],p.19).

Remark 3.4 *When $R(s)$ is constant we have $\kappa(s) < 1/R$. This is the setting of tubular surfaces and thick knots [Stasiak 98] where (locally) the thickness of the tubular surface is bounded from above by the curvature of the generating curve. Recently this has extended to include non-uniform thickness [Durumeric 07].*

II.1 How much can the generating curve of a cone bend?

With an upper bound on the curvature it is natural to ask what *shape* the generating curve of the maximally bend cone has. As mentioned in Remark 3.3 cone regularity does not depend on the torsion of the generating curve so we can restrict our attention to arc-length parametrized *planar* curves.

Theorem 3.2 *Let $\boldsymbol{\gamma} : I \to \mathbb{R}^2$ be a smooth arc-length parametrized regular curve with nowhere vanishing curvature κ. Now assume that the curvature of $\boldsymbol{\gamma}$ at all points is equal to the supremum of the curvatures of the generating curves of regular cones, given in Eq. (3.5). The shape of $\boldsymbol{\gamma}$ is then a logarithmic spiral.*

Proof: We make an educated guess and take the following explicit parametrization of the logarithmic spiral

$$\boldsymbol{\gamma}(s) = R(s)(\cos\beta(s), \sin\beta(s)), \quad R(s) = as + 1 > 0, \tag{3.12}$$

as the candidate curve. Here $\beta(s)$ is the turning angle between tangents to $\boldsymbol{\gamma}$ at 0 and s respectively. As we only look for the shape of $\boldsymbol{\gamma}$, the most general radial function is $R(s)$, equivalent to Eq. (3.4) up to a scale factor.

With $\boldsymbol{\gamma}$ arc-length parametrized we have

$$1 = \|\boldsymbol{\gamma}'(s)\|^2 = a^2 + R^2(s)\beta^2(s), \quad \beta'(s) = \frac{\sqrt{1-a^2}}{R(s)}, \tag{3.13a}$$

which we recognize as the upper bound on the curvature in Eq. (3.5). Then

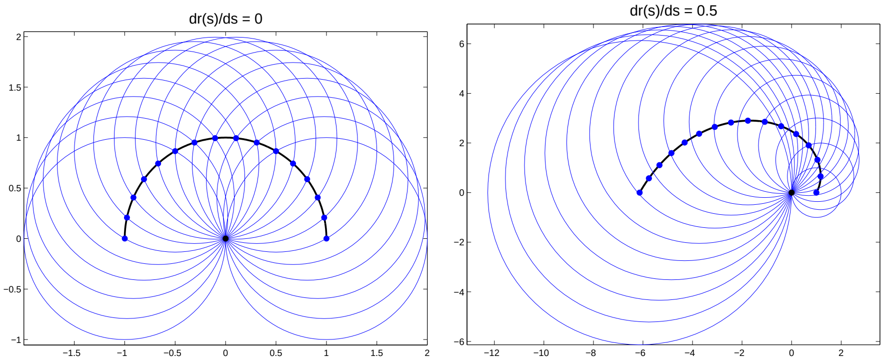
$$\beta(s) = \int_0^s \beta'(s)\mathrm{d}s = \int_0^s \frac{\sqrt{1-a^2}}{as+1}\,\mathrm{d}s = \sqrt{\frac{1}{a^2} - 1}\,\ln R(s), \tag{3.13b}$$

using $\beta(0) = 0$ as initial condition. After repeated differentiation of Eq. (3.12) and using Eq. (3.13) we have

$$\|\boldsymbol{\gamma}''(s)\|^2 = \kappa(s)^2 = \frac{1-\alpha^2}{R^2(s)}, \tag{3.14}$$

which is precisely the supremum of the curvature bound in Eq. (3.5).

□



(a) When the radial function is constant this is a semi-circle...

(b) and otherwise a logarithmic spiral.

Figure 3.3: Cross-sections of maximally bend cones and their generating curve. These are curves with a curvature that at all points equals the supremum of the curvatures of generating curves of regular cones.

The generating curve of a cone, that at all points have a curvature that is the supremum of the curvatures of the generating curves of regular cones, is then a logarithmic spiral and defined by Eqs. (3.12) and (3.13). This reduces to a semi-circle when the radius is uniform [Litherland 99].

Fig. 3.3 shows the cross-section of maximally bend cones for constant and linearly varying radial function. We see that the upper bound on the curvature Eq. (3.5) is a *bona fide* geometrical constraint, and not some artifact of the parametrization in Eq. (3.12). For a maximally bend curve all discs of the cone meet in a focal point at the origin. A larger curvature at *any* point along the curve would cause the discs to contain the origin and intersect along a line.

III Conclusion

We have introduced the notion of a *canal surface*, the envelope of a 1-parameter family of spheres, and elaborated on the properties of this construction relevant for the coming chapters.

Special attention was given to the class of canal surfaces with a radial function that varies linearly with arc-length also called *cones*. This was in anticipation of two later applications, first in Chapter 4 where we give an exact solution to the cone-cone distance problem, and then in Chapter 5 where cones are used in the formulation of a self-avoiding tube.

We found that for a cone to be a regular surface the curvature, κ, must satisfy

$$\kappa(s) < \frac{\sqrt{1-(R')^2}}{R(s)},$$

at each point of the generating curve. We proceeded to demonstrate, that the shape of the curve with a curvature, that at all points is equal to the supremum of the curvatures of the generating curves of regular cones, is a logarithmic spiral.

Bibliography

[do Carmo 76] M.P. do Carmo. Differential geometry of curves and surfaces. Prentice Hall, London, 1976.

[Durumeric 07] O.C. Durumeric. *Nonuniform Thickness and Weighted Distance*, 2007. eprint arXiv:0705.2407.

[Farouki 96] R.A.M.T. Farouki & R. Sverrisson. *Approximation of rolling-ball blends for free-form parametric surfaces.* CAD, vol. 28, no. 11, pages 871–878, 1996.

[Garcia 06] R. Garcia, J. Llibre & J. Sotomayor. *Line of principal curvature on canal surfaces in* $\mathbb{R}^3$. Ann. Braz. Acad. of Sci., vol. 78, no. 3, pages 405–415, 2006.

[Kim 03] K-J. Kim. *Minimum distance between a canal surface and a simple surface.* CAD, vol. 35, pages 871–879, 2003.

[Langevin 06] R. Langevin & G. Solanes. *Conformal geometry of curves and osculating canals.* Lectures notes, Sao Paulo, 2006.

[Lee 07] K. Lee, J-K. Seong, K-J. Kim & S. J. Hong. *Minimum distance between two sphere-swept surfaces.* CAD, vol. 39, pages 452–459, 2007.

[Litherland 99] R.A. Litherland, J. Simon, O. Durumeric & E. Rawdon. *Thickness of knots.* Topology and its Applications, vol. 91, pages 233–244, 1999.

[Monge 50] G. Monge. Application de l'analyse a la géométrie. Bachelier, Paris, fifth edition, 1850.

[Stasiak 98] A. Stasiak, V. Katritch & L.H. Kauffman, editeurs. Ideal knots. World Scientific, London, 1998.

[Tornero 91] J. Tornero, J. Hamlin & R.B. Kelley. *Spherical-Object Representation and Fast Distance Computation for Robotic Applications.* In Int. Conf. Robot. Aut., Sacramento, California, pages 1602–1608. IEEE, April 1991.

Chapter 4

The shortest distance between two cone segments

In this chapter we demonstrate how to compute the shortest distance between a pair of cone segments. This is a concrete application of the canal surfaces introduced in the previous chapter. In the next chapter the distance computation is used in the construction of a self-avoiding non-uniform tube.

The use of parameter families of spheres to represent more complex objects vastly simplifies collision detection and distance computations. It has therefore received some attention in the robotics and computer graphics literature [Tornero 91, Kim 03, Lee 07]. After spheres and cylinders, the simplest objects in this class of surfaces are spherically generated cone segments. It was therefore a surprise that the only attempt to solve the problem of finding the shortest distance between a pair of cone segments in 3-space, presented in [Tornero 91], is incomplete. The principal result in [Tornero 91] is the solution of a set of equations defining a vector along the shortest distance between the segments. Unfortunately, two non-trivial problems arise:

1. There may be no solution.
2. If there is a solution it is not unique. The solution in [Tornero 91] provides the correct distance in exactly half of the cases.

We here give an exact solution, also based on the identification of a vector normal to both cone segments. This approach is known as normal matching in the geometric modeling literature. Normal vector fields are best understood in terms of the Gauss map on a surface, and we used it to provide a clarification of the problems in [Tornero 91].

The use of Gauss maps for distance computations is not new. A similar approach has been used to compute the distance between two surfaces of linear extrusion generated by slope-monotone closed curves [Seong 02]. In [Thomas 00] is provided a numerical distance algorithm for objects described by Bézier patches, also based on normal matching.

In Section I we fix the notation and introduce the distance function. The Gauss map is defined. Section II concerns a closed form solution to a set of equations similar to those of [Tornero 91] supplemented with a binary relation to make a solution unique. Providing a criteria to determine when a solution *does not exist*, we show how the problem in this case reduces to finding the shortest distance between a cone segment and a sphere. Finally we conclude in Section III.

I Notation and a few geometric concepts

A cone segment segment, $\mathcal{S}$, is the envelope of a 1-parameter family of spheres of radius

$$R(s) = (1-s)r_1 + s\,r_2, \text{ centered at } \mathbf{e}(s) = (1-s)\mathbf{v}_1 + s\,\mathbf{v}_2, \tag{4.1}$$

where the vertices $\mathbf{v}_1$ and $\mathbf{v}_2$ are points in $\mathbb{R}^3$, r_1 and r_2 points in $\mathbb{R}_+$, and $s \in [0,1]$. For $\mathcal{S}$ to be well-defined, vertices and radii must satisfy the regularity condition

$$\frac{|r_2 - r_1|}{\|\mathbf{v}_2 - \mathbf{v}_1\|} < 1. \tag{4.2}$$

This was initially introduced in Section I of Chapter 3 where it came from the requirement that the surface should be regular.

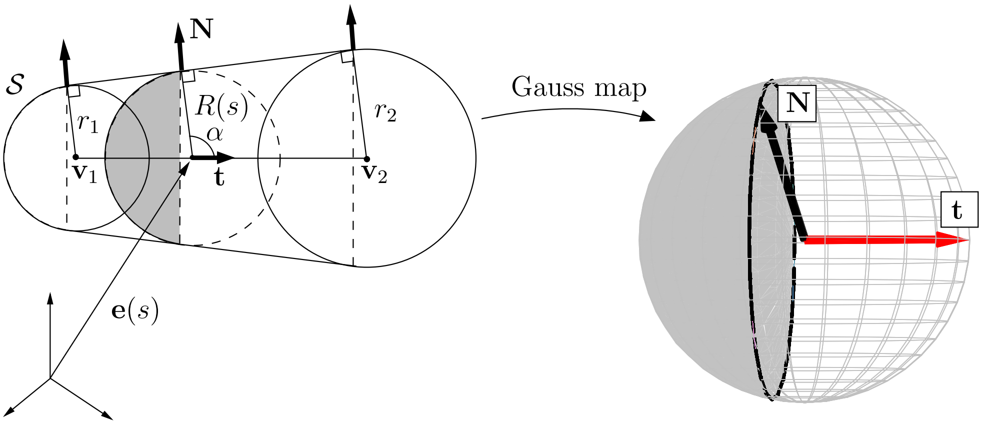


Figure 4.1: **Left:** *A cone segment $\mathcal{S}$ is the envelope of a 1-parameter family of spheres of radius $R(s)$ centered at $\mathbf{e}(s)$. The envelope is a cone frustum sandwiched by parts of a sphere.* **Right:** *The image of the Gauss map* $\mathbf{N}$ *is the set of oriented unit normal vectors of $\mathcal{S}$ seen on the 2-sphere. On the cone frustum normal vectors span a circle dividing the 2-sphere into disjoint regions. On the end-spheres each set of normal vectors cover a disjoint regions.*

An example of a cone segment is given in Fig. 4.1. It consists of a cone frustum sandwiched at each end by a part of a sphere, called an *end-sphere* in the following.[1]

I.1 The Gauss map on a cone segment

Let κ be the curvature of the generating curve of a cone segment $\mathcal{S}$. In Section II of Chapter 3 we saw, that for $\mathcal{S}$ to be a regular $\mathcal{C}^1$-surface the curvature must satisfy

$$\kappa(s) < \frac{\sqrt{1-(R')^2}}{R(s)}. \tag{4.3}$$

For a line segment we have $\kappa \equiv 0$ so a cone segment satisfying Eq. (4.2) and $R(s) > 0$ is always a regular. Furthermore, since $\mathcal{S}$ is orientable there is a differentiable field of unit normal vectors, $\mathbf{N}$, defined on all of $\mathcal{S}$.

Interpreting the unit vector-field on $\mathcal{S}$ as a map from the surface to the 2-sphere, $\mathbf{N} : \mathcal{S} \to S^2$, defines the *Gauss map* [do Carmo 76]. From Fig. 4.1 we see that the Gauss map takes the frustum part to the circle defined by

$$\mathbf{N}\cdot\mathbf{t} = \cos\alpha = -\frac{r_2 - r_1}{\|\mathbf{v}_2 - \mathbf{v}_1\|} \equiv -\eta, \quad \text{where} \quad \mathbf{t} = \frac{\mathbf{v}_2 - \mathbf{v}_1}{\|\mathbf{v}_2 - \mathbf{v}_1\|}, \tag{4.4}$$

and α is the opening angle of the segment. We call this circle-image on the 2-sphere a *Gauss circle*. The image of the Gauss map on the end-spheres each cover one of the two disjoint regions separated by the Gauss circle. Together the three families of normal vectors completely cover the 2-sphere.

I.2 The distance function between two cone segments

To find the shortest distance between two cone segments, a and b, we consider the function $\mathrm{d} : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\mathrm{d}(s,t) = \|\mathbf{e}_a(s) - \mathbf{e}_b(t)\| - R_a(s) - R_b(t), \tag{4.5}$$

that for a pair of (s,t)-values returns the *signed* distance between spheres of radius $R_a(s)$ and $R_b(t)$ centered at $\mathbf{e}_a(s)$ and $\mathbf{e}_b(t)$ respectively. The *shortest* signed distance is then

$$\mathrm{d}_{ab} = \min_{(s,t)\in[0,1]^2} \mathrm{d}(s,t) \equiv \mathrm{d}(s^*,t^*), \tag{4.6}$$

where s^* and t^* are minimizers. In order to determine d_{ab} we use that the shortest distance between two regular surfaces is realized between points where normal vectors are anti-parallel. The distance problem then becomes one of realizing a vector, $\mathbf{N}^c$, normal to both cone segments, or *normal matching* for short.

Remark 4.1 *The definition in Eq. (4.6) differs slightly from that of [Tornero 91] where the distance is set to zero when* $\mathrm{d}_{ab} \le 0$.

[1] A frustum is a truncated cone, with the top cut off parallel to the base.

II Normal vector along the shortest distance

The shortest distance between two cone segments is given by the minimum of the continuous function, d, which maps a compact subset $D = [0,1]^2 \subset \mathbb{R}^2$ into the real numbers. From Analysis we then know that a minimum exists. This can also be from the fact that the Gauss map on a cone segment cover the 2-sphere. This defines a set of rays passing through all points in 3-space (see Fig. 4.1).

The normal matching comes in two distinct flavors. Which one depends on the shortest distance occurring

1. entirely between cone frustums with $(s^*, t^*) \in \overset{\circ}{D} =]0,1[^2$, or
2. between the end-sphere of at least one segment with $(s^*, t^*) \in \partial D$.

As we look for an *anti*-parallel normal vector, the two normal vector fields should have opposite orientation. We choose to look for a matching normal amongst the vectors oriented outward on segment a and inward on segment b.

II.1 Normal matching between cone frustums

We first look for a matching normal over $\overset{\circ}{D}$. This is best done by considering matching normals on the double cones defined by letting the axis parameter s take values in all of $\mathbb{R}$. An example of a double cone is shown in Fig. 4.2.

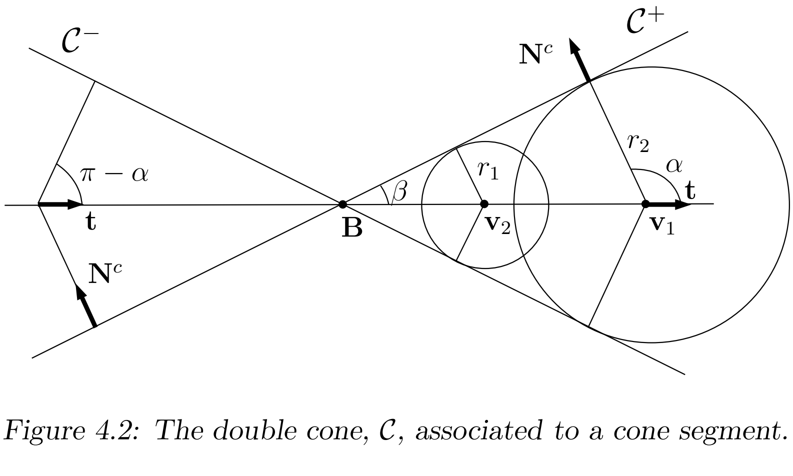


Figure 4.2: The double cone, $\mathcal{C}$, associated to a cone segment.

We denote the double cone associated to segment a by $\mathcal{C}_a$. It is characterized by a cone vertex $\mathbf{B}_a$, an axis vector $\mathbf{t}_a$, and an angle $\beta_a = |\alpha_a - \pi/2|$.[2] The cone that contains segment a is called the *positive cone* and denoted by

[2] Usually the angle β_a is called the *opening angle* of the cone $\mathcal{C}_a^+$ but as the reader may have noticed we reserve this name for α_a defined in Eq. (4.4).

$\mathcal{C}_a^+$. The remaining part of the double cone is the negative cone $\mathcal{C}_a^-$. Similar quantities define the double cone of segment b.

Normal matching between positive cones

By construction we look for a matching normal $\mathbf{N}^c$ between $\mathcal{C}_a^+$ and $\mathcal{C}_b^+$. Using Eq. (4.4) the relevant Gauss circles are then given by

$$\mathbf{N}^c{\cdot}\mathbf{t}_a = \cos\alpha_a = -\eta_a, \qquad \mathbf{N}^c \cdot \mathbf{t}_b = \cos(\pi - \alpha_b) = +\eta_b, \tag{4.7a}$$

$$\mathbf{N}^c{\cdot}\mathbf{N}^c = 1, \tag{4.7b}$$

In Fig. 4.4(a) we see how the choice of orientation of $\mathbf{N}^c$ introduces an asymmetry in the problem. For segment b it is then the Gauss circle of the negative cone $\mathcal{C}_b^-$ we must consider and this is the reason for the sign difference in Eq. (4.7a).

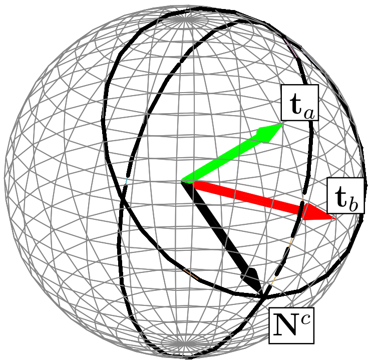


Figure 4.3: Gauss circles of the cones $\mathcal{C}_a^+$ and $\mathcal{C}_b^-$ used to find the matching normal vector $\mathbf{N}^c$ (see text). $\mathbf{t}_a$ and $\mathbf{t}_b$ is the axis vector of segment a and b respectively.

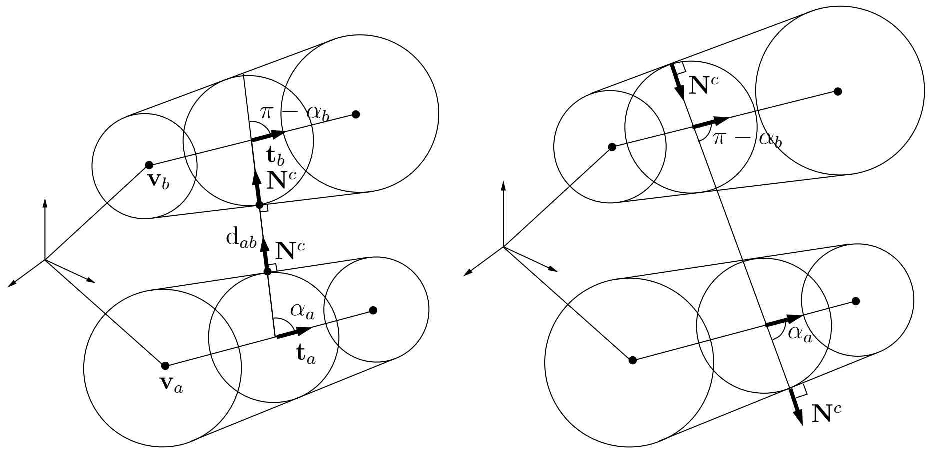
Now assume that a matching normal between the cones exists. In this case there are *two* solutions to the set of equations in Eq. (4.7) which can be seen as two intersections between the Gauss circles in Fig. 4.3 given by Eq. (4.7a). To find the solution realizing the shortest distance we expand $\mathbf{N}^c$ in terms of the non-orthogonal basis

$$(\mathbf{t}_a, \mathbf{t}_b, \delta\mathbf{t}_a \times \mathbf{t}_b), \quad \text{where} \quad \delta = \operatorname{sign}\left[\det\left(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b\right)\right]. \tag{4.8}$$

Again the correct choice of sign is related to the orientation of $\mathbf{N}^c$. Details are provided in Appendix A and Remark 4.3. Now $\mathbf{N}^c$ can be written

$$\mathbf{N}^c = c_1\mathbf{t}_a + c_2\mathbf{t}_b + \delta c_3\mathbf{t}_a \times \mathbf{t}_b, \tag{4.9}$$

where c_1, c_2, and c_3 are scalar coefficients. Inserting into Eq. (4.7) and using



(a) The appropriate choice of sign gives a matching normal vector, $\mathbf{N}^c$, along the direction of the shortest distance.

(b) The wrong choice of sign gives rise to incorrect footpoints.

Figure 4.4: In the generic case the system of equations in (4.7) defining the matching normal vector, $\mathbf{N}^c$, has two solutions. These are distinguished by the choice of sign δ in the basis Eq. (4.8).

$\mathbf{t}_a \cdot \mathbf{t}_b = \cos\theta_{ab}$ gives

$$\begin{aligned} -\eta_a = \mathbf{N}^c \cdot \mathbf{t}_a = c_1 + c_2\cos\theta_{ab}, \qquad \eta_b = \mathbf{N}^c \cdot \mathbf{t}_b = c_1\cos\theta_{ab} + c_2, \\ 1 = \mathbf{N}^c \cdot \mathbf{N}^c = c_1(c_1 + c_2\cos\theta_{ab}) + c_2(c_1\cos\theta_{ab} + c_2) + \delta^2 c_3^2 \sin^2\theta_{ab}. \end{aligned} \tag{4.10}$$

With the choice of sign, δ, in Eq. (4.8) the positive square-root solution for c_3 gives the normal vector realizing the shortest distance and finally we get the coefficients for $\mathbf{N}^c$

$$\begin{aligned} c_1 &= -\frac{\eta_a + \eta_b\cos\theta_{ab}}{\sin^2\theta_{ab}}, \qquad c_2 = \frac{\eta_b + \eta_a\cos\theta_{ab}}{\sin^2\theta_{ab}}, \\ c_3 &= \frac{1}{\sin\theta_{ab}^2}\sqrt{\sin^2\theta_{ab} - (\eta_a^2 + \eta_b^2 + 2\eta_a\eta_b\cos\theta_{ab})}. \end{aligned} \tag{4.11}$$

Remark 4.2 *In [Tornero 91] the coefficients c_1, c_2, and c_3 are written in terms of trigonometric functions. This is avoided in the new formulation of the non-orthogonal basis Eq. (4.8) and makes the appropriate choice of solution to c_3^2 transparent.*

Remark 4.3 *The correct choice of Gauss circles, and hence the correct combination of signs in Eq. (4.7), depends on the orientation of $\mathbf{N}^c$. Situated at segment a the correct choice of sign, δ, will make the vector $\delta(\mathbf{t}_a \times \mathbf{t}_b)$*

point in the direction of segment b in the sense that

$$(\mathbf{v}_b - \mathbf{v}_a) \cdot \delta(\mathbf{t}_a \times \mathbf{t}_b) > 0.$$

Alternatively, situated at segment b the product $-\delta(\mathbf{t}_a \times \mathbf{t}_b)$ points in the direction of segment a. However, in this case the combination of signs in Eq. (4.7) does not give rise to a normal vector along the direction of shortest distance. This is illustrated in Fig. 4.4(b).

No matching normal vector on double cones

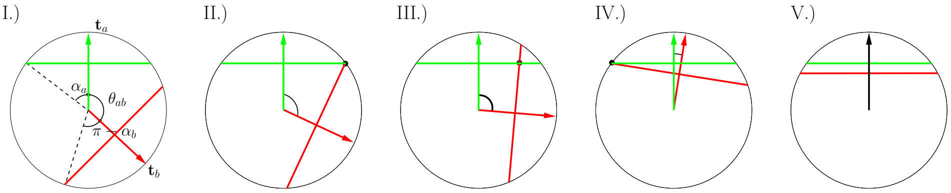


Figure 4.5: The question of existence and number of solutions is determined by the sign of $H(\alpha_a, \alpha_b, \theta_{ab})$ and relates to the number of intersections between Gauss circles (see text). I: No solutions ($H < 0$). II: Degenerate solutions ($H = 0$). III: Two solutions given by $\pm c_3$ ($H > 0$). IV: Degenerate solutions ($H = 0$). V: No solutions ($D < 0$).

We now have an expression for the matching normal normal vector, $\mathbf{N}^c$, *when a solution exists.* Whether or not this is the case can be determined by looking at the sign of

$$H(\alpha_a, \alpha_b, \theta_{ab}) = \sin^2\theta_{ab} - (\eta_a^2 + \eta_b^2 + 2\eta_a\eta_b\cos\theta_{ab}), \tag{4.12}$$

the expression under the square-root in Eq. (4.11). When H is positive there are two non-degenerate solutions, when $H = 0$ the solution is degenerate, and when H is negative there are no solutions . In Fig. 4.5 this information is coupled to the number of intersections between Gauss circles. There are two distinct situations with no solutions

1. There is no matching normal between $\mathcal{C}_a^+$ and $\mathcal{C}_b^+$. However, there is one between $\mathcal{C}_a^+$ and $\mathcal{C}_b^-$, and another between $\mathcal{C}_a^-$ and $\mathcal{C}_b^+$. This is case I in Fig. 4.5.

2. There are no matching normal between *any* pair of cones. This can be further subdivided: (i) $\mathbf{t}_a \cdot \mathbf{t}_b \approx \pm 1$ which is case V in Fig. 4.5. This is also a situation where the basis Eq. (4.8) becomes ill-defined. An assessment of the robustness of the distance algorithm is provided in Appendix B. (ii) $\mathbf{t}_a \cdot \mathbf{t}_b \approx 0$ which is not shown.

In the case there are no matching normal between the cones $\mathcal{C}_a^+$ and $\mathcal{C}_b^+$, the shortest distance necessarily involves at least one end-sphere. We consider this question in Section II.2. First we proceed to find the minimizers s^* and t^* when a matching normal *do* exist.

Footpoints of normal vectors

To determine the minimizers s^* and t^* in Eq. (4.6) we follow [Tornero 91] and consider the relation between points on the cone axes

$$\mathbf{e}_a(s^*) + (\mathrm{d}(s^*, t^*) + R_a(s^*) + R_b(t^*))\mathbf{N}^c = \mathbf{e}_b(t^*). \tag{4.13}$$

Taking the scalar product on both sides with $\mathbf{N}^c \times \mathbf{t}_b$ and $\mathbf{N}^c \times \mathbf{t}_a$, we get

$$s^* = \frac{(\mathbf{v}_b - \mathbf{v}_a)\cdot(\mathbf{N}^c \times \mathbf{t}_b)}{\|\Delta\mathbf{v}_a\|\mathbf{t}_a \cdot (\mathbf{N}^c \times \mathbf{t}_b)}, \qquad t^* = -\frac{(\mathbf{v}_b - \mathbf{v}_a)\cdot(\mathbf{N}^c \times \mathbf{t}_a)}{\|\Delta\mathbf{v}_b\|\mathbf{t}_b \cdot (\mathbf{N}^c \times \mathbf{t}_a)}, \tag{4.14}$$

respectively. If $(s^*, t^*) \in \overset{\circ}{D}$, the distance $\mathrm{d}_{ab} = \mathrm{d}(s^*, t^*)$ is realized between the frustum parts of the cone segments. If not, then it is given by parameters on ∂D, the boundary of D. We address this question in the next section.

Remark 4.4 *Only when the footpoints $\mathbf{e}_a(s^*)$ and $\mathbf{e}_b(t^*)$ lie on the axis of both $\mathcal{C}_a^+$ and $\mathcal{C}_b^+$, is $\mathbf{N}^c$ a vector in the direction of shortest distance. This is because the normal vector field changes orientation when we pass a cone vertex as shown in Fig. 4.2 (see also Remark 4.3).*

II.2 Normal matching involving end-spheres

Now assume there is no matching normal for $(s^*, t^*) \in \overset{\circ}{D}$. This is either because the minimizers lie outside the domain or because no matching normal exists. The shortest distance between cone segments must then involve the end-sphere of *at least* one segment.

If a matching normal does exist the appropriate choice of end-sphere depends on the position of the minimizers relative to $\overset{\circ}{D}$

1. If exactly one of the minimizers is outside $[0, 1]$ the closest end-sphere on the corresponding segment should be chosen.

2. If both minimizers are outside $[0, 1]$ then both, s fixed with $t \in [0, 1]$, and ,t fixed with $s \in [0, 1]$, have to be computed and distances compared.

When no normal vector exists there is nothing to guide us. All four end-spheres must be considered and distances compared.

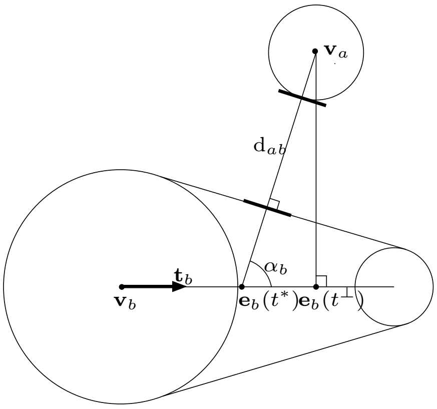


Figure 4.6: When there is no matching normal between cone frustums the problem reduces to finding the shortest distance between a sphere and a cone segment. Here is the situation for $s = 0$ and $\Delta r_b < 0$.

In the sphere-segment distance calculation we follow [Tornero 91] and illustrate it for the case $s = 0$ shown in Fig. 4.6. The minimum distance between a sphere of radius r_a centered at $\mathbf{v}_a$ and segment b is then given by

$$\mathrm{d}_{ab} = \min_{t \in [0,1]} \mathrm{d}(0, t). \tag{4.15}$$

First we find the parameter minimizing the distance between the sphere and the cone axis

$$t^\perp = -\frac{(\mathbf{v}_b - \mathbf{v}_a) \cdot \mathbf{t}_b}{\|\Delta \mathbf{v}_b\|}, \tag{4.16}$$

and then the true minimizer by adding a correction factor

$$t^* = t^\perp + \operatorname{sign}(\eta_b)\Big|t^* - t^\perp\Big| = t^\perp - \frac{\|\mathbf{e}_b(t^\perp) - \mathbf{v}_a\|}{\|\Delta \mathbf{v}_b\| \tan \alpha_b}. \tag{4.17}$$

If $t^* \in [0,1]$, the shortest distance is $\mathrm{d}_{ab} = \mathrm{d}(0, t^*)$. If not, then the shortest distance is between two spheres.

III Discussion

We have provided a complete solution to the problem of finding the shortest distance between two cone segments in 3-space using the method of normal matching. Initially, our intent was to clarify the problems of

1. existence and

2. uniqueness,

encountered in [Tornero 91] but we ended up with a complete reformulation of the solution.

The present chapter bridges the gap between the fairly general considerations on smooth canal surfaces in Chapter 3 and the applications in Chapter 5. Here, a sequence of cone segments is used to represent the 3-dimensional structure of a protein. For this application the ability to do fast and reliable distance calculations between cone segments is crucial in order to preserve the initial shape of the structure. As an added benefit, compared to an approximate distance calculation, the exact form of the solution allows for implementation in a gradient based optimization scheme.

A Details on the choice of sign in Eq. (4.8)

The choice of the sign in the basis (Eq. (4.8) in Section II.1)

$$(\mathbf{t}_a, \mathbf{t}_b, \delta\mathbf{t}_a \times \mathbf{t}_b), \quad \text{where} \quad \delta = \text{sign}\left[\det\left(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b\right)\right], \tag{A.1}$$

is not addressed in [Tornero 91] so when a matching normal exists it is not uniquely defined. Here we briefly explain the appropriate choice in a situation where cone axes do not intersect.

The coordinate system $(\mathbf{t}_a, \mathbf{t}_b, \mathbf{t}_a \times \mathbf{t}_b)$, shown in Fig. A.1, divides 3-space in halves that are distinguished by the sign of $\det(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b)$. One half-space contains the axis of segment b, the other not. Having the matching normal point from segment a to segment b we should use the half-space defined by

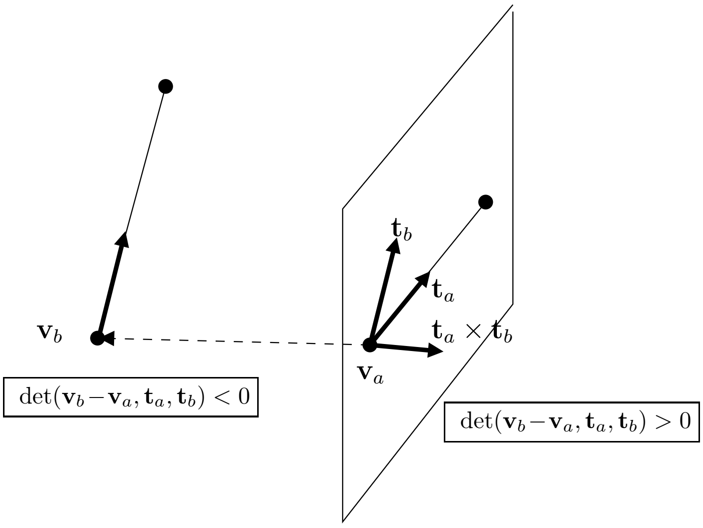

$\mathbf{v}_b$
$\mathbf{t}_b$
$\mathbf{t}_a$
$\mathbf{t}_a \times \mathbf{t}_b$
$\mathbf{v}_a$
$\det(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b) < 0$
$\det(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b) > 0$

Figure A.1: The basis, $(\mathbf{t}_a, \mathbf{t}_b, \delta\mathbf{t}_a \times \mathbf{t}_b)$, divides 3-space into halves that are distinguished by the sign δ.

$$\delta \det(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b) = \delta(\mathbf{v}_b - \mathbf{v}_a) \cdot (\mathbf{t}_a \times \mathbf{t}_b) > 0, \tag{A.2}$$

and hence the coordinate system $(\mathbf{t}_a, \mathbf{t}_b, \delta\mathbf{t}_a \times \mathbf{t}_b)$ with

$$\delta = \text{sign}[\det(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b)]. \tag{A.3}$$

B Numerical robustness

To assess the numerical robustness of the distance algorithm we consider the situations where the non-orthogonal basis

$$(\mathbf{t}_a, \mathbf{t}_b, \delta\mathbf{t}_a \times \mathbf{t}_b), \quad \text{with} \quad \delta = \text{sign}\left[\det\left(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b\right)\right], \tag{B.1}$$

becomes ill-defined. This is the case when one or more of the following situations is approximately realized

1. $\mathbf{v}_a = \mathbf{v}_b$. Identical cone vertices.
2. $\mathbf{t}_a \parallel \mathbf{t}_b$. Parallel cone axes.
3. $(\mathbf{v}_b - \mathbf{v}_a) \parallel \mathbf{t}_a$ or $(\mathbf{v}_b - \mathbf{v}_a) \parallel \mathbf{t}_b$. One cone vertex lies on the axis of the other cone segment.
4. $(\mathbf{v}_b - \mathbf{v}_a) \perp \mathbf{t}_a \times \mathbf{t}_b$. Both cone axes lie in a single plane.

Case 1-3 are all special instances case 4, shown in Fig. 4.1(a), where cone axes lie in a single plane.

The signed distance between a cone segment and an end-sphere is always well-defined. The question of robustness can therefore be stated in the following way: Assume that a matching normal exists between two cone segments. In the situation where the coordinate system becomes ill-defined, is there a significant error in doing the distance calculation at an end-sphere?

The situation that cone axes approximately lie in a single plane severely limits the configurations where a matching normal exists. In fact, as illustrated in Fig. 4.1(b), these are precisely the configurations where all vectors, normal to one segment and intersecting the other cone axis, give rise to almost identical distances. The error in doing the distance calculation at an end-sphere, when the basis Eq. (B.1) becomes ill-defined, is therefore small. In this very loose sense the distance algorithm is robust.

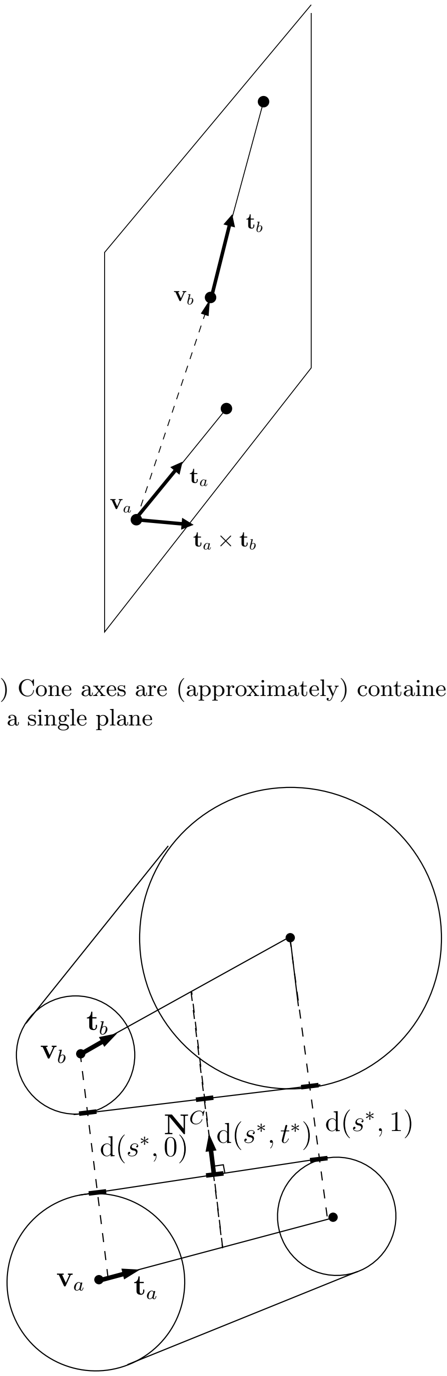
Remark 4.5 *A better measure of the robustness would be, for a given bound*

$$\det\left(\mathbf{v}_b - \mathbf{v}_a, \mathbf{t}_a, \mathbf{t}_b\right) < \varepsilon, \tag{B.2}$$

to have an estimate of the error

$$\left| \mathrm{d}(s^*, t^*) - \mathrm{d}(s,t)|_{(s,t)\in\partial D} \right| < f(\varepsilon), \tag{B.3}$$

where $f \in \mathbb{R}_+$ is some function of ε with the property that $f \to 0$ for $\varepsilon \to 0$. In other words, we should be able to give an upper bound on the error involved in doing the distance calculation at an end-sphere when Eq. (B.1) becomes ill-defined.



(a) Cone axes are (approximately) contained in a single plane

(b) The difference between the true distance, given by a matching normal $\mathbf{N}^c$, and the minimum distance involving end-spheres is small in this case.

Figure B.1: When cone axes approximately lie in single plane the sign δ, and hence the basis $(\mathbf{t}_a, \mathbf{t}_b, \delta\mathbf{t}_a \times \mathbf{t}_b)$, becomes ill-defined.

Bibliography

[do Carmo 76] M.P. do Carmo. Differential geometry of curves and surfaces. Prentice Hall, London, 1976.

[Kim 03] K-J. Kim. *Minimum distance between a canal surface and a simple surface.* CAD, vol. 35, pages 871–879, 2003.

[Lee 07] K. Lee, J-K. Seong, K-J. Kim & S. J. Hong. *Minimum distance between two sphere-swept surfaces.* CAD, vol. 39, pages 452–459, 2007.

[Seong 02] J-K. Seong, M-S. Kim & K. Sugihara. *The Minkowski Sum of Two Simple Surfaces Generated by Slope-Monotone Closed Curves.* In Geometric Modeling and Processing - Theory and Applications (GMP'02), page 33, 2002.

[Thomas 00] F. Thomas, C. Turnbull, L. Ros & S. Cameron. *Computing signed Distances between Free-Form Objects.* In Int. Conf. Rob. Aut., San Francisco, California, pages 3713–3718. IEEE, April 2000.

[Tornero 91] J. Tornero, J. Hamlin & R.B. Kelley. *Spherical-Object Representation and Fast Distance Computation for Robotic Applications.* In Int. Conf. Robot. Aut., Sacramento, California, pages 1602–1608. IEEE, April 1991.

Chapter 5

Optimal tube representations of protein structures

The aim of this chapter is twofold:

1. First, to present a geometrical formulation of a discrete self-avoiding non-uniform tube.

2. Second, to assign self-avoiding tubes to proteins as a geometrical way to represent the structure.

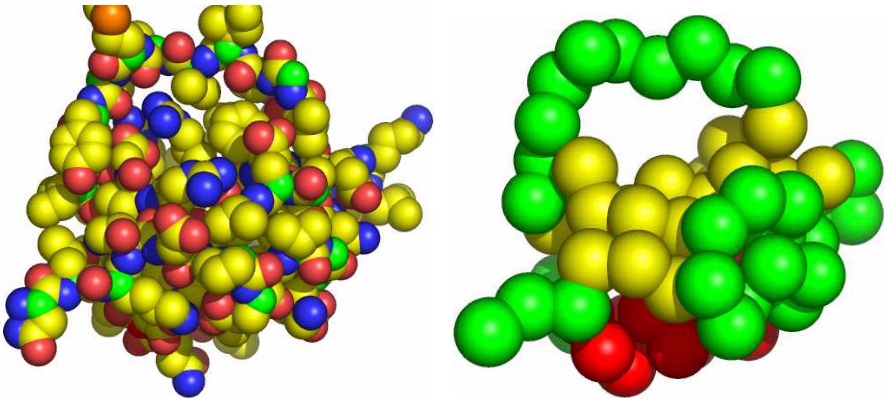
This is a departure from the simplest model where a single radius is assigned to the structure [Banavar 00]. Our intent is to study the geometrical basis of protein structure with a more detailed representation of shape and volume distribution. In this context, it is a large conceptual difference, compared with a uniform tube, that some of the geometrical information contained in the sidechains is retained in the model.

In Section I we briefly look at some successful applications of geometry in a protein context, and comment on the previous work using tube representations [Banavar 00, Banavar 02, Banavar 03a, Banavar 03b, Hoang 04]. Together with some necessary notation we present the constraints defining a self-avoiding non-uniform tube in Section II. In Section III we show how a protein tube can be found as the solution to a volume optimization problem. In Section IV we present some preliminary results. Finally we conclude and look to possible future work in Section V.

So far we have only considered protein structures but the model applies equally well to other biomolecules, e.g. RNA [Berman 92].

I Introduction

The realization that the *shape* of a biomolecule plays a large role - in some situations even larger than the chemical details [Anfinsen 73] - has lead a part of the structural biology community to take a more geometrical approach to the study of biomolecules, and in particular globular proteins [Taylor 01, Koehl 02, Kolodny 05, Edelsbrunner 05, Lindorff-Larsen 05]. The underlying philosophy is, that properties such as *shape*, *volume* and *mass distribution*, *surface area*, *mass distribution*,..., determine a significant part of the experimentally observed behavior. Their role must therefore be understood to fully appreciate (and model) the working protein. Fruits of this approach are, e.g., the use of the accessible surface area to describe the solvation energy implicitly [Eisenberg 86, Edelsbrunner 05], the application of topological [Arteca 99] and geometrical [Røgen 03] shape descriptors in protein structure classification, or the observed correlation between the folding rate and contact order in single-domain proteins [Plaxco 98].



(a) All-atom representation. Atoms colored by element.

(b) Representation using spheres of radius 2.7Å centered at the positions of the C_α-atoms. Atoms colored by secondary structure: helix (red) coil (yellow).

Figure 5.1: Two different representations of a protein (`2ci2`*, [McPhalen 87]). The two representations are identically oriented.*

The geometrical basis of protein structure and function goes hand in hand with the need to perform some form of coarse-graining in any kind of modelling scheme. The native protein structure is only marginally stable, the product of a careful balancing of a huge number of small electrostatic and entropic contributions (see Section 2 of Chapter III). For this reason, brute force approaches such as Molecular Dynamics (MD) are computationally expensive [Karplus 02] and typically some coarse-graining of both structure and force-field is used [Kolinski 04]. The C_α-representation of a protein, shown in Fig. 5.1(b), is a widely used structural model because it is simple and captures the shape of the backbone.

Our use of tubes to represent protein structure is partly inspired by atomic density plots of ensembles of NMR structures as found in [Lindorff-Larsen 04, Lindorff-Larsen 05] and partly by the work in [Banavar 00, Banavar 02] where a uniform tube is used to represent the structure of a protein. We go a step further, to a tube of non-uniform radius. Here the local radius can be used to retain geometrical information concerning sidechains. This allows for a more detailed representation of the protein geometry, with only a small increase in the number of variables.

Previous work using a tube representation

In [Banavar 00] uniform tubes were demonstrated to fold into compact secondary structure-like elements. The underlying hypothesis is, that part of the folding of proteins is only directed by geometrical considerations. Based on a set of helices and sheets it was found that a protein can approximately be described by a uniform tube of radius 2.7Å [Banavar 02].

In [Banavar 03b] the authors argue, that an important aspect of the tube, and not present in the C_α-representation, is the implicit local distinction between the direction *along* the tube and those perpendicular to it. However, it is not obvious that the model of [Gonzalez 99] applied to proteins in [Banavar 00, Banavar 02] can be called a tube. The reason is the following. A radius is imposed on a set of vertices, $\{C_\alpha\}$, by requiring that R_{ij}, the radius of the circle going through $C_{\alpha,i}, C_{\alpha,i+1}$, and $C_{\alpha,j}$ must satisfy[1]

$$R_{ij} < R, \tag{5.1}$$

where R is the ideal radius [Banavar 03a]. The simplest possible configuration, that of two straight backbone segments lying in parallel, is shown in Fig. 5.2. We see that the *actual* backbone distance can vary significantly, without changing R_{ij}. Other configurations allow for more complicated motion keeping R_{ij} fixed but in any case, the minimum distance between points on the backbone is not uniform.

In the next section we present a formulation of a self-avoiding tube of non-uniform radius that faithfully preserves the ideal radius along a segment. Though different from the model in [Gonzalez 99] it still originates from the theory of ideal knots [Stasiak 98, Rawdon 00].

Remark 5.1 *When $\|C_{\alpha,i} - C_{\alpha,i+i}\| \to 0$ then $2R_{min} \to 2R$ (see Fig. 5.2). This is the case in the ideal knot application for which the model was initially intended [Gonzalez 99].*

[1] Initially the radius check involved all triplets of points [Banavar 03a] but it can be shown that it is sufficient to consider triplets where two points are neighbors [Røgen 07]. This reduces the number of intersection checks from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ for a tube with N vertices.

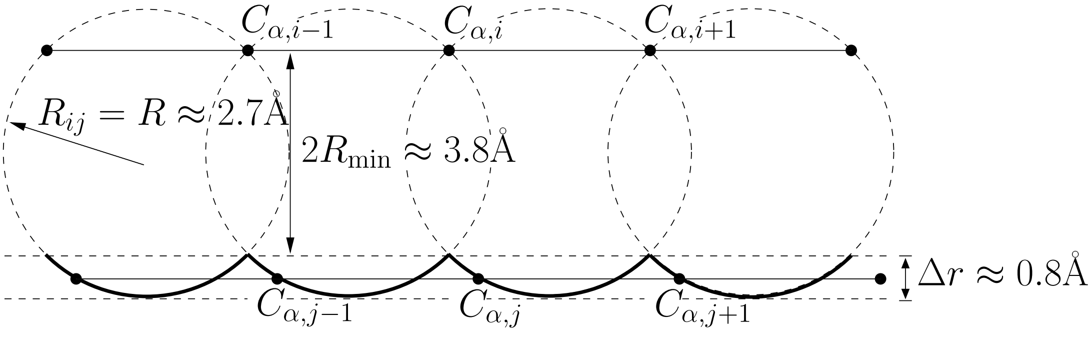

$C_{\alpha,i-1}$
$C_{\alpha,i}$
$C_{\alpha,i+1}$
$R_{ij} = R \approx 2.7\text{Å}$
$2R_{\min} \approx 3.8\text{Å}$
$\Delta r \approx 0.8\text{Å}$
$C_{\alpha,j-1}$
$C_{\alpha,j}$
$C_{\alpha,j+1}$

Figure 5.2: The simplest configuration of two straight backbone segments in the tube model used in [Banavar 00, Banavar 02, Banavar 03a]. Segment j can move along the thick line without changing R_{ij} and violating the ideal radius R. However, the actual distance between the backbone segments varies significantly.

II The self-avoiding non-uniform tube

We now turn to the mathematical formulation of a self-avoiding tube build from a collection of tube segments. Traces of inspiration drawn from the computational studies of ideal knots [Stasiak 98] are scattered throughout the section.

II.1 Defining the tube building blocks

An $(N-1)$-segment tube is given by $(3+1)N$ coordinates

$$(\mathbf{v}_1, r_1, \ldots, \mathbf{v}_i, r_i, \ldots, \mathbf{v}_N, r_N) \equiv (\mathbf{v}, \mathbf{r}), \tag{5.2}$$

where $\mathbf{v}_i$ is a 3-dimensional vector containing the spatial coordinates of vertex i and r_i is the associated radius. Each pair of consecutive points gives rise to a 1-parameter family of spheres of radius

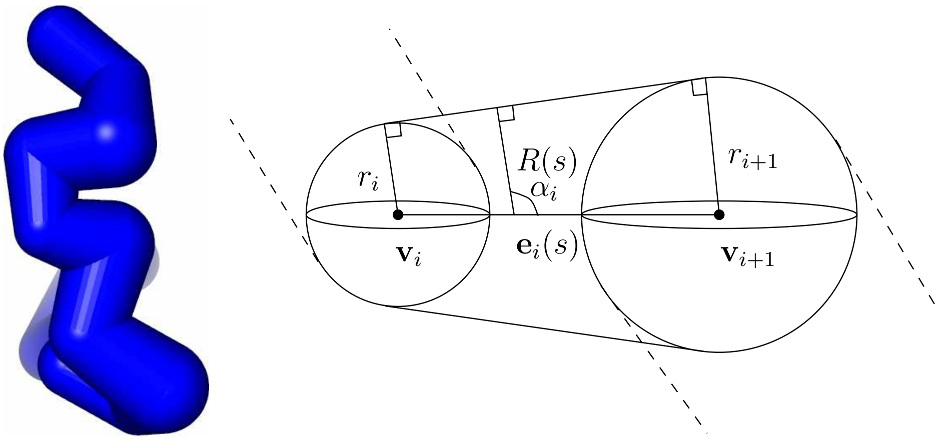
$$R_i(s) = (1-s)r_i + sr_{i+1}, \text{ centered at } \mathbf{e}_i(s) = (1-s)\mathbf{v}_i + s\mathbf{v}_{i+1}, \tag{5.3}$$

where $s \in [0,1]$. Furthermore the *opening angle* α_i is given by

$$\cos\alpha_i = -\frac{r_{i+1} - r_i}{\|\mathbf{v}_{i+1} - \mathbf{v}_i\|} \equiv -\frac{\Delta r_i}{\|\Delta \mathbf{v}_i\|}. \tag{5.4}$$

All together this defines a *tube segment*[2]. Spheres at vertices are called *end-spheres*. By a *tube* we then understand a sequence of tube segments and refer to the polygonal curve given by the line segments as the *tube axis*. An illustration of a tube and a tube segment is given in Fig. 5.3. The special

[2] The same object was referred to as a *cone* segment in Chapter 4.



(a) A non-uniform tube.

(b) Tube segment i in a non-uniform tube. $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ are vertex coordinates and an associated radius r_i and r_{i+1} respectively. α_i is the opening angle and $\mathbf{e}_i(s)$ a point along the tube axis with radius $R_i(s)$.

Figure 5.3: Example of a tube and a tube segment.

case $r_1 = r_2 = \cdots = r_N$ is identical to the tube defined in [Rawdon 00] which we call a *uniform tube* (UT) to distinguish it from the general *non-uniform tube* (NUT).

To be well-defined a segment must satisfy the constraint

$$\mathcal{C}^{\text{bound}}(\mathbf{v}, \mathbf{r}) = 1 - \frac{|\Delta r_i|}{\|\Delta \mathbf{v}_i\|} \geq 0, \qquad i = 1, \ldots, N-1, \tag{5.5}$$

where the limiting case $|\Delta r_i|/\|\Delta \mathbf{v}_i\| = 1$ is a configuration where the smaller end-sphere is completely contained in the larger.[3]

II.2 Stiffness: A local angle constraint

In an infinitely flexible rope the finite thickness imposes a stiffness limiting how much the rope can bend without deformation. This is a property of any polymer structure and should therefore be included in the self-avoiding tube. Here we formulate it as a set of upper bounds on the turning angles, $\{\theta_i\}_{i=2,\ldots,N-1}$, given by

$$\cos\theta_i = \mathbf{t}_{i-1} \cdot \mathbf{t}_i, \text{ where } \mathbf{t}_{i-1} = \frac{\Delta \mathbf{v}_{i-1}}{\|\Delta \mathbf{v}_{i-1}\|} \text{ and } \mathbf{t}_i = \frac{\Delta \mathbf{v}_i}{\|\Delta \mathbf{v}_i\|}, \tag{5.6a}$$

[3] Eqs. (5.4) and (5.5) are the discrete formulations of Eqs. (3.3) and (3.2) in Section II of Chapter 3.

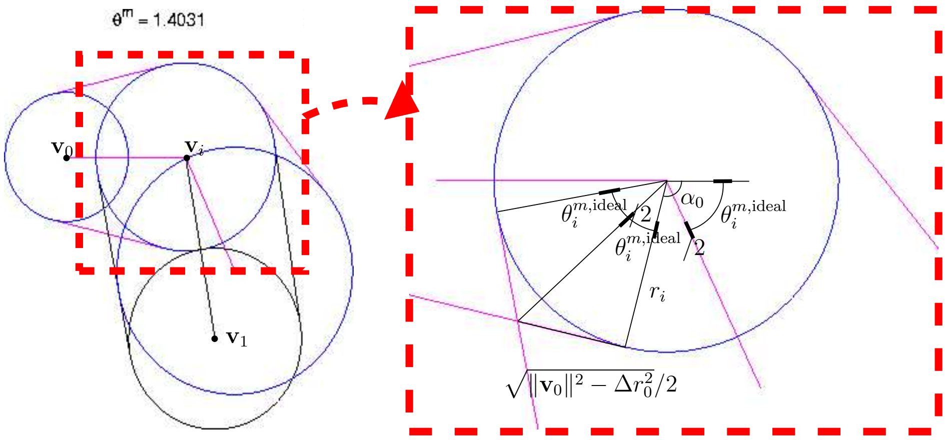

θ^m = 1.4031
$\mathbf{v}_0$
$\mathbf{v}_i$
$\mathbf{v}_1$
$\theta_i^{m,\text{ideal}}/2$
α_0
$\theta_i^{m,\text{ideal}}$
$\theta_i^{m,\text{ideal}}/2$
r_i
$\sqrt{\|\mathbf{v}_0\|^2 - \Delta r_0^2/2}$

Figure 5.4: The tube radius leads to an upper bound on the turning angle, $\theta_i^{m,\text{ideal}}$. Different opening angles in the segments leads to a correction of the ideal maximal angle (see text).

is a unit vector along the axis of tube segment $i-1$ and i respectively. We introduce the *ideal maximal turning angle* of vertex i, $\theta_i^{m,\text{ideal}}$, given by

$$\tan(\theta_i^{m,\text{ideal}}/2) = \min_{j=i-1,i} \left(\sqrt{\|\Delta \mathbf{v}_j\|^2 - \Delta r_j^2} \right) / 2r_i, \tag{5.6b}$$

which is identical to the true maximal turning angle, θ_i^m, if and only if the opening angles of the two segments, α_{i-1} and α_i, are identical.[4] When this is not the case, there is a correction factor

$$\theta_i^{m,\text{ideal}} \to \theta_i^m = \theta_i^{m,\text{ideal}} + \alpha_0 - \alpha_1, \tag{5.6c}$$

where α_0 is the opening angle of the minimizing segment in Eq. (5.6) and and α_1 the opening angle of the remaining segment.

Remark 5.2 *Consecutive segments can be "locked", i.e. $\theta_i^m = 0$, when $\Delta r_0 < 0$ and $\Delta r_1 > 0$. Even worse, the correction term $\alpha_0 - \alpha_1$ in Eq. (5.6c) can lead to $\theta^m < 0$. To avoid this we impose $\theta_i^m \geq 0$ as a set of additional constraints on the tube coordinates.*

Formulated as a set of constraints on the turning angles, and hence the tube coordinates, we have

$$\mathcal{C}^{\text{angles}}(\mathbf{v}, \mathbf{r}) = \begin{pmatrix} \theta_i^m - \theta_i \\ \theta_i^m \end{pmatrix} \geq 0, \qquad i = 2, \ldots, N-1, \tag{5.7}$$

defined in terms of Eq. (5.6).

[4] When $\alpha_i = \alpha_{i-1} \equiv \pi/2$ Eq. (5.6) is the discrete radius of curvature of a uniform tube [Rawdon 97, Rawdon 00].

Remark 5.3 *The set of upper bounds on the turning angles, given by Eq. (5.6b), is the discrete version of the point-wise upper bound*

$$\kappa < \frac{\sqrt{1-(R')^2}}{R^2},$$

on the curvature, $\kappa(s)$, of the generating curve of a canal surface with a radial function that varies linearly with arc-length. This was the subject of Section II of Chapter 3.

II.3 Shape: A global overlap constraint

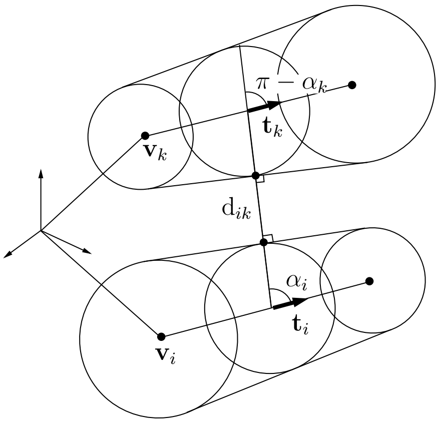


Figure 5.5: Testing for self-intersections involves finding the signed distance between tube (or cone) segments. This question was addressed in detail in Chapter 4.

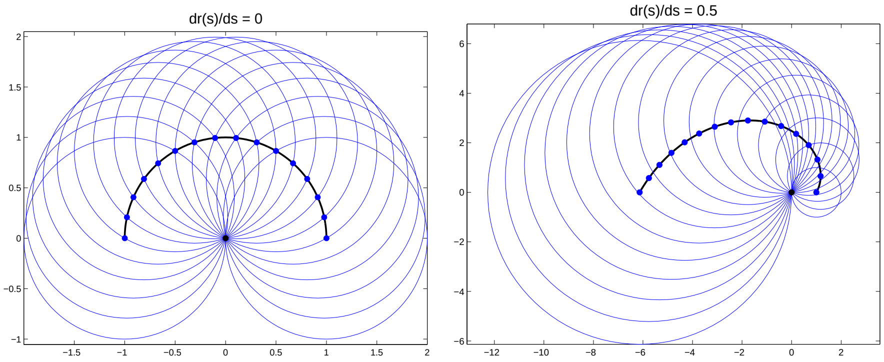
The set of local constraints Eq. (5.7) is not sufficient to preserve the shape of a tube. Instead this is done by imposing a constraints on the pairwise overlap between segments. This is also called self-intersection when two segments belong to the same tube.

For a uniform tube, build from a collection of cylinder segments and spheres, the self-intersection check is straightforward [Eberly 00, Ashton 05] but not so when radii can vary. In this case we have to find the *signed cone-cone distance* between tube segments i and k, given by[5]

$$\mathrm{d}_{ik} = \min_{(s,t)\in[0,1]^2} \Big\{ \|\mathbf{e}_i(s) - \mathbf{e}_k(t)\| - R_i(s) - R_k(t) \Big\}. \tag{5.8}$$

A detailed solution to this problem, illustrated in Fig. 5.5, is given in Chapter 4.

[5] d_{ik} is a $\mathcal{C}^1$ function when cone axes do not intersect.



(a) Uniform tube.

(b) Non-uniform tube.

Figure 5.6: Cross-sections of maximally bend smooth tubes. See Remark 5.3 and Section II.1 of Chapter 3 for details.

It is obvious that neighboring segments *should* be allowed to intersect (see Fig. 5.4) but also that segments distant along the tube axis *should not*. What about next-nearest neighbors, should they be allowed to intersect? [Rawdon 00] provides a solution to this problem, for the case of a uniform tube, in the following way. Consider the maximally bend curve for a smooth tube of uniform radius in Fig. 5.6(a). Now take each pair of consecutive circles to be end-spheres defining a tube segment. Then a collection of connected segments for which the turning angles sum to less than π should be allowed to intersect. From Fig. 5.6(b) we see that a similar construction can be made for a non-uniform tube.

Remark 5.4 *The explanation given above is in fact not correct. [Rawdon 00] demonstrates, that to ensure proper convergence of the tube radius, the generating curve of a smooth uniform tube should be inscribed in the tube axis, and not the other way around as done here. However, for our purposes the intuition provided by Fig. 5.6 is sufficient.*

We use the maximal turning angles, $\{\theta_i^m\}_{i=2,\dots N-1}$, to define a neighborhood of a vertex i where pairwise overlaps are accepted. This is done in the following way

$$\begin{aligned} \theta_{i+1}^m + \cdots + \theta_{i+n-1}^m + u_i\theta_{i+n}^m &= \pi, \\ \theta_i^m + \cdots + \theta_{i-h+1}^m + l_i\theta_{i-h}^m &= \pi, \end{aligned} \qquad 0 < u_i, l_i \leq 1, \tag{5.9}$$

and with this notion of a vertex neighborhood we can define the set of overlap constraints

$$\begin{aligned} &\mathcal{C}^{\text{distance}}(\mathbf{v},\mathbf{r}) = \mathrm{d}_{ik} \geq 0, \\ &\qquad \text{for} \quad k < i-h \text{ or } k > i+n \quad \text{and} \quad i = 1, 2, \dots N-1. \end{aligned} \tag{5.10}$$

The parameters u_i and l_i are introduced to ensure piece-wise differentiability of the objective function used to assign tubes to proteins. This is the subject of Section III.

Remark 5.5 *For a non-uniform tube there are configurations where the maximal turning angles never sum up to π. This would be the case for a sequence of locked segments (see Remark 5.2) where $\theta_i^m = 0$. However, this is not a problem since the angle constraint in Eq. (5.7) then reduces to*

$$\mathcal{C}^{angle} = \theta_i^m - \theta_i = -\theta_i \geq 0,$$

i.e. $\theta_i = 0$, which keeps the tube from turning at vertex i and hence from self-intersecting.

Remark 5.6 *We set $\theta_2^m = \theta_{N-1}^m \equiv \pi$ to ensure proper behavior at the tube terminals.*

III Assigning tubes to protein structures

The discrete self-avoiding non-uniform tube of Section II is defined by the choice of building blocks, in the form of tube segments, and then the constraints

$$\mathcal{C} = \begin{pmatrix} \mathcal{C}^{\text{ bound}} \\ \mathcal{C}^{\text{ angle}} \\ \mathcal{C}^{\text{ distance}} \end{pmatrix} \geq 0, \tag{5.11}$$

given by Eqs. (5.5), (5.7), and (5.10) respectively. However, without some means of specifying vertex positions and radii the model is of little interest. In [Rawdon 00] this is done by minimizing the length a closed polygonal curve, while radius and knot type are kept fixed, to obtain the ideal form of a knot. In [Banavar 00] it is done via an attractive contact potential and the desire to obtain local compact configurations that resemble secondary structure elements in proteins. Here we find a tube that reflects the shape and volume of a protein by maximizing the intersection volume between protein atoms and tube. This is the subject of the remaining part of the chapter.

Remark 5.7 *Other geometrically motivated objectives can be constructed but volume-exclusion and shape are two of the most fundamental properties of a protein structure. This makes intersection volume the most natural choice.*

Assignment of protein atoms and weighting of intersection volumes

We use the all-atom sphere representation of a protein structure. Here an atom is represented by a sphere of van der Waals radius centered at the coordinates provided by the crystal structure. Given a protein structure the number of tube vertices, N, is set equal to the number of residues. This gives a canonical identification between the atoms of residue i, denoted by $\mathcal{A}_i$, and vertex i. We use $\mathbf{a}_i^k$ to denote the ball of radius r_i^k representing the k'th atom in $\mathcal{A}_i$.

In the following we do not distinguish, in words or in notation, between protein atoms and their representations.

Allowed and non-allowed protein atoms

To retain the shape of a tube, a given segment must not be allowed to overlap segments outside a well-defined vertex neighborhood. This was explained in Section II.3. Similarly, an atom cannot not be allowed inside just any segment, if the tube should reflect the shape of the protein. Instead, we need an intermediate formulation, where only atoms from residues close to a segment along the protein backbone are allowed inside.

The notion of a vertex neighborhood defined in Section II.3 is precisely the required coarse-grained assignment of atoms: When two segments are allowed intersect according to Eq. (5.10), then the atoms of one segment are allowed inside the other, and *vice versa.* Using Eq. (5.9), we then define three sets of indices for a vertex i

$$\begin{aligned}\mathcal{L}_i &= \{1,\ldots,i-h-1\}, && (5.12)\\ \mathcal{M}_i &= \{i-h,\ldots,i+n\}, \text{ and} && (5.13)\\ \mathcal{U}_i &= \{i+n+1,\ldots,N\}, && (5.14)\end{aligned}$$

where $\mathcal{M}_i$ contains the indices for the sets of allowed atoms, $\mathcal{A}_{i-h},\ldots,\mathcal{A}_{i+n}$ of segments $i-1$ and i (joined at vertex i). Together $\mathcal{L}_i \cup \mathcal{U}_i$ define the set of non-allowed atoms.

Fractional allowedness of protein atoms

In the present situation an atom is either allowed inside a given segment, or it is not. This strict assignment is unfortunate since a small perturbation, such as moving a tube vertex slightly, can lead to a change in the status of an atom. Instead we should have a notion of "fractional allowedness". This is provided by the parameters u_i and l_i in Eq. (5.9), which give a measure of how much the segments farthest away should contribute.

With the fractional allowedness we can construct a differentiable weighting of the tube-atom intersection volumes in the following way. With a slight abuse of notation we define, $\partial M_i = \{i-h, i+n\}$ and consider an atom $\mathbf{a}_j^k$ from the set $\mathcal{A}_j$. The intersection volume between the atom and tube seg-

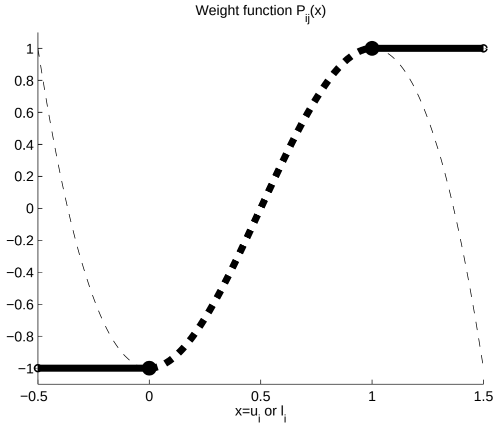


Figure 5.7: Differentiable weight function defined terms of the parameters of fractional allowedness $0 < u_i, l_i < 1$ *(see text).*

ment i is then weighted by

$$P_{ij} = \begin{cases} \quad +1, & j \in (\mathcal{M}_i \setminus \partial M_i), \\ \quad -1, & j \in (\mathcal{L}_i \cup \mathcal{R}_i), \\ P(x)|_{x=l_i \text{ or } u_i}, & j \in \partial\mathcal{M}_i, \end{cases} \tag{5.15}$$

where

$$\begin{aligned} &P(x) = 2(3x^2 - 2x^3) - 1, \\ &\qquad \text{satisfying } P(0) = -1, P(1) = 1 \text{ and } P'(0) = P'(1) = 0, \end{aligned} \tag{5.16}$$

is the lowest order polynomial that ensures differentiability of the weight function P_{ij}. A plot of P_{ij} is given in Fig. 5.7.

III.1 Objective function: Atom-segment intersection volume

Using the (segment dependent) partition of protein atoms into allowed, non-allowed, and fractionally allowed atoms, we define a piecewise differentiable objective function $F(\mathbf{v}, \mathbf{r})$ that counts the tube-atom intersection volume, in the following way. Let $V_i(\mathbf{a}_j^k)$ be the fraction of atom $\mathbf{a}_j^k$ inside segment i, then

$$F(\mathbf{v}, \mathbf{r}) = \sum_{\substack{\text{segment} \\ i}} \left(\sum_{\substack{\text{protein} \\ \text{atom } \mathbf{a}_j^k}} P_{ij} V_i(\mathbf{a}_j^k) - \sum_{\substack{\text{solvent} \\ \text{atom } \mathbf{a}^k}} V_i(\mathbf{a}^k) \right) / \mathcal{N}_{\text{atoms}}, \tag{5.17}$$

where P_{ij} is the weight function in Eq. (5.15) and $\mathcal{N}_{\text{atoms}}$ the number of (non-hydrogen) atoms in the protein. To ensure proper behavior of the tube at the protein surface we have introduced solvent atoms into the model, always weighted by -1. The objective function is normalized such that, for a value of 1, the tube has all allowed atoms and no non-allowed or solvent atoms inside.

A tube assigned to a protein structure is then a set of tube coordinates, $(\mathbf{v}, \mathbf{r})$, providing a (local) solution to the non-linear optimization problem

$$\max_{(\mathbf{v},\mathbf{r})\in \mathrm{R}^{(3+1)N}} F(\mathbf{v}, \mathbf{r}) \text{ subject to } \mathcal{C}(\mathbf{v}, \mathbf{r}) \geq 0, \tag{5.18}$$

where F is the objective function and $\mathcal{C}$ the sets of constraints in Eq. (5.11).

Remark 5.8 *For simplicity we do not take the overlap between atoms into account, i.e. the objective function only sees one atom at a time. This is of no major consequence since overlap regions are (almost) uniformly distributed over the interior of the protein. However, it could have a small effect in shifting the tube away from the surface of the protein.*

Remark 5.9 *It could be argued, that one should consider tubes for an ensemble of solvent configurations. However, we are not yet in a position to appreciate the finer details of the tube representations and the actual solvent configuration should have no bearing on the results reported in Section IV.*

Details on the intersection volume The intersection volume $V_i(\mathbf{a}_j^k)$ in Eq. (5.17) consists of two terms

$$V_i(\mathbf{a}_j^k) = V_i^{\text{single}}(\mathbf{a}_j^k) - V_i^{\text{double}}(\mathbf{a}_j^k), \tag{5.19}$$

and is only an approximation of the true intersection volume. In the following we provide a few details on this matter.

The true intersection volume between an atom and a tube segment is the intersection of a sphere and a cone frustum with end-spheres. Unfortunately even the simpler problem of a sphere-cylinder intersection volume has a solution in terms of elliptical integrals [Lamarche 90] and an exact solution in the present situation *must* be elaborate, perhaps even impossible, and definitely numerically intractable. Instead, the true volume is approximated by the volume of a sphere cut by the tangent plane of the tube segment given by the vector along the shortest sphere-cone distance,

$$\mathrm{d}_i(\mathbf{a}_j^k) = \min_{s\in[0,1]} \left\{ \|\mathbf{e}_i(s) - \mathbf{a}_j^k\| - R_i(s) - r_j^k \right\}. \tag{5.20}$$

This is illustrated in Fig. 5.8 and how to determine d_i is explained in Section II.2 of Chapter 4. The approximate intersection volume, normalized by the

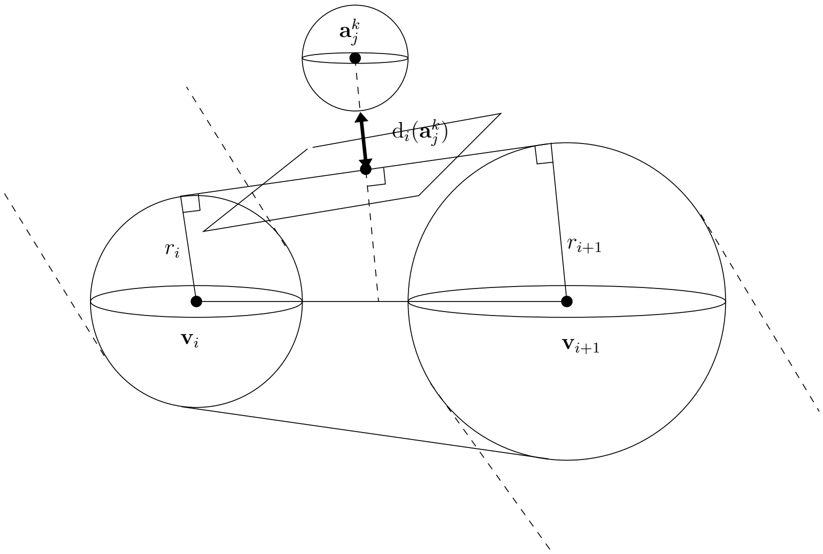


Figure 5.8: Tube segment i and atom $\mathbf{a}_j^k$ from $\mathcal{A}_j$. The vector along the shortest sphere-cone distance is perpendicular to both surfaces. The true intersection volume is approximated by the volume of a sphere cut by the tangent plane of the tube segment defined by the vector along the shortest distance.

sphere volume, is then

$$V_i^{\text{single}}(\mathbf{a}_j^k) = \begin{cases} 0, & \text{for} \quad 0 \leq \tilde{\mathrm{d}}_i, \\ \tilde{\mathrm{d}}_i^2(2+\tilde{\mathrm{d}}_i)/4, & \text{for} \quad -1 < \tilde{\mathrm{d}}_i < 0, \\ 1-(1+\tilde{\mathrm{d}}_i)^2(2-\tilde{\mathrm{d}}_i)/4, & \text{for} \quad -2 < \tilde{\mathrm{d}}_i \leq -1, \\ 1 & \text{for} \quad \tilde{\mathrm{d}}_i \leq -2, \end{cases} \tag{5.21}$$

where $\tilde{\mathrm{d}}_i(\mathbf{a}_j^k) = \mathrm{d}_i(\mathbf{a}_j^k)/r_j^k$.

The single segment volume Eq. (5.21) overestimate contributions from atoms lying in the overlap regions between segments. To avoid this, the doubly counted volume, $V_i^{\text{double}}(\mathbf{a}_j^k)$, must be subtracted. In the present approximation this volume has the form of a sphere cut by two planes. The same intersection volume arises in a volume calculation of the all-atom sphere representation (see [Edelsbrunner 05] and references therein). An exact solution can be found in [Dodd 91].

IV Results

We now report on the results from a numerical solution of the tube-assignment problem. Details on the implementation are given in Appendix A. An ex-

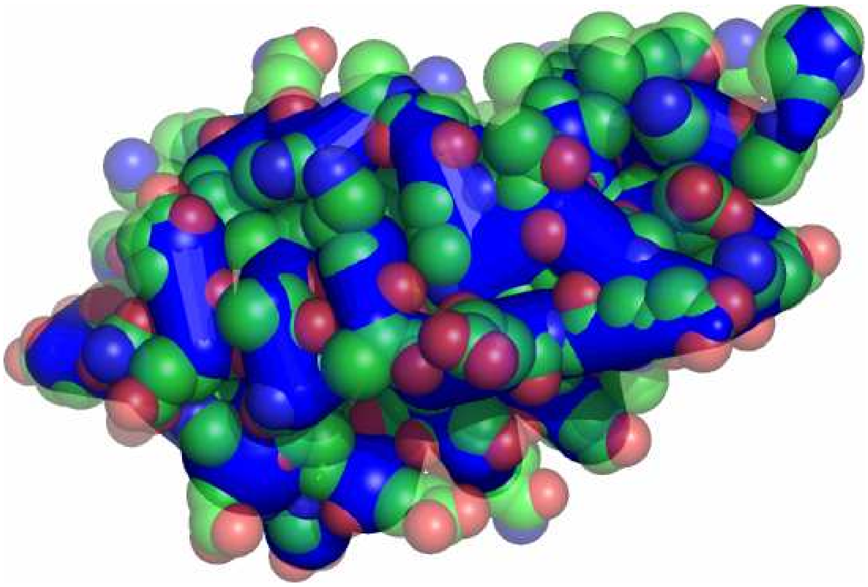


Figure 5.9: Non-uniform tube assigned to the $\alpha-\beta$ CATH domain (`1f60B0`) superimposed on the protein atoms.

ample of a non-uniform tube assigned to a protein is given in Fig. 5.9.

IV.1 Intersection volume and radius distribution in protein tubes

We have looked at the intersection volumes and radius distributions of tubes assigned to a set of 31 protein domains.[6] The domains are all from different homology classes in the CATH database [Orengo 97] and almost evenly distributed amongst the mostly-α, mostly-β, and $\alpha-\beta$ classes [Levitt 76].

We have consider three version of a tube: A uniform tube of radius 2.25Å and vertices at C_α coordinates (UT fixed), the uniform tube (UT) and the non-uniform tube (NUT).

Intersection volume

To evaluate the quality of a tube model we split the objective function into contributions from allowed and non-allowed protein atoms, and then solvent atoms. The results are shown in Fig. 5.10 where we see that the volumes are roughly identical across structures and classes.

Surprisingly the non-uniform tube is only slightly better at capturing the correct intersection volume than the uniform tube - where design variables

[6] This is a very limited data set. The results in this sections should be read with this in mind.

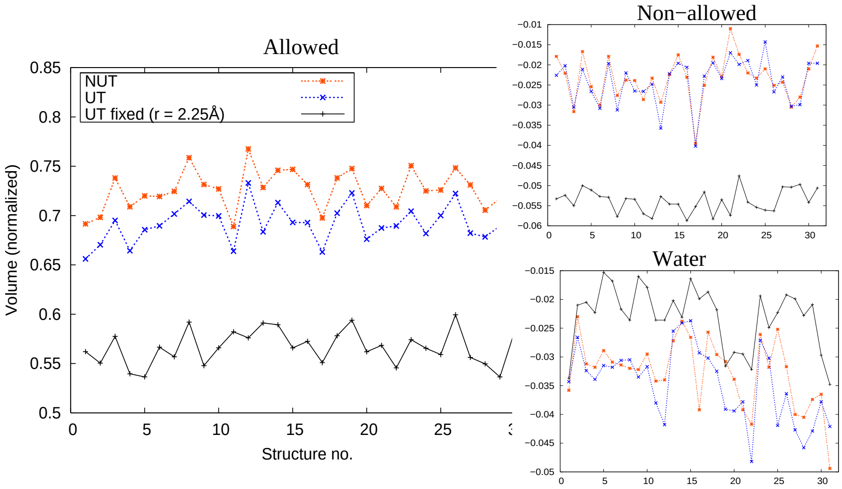


Figure 5.10: Contributions to the objective function from allowed and non-allowed protein atoms, and then solvent atoms. The values for three tube models assigned to a set of 31 non-homologous protein domains are shown. The models are: A uniform tube of radius 2.25Å and vertices at the C_α coordinates (UT fixed), the uniform tube (UT) and the non-uniform tube (NUT).

are vertex position and a single radius. The only significant difference lies in the fraction of allowed atoms included.

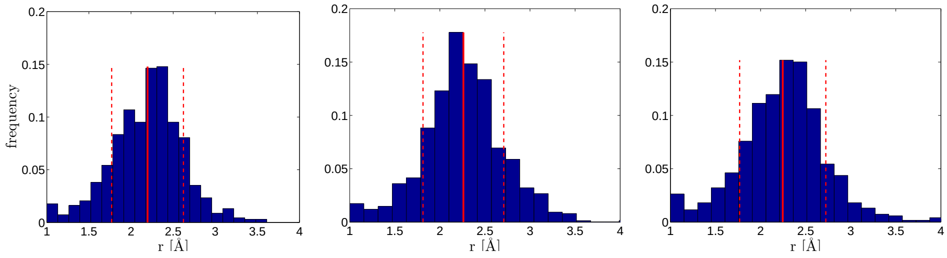
Radius distribution

We have also looked at the radius distributions within the three structure classes. The results are presented in Fig. 5.11 and Table 5.1. We see that, within the set of non-uniform tube, the average radius - even the variations - over the entire set of structures is almost identical to the values within each class.

Class	$\langle r_i \rangle_{\mathrm{NUT}} \pm \sigma(r_i)_{\mathrm{NUT}}$	$\langle r_i \rangle_{\mathrm{UT}} \pm \sigma(r_i)_{\mathrm{UT}}$
Mostly-α	2.20 ± 0.43	2.20 ± 0.04
Mostly-β	2.26 ± 0.42	2.24 ± 0.02
$\alpha - \beta$	2.25 ± 0.48	2.20 ± 0.04
Total	2.24 ± 0.46	2.21 ± 0.04

Table 5.1: Average radii in tubes assigned to 31 protein domains and separated in terms of class. For non-uniform tubes the average is over segments and structures, for uniform tubes only over structures.

Within the set of uniform tubes we find radii close to the average values



(a) Radius distribution within the mostly-α class.

(b) Radius distribution within the mostly-β class.

(c) Radius distribution within the $\alpha - \beta$ class.

Figure 5.11: Distributions of radii in non-uniform tubes assigned to 32 protein domains and separated into classes. A full red line gives the position of $\langle r \rangle$ and hatched red lines a standard deviation on each side of this. The increase in frequency of radii at 1Å is due to an absolute constraint on the vertex radii, $r_i > 1$Å.

for non-uniform tubes. The difference between the models then lies in the radius *variations* within the non-uniform tubes, as seen by the standard deviation $\sigma(r)_{\text{NUT}} \sim 0.4$. It should be kept in mind, that volume variations are proportional to r^3, so much is gained by a small variation in radius. This is the reason for additional allowed atoms included by the non-uniform tubes (see Fig. 5.10).

Remark 5.10 *In the numerical implementation of Eq. (5.18) we have imposed a set of absolute constraints on the radii, namely $\{r_i > 1\text{Å}\}_{i=1,\ldots N}$. This is the reason for the increase in the frequency of radii around 1Å in Fig. 5.11.*

IV.2 Secondary structures in protein tubes

Walking along a non-uniform tube we gain geometrical information about the protein environment by considering the radius variations. In other words, by the transition to a non-uniform radius we have retained some level of sequence information. This could possible be used to distinguish semi-local structures such as α-helices and β-sheets.

To allow for a larger radius variation than observed in the previous section, we use a locally smoothed version of the C_α-backbone. This is given by substituting all $C_{\alpha,i}$ with the weighted average

$$C^{\text{new}}_{\alpha,i} = \frac{C_{\alpha,i-2} + aC_{\alpha,i-1} + bC_{\alpha,i} + aC_{\alpha,i+1} + C_{\alpha,i+2}}{2 + 2a + b}, \tag{5.22}$$

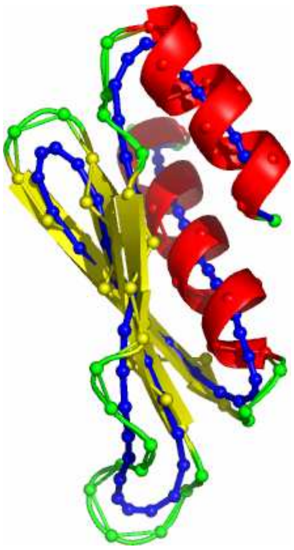


Figure 5.12: The consequences of a smoothed backbone (in blue) are much more severe for helices (red) than sheets (green).

which, with the choice $a = 2.4$ and $b = 2.1$, minimize the curvature for all the most frequent local structures found in proteins [Røgen 05]. An example of a smoothed protein backbone is given in Fig. 5.12. We see that the consequences are much more severe for helices than sheets. We now briefly describe what consequences this has for the non-uniform tube.

Reforming of β-sheets

In Fig. 5.13 is shown the results of an optimization with initial vertices given by Eq. (5.22). We see that the β-sheets are reformed, and this to the extent, that the tube axis resembles the C_α-backbone of the crystal structure (RMSD 1.9Å) more than the smoothed backbone from whence it came (RMSD 2.4Å). This reforming of sheets is not too surprising given the alternating position of the sidechains on opposite sides of a strand.

α-helix "troubles"

It was hoped that barrel-like representations of helices would emerge from the use of the smoothed backbone Eq. (5.22). Instead we found the structures seen Fig. 5.14. Here shorter helices are well represented by a short sequence of fat tube segments, or blobs. Longer helices are apparently represented by a sequence of such blobs as seen in Fig. 5.14(b).

We thought that, by using DSSP [Kabsch 83] to assign secondary structure to residues and subsequently assign larger values of radii in helix regions,

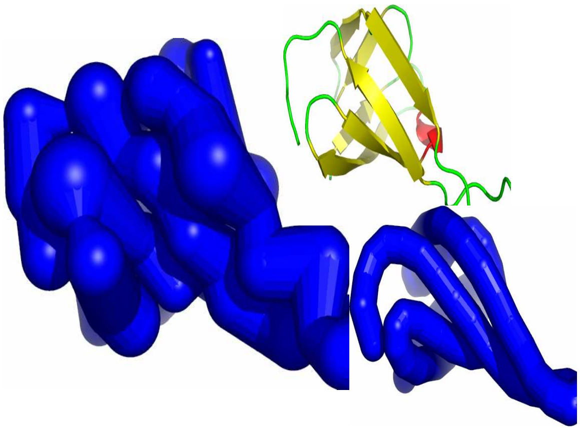
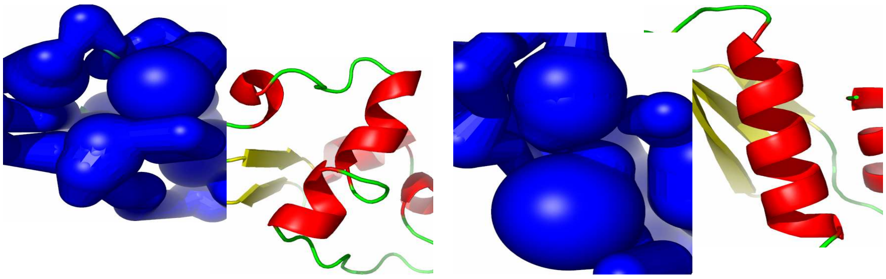


Figure 5.13: Reformation of β sheets in (`1cqqA1`). The initial tube (lower insert) has a smoothed backbone as axis and a uniform radius of 1.9Å. All structures are identically oriented.



(a) 3 short and relatively well-formed helices in CATH domain (`1hh5A0`).

(b) A longer helix in CATH domain (`1stu00`) represented by two globular parts instead of a single barrel-like segment.

Figure 5.14: A better choice of initial tube in the optimization, could possible solve the problem.

would lead to the desired barrel structures. However, all values of the radii, not blatantly violating the constraints ($r_i \approx 2 - 3.5$Å), give rise to similar local tube structures. Again the reason is the sidechains. They make helices considerably less barrel-like, and more like a sequence of blobs linked by shorter thin segments responsible for the inter-blob orientation. This is observed in Fig. 5.14(b).

Remark 5.11 *A way of producing the desired barrel structures could be to merge segments in helix regions. This would also remove a part of the short* *segments, a consequence of the projection of C_α coordinates onto the shorter smoothed backbone, as seen in Fig. 5.12. This is a source of considerable numerical instability.*[7]

IV.3 Stability of tube axes

For a tube assigned to a protein, the position of tube segments depends on the atoms in a whole neighborhood. Therefore, a tube axis is very robust against perturbations of the protein structure. To see if this robustness could be observed within a family of related structures, we have considered a set structures from the transition state ensemble of the SH3 domain of α-spectrin (`1kb2`, [Shaffer 02]).[8]

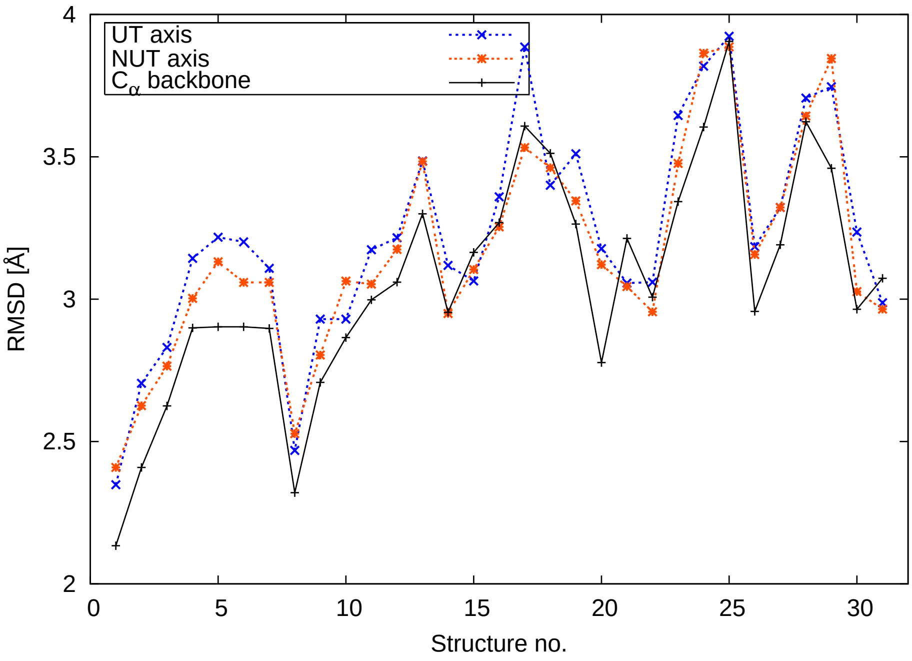


Figure 5.15: Root mean square distance between the backbone of structures from the transition state ensemble and then the native structure. We consider both the C_α-backbone, and the axes of the uniform (UT) and non-uniform tube (NUT).

Fig. 5.15 shows the results of a pairwise comparison between each of

[7] Unfortunately SNOPT works with a fixed number of constraints. This is the principal reason why segment merging has not been implemented.

[8] The structures were kindly made available to us by K. Lindorff-Larsen. The data have previously been used to demonstrate that structures in the transition state ensemble resemble the native structure, despite high structural variability [Lindorff-Larsen 04].

the structures in the ensemble, and then the native state. This has been done for the C_α-backbone, the axis of the uniform tube, and the axis of the non-uniform tube. We see that tube axes display the a higher variability (relative to the native structure), than the C_α backbone. The same pattern is observed in a pairwise comparison between all structures.

In hindsight, the result may not be so surprising. The tubes axes reflect sidechain positions, which are more prone to variations than the protein backbone. In the "looser" structures of the transition state, this is even more so, which is observed in the larger variability amongst tube axes.

V Conclusion & future work

We have reported on the first formulation of a self-avoiding tube of non-uniform radius. Contrary to the previous (uniform) tube models used for protein studies [Banavar 00, Banavar 02, Banavar 03a, Hoang 04] the new formulation maintains its promised radius across a tube segment. The self-avoiding tube is defined by constraints imposed on individual tube segments:

- Stiffness is introduced in the form of an upper bound on turning angles.
- The shape is preserved by preventing self-intersections. This is done using the solution to the cone-cone distance problem in Chapter 4.

In a uniform tube, local properties have global implications via the radius. This local-global aspect is not present in the non-uniform tube. Here, local events only directly affect their immediate environment. This again implies, that more information is needed to determine appropriate values for the additional radial degrees of freedom.

The local-global aspect has disappeared but we have gained the ability to retain local information. Precisely in the context of proteins, this is a large conceptual improvement over the uniform tube. We have used protein atoms to provide the required information. In this way we assign tubes to protein structures by maximizing the tube-protein intersection volume. In this case, the non-uniform tube retains some measure of geometric sequence information.

To examine this property, we have looked at intersection volumes and radius distributions in uniform and non-uniform tubes assigned to a set of 31 non-homologous protein domains. We found a remarkable agreement between radii of uniform tubes and then the average radii of non-uniform tubes; both within individual structures and within classes based on secondary structure content [Levitt 76]. The difference between the two tube models then lies in the radius variation *within* a structure. This enables the non-uniform tube pack closer and thus include more of the correct protein atoms.

Future applications

It would be interesting to explore the connection between the active distance constraints in a protein tube, and then a biological measure of contact, for example as given by Φ-value analysis.[9] The connection is plausible because the set of contacts formed in the transition state closely resembles the set of native contacts [Lindorff-Larsen 04]. With a connection established, we would be able to gain information about the transition state by looking at tubes assigned to the native structure.

Another interesting perspective is to implement a simple attractive contact potential along the lines of [Banavar 00, Banavar 02]. Working on the level of secondary structure, or even larger structures, we could determine if the addition of geometric sequence information influences the folding patterns.

On the mathematical side, we would like to extend the discrete radius of curvature for uniform tubes, given in [Rawdon 00], to the case of linearly varying radius. This would involve the ideal maximal turning angles, used to model stiffness in the non-uniform tube, and then the generating curve of a maximally bend smooth tube. In Chapter 3 we demonstrated that the shape of this curve is a logarithmic spiral. The question is then, how tube segments can be used to approximate the smooth tube.

[9] In Φ-value analysis the effect of point mutations throughout a protein is evaluated in terms of changes in the free energy differences between the unfolded and native state [Matouschek 89].

A Numerical implementation

To solve the nonlinear tube-assignment problem

$$\max_{(\mathbf{v},\mathbf{r})\in \mathrm{R}^{(3+1)N}} F(\mathbf{v},\mathbf{r}) \text{ subject to } \mathcal{C}(\mathbf{v},\mathbf{r}) \geq 0,$$

we use the commercial software SNOPT 7.0 written for large-scale linear and non-linear optimization problems [Gill 06]. The objective function and constraints, together with the gradients with respect to the tube coordinates, $\nabla_{(\mathbf{v},\mathbf{r})}F$ and $\nabla_{(\mathbf{v},\mathbf{r})}\mathcal{C}$, has been implemented in MATLAB. To generate solvent coordinates around the crystal structure we used Solvate 1.0 [Grubmüller 96].

Initial condition

For the gradient based method used by SNOPT a good choice of initial condition is crucial to ensure proper convergence. A good starting point is provided by the C_α coordinates. To have an automatic and fast radius assignment we use the maximum radius satisfying the constraints on a uniform tube. This value is severely restricted at turns but it is better to start with a small radius than too large. In the latter case a host of constraints becomes activate which cause a significant slowing down of SNOPT.[10] Results from a few runs with a different radius assignment scheme are reported in Section IV.2.

Computational complexity

The number of angle constraints is linear in the number of vertices N. Naïvely checking for self-intersections between pairs of cone segments is $\mathcal{O}(N^2)$ and the most costly part of the algorithm.[11] As a crude but efficient way to lower the number of self-intersection checks we only consider segments within a distance of ~ 20Å of each other in the starting tube. This works because we are able to provide a good starting point.

Assigning a non-uniform tube to a protein is a computationally and takes $\sim$ 8h on a standard laptop computer. A typical small protein has around 125 residues which gives 600 tube coordinates and around 1000 atoms. Solvating the crystal structure introduces a further $\sim 6-7$ times this number atoms. For both sets of atoms, the number of atom-segment distance computations can be reduced significantly by only considering atoms within a distance

[10] At each major iteration SNOPT divides the set of constraints into active and inactive constraints.

[11] Checking for self-intersections is also the limiting step in the algorithms used to tighten ideal knots and ways of lowering the number has received some attention in the literature [Ashton 05].

R_0 of a segment in the starting tube. We have, somewhat arbitrarily, used $R_0 \sim 12$Å,

The size of the proteins we can consider is limited by memory capacity, which in our case is equivalent to ~ 100 residues. This number could most likely be significantly increased by a better filtering of atoms and constraints.

Bibliography

[Anfinsen 73] C. Anfinsen. *Principles that Govern the Folding of Protein Chains (Nobel lecture in chemistry 1972).* Science, vol. 181, no. 4096, pages 223–230, 1973.

[Arteca 99] G.A. Arteca. *Path-Integral Calculation of the Mean Number of Overcrossings in an Entangled Polymer Network.* J. Chem. Inf. Comput. Sci., vol. 39, pages 550–557, 1999.

[Ashton 05] T. Ashton & J. Cantarella. *A fast octree-based algorithm for computing ropelength.* In Physical and Numerical Models in Knot Theory and their Application to the Life Sciences, pages 323–341. World Scientific Press, 2005.

[Banavar 00] J.R. Banavar, A. Maritan, C. Micheletti & A. Trovato. *Optimal Shapes of Compact Strings.* Nature, vol. 287, pages 287–290, 2000.

[Banavar 02] J.R. Banavar, A. Maritan, C. Micheletti & A. Trovato. *Geometry and Physics of Proteins.* Proteins, vol. 47, pages 315–322, 2002.

[Banavar 03a] J.R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan & A. Trovato. *Tubes near the edge of compactness and folded protein structures.* J. Phys.: Condens. Matter, vol. 15, pages 1787–1796, 2003.

[Banavar 03b] J.R. Banavar & A. Maritan. *Geometrical approach to protein folding: a tube picture.* Rev. Mod. Phys., vol. 75, pages 23–34, 2003.

[Berman 92] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S-H. Hsieh, A.R. Srinivasan & B. Schneider. *The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structure of Nucleic Acids.* Biophys. J., vol. 63, pages 751–759, 1992.

[Dodd 91] L.R. Dodd & D.N. Theodorou. *Analytical treatment of the volume and surface area of molecules formed by an arbitrary collection of* *unequal spheres intersected by planes.* Mol. Phys., vol. 72, no. 6, pages 1313–1345, 1991.

[Eberly 00] D. Eberly. 3d game engine design. Morgan Kaufmann Publishers, 2000.

[Edelsbrunner 05] H. Edelsbrunner & P. Koehl. *The Geometry of Biomolecular Solvation.* Discrete and Computational Geometry, vol. 52, pages 241–273, 2005.

[Eisenberg 86] D. Eisenberg & A.D. McLachlan. *Solvation energy in protein folding and binding.* Nature, vol. 319, pages 199–203, 1986.

[Gill 06] P.E. Gill, W. Murray & M.A. Saunders. *User's Guide For Snopt Version 7: A Fortran Package for Large-Scale Nonlinear Programming*, 2006.

[Gonzalez 99] O. Gonzalez & J.H. Maddocks. *Global curvature, thickness and the ideal shapes of knots.* Proc. Nat. Acad. Sci. USA, vol. 96, no. 9, pages 4769–4773, 1999.

[Grubmüller 96] H. Grubmüller. *SOLVATE 1.0*, 1996.

[Hoang 04] T.X. Hoang, A. Trovato, F. Seno, J.R. Banavar & A. Maritan. *Geometry and symmetry presculpt the free-energy landscape of proteins.* Proc. Nat. Acad. Sci. USA, vol. 101, no. 21, pages 7960–7964, 2004.

[Kabsch 83] W. Kabsch & C. Sander. *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.* Biopolymers, vol. 22, pages 2577–2637, 1983.

[Karplus 02] M. Karplus & J.A. McCammon. *Molecular dynamics simulations of biomolecules.* Nat. Struct. Bio., vol. 9, no. 9, pages 646–652, 2002.

[Koehl 02] P. Koehl & M. Levitt. *Protein topology and stability define defines the space of allowed sequences.* Proc. Nat. Acad. Sci. USA, vol. 99, pages 1280–1285, 2002.

[Kolinski 04] A. Kolinski & J. Skolnick. *Reduced models of proteins and their applications.* Polymer, vol. 45, pages 511–524, 2004.

[Kolodny 05] R. Kolodny, P. Koehl & M. Levitt. *Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures.* J. Mol. Biol., vol. 346, pages 1173–1188, 2005.

[Lamarche 90] F. Lamarche & C. Leroy. *Evaluation of the volume of intersection of a sphere with a cylinder by elliptic integrals.* Comp. Phys. Comm., vol. 59, pages 359–369, 1990.

[Levitt 76] M. Levitt & C. Chothia. *Structural patterns in globular proteins.* Nature, vol. 17, no. 261, pages 552–558, 1976.

[Lindorff-Larsen 04] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C.M. Dobson, F.M. Poulsen & M. Vendruscolo. *Transition states for protein folding have native topologies despite high structural variability.* Nat. Struct. Bio., vol. 11, pages 443–449, 2004.

[Lindorff-Larsen 05] K. Lindorff-Larsen, P. Røgen, E. Paci, M. Vendruscolo & C.M. Dobson. *Protein folding and the organization of the protein topology universe.* Trends Biochem. Sci., vol. 30, no. 1, pages 13–19, 2005.

[Matouschek 89] A. Matouschek, J.T. Kellis Jr, L. Serrano & A.R. Fersht. *Mapping the transition state and pathway of protein folding by protein engineering.* Nature, vol. 340, pages 122–126, 1989.

[McPhalen 87] C.A. McPhalen & M.N. James. *Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds.* Biochemistry, vol. 26, pages 261–269, 1987.

[Orengo 97] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells & J.M. Thornton. *CATH- A Hierarchic Classification of Protein Domain Structures.* Structure, vol. 5, no. 8, pages 1093–1108, 1997.

[Plaxco 98] K.W. Plaxco, K.T. Simons & D. Baker. *Contact order, transition state placement and the refolding rates of single domain proteins.* J. Mol. Biol., vol. 277, pages 985–994, 1998.

[Rawdon 97] E.J. Rawdon. *Thickness of polygonal knots.* PhD thesis, University of Iowa, 1997.

[Rawdon 00] E.J. Rawdon. *Approximating smooth thickness.* J. Knot Theory Ramifications, vol. 9, no. 1, pages 113–145, 2000.

[Røgen 03] P. Røgen & B. Fain. *Automatic classification of protein structure by using Gauss integrals.* Proc. Natl. Acad. Sci. USA, vol. 100, no. 1, pages 119–124, 2003.

[Røgen 05] P. Røgen. *Evaluating protein structure descriptors and tuning Gauss integral based descriptors.* J. Phys.: Cond. Mat., vol. 17, pages 1523–1538, 2005.

[Røgen 07] P. Røgen, 2007. Personal communication.

[Shaffer 02] P.L. Shaffer & D.T. Gewirth. *Structural basis of VDR-DNA interactions on direct repeat response elements.* EMBO J., vol. 21, pages 2242–2252, 2002.

[Stasiak 98] A. Stasiak, V. Katritch & L.H. Kauffman, editeurs. Ideal knots. World Scientific, London, 1998.

[Taylor 01] W.R. Taylor, A.C.W. May, N.P. Brown & A. Aszódi. *Protein structure: geometry, topology, and classification.* Rep. Prog. Phys., vol. 64, pages 517–590, 2001.

Chapter 6

Flexibility and conformational change in proteins

The present chapter was initially motivated by a question posed in [Petrone 06]: Can conformational change be described by only a few normal modes? Their answer was, that for the four proteins considered '... the first 20 modes only contribute 50% or less of the total conformational change...'. Perhaps this would change if the model was restricted to the degrees of freedom most relevant at low energies, namely, dihedral angles? To attempt an answer, we have implemented an all-atom model in dihedral angle coordinates, thus "freezing" the stretching and angle bending modes of the covalent bonds. We follow the elegant formulation of Gō and coworkers [Noguti 83b, Abe 84, Sunada 95] and use the single-parameter harmonic potential introduced in [Tirion 96].

We first examine if a few dihedral normal modes are sufficient to represent the observed conformational differences in the set of proteins used by [Petrone 06]. We then proceed to consider the stereochemical properties of protein structures as they undergo deformations along normal mode vectors. In both cases we compare with results obtained by normal mode analysis in Cartesian coordinates.

Finally, the chapter is a basic introduction to the general theory of normal mode analysis in a biological context. Special attention is given to the formulation in dihedral angles.

In Section I we survey the literature on normal mode analysis in biology and introduce some terminology. We address some general concerns about the model and give an overview of some applications. In Section II we see how, imposing the so-called Eckart-Sayvetz conditions, can decouple the rotation and translation from the vibrational motion (in the limit of small vibrations). In Section III & IV we present the energy terms for normal mode analysis in dihedral angles and proceed to solve the equations of motion. In Section V we introduce a new method for finding the best approximation to a conformational difference in a span of non-linear motion. The results of our investigations concerning dihedral angle NMA are presented in Section VI. Finally we end with a discussion and a look to the future in Section VII.

I Introduction

The function of proteins are often chemical in nature and for the chemical reactions to take place, the components involved must be close in space(-time). An efficient way of inhibiting function is then a separation of the components. This is indeed what is observed, conformational changes and flexibility are intimately linked to functionality [Frauenfelder 91, Benkovic 03]. Thus the life of a protein - and of biological macromolecules in general - is inherently dynamic, a fact already appreciated more than two decades ago [Petsko 83].

The energy-landscape of a protein is extremely complex and a model always involves some level of coarse-graining, either structural, energetic, or both. Even the elaborate semi-empirical potentials presented in Section III of Chapter 2 are vast simplifications compared to the ”true” interactions in a protein *in vivo*. However, the computational work of the last two decades show that averaging over some of the many details can lead to useful biological insight.

The structural hierarchy of proteins (see Section II of Chapter 2) suggests that thermal fluctuations, and other low-energy excitations, can take the form of a coherent motion involving many atoms.[1] *Normal mode analysis* is an extreme version of energetic coarse-graining ideally suited for studying low-energy collective motions in proteins. Here, the (pairwise) interactions are mediated by springs and we forget about the solvent. It is a purely mechanical model, that represents the protein as a set of connected point masses. The properties of the system then depend on the spatial distribution of the masses and the way in which they are connected.

In the following sections we review a small part of what this mechanical description of the protein can provide in terms of dynamical and time-averaged information.

I.1 Vibrations and collective motion in molecules

The extreme case of collective motion is the rigid body in which point masses are situated at fixed relative distances and the dynamics given by the motion

[1] That low-energy perturbations of a system lead to collective motion is observed in many physical systems, e.g. in many-particle systems as *phonons* (vibrations) or *polarons* (charge displacements) [Mahan 00].

of the center of mass and the axes of inertia [Landau 88]. In this case the particles obviously move collectively but what if the particles start vibrating slightly? Is the motion still coherent?

Normal mode, or harmonic, analysis (NMA) provides an answer to this question in the form of an analytical solution to the (classical) equations of motion in the limit of (very) small (frictionless) vibrations. It assumes that the potential energy , V, can be approximated by a quadratic function of some set of generalized coordinates $\mathbf{q} = (q_1, \ldots, q_N)$ in the vicinity of a stable equilibrium $\mathbf{q}^0$, in other words

$$V(\mathbf{q}) = (\mathbf{q} - \mathbf{q}^0)^T \nabla^2 V\big|_{\mathbf{q}=\mathbf{q}^0} (\mathbf{q} - \mathbf{q}^0). \tag{6.1}$$

I.2 Normal mode analysis in biology

Normal mode analysis was largely developed by the spectroscopy community where it is routinely used to back-calculate physical parameters, e.g. force constants, from vibrational spectra obtained by Raman or infrared spectroscopy [Califano 76]. The generic system is therefore a molecule at finite temperature where the atoms vibrate around a stable equilibrium configuration which then serves as "rigid body" of the previous section.

NMA was first applied to proteins in the 1980's by the groups of Gō [Gō 83], Karplus [Brooks 83], and Levitt [Levitt 85].[2] In order to obtain a minimum energy configuration they all performed a regularization of the crystal structure using semi-empirical energy potentials (see Section III) and it was observed that the new structure often deviated considerably from the initial crystal structure. This seemingly important energy minimization comes at a considerable computational cost which increases with the level of detail desired [Wako 04].

It therefore attracted considerable attention when it was shown in [Tirion 96] that a simple single-parameter potential

$$V = \frac{K}{2} \sum_{\alpha<\beta} \left(\|\mathbf{r}_\alpha - \mathbf{r}_\beta\| - \|\mathbf{r}_\alpha^0 - \mathbf{r}_\beta^0\| \right)^2, \tag{6.2}$$

was able to qualitatively reproduce the spectrum of atomic fluctuations for a number of well-known protein structures. Here K is a universal spring constant and the sum runs over atoms less than a distance $C_{\text{cut-off}} \sim 8$Å apart. Models using phenomenological potentials such as Eq. (6.2) have since become known as *elastic network models* (ENM) [Bahar 05]. Often they are combined with some sort of structural coarse-graining, e.g. using only C_α atoms [Bahar 97] or collecting atoms into rotational-translational blocks [Durand 94].

[2] [Gō 83] and [Levitt 85] both used dihedral angles as generalized coordinates.

Remark 6.1 *The Tirion potential Eq. (6.2) involving only C_α atoms is sometimes called an anisotropic network model (ANM) while*

$$V = \frac{K}{2} \sum_{\alpha} \left\| \mathbf{r}_\alpha - \mathbf{r}_\alpha^0 \right\|^2 ,$$

is the Gaussian network model (GNM) introduced in [Bahar 97].

A related method is to use principal component analysis of the second moment matrix of coordinate fluctuations obtained from a molecular dynamics trajectory. This it then used to calculate force constants which then serve as input to a quadratic approximation of the full semi-empirical potential. This approach is the quasi-harmonic approximation [Levy 84].

The advent of elastic network models has made large-scale computations of normal modes possible in mainly two directions [Bahar 05]

(i) Larger structures [Tama 06]. Using a hierarchy of coarse-graining techniques [Doruker 04, Gohlke 06] and then applied to large biomolecules such as the functional parts of the ribosomes [Tama 03b].

(ii) More structures. Several groups have developed publicly accessible databases. They come in many flavors such as using an all-atom model with a semi-empirical potential and dihedral angle coordinates [Wako 04] or only C_α atoms [Hollup 05]. Others focus on function [Yang 05] or on classifying the observed motions [Flores 06].

Implementing NMA on a computer is fairly easy and much less CPU-demanding than the other principal sources of dynamical information: Molecular dynamics (MD) and Monte Carlo simulations (MC). Within the harmonic approximation NMA provides exact information about time-averaged properties of the system and the ability to study large conformational changes typically occurring on long time-scales ($> 1\mu$s). In this way it complements the aforementioned methods where sampling on time-scales $> 10 - 100ns$ for larger systems becomes a problem and MD is still only done for carefully selected structures [Karplus 02, Snow 05]. For these reasons NMA has become a standard tool in the interpretation and application of the huge amount of structural data obtained by X-ray crystallography, nuclear magnetic resonance (NMR), or cryo-electron microscopy (cryo-EM).

Applications of normal mode analysis

Structural flexibility A main application of normal mode analysis has been to identify flexible domains [Hinsen 98] and motions related to function [Bahar 05]. Specific examples are (the hinge-bending of) lysozyme [Levitt 85, Brooks 85, Horiuchi 91], (the inter-domain motions of) GroEL [Ma 98, Ma 00], the retinol-binding protein [Atilgan 01], (functional parts of) the ribosome [Tama 03b], F_1ATPase [Cui 04], and potassium channels [Shrivastava 06] to name but a few. The general observation has been, that a single or a few low-frequency mode(s) is(are) sufficient and a similar conclusion is reached using eigenmodes to construct a linear map between configurations of the same protein [Tama 01].

Important subspace With the gross simplifications involved in elastic network models - and to some extent this applies to NMA with realistic potentials - it makes little sense to see the eigenvalues as providing quantitative information about the vibrational frequencies of the protein. Many authors instead use the eigenmodes to determine if the protein dynamics or conformational changes are contained in a low-dimensional part of the configuration space [Horiuchi 91, van Vlijmen 99, Tama 01, Petrone 06]. In this case the eigenvalues simply impose an ordering of the normal modes.

Several studies compare the *important subspace* spanned by a small number of low-frequency normal modes with the same subspace based on data from molecular dynamic trajectories. In the latter approach vibrational modes are found, either by the quasi-harmonic approximation [Horiuchi 91, Kitao 91] or by *essential dynamics* [Amadei 93] which use a principal component analysis of the covariance matrix of atomic fluctuations to obtain force constants [van Aalten 97, Rueda 07].[3] The studies indicate that there is not a one-to-one correspondence between NMA modes and the modes of the quasi-harmonic approximation but there is an overlap on the level of subspaces spanned by the 5-10 lowest frequency modes [Kitao 91, Hayward 94, van Aalten 97, Rueda 07].

Structure refinement The idea of a dynamically important subspace has been successfully applied in several refinement schemes where it is used restrict the search space. Examples are refinement based on data from X-ray structures [Kidera 90, Kidera 92, Delarue 04] or cryo-electron microscopy [Tama 03a, Hinsen 05], the calculation of NMR order parameters [Sunada 96], and applied to protein-docking by refining the energy at the protein-ligand interface [Lindahl 05].

Normal modes and evolution Identifying a sparse network of evolutionary conserved residues involved in allosteric[4] signalling/regulation [Lockless 99, Süel 03], Zheng and coworkers find that the functionally relevant low-frequency modes are the most robust against sequence variations [Zheng 06]. They use this to explain the conservation of motions related to function despite large variations in sequence.

[3] The terminology *important subspace* to denote the subspace in which large-scale conformational changes take place was introduced by Gō *et al* [Hayward 95]. Amadei *et al* instead use *essential subspace* [Amadei 93].

[4] From Greek *allos* other and *steric* shape/three dimensional.

Corroborating evidence for the "mode conservation" is given in [Qian 04]. Here quasi-harmonic modes of a family of homologous proteins are used to define evolutionarily favored sampling directions which are then successfully used in an energy refinement scheme.

Remark 6.2 *Basically normal modes are just another set of coordinates and as such, the full set can be used to describe any change in Cartesian coordinates, which then just amounts to a change of basis. NMA is only useful if a few eigenmodes are sufficient to model the conformational changes we are interested and, as indicated above, all applications require this. Whether or not this is the typical situation is therefore a pertinent question, one that we return to in the following sections.*

Normal mode analysis: (Why) does it make sense to use it?

There are numerous reasons why a quadratic potential *in vacuo* using a crystal structure as equilibrium configuration would not provide any biologically relevant information about a protein. Nevertheless, the large body of work appearing since the early 80's indicate that it does but we should be aware of its limitations and careful when interpreting the results.

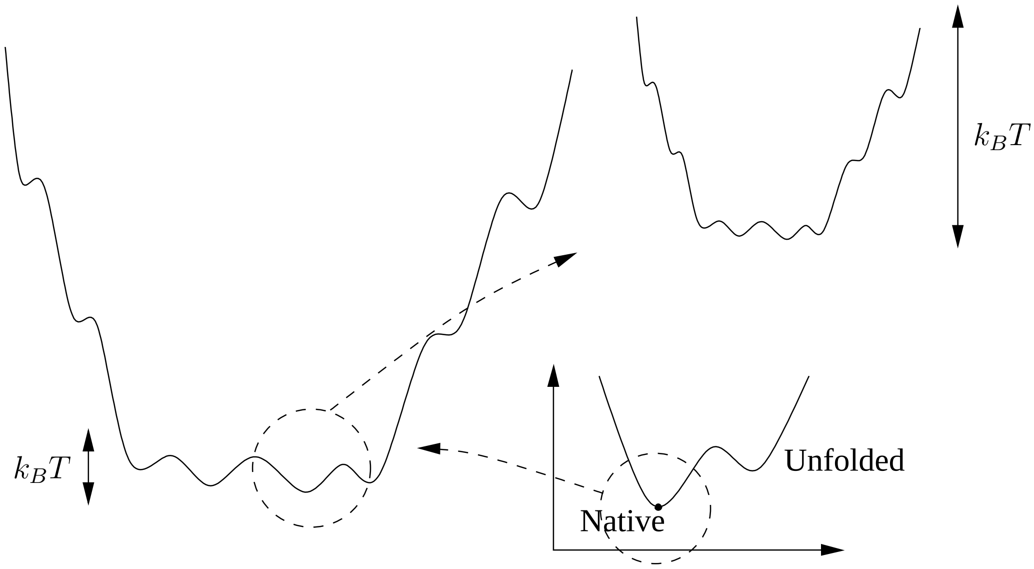


Figure 6.1: Already in a simple two-state model of protein folding the protein energy landscape is obviously not globally harmonic. Here we mostly consider conformational variations in the structure of the native protein but even in this local setting the valleys are not harmonic. Finally, a further host of substates is introduced due to viscous effects. See text for details.

We now address the most pertinent questions concerning the biological/physical relevance of the model.

Energy landscape with multiple minima For small molecules at room temperature a quadratic approximation of the potential energy is quite reasonable but it is not obvious why it should hold for biological macromolecules. The harmonic energy landscape in Eq. (6.1) is a very strong assumption and far from correct, even in vacuum, as both experimental [Nevo 06] and computational [Noguti 82, Elber 87, Frauenfelder 91] work demonstrate.

The energy landscape of a native protein is rugged, with a multitude of nearly iso-energetic conformational substates thermally accessible at room temperature. This is illustrated in Fig. 6.1. The multiple minima are related to structural changes such as the secondary structure movements observed in myosin which are accompanied by side chain rearrangements (to preserve compactness of the core) [Elber 87], or more generally, local deformations of the backbone [Gō 89] and changes in rotamer states [Kitao 98].

Nevertheless, computational studies indicate that the important subspace spanned by low-frequency modes is largely conserved as the protein moves between conformational substates [Hayward 94, van Vlijmen 99]. The expected anharmonic motion, involving large parts of the protein backbone, then takes place within this subspace [Kitao 98, Kitao 99].

NMA in vacuum Solvent leads to viscous effects such as protein-water collision and hydrogen bonding between water molecules and atoms on the surface of the protein. Both effects introduce a further host of conformational substates (see previous paragraph and Fig. 6.1).

By introducing diffusion into an effective harmonic potential it has been possible to reproduce experimental observables in several cases[5]: The time-dependent neutron scattering function based on data from a C-phycocyanin dimer by including Brownian motion [Hinsen 00], or the spectral densities for melittin, bpti, and lysozyme using Langevin modes [Kitao 91, Hayward 93, Kitao 98] as introduced in [Lamm 86]. The picture that emerges is the following

- The envelope of minima with solvent-induced substates is well approximated by the corresponding normal mode (vacuum) potential [Kitao 91, Hinsen 00].
- The conformational substates are nearly harmonic and almost identical [Nishikawa 87, Kitao 98, van Vlijmen 99].
- Protein conformational dynamics can be understood as a superposition of intra-minimum vibrational motion, typically governed by a "small" set of (vacuum) normal modes on a short time-scale, and long term inter-minima diffusive motion [Kitao 91, Kitao 98, Hinsen 00].

[5] The intra-minima conformational substates introduced by the presence of solvent act as an effective friction force on longer time-scales [Hinsen 00].

The important subspaces found in the solvent and in the vacuum model thus appear to be largely identical [Hayward 95].

The protein crystal's effect on collective motion Some proteins are much affected by the crystal environment and lose their functionality, e.g. hemoglobin where the cooperative nature of oxygen binding is lost [Mozzarelli 91] or the loss of flexibility in the central helix of calmodulin [Ikura 92]. Furthermore regularization of the crystal structure found in the Protein DataBank [Berman 03] often leads to considerable RMSD indicating that it is not the *in vivo* (or *vitro*) equilibrium configuration [Wako 04].[6] On the other hand, many proteins *do* retain functionality in a crystal and in some applications, such as refinement from X-ray data, it is not even an issue. The documented exploits of the Tirion potential makes the crystal structure a suitable choice of equilibrium configuration. It provides an automatic assignment of equilibrium structure and at the same time removes the most CPU-demanding part of traditional normal mode analysis.

II Separating internal and external degrees of freedom

To study the intrinsic properties of a protein the underlying model must be formulated in terms of internal coordinates. In other words, the external degrees of freedom in the form of global translation and rotation of the entire structure, think of a molecule suspended in solution, must be removed from the description.

Typically internal coordinates in molecules are separated into the bond stretching, bond bending, and dihedral (or torsion) angles shown in Fig. 6.2. For a nonlinear chain molecule consisting of N atoms this amounts to

$$\underbrace{(N-3)}_{\text{Dihedral angles}} + \underbrace{(N-2)}_{\text{Bond bending}} + \underbrace{(N-1)}_{\text{Bond stretching}} = 3N-6, \tag{6.3}$$

internal degrees of freedom. This is precisely the number of Cartesian coordinates necessary to specify the position of the N atoms, minus the 6 used to define a global rigid body transformation of the structure.

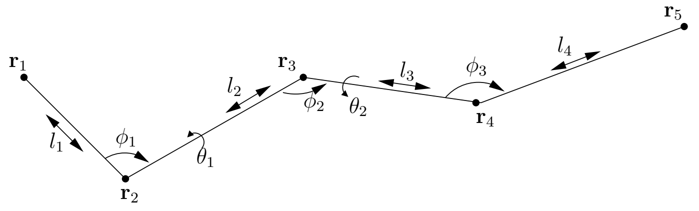
Introducing the moving coordinate system

In a rigid body the principal axes of the inertia tensor can be used to define a moving coordinate system that separate translational and rotational terms in

[6] In [Wako 04] is reported the following RMSD between regularized structures and crystal structures in the Protein DataBank

% of structures	60	25	15
r.m.s.d (Å)	<2	2-4	>4

.



Euclidean coordinates - rotation and translation: $3 \cdot 5 - 6 = 9$

Internal coordinates: $4 + 3 + 2 = 9$

Figure 6.2: A typical separation of internal coordinates into bond stretching $\{l_i\}$, bond bending $\{\phi_j\}$, and dihedral (or torsion) angles $\{\theta_k\}$ defined by 2,3 and 4 consecutive atoms respectively. The number of internal degrees of freedom is equal to the number of Cartesian coordinates necessary to specify the position of N atoms, minus the 6 defining a rigid body transformation the entire molecule.

the kinetic energy [Landau 88]. Alas, vibrations couple to both translation and rotation and it is not obvious that a similar set of coordinates can be defined in the presence of vibrations. In fact, it is *not* possible in general, but for a system performing small oscillations about a stable equilibrium a set of coordinates *do* exist [Eckart 35, Sayvetz 39]. The situation can be seen as an expansion in the displacement vector

$$\boldsymbol{\eta}_\alpha = \mathbf{r}_\alpha - \mathbf{r}^0_\alpha, \quad \alpha = 1, \ldots, N, \tag{6.4}$$

around the rigid body problem. Here $\mathbf{r}^0_\alpha$ and $\mathbf{r}_\alpha$ is the equilibrium and instantaneous position, respectively, of the α'th atom.

We take the center of mass of the molecule in equilibrium

$$\mathbf{Y}^0 = \frac{\sum_\alpha m_\alpha \mathbf{r}^0_\alpha}{\sum_\alpha m_\alpha}, \tag{6.5}$$

to be the center of our fixed-space (inertial) coordinate system. The moving coordinate system is given by a set of (positive) orthonormal basis vectors, $\{\boldsymbol{\varepsilon}_i\}_{i=1,2,3}$, situated at $\mathbf{Y}$, the center of mass of the vibrating molecule, and oriented by the three Euler angles. This is illustrated in Fig. 6.3.

Specification of instantaneous position of the moving coordinate system requires 6 scalar relations amongst the $3N$ coordinates. What is left are the $3N - 6$ internal coordinates as mentioned above. We take the 6 relations to be the so-called first and second Eckart-Sayvetz conditions presented in Section II.1-II.1. These are relations that define a molecule-fixed coordinate system, minimizing the coupling between rotational and vibrational degrees of freedom, and completely separating them in the limit of infinitesimal vibrations [Eckart 35, Sayvetz 39].

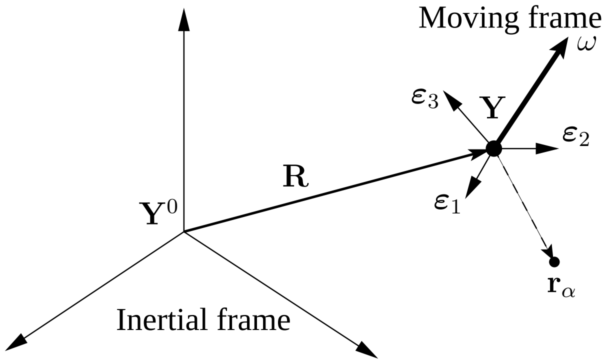


Figure 6.3: Instantaneous configuration of the coordinate system, attached to and moving with the molecule, relative to its inertial frame. $\{\varepsilon_i\}$ *is a set of orthonormal basis vectors defining the moving coordinate system,* ω *is the instantaneous angular velocity.* $\mathbf{Y}$ *is the center of mass of the molecule,* $\mathbf{r}^0_\alpha$ *and* $\mathbf{r}_\alpha$ *the equilibrium position and instantaneous position of the* α*'th atom respectively.*

II.1 Total kinetic energy

Given that no external forces are acting on the molecule, the potential energy depends only on internal coordinates. A separation of internal and external coordinates is therefore only necessary for the kinetic energy. On most points our exposition follows that of [Califano 76].

Let $\boldsymbol{\omega}$ be the instantaneous angular velocity of the moving coordinate system. With $\mathbf{r}_\alpha = \sum_i c_i \varepsilon_i$ we have, in the fixed-space coordinate system,

$$\frac{\mathrm{d}\mathbf{r}_\alpha}{\mathrm{d}t} = \sum_{i=1}^{3} \left(\frac{\mathrm{d}c^i_\alpha}{\mathrm{d}t}\varepsilon_i + c^i_\alpha \frac{\mathrm{d}\varepsilon_i}{\mathrm{d}t} \right) = \sum_{i=1}^{3} \frac{\mathrm{d}c^i_\alpha}{\mathrm{d}t}\varepsilon_i + \boldsymbol{\omega} \times \mathbf{r}_\alpha, \tag{6.6}$$

explicitly showing the two contributions to the velocity. The first is from the motion of the atom relative to the moving coordinate system, the second from the rotation of the moving coordinate system. With Eq. (6.6) the total velocity of the α'th atom in the fixed-space coordinate system can be written

$$\begin{aligned} \mathbf{V}_\alpha = \frac{\mathrm{d}(\mathbf{R}+\mathbf{r}_\alpha)}{\mathrm{d}t} &= \sum_{i=1}^{3} \left(\frac{\mathrm{d}C^i}{\mathrm{d}t}\mathbf{e}_i + \frac{\mathrm{d}c^i_\alpha}{\mathrm{d}t}\varepsilon_i \right) + \boldsymbol{\omega} \times \mathbf{r}_\alpha \\ &\equiv \dot{\mathbf{R}} + \dot{\mathbf{r}}_\alpha + \boldsymbol{\omega} \times \mathbf{r}_\alpha, \end{aligned} \tag{6.7}$$

where a dot represents the time-derivative in the appropriate coordinate system, i.e. fixed-space for $\mathbf{R} = \sum_i C^i \mathbf{e}_i$ and moving for $\mathbf{r}_\alpha$. The total kinetic energy in the inertial system is then

$$\begin{aligned}
2\tilde{T} &= \sum_\alpha m_\alpha \left(\mathbf{V}_\alpha \cdot \mathbf{V}_\alpha\right) \\
&= \Big(\sum_\alpha m_\alpha\Big)\|\dot{\mathbf{R}}\|^2 + \sum_\alpha m_\alpha \|\dot{\mathbf{r}}_\alpha\|^2 + \sum_\alpha m_\alpha \|\boldsymbol{\omega} \times \mathbf{r}_\alpha\|^2 \\
&\quad + 2\dot{\mathbf{R}} \cdot \Big(\sum_\alpha m_\alpha \dot{\mathbf{r}}_\alpha\Big) + 2\dot{\mathbf{R}} \cdot \Big(\boldsymbol{\omega} \times \sum_\alpha m_\alpha \mathbf{r}_\alpha\Big) + 2\sum_\alpha m_\alpha \Big(\boldsymbol{\omega} \times \mathbf{r}_\alpha\Big) \cdot \dot{\mathbf{r}}_\alpha .
\end{aligned} \tag{6.8}$$

The first three terms are the translational, vibrational energy, and rotational energy respectively. The remaining terms are the vibro-translational, rotranslational, and rovibrational coupling energy respectively.

We are now in a position where we can introduce the Eckart-Sayvetz conditions. Each condition comes in two flavors: The conservation law itself and its time derivative.

Remark 6.3 *With a dotted variable denoting a time derivative in the appropriate coordinate system we have, using Eq. (6.4),*

$$\dot{\mathbf{r}}_\alpha = \left(\frac{\mathrm{d}\mathbf{r}_\alpha}{\mathrm{d}t}\right)_{\mathit{moving}} = \left(\frac{\mathrm{d}(\mathbf{r}_\alpha^0 + \boldsymbol{\eta}_\alpha)}{\mathrm{d}t}\right)_{\mathit{moving}} = \left(\frac{\mathrm{d}\boldsymbol{\eta}_\alpha}{\mathrm{d}t}\right)_{\mathit{moving}} = \dot{\boldsymbol{\eta}}_\alpha . \tag{6.9}$$

The first Eckart-Sayvetz condition

Translation of the molecule has no effect on either rotations or vibrations. By requiring that the center of mass should be held fixed

$$\sum_\alpha m_\alpha \mathbf{r}_\alpha \equiv 0, \tag{6.10}$$

the coupling terms of Eq. (6.8) that involve translational degrees of freedom, i.e. $\dot{\mathbf{R}}$, vanish. This is most clearly seen by taking the derivative of Eq. (6.10) with respect to time (in the inertial frame)

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_\alpha m_\alpha \mathbf{r}_\alpha = \sum_\alpha m_\alpha (\dot{\mathbf{r}}_\alpha + \boldsymbol{\omega} \times \mathbf{r}_\alpha) \\
&= \sum_\alpha m_\alpha \dot{\mathbf{r}}_\alpha + \boldsymbol{\omega} \times \sum_\alpha m_\alpha \mathbf{r}_\alpha = \sum_\alpha m_\alpha \dot{\mathbf{r}}_\alpha .
\end{aligned} \tag{6.11}$$

Eq. (6.10) and Eq. (6.11) constitute the first Eckart-Sayvetz condition which ensures that there is no linear momentum from the molecular vibrations [Califano 76]. Together they cancel the vibro-translational and rotranslational coupling terms in the total kinetic energy.

Remark 6.4 *The first Eckart-Sayvetz condition is exact. Nowhere have we used that the displacement $\boldsymbol{\eta}$ is small.*

Second Eckart-Sayvetz condition

As mentioned in Section II, rotational and vibrational degrees of freedom cannot be separated in general, but with

$$\sum_\alpha m_\alpha \mathbf{r}_\alpha^0 \times \mathbf{r}_\alpha = \sum_\alpha m_\alpha \mathbf{r}_\alpha^0 \times (\mathbf{r}_\alpha^0 + \boldsymbol{\eta}_\alpha) = \sum_\alpha m_\alpha \mathbf{r}_\alpha^0 \times \boldsymbol{\eta}_\alpha \equiv 0, \tag{6.12}$$

the coupling is minimized and vanish in the limit of infinitesimal (or linear) displacements $\boldsymbol{\eta}_\alpha$. Differentiation of Eq. (6.12) with respect to time and using Eq. (6.9) gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} \sum_\alpha m_\alpha \mathbf{r}_\alpha^0 \times \mathbf{r}_\alpha = \sum_\alpha m_\alpha \mathbf{r}_\alpha^0 \times \dot{\mathbf{r}}_\alpha = \sum_\alpha m_\alpha \mathbf{r}_\alpha^0 \times \dot{\boldsymbol{\eta}}_\alpha. \tag{6.13}$$

Together Eq. (6.12) and Eq. (6.13) constitute the second Eckart-Sayvetz condition which ensures that, to zeroth order in $\boldsymbol{\eta}$, the angular momentum in the moving frame due to molecular vibrations vanish [Califano 76].

Coriolis energy Using Eq. (6.9) and Eq. (6.13) the rovibrational coupling energy in Eq. (6.8) reduces to

$$\sum_\alpha m_\alpha (\boldsymbol{\omega} \times \dot{\mathbf{r}}_\alpha) \cdot \mathbf{r}_\alpha = \boldsymbol{\omega} \cdot \sum_\alpha m_\alpha \dot{\boldsymbol{\eta}}_\alpha \times (\mathbf{r}_\alpha^0 + \boldsymbol{\eta}_\alpha) = \boldsymbol{\omega} \cdot \Big(\sum_\alpha m_\alpha \dot{\boldsymbol{\eta}}_\alpha \times \boldsymbol{\eta}_\alpha \Big), \tag{6.14}$$

and what remains is the so-called *Coriolis* energy for motion in a rotating coordinate system. With both the angular velocity $\boldsymbol{\omega}$, and the displacements $\boldsymbol{\eta}_\alpha$, being small compared to the vibrational velocities $\dot{\boldsymbol{\eta}}$, the rovibrational coupling can safely be disregarded in a system with small vibrations.

Equivalent formulation of the Eckart-Sayvetz conditions Imposing the conditions Eq. (6.10) and Eq. (6.12) is equivalent to applying a rigid body transformation to $\{\mathbf{r}_\alpha\}$ that minimize the weighted least-squares sum

$$\sum_\alpha m_\alpha \left\| \mathbf{r}_\alpha - \mathbf{r}_\alpha^0 \right\|^2 \quad \text{for} \quad \left\| \mathbf{r}_\alpha - \mathbf{r}_\alpha^0 \right\| \to 0. \tag{6.15}$$

This equivalence is mentioned in [Noguti 83b] and based on [McLachlan 79] but was already demonstrated in [Jørgensen 78].

With the Eckart-Sayvetz conditions we have a molecule-fixed coordinate system where, whenever vibrations result in a translation or rotation of the molecule, the axes reorient themselves so as to eliminate this part of the motion in the best way possible (optimal in the sense defined in the previous sections).

Table 6.1: Characteristic vibrational frequencies in biomolecules ([Schlick 02],p.230). Data are derived from vibrational spectra of alkane molecules.

Vibrational mode	Frequency [cm^{-1}]
Bond stretching[1]	1000-1800
Bond bending[1]	500-600
Dihedral angles[1,2]	300-600

[1] No modes involve hydrogen or sulfide atoms
[2] Only single bonds considered

III Normal mode analysis in dihedral angles

There is no consensus in the literature as to the number of normal modes required to capture the large-amplitude changes in flexible proteins. So *a few* can mean:

- $1-10$ mode(s) are sufficient to represent the bulk of observed conformational variations [Brooks 83, Levitt 85, Tama 01, Cui 04].
- > 100 modes must be used to represent the important subspace of motions potentially involved in function [Hayward 95, Petrone 06].

The structural refinement schemes in [Kidera 90, Kidera 92] use > 100 modes whereas [Delarue 04, Lindahl 05] find $5-15$ to be sufficient. Not surprisingly, the number of modes depends on the protein and application in question. However, it is always significantly smaller than the total number of degrees of freedom in the system.

Independent of the above considerations, there is an obvious way of lowering the number of required modes. Namely, by an appropriate choice of coordinates. In Table 6.1 are listed characteristic vibrational frequencies for the typical degrees of freedom in a biomolecule (see Fig. 6.2). We see, that the dihedral angles, defined in Fig. 6.4, in are the prime movers in the low-energy dynamics.

In the Cartesian coordinate formulation of NMA atoms vibrate freely, irrespective of the underlying (covalent) bond structure. It thus includes variations in bond angles and lengths, typically negligible at low-frequencies. In therefore makes makes sense to formulate NMA in dihedral angles.

Remark 6.5 *In the semi-empirical potentials (see Section III of Chapter 2 and Eq. (6.47)) the different force constants ensure a weighting of the various degrees of freedom. This is not the case in Elastic Network Models, such as Eq. (6.2), using a single-parameter. One might therefore suspect a dihedral angle formulation to be even more relevant in this context.*

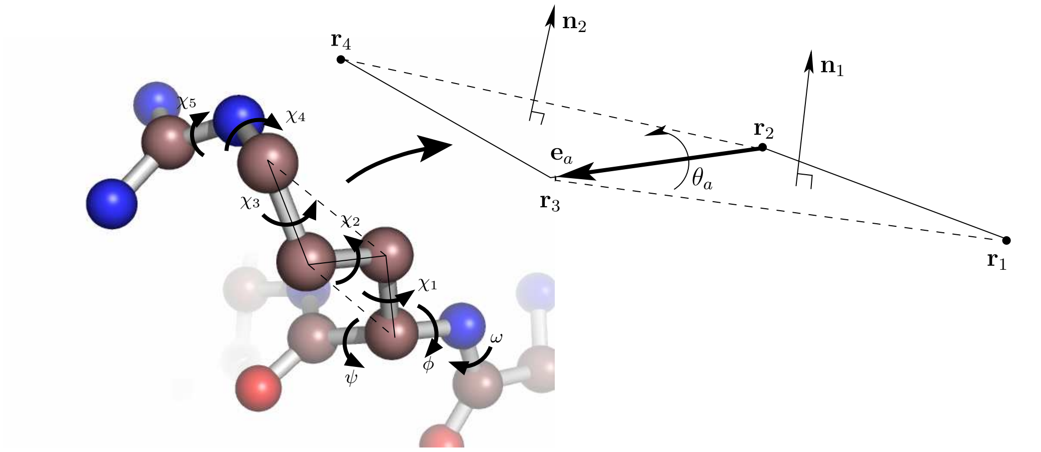


Figure 6.4: In general a dihedral angle, of magnitude θ_a, is defined by four consecutive atoms $\mathbf{r}_1, \ldots, \mathbf{r}_4$. The bond vector $\mathbf{e}_a$ is a unit vector along the bond separating the dihedral planes. Here is shown the dihedral angle structure of the amino acid arginine with θ_a corresponding to χ_2. Hydrogen atoms are left out for simplicity.

Normal mode analysis in dihedral angles has several advantages. On average it reduces the number of degrees of freedom by a factor of 8 [Kitao 94]. Furthermore, as we will see in Section VI.3, it has the very nice property of preserving the stereochemistry of the protein significantly better than its Cartesian cousin. The implementation is somewhat more elaborate than its Cartesian equivalent but all technical aspects has been worked out in a series of articles by Gō and coworkers [Noguti 83a, Abe 84, Sunada 95].

We can now give a more concise formulation of our attempt to answer the question initially posed: Can conformational change be described by only a few normal modes?

- First of all, it depends on the application.
- Second, it depends on the protein which was as also found in [Petrone 06]. Normal mode analysis is not well-suited to study localized conformational changes.
- Third, the model should use the appropriate degrees of freedom in terms of coordinates and level of coarse-graining.

[Petrone 06] use Cartesian normal modes, based on the Tirion potential Eq. (6.2) and using the rotational-translational block formulation of [Durand 94], to map between open and closed configurations of four proteins (See Section VI.1). They find in all four cases, that the first 20 normal modes contribute less than 50% to the observed conformational changes.

Here we reproduce their study in dihedral angles to see if this, seemingle more suitable choice of coordinates, will yield better results.

To the best of our knowledge we report on the first implementation of a Tirion potential in dihedral coordinates. To evaluate our result in the more familiar Cartesian coordinate space (CCS), and to compare with the standard formulation in mass-weighted Cartesian coordinates, we have also implemented the expansion of Cartesian coordinates to second order in dihedral angles presented in [Sunada 95].

III.1 Kinetic energy in dihedral angles

Now consider a molecular system at energies characteristic for small vibrations such that Coriolis term in Eq. (6.14) can safely be disregarded. In this case the Eckart-Sayvetz conditions define a molecule-fixed coordinate system that separates the (internal) vibrational degrees of freedom from the (external) translational and rotational degrees of freedom. The internal kinetic energy in the molecule-fixed coordinate system is then given by

$$T = \frac{1}{2}\sum_{\alpha} m_\alpha \|\dot{\mathbf{r}}_\alpha\|^2, \tag{6.16}$$

where the sum runs over atoms in the molecule.

From hereon, we restrict ourselves to dihedral angles, $\theta_1, \ldots, \theta_M$, and small displacements of the form

$$\theta_a = \theta_a^0 + \Delta\theta_a, \quad a = 1, \ldots, M. \tag{6.17}$$

In this case Eq. (6.16) takes the form

$$T = \frac{1}{2}\sum_{\alpha} m_\alpha \left(\dot{\mathbf{r}}_\alpha \cdot \dot{\mathbf{r}}_\alpha\right) = \frac{1}{2}\sum_{a,b} \underbrace{\left(\sum_{\alpha} m_\alpha \frac{\partial \mathbf{r}_\alpha}{\partial \theta_a} \cdot \frac{\partial \mathbf{r}_\alpha}{\partial \theta_b}\right)}_{\mathbf{H}_{a,b}} \Delta\dot{\theta}_a \Delta\dot{\theta}_b, \tag{6.18}$$

where we have used the dihedral angle equivalent of Eq. (6.9)

$$\dot{\theta}_a = \frac{\mathrm{d}(\theta_a - \theta_a^0)}{\mathrm{d}t} = \Delta\dot{\theta}_a, \tag{6.19}$$

to obtain a kinetic energy in terms of dihedral angle *displacements.*

Cartesian displacements in terms of dihedral angles

Typically the kinetic energy matrix $\mathbf{H}$, defined in Eq. (6.18), is determined either: (i) analytically, using the so-called Wilson s-vector method, where the Eckart-Sayvetz coordinate system is constructed explicitly [Wilson Jr. 55];

or (ii) numerically, by minimizing Eq. (6.15) [Levitt 85].[7] With the formulation in dihedral angles we do neither, and instead implement the elegant approach presented [Noguti 83a, Sunada 95]. Here the Eckart-Sayvetz conditions are used to find the appropriate translation and rotation

$$\Delta\mathbf{r}_\alpha = \Delta\mathbf{T}\bigg|_{\mathbf{r}_\alpha=\mathbf{r}_\alpha^0} + \mathbf{\Omega}\bigg|_{\mathbf{r}_\alpha=\mathbf{r}_\alpha^0} \times \mathbf{r}_\alpha + \mathcal{O}(\Delta\theta^3), \tag{6.20}$$

minimizing Eq. (6.15). This is subsequently used, not only determine an expression for $\partial\mathbf{r}_\alpha/\partial\theta_a$ (and hence $\mathbf{H}$), but also to convert dihedral eigenvectors into Cartesian coordinates. This is done by a second order Taylor expansion in $\Delta\theta$ around the equilibrium structure,

$$\begin{aligned}\mathbf{r}_\alpha\{\boldsymbol{\theta}^0 + \Delta\boldsymbol{\theta}\} =& \mathbf{r}_\alpha^0 + \sum_a \frac{\partial\mathbf{r}_\alpha}{\partial\theta_a}\bigg|_{\mathbf{r}_\alpha=\mathbf{r}_\alpha^0}\Delta\theta_a + \frac{1}{2}\sum_{a,b}\frac{\partial^2\mathbf{r}_\alpha}{\partial\theta_a\partial\theta_b}\bigg|_{\mathbf{r}_\alpha=\mathbf{r}_\alpha^0}\Delta\theta_a\Delta\theta_b + \mathcal{O}(\Delta\theta^3)\\ \approx& \mathbf{r}_\alpha^0 + (\mathbf{K}\boldsymbol{\Delta\theta})_\alpha + \frac{1}{2}(\boldsymbol{\Delta\theta}^T\mathbf{L}\boldsymbol{\Delta\theta})_\alpha\end{aligned} \tag{6.21}$$

where

$$(\mathbf{K})_{ia} = \frac{\partial(\mathbf{r}_\alpha)_k}{\partial\theta_a}\bigg|_{\mathbf{r}_\alpha=\mathbf{r}_\alpha^0}, \quad (\mathbf{L})_{iab} = \frac{\partial^2(\mathbf{r}_\alpha)_k}{\partial\theta_a\partial\theta_b}\bigg|_{\mathbf{r}_\alpha=\mathbf{r}_\alpha^0}, \\ i = 3(\alpha-1)+k, \quad k = 1,2,3 \tag{6.22}$$

is a $(3N \times M)$-matrix and a $(3N \times M \times M)$-tensor respectively. We return to this in Section IV when we solve the classical equations of motions.

Remark 6.6 *[Sunada 95] mentions, that using*

$$T = \frac{1}{2}\sum_{a,b}\left(\sum_\alpha m_\alpha \frac{\partial\mathbf{r}_\alpha}{\partial\theta_a}\bigg|_{\mathbf{r}_\alpha=\mathbf{r}_\alpha^0} \cdot \frac{\partial\mathbf{r}_\alpha}{\partial\theta_b}\bigg|_{\mathbf{r}_\alpha=\mathbf{r}_\alpha^0}\right)\Delta\dot{\theta}_a\Delta\dot{\theta}_b,$$

with $\mathbf{r}_\alpha$ *is expanded to first order in* $\Delta\theta$*, is inconsistent with the use of Eq. (6.21). However, they show numerically that it is still an improvement over a purely first order implementation [Sunada 95].*

Remark 6.7 *The expressions for* $\Delta\mathbf{T}_\alpha$ *and* $\boldsymbol{\omega}_\alpha$ *in Eq. (6.20) are somewhat involved and are not presented here. Instead we refer the interested reader to [Noguti 83b, Sunada 95].*

[7] For example by singular value decomposition, as described in [Koehl 06].

III.2 The Tirion potential

Following [Tirion 96] we consider the single-parameter quadratic potential

$$V = \frac{K}{2} \sum_{\alpha<\beta} \left(\|\mathbf{r}_\alpha - \mathbf{r}_\beta\| - \|\mathbf{r}^0_\alpha - \mathbf{r}^0_\beta\| \right)^2 . \tag{6.23}$$

Here $\mathbf{r}^0_\alpha$ and $\mathbf{r}^0_\beta$ are the coordinates of the α'th and β'th atom in the protein crystal structure. K is a universal force constant. Only atoms satisfying

$$\|\mathbf{r}^0_\alpha - \mathbf{r}^0_\alpha\| < C_{\text{cut-off}} \sim 5 - 10\text{Å}, \tag{6.24}$$

are connected by springs which mimics the decay of interactions with increasing distance. Following most authors we use a cut-off in the form of a step-function [Tirion 96, Tama 01] but also exponential [Hinsen 98] and polynomial [Hinsen 00] functions have been used.

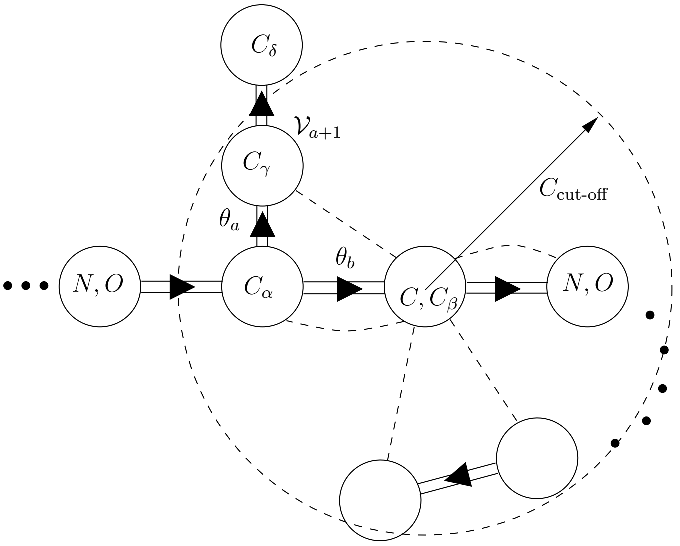


Figure 6.5: Spring network defined by the covalent bonds of the amino acid glutamine. Only heavy (non-hydrogen) atoms are considered. Atoms connected by hatched lines are joined by springs. $C_{\text{cut-off}}$ define a sphere around each atom that determining which atoms it should be connected to. The network is used to construct the hessian, $\mathbf{F}$, of the potential energy in the iterative algorithm of [Abe 84].

Only allowing for variations in dihedral angles, and with the cut-off given in Eq. (6.24), we get the setup illustrated in Fig. 6.5. The covalent bonds of the molecule define a "canonical graph" (more precisely a tree) that indicates how the spring network is disturbed by dihedral angle variations.

Potential Energy in Dihedral Angles Using the notation $\mathbf{r}_{\alpha\beta} \equiv \mathbf{r}_\alpha - \mathbf{r}_\beta$ we get

$$\begin{aligned}
&\|\mathbf{r}_{\alpha\beta}(\theta_1+\Delta\theta_1,\ldots,\theta_M+\Delta\theta_M)\| &-& \quad \|\mathbf{r}^0_{\alpha\beta}\| \\
\simeq\ &\|\mathbf{r}^0_{\alpha\beta}\|\sqrt{1+\frac{\mathbf{r}^0_{\alpha\beta}}{\|\mathbf{r}^0_{\alpha\beta}\|^2}\cdot\sum_a\frac{\partial\mathbf{r}^0_{\alpha\beta}}{\partial\theta_a}\Delta\theta_a} &-& \quad \|\mathbf{r}^0_{\alpha\beta}\| \simeq \frac{\mathbf{r}^0_{\alpha\beta}}{\|\mathbf{r}^0_{\alpha\beta}\|}\cdot\sum_a\frac{\partial\mathbf{r}^0_{\alpha\beta}}{\partial\theta_a}\Delta\theta_a,
\end{aligned}$$

and expanding Eq. (6.23) to second order in dihedral angle displacements we have

$$V=\frac{1}{2}\sum_{a,b}\underbrace{K\left(\sum_{\alpha<\beta}\frac{\mathbf{r}^0_{\alpha\beta}}{\|\mathbf{r}^0_{\alpha\beta}\|}\cdot\frac{\partial\mathbf{r}^0_{\alpha\beta}}{\partial\theta_a}\right)\left(\sum_{\alpha<\beta}\frac{\mathbf{r}^0_{\alpha\beta}}{\|\mathbf{r}^0_{\alpha\beta}\|}\cdot\frac{\partial\mathbf{r}^0_{\alpha\beta}}{\partial\theta_b}\right)}_{\mathbf{F}_{a,b}}\Delta\theta_a\Delta\theta_b, \tag{6.25}$$

where the $(M \times M)$-matrix $\mathbf{F}$ is the Hessian of V.

IV Solving the Euler-Lagrange equations: Normal mode coordinates

We are now in a position to solve the classical equations of motion. This is best done in the Lagrangian formulation and our derivation closely follows that of [Levitt 85]. Using Eq. (6.18) and Eq. (6.25) we get a Lagrangian

$$\begin{aligned}
\mathcal{L}=T-V&=\frac{1}{2}\sum_{a,b}\left(\mathbf{H}_{a,b}\Delta\dot\theta_a\Delta\dot\theta_b-\mathbf{F}_{a,b}\Delta\theta_a\Delta\theta_b\right)\\
&=\frac{1}{2}\Delta\dot{\boldsymbol\theta}^T\mathbf{H}\Delta\dot{\boldsymbol\theta}-\frac{1}{2}\Delta\boldsymbol\theta^T\mathbf{F}\Delta\boldsymbol\theta,
\end{aligned}\tag{6.26}$$

which is symmetric in a and b, and quadratic in dihedral angle displacements and their time derivatives. Substituting this into the Euler-Lagrange equations and making use of the symmetry, we get

$$\begin{aligned}
0&=\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial\mathcal{L}}{\partial(\Delta\dot\theta_i)}\right]-\frac{\partial\mathcal{L}}{\partial(\Delta\theta_i)}\\
&=\sum_a\left(\mathbf{H}_{i,a}\Delta\ddot\theta_a+\mathbf{F}_{i,a}\Delta\theta_a\right),\quad i=1,\ldots,M.
\end{aligned}\tag{6.27}$$

A general solution to this set of linear equations is a superposition of simple oscillations of the form

$$\Delta\theta_a(t)\propto(\mathbf{w})_a\cos(\omega t+\phi), \tag{6.28}$$

where $(\mathbf{w})_a$ determines the relative amplitude of each coordinate, $\Delta\theta_a$, in the oscillation of frequency ω. The absolute amplitude and the phase factor

ϕ depend on the initial conditions. Substituting into the Euler-Lagrange equations Eq. (6.27) gives

$$0 = \sum_a \left(\mathbf{F}_{i,a}(\mathbf{w})_a - \omega^2 \mathbf{H}_{i,a}(\mathbf{w})_a\right) \cos(\omega t + \phi), \quad i = 1, \ldots, M. \tag{6.29}$$

This holds for all values of t and we get a generalized eigenproblem

$$\mathbf{F}\mathbf{w} = \boldsymbol{\omega}^2 \mathbf{H}\mathbf{w}, \tag{6.30}$$

with non-trivial solutions only when

$$\det(\mathbf{F} - \boldsymbol{\omega}^2 \mathbf{H}) = 0, \tag{6.31}$$

which is an M'th order polynomial in ω^2 with M distinct roots, $\{\lambda_i = \omega_i^2\}_{i=1,\ldots,M}$, in the absence of any symmetries in the model. The eigenvalue equation 6.30 can then be written as

$$\begin{gathered} \mathbf{F}\mathbf{W} = \boldsymbol{\Omega}^2 \mathbf{H}\mathbf{W}, \\ \mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_N], \quad \boldsymbol{\Omega}^2 = \mathrm{Diag}(\omega_1^2, \ldots, \omega_M^2), \end{gathered} \tag{6.32}$$

where $\mathbf{w}_i$ is the eigenvector associated to the eigenvalue λ_i. An equivalent formulation of the problem is as the simultaneous digonalization of two bilinear forms

$$\mathbf{W}^T \mathbf{F}\mathbf{W} = \boldsymbol{\Omega}^2, \quad \mathbf{W}^T \mathbf{H}\mathbf{W} = \mathbb{I}, \tag{6.33}$$

where $\mathbb{I}$ is the identity matrix.

Remark 6.8 *In Cartesian coordinates, where the kinetic energy matrix* $\mathbf{H}$ *is diagonal, the second term in Eq. (6.33) implies mass-weighted coordinates. Otherwise*

$$\mathbf{w}^T \mathbf{H}\mathbf{w} = \mathbf{M}, \quad \mathbf{M} = \mathit{Diag}(m_1, \ldots, m_N).$$

Introducing normal coordinates The complete solution to Eq. (6.27) in dihedral angle coordinates is given by

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}^0 + \sum_{k=1}^{M} A_k \mathbf{w}_k \cos(\omega_k t + \phi_k), \tag{6.34}$$

where again, ϕ_k and the absolute amplitude A_k depend on the initial conditions. Using the eigenvectors, $\{\mathbf{w}_i\}_{i=1,\ldots,M}$, we define a transformation from dihedral angle displacements into *normal coordinates* $\{Q_i\}_{i=1,\ldots,M}$, given by

$$\Delta\boldsymbol{\theta}(t) = \mathbf{W}\mathbf{Q}(t), \quad Q_i(t) = A_k \cos(\omega_k t + \phi_k), \tag{6.35}$$

which follows immediately from Eq. (6.34). Introducing the coordinate transformation into Eq. (6.26), and using Eq. (6.33), we get

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}(\mathbf{W}\dot{\mathbf{Q}})^T\mathbf{H}(\mathbf{W}\dot{\mathbf{Q}}) - \frac{1}{2}(\mathbf{W}\mathbf{Q})^T\mathbf{F}(\mathbf{W}\mathbf{Q}) \\ &= \frac{1}{2}\dot{\mathbf{Q}}^T\dot{\mathbf{Q}} - \frac{1}{2}\mathbf{Q}^T\mathbf{\Omega}^2\mathbf{Q} \\ &= \frac{1}{2}\sum_k\left(\dot{Q}_k^2 - \omega_k^2 Q_k\right),\end{aligned} \tag{6.36}$$

i.e. a complete separation of the Lagrangian into simple vibrational modes of a single frequency.

Remark 6.9 *Higher order terms in the potential energy leads to a coupling of the individual modes [Moritsugu 00],([Landau 88],p.136).*

Thermal amplitudes

The general solution to the eigenvalue problem Eq. (6.34), and the coordinate transformation Eq. (6.35), both contain a scale-factor, A_k, that depends on the initial conditions. As we consider thermal vibrations in a protein, A_k should somehow depend on the system temperature.

The precise form of the dependence can be derived from the equipartition law of classical statistical mechanics ([Landau 63],p.129). It states that a degree of freedom, that appears only in the form of a quadratic term in the Lagrangian (see Eq. (6.26)), contributes $k_BT/2$ to the average energy. The time-averaged potential energy of the k'th mode is then

$$\begin{aligned}\frac{1}{2}k_BT &= \frac{1}{2}\omega_k^2\langle Q_k^2(t)\rangle_t = \frac{1}{4}\omega_k^2 A_k^2, \\ A_k &= \sqrt{\frac{2k_BT}{\omega_k^2}}.\end{aligned} \tag{6.37}$$

where $\langle\cdot\rangle_t$ denotes averaging with respect to time. Besides the explicit temperature dependence we also see that low-frequency modes have the largest amplitudes, with $A_k \sim 1/\omega_k$.

Eigenmodes in Cartesian coordinates

(Mass-weighted) Cartesian coordinate eigenvectors are orthonormal since $\mathbf{w}^T\mathbf{H}\mathbf{w} = \mathbf{w}^T\mathbf{w} = \mathbb{I}$ (see Remark 6.8). Alas, this is not true for dihedral angles eigenvectors transformed to Cartesian coordinate space. With the Taylor expansion in Eq. (6.21) we have an approximate representation of dihedral angle eigenvector $\mathbf{w}_i$, given by

$$\begin{aligned}\mathbf{u}_i &= \mathbf{r}\{\boldsymbol{\theta}^0 + \mathbf{w}_i\} - \mathbf{r}\{\boldsymbol{\theta}^0\} = \sum_a \frac{\partial \mathbf{r}}{\partial \theta_a}(\mathbf{w}_i)_a + \frac{1}{2}\sum_{a,b} \frac{\partial^2 \mathbf{r}}{\partial \theta_a \partial \theta_b}(\mathbf{w}_i)_a(\mathbf{w}_i)_b \\ &= \mathbf{K}\mathbf{w}_i + \frac{1}{2}\mathbf{w}_i^T \mathbf{L}\mathbf{w}_i, \\ \mathbf{u}_i^T \mathbf{u}_j &= \underbrace{\mathbf{w}_i^T \mathbf{H}\mathbf{w}_j}_{\delta_{ij}} + \mathcal{O}(\|\mathbf{w}\|^3),\end{aligned} \tag{6.38}$$

where $\mathbf{r}$ is the $3N$-dimensional vector of atomic coordinates, $\boldsymbol{\theta}^0$ the M-vector of dihedral angles in the crystal structure, and $\mathbf{K}$ and $\mathbf{L}$ are defined in Eq. (6.22). We see that in CCS the eigenvectors are only orthogonal up to first order in dihedral angles.

V The normal mode spectrum as a min-value problem

Following [Petrone 06] we consider the normal mode spectrum. This says how much each normal mode contributes in a representation of a conformational difference. For reasons that will become clear soon, we determine the spectrum in a manner different from [Petrone 06].

The remainder of the chapter contains repeated comparison between normal mode analysis in mass-weighted Cartesian coordinates and in dihedral angles. Following [Kitao 94, Sunada 95] we introduce the notation: (DAS) Dihedral Angle Space and (CCS) Cartesian Coordinate Space, to distinguish the two methods.

The normal mode spectrum: General formulation

Let A and B be two different configurations of the same protein, called the *reference* and the *target* structure with coordinates $\mathbf{r}^A$ and $\mathbf{r}^B$ respectively. By the *conformational difference* between A and B we mean

$$(\mathbf{r}^A_\alpha - \mathbf{r}^B_\alpha)|_{\text{int}}, \tag{6.39}$$

where we compare the atoms present in both crystal structures. This collection of atoms is called the *set of intersection atoms* of order N_{int}. To quantify the conformational difference we consider

$$RMSD^2(A,B) \equiv \frac{1}{N_{\text{int}}}\sum_{\alpha=1}^{N_{\text{int}}} \|\mathbf{r}^A_\alpha - \mathbf{r}^B_\alpha\|^2_{\text{int}}, \tag{6.40}$$

where the sum runs over atoms in the set of intersection atoms.

The question is now to what extent A can be mapped onto B along the eigenmodes of structure A? Let $T_{\mathbf{t}} : \mathbb{R}^L \times \mathbb{R}^{3N_{\text{int}}} \rightarrow \mathbb{R}^{3N_{\text{int}}}$ be a map deforming A using the first L eigenmodes. The problem of finding the individual eigenmode contributions can then be formulated as a min-value problem:

Definition 6.1 *Given a map $T_{\mathbf{t}}$, and two structures of a same protein, A and B, we solve*

$$D_0^2 = \min_{\mathbf{t}\in\mathbb{R}^L} RMSD^2(T_{\mathbf{t}}(A), B) \equiv RMSD^2(T_{\mathbf{t}_0}(A), B), \tag{6.41}$$

where L is the number of normal modes included and $RMSD^2$ is given by Eq. (6.40). The L-dimensional minimizing vector $\mathbf{t}_0$ is then the normal mode spectrum.

If low-frequency normal modes can be used to represent the observed conformational differences between A and B this can be seen observed in the relative magnitudes of the elements in $\mathbf{t}_0$. We call $T_{\mathbf{t}_0}(A)$ the model structure of B denoted by B_{model}.

Remark 6.10 *As both A and B constitute equilibrium structures an obvious generalization of the normal mode spectrum in Definition 6.1 is to include the eigenmodes of both structures. This would resemble the multiple-basin approach of [Maragakis 05]. However, this has not been done. Instead we follow [Petrone 06] in our choice of reference and target structures. This is in line with the results of [Tama 01], where normal modes of the open structures were found to be best at capturing conformational differences.*

Remark 6.11 *Though A and B are two versions of the same protein, there can be quite a difference in the number of atoms in the PDB files. Either because one structure forms a complex or because a structure is not be fully resolved. We use a Needleman-Wunsch algorithm, known from sequence alignment, to find the set of intersection atoms [Needleman 70]. It works on atoms instead of residues, and with a scoring matrix that makes it preferable to have gaps, rather than misalignment.*

Different normal mode spectra

We now consider the two simplest version of a map, $T_{\mathbf{t}}$, between two structures and restrict ourselves to minimization over vectors in Cartesian coordinates.

Linear map between structures We take

$$T_{\mathbf{t}}(\mathbf{A}) = \mathbf{r}_\alpha^A|_{\text{int}} + \sum_{i=1}^{L} \mathbf{u}_i|_{\text{int}} t_i = \mathbf{r}_\alpha^A|_{\text{int}} + \mathbf{U}|_{\text{int}}\mathbf{t}, \tag{6.42}$$

where $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_L]$ is a $(3N \times L)$ matrix consisting of eigenvectors either from CCS normal mode analysis or DAS using a first order transformation into Cartesian coordinates. The minimizer is determined by solving

$$\mathbf{U}|_{\mathrm{int}}\mathbf{t} = (\mathbf{r}^B_\alpha - \mathbf{r}^A_\alpha)|_{\mathrm{int}}, \tag{6.43}$$

which has an exact solution only when we consider the full set of (mass-weighted) eigenmodes from the CCS NMA. In all other cases it is a least-squares problem and still have an unique solution.

Quadratic map between structures Including the second order term in the Taylor expansion Eq. (6.21) we get a map, $T_{\mathbf{t}}$, quadratic in $\mathbf{t}$. Linear motion in dihedral angle space results in curvilinear motion in Cartesian coordinates, as $t_i\mathbf{w}_i$ to second order goes into

$$\begin{aligned} \mathbf{u}_\alpha(t_i) &= \sum_a \frac{\partial \mathbf{r}^A_\alpha}{\partial \theta_a}(\mathbf{w}_i)_a t_i + \frac{1}{2}\sum_{a,b} \frac{\partial^2 \mathbf{r}^A_\alpha}{\partial \theta_a \partial \theta_b}(\mathbf{w}_i)_a t_i (\mathbf{w}_i)_b t_i \\ &= \mathbf{K}^A \mathbf{w}_i t_i + \frac{1}{2}\mathbf{w}_i^T \mathbf{L}^A \mathbf{w}_i t_i^2, \end{aligned} \tag{6.44}$$

so

$$T_{\mathbf{t}}(A) = A + \sum_{i=1}^{L}(\mathbf{K}^A \mathbf{w}_i t_i + \frac{1}{2}\mathbf{w}_i^T \mathbf{L}^A \mathbf{w}_i t_i^2), \tag{6.45}$$

in which case Eq. (6.41) is a fourth order polynomial in L variables.

Remark 6.12 *The map $T_{\mathbf{t}}(A)$ in Eq. (6.45) is not truly quadratic in $\mathbf{t}$, as we ignore cross-terms of the form $t_i t_j$. A more general map would be*

$$T_{\mathbf{t}}(A) = A + \sum_{i=1}^{L}(\mathbf{K}^A \mathbf{w}_i t_i + \frac{1}{2}\sum_j \mathbf{w}_i \mathbf{T}^A \mathbf{L} \mathbf{w}_j t_i t_j),$$

leading to a coupling of dihedral angle vectors in CCS.

Computing the normal mode spectrum

In both the linear and quadratic case, the gradient of the objective function Eq. (6.40) can be calculated. The min-value problem in Eq. (6.41) was then solved using the conjugate-gradient method described in ([Press 97],Chap.10-6).

For a linear map with orthonormal eigenvectors, the minimizer for L modes, $\mathbf{t}_0$, is left unchanged as more modes are added. For a quadratic map, changing the number of eigenvectors change the minimizer. The min-value problem therefore has to be solved for each value of L. In this case, an interpretation of the minimizer as a normal mode spectrum, is less obvious. Instead we look at the decrease in RMSD, i.e. D_0 in Eq. (6.41), as more eigenvectors are added to the subspace.

Remark 6.13 *[Petrone 06] uses the orthogonality of eigenvectors to solve Eq. (6.43), that is*

$$\mathbf{t} = \mathbf{U}|_{int}^{T}(\mathbf{r}_{\alpha}^{B} - \mathbf{r}_{\alpha}^{B})|_{int}.$$

However, orthogonality is not preserved on the set of intersection atoms and so $\mathbf{U}^T\mathbf{U}|_{int}$ has off-diagonal elements. Instead one could: (i) Find an orthonormal basis for the space spanned by $\mathbf{U}|_{int}$. (ii) Treat Eq. (6.43) as a special case of the general min-value problem in Eq. (6.41).

VI Results

In the following we compare normal mode analysis in dihedral angles (DAS) with the formulation in mass-weighted Cartesian coordinates (CCS). First, we examine the proficiency of eigenmodes in representing conformational differences in a set of four proteins. This is done using the normal mode spectrum of the previous section. Second, we study the change in stereochemistry as structures undergo deformations along eigenmode directions. Details on the numerics are provided in Appendix A.

VI.1 Proteins considered

Following [Petrone 06] we have considered conformational differences in a set of four proteins: calmodulin, the NtrC switch, hemoglobin, and myosin. The proteins were chosen because of their diversity in size, function and motion as illustrated by the information listed in Table VI.1. We start this section by giving a brief presentation of each structure.

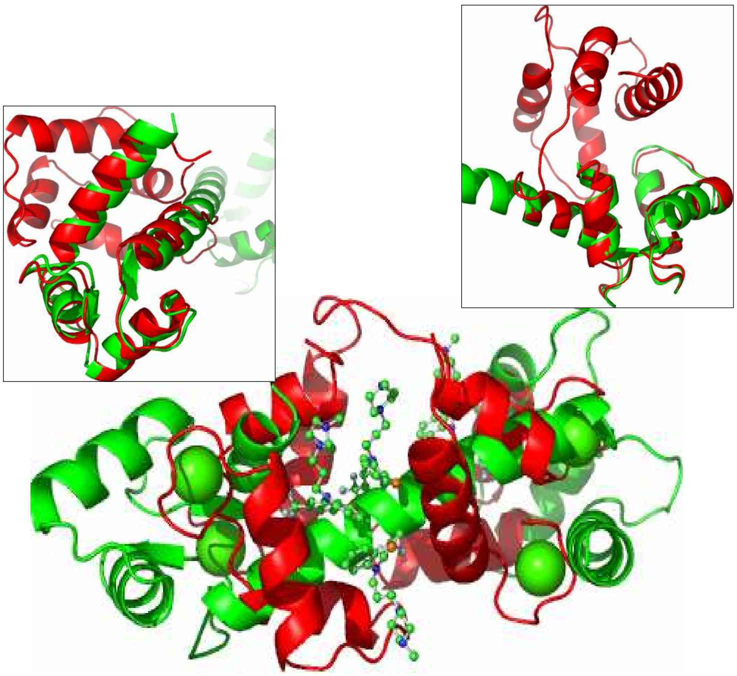
Remark 6.14 *Unfortunately the data for myosin came too late to be included here. Nevertheless, we have chosen to present the protein since it is a natural part of our study.*

Table 6.2: Details on the proteins previously used in [Petrone 06]. Atoms and residues refer to the number present in both reference (A) and (B) target structure.

Protein name	**Calmodulin**	**NtrC**	**Hemoglobin**[1]	**Myosin**
Reference (A)	`1cll`	`1dc7`	`1a3n`	`1fmw`
Target (B)	`1lin/1xa5`	`1dc8`	`1bbb`	`1vom`
No. residues ($A \cap B$)	146	126	288	740
No. atoms ($A \cap B$)[a]	1135/1118	956	2184	5936
No. dihedral angles (A)	717	583	1305	3654
RMSD (A,B)[b]	14.9Å/14.2Å	4.1 Å	2.2Å	5.4Å
RMSD (B,B_{model})[b,c]	4.5Å/3.3Å	3.7Å	1.9Å	-
Movement[d]	Hinge	Shear	Allosteric	Unclassified

[1] Only chain A and B considered. [a] Hydrogen atoms not included. [b] In mass-weighted coordinates.
[c] 100 DAS normal modes used in B_{model}. [d] According to the Molecular Movement Database [Flores 06].

Calmodulin is a **cal**cium **modul**ated prote**in** which acts as a mediator of signals to other proteins based on the calcium concentration. It is a small, 146 residue, protein composed of two globular domains[8] connected by a flexible linker in the form of an α-helix



Reference structure (`1cll`, [Chattopadhyaya 92]) is ligand free (green) whereas each of the targets (red) is complexed with an inhibitor either: (i) trifluoroperazin (TRP), a drug used to treat schizophrenia (`1lin`, [Vandonselaar 94]) or (ii) KAR-2, an anti-tumor drug (`1xa5`, [Horvath 05]).

With an initial RMSD between reference and targets of $\sim 14 - 15$Å it makes *a priori* no sense to compare the structures. However, for the domains we find (see inserts)

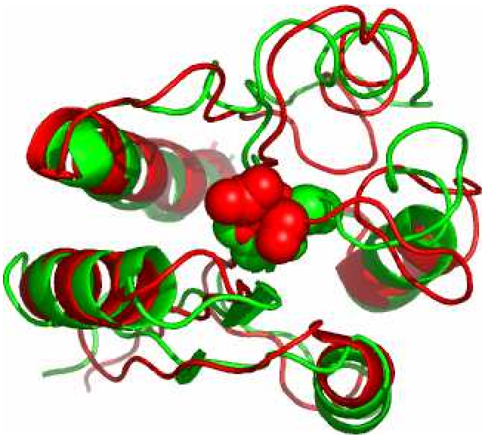
- RMSD 3.3Å and 0.8Å for the first domain, and
- RMSD 0.8Å and 0.9Å for the second domain,

of the TRP and KAR-2 complex respectively. The conformational change upon ligand binding mainly involves the central helix. We shall see, that this makes calmodulin an ideal case for normal mode analysis.

NtrC short for **nit**rogen **r**egulatory protein **C**, is a molecular switch involved in bacterial signal transduction. The switch is activated via the

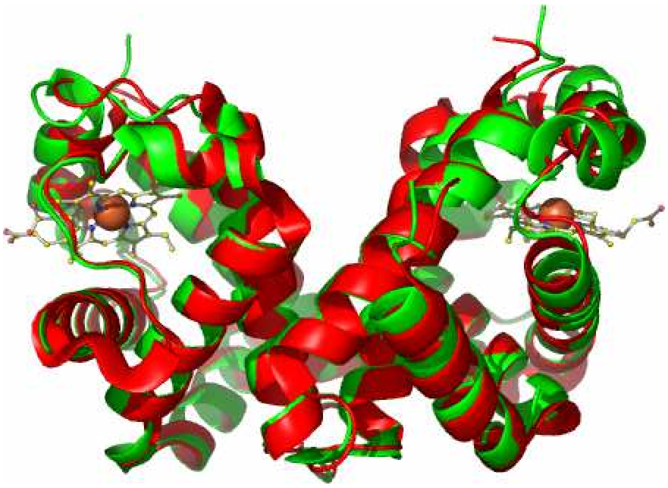
[8] The notion of a protein domain was introduced in Section II.3 of Chapter 2.

phosphorylation[9] of a single residue, *aspartic acid* 54 (here represented by spheres)



Reference (green) is the (unphosphorylated) receiver domain (`1dc7`) and the target (red) the activated, transiently phosphorylated, switch (`1dc8`,[Kern 99]). The difference between the structures is ∼4Å.

Hemoglobin is an iron-containing protein responsible for oxygen transport in the red blood-cells of vertebrates.[10] The reference structure is deoxyhemoglobin (`1a3n`, [Tame 98]), and the target carbonmonoxy-hemoglobin (`1bbb`, [Silva 92]). In either structure we only consider two of four chains



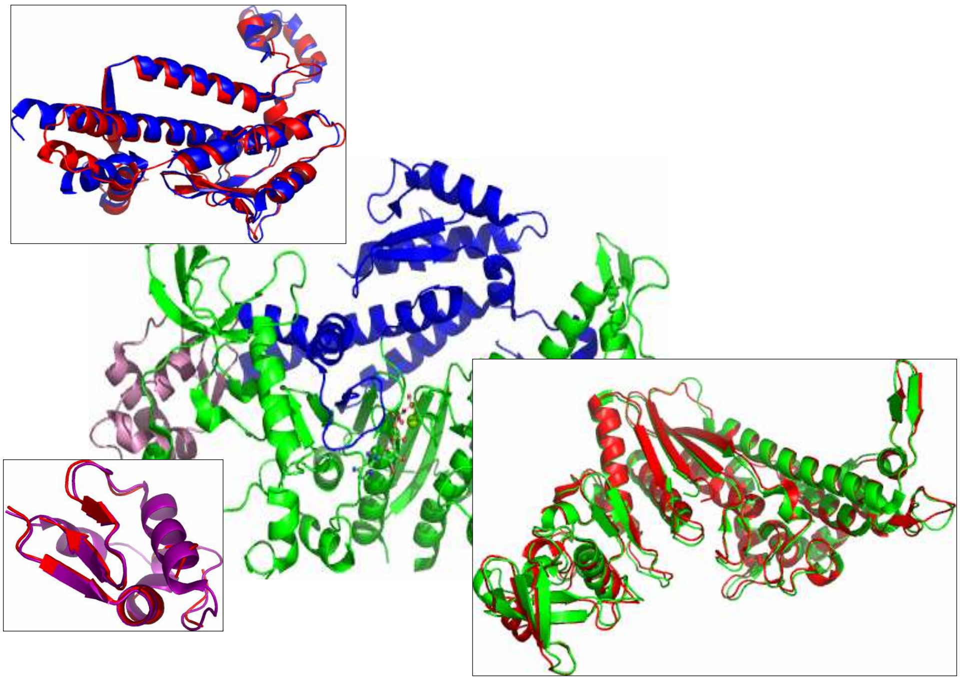
which also shows the position of the heme groups in the ligand free reference structure (green). The target structure is believed to act as a stable

[9] *Phosphorylation* is the addition of a phosphate group, PO_4, to a protein molecule or a small molecule.

[10] The *vertebrates* include animals such as fish, amphibians, reptiles, birds, and mammals.

intermediate between the de-oxidized state and a more common oxidized quaternary structure [Silva 92].

Myosin is a motor protein found in eukaryotic tissue. It transforms energy in the form of ATP molecules into motion and is responsible for muscle movement. It consists of four light and two heavy chains in total. We only consider a single single heavy chain consisting of 6 domains. Reference is the *magnesium-ATP* complex (`1fmw`, [Bauer 00]) and target the *Mg-ADP* complex (`1vom`, [Smith 96]). The domains are colored as follows: 1-4 (green), 5 (blue), and 6 (purple). The same colors are used in the inserts which show alignments with the corresponding domains in `1vom` (red).



RMSD between the two structures is around 5.4Å. However, for the domains we find

- RMSD 1.7Å for domain 1 to 4,
- RMSD 2.9Å for domain 5, and
- RMSD 1.7Å for domain 6.

The structural change is seen to mainly involve inter-domain regions.

VI.2 Structural approximation: Normal mode spectrum

We now consider the proficiency of eigenmodes in representing conformational differences. To do this we use the normal mode spectrum introduced in Section V, and consider

$$D_0^2 = RMSD^2(B_{\mathrm{model}}, B), \qquad B_{\mathrm{model}} = T_{\mathbf{t}_0}(A), \tag{6.46}$$

as a function of the number of normal modes used. This is done for the following five cases: DAS using (i) first and (ii) second order terms in the transformation of vectors into into Cartesian coordinates, (iii) applying a RMSD-minimizing rigid-body transformation *after* having found the minimizer $\mathbf{t}_0$; CCS (iv) pure, and (v) applying a rigid-body transformation as in (iii).

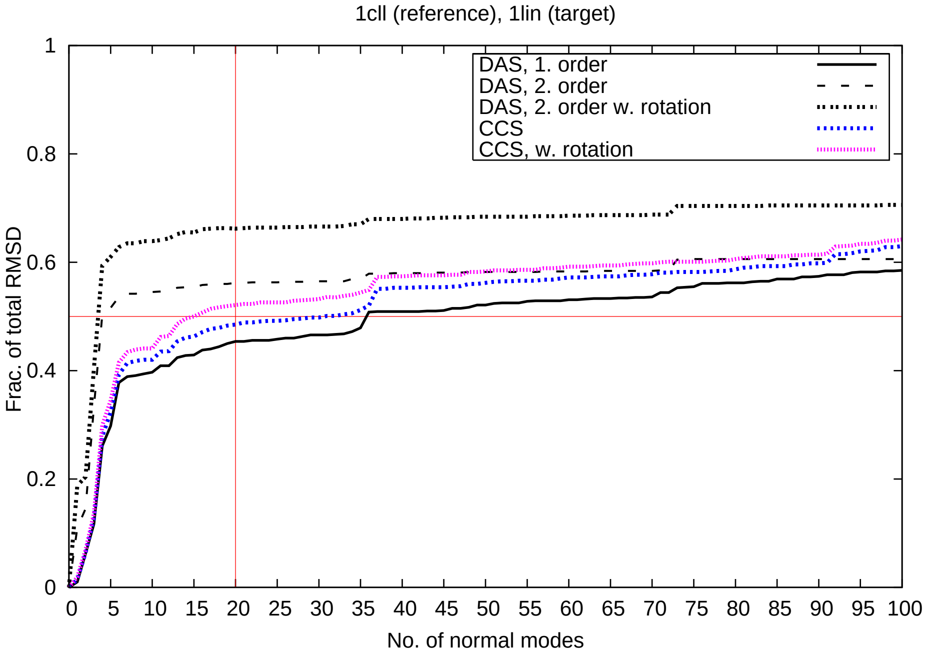
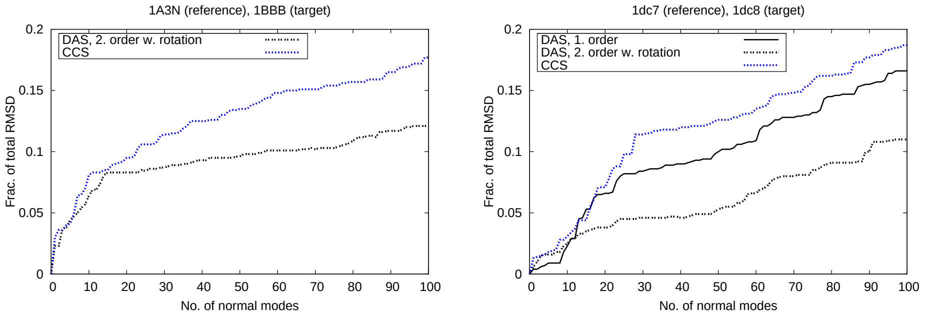


Figure 6.6: Calmodulin. RMSD between model (based on `1cll`*) and target structure* `1lin` *as a function of the number of normal modes included. Straight lines define the benchmark '20 modes and 50% of the total RMSD' of [Petrone 06]. The DAS formulation clearly improves on this (see text). RMSD values are given relative to the reference/target difference.*

For calmodulin, DAS normal modes better represent the conformational difference than CCS modes. This is observed in Fig. 6.6 for the target `1lin`. The bulk of the difference is captured by a few low-frequency modes (the first 4 capture 60% of the conformational difference). Similar results are obtained with `1xa5` as the target. [Tama 01] mention that, when going from the dumbbell-shape of the (open) reference structure to the globular form of the target, atoms most likely follow a curvilinear path. Our results corroborate this.

For hemoglobin and NtrC the results are much less encouraging. Fig. 6.7 contains the best results from each of the NMA formulations. We see that DAS fares worse than CCS, and neither provides a set of low-frequency modes suited to represent the conformational differences. This is in line with [Petrone 06]. It is worthy of notice, that for NtrC the first order version of DAS works better than second order. A reason could be that NtrC is quite different from most protein structures as shall in the next section.



(a) Hemoglobin.

(b) NtrC. Notice that the linear transformation into Cartesian coordinates works better.

Figure 6.7: RMSD between model and target structure (relative to the reference/target difference) as a function of the number of included normal modes. Notice the change of scale compared to Fig. 6.6.

VI.3 Evaluating the stereochemistry of a model structure

The DAS formulation of normal analysis comes with the advantage of preserving the stereochemistry of a protein structure. We now take a closer look at this question by considering deformations along normal mode directions

Secondary structure and C_α-geometry

Many applications rely on the backbone geometry of the protein being preserved. Here the C_α-atoms play a special role, as the loci where sidechains join the backbone [**?**]. It is therefore an important question whether the C_α-geometry is preserved under normal mode deformations.

The Ramachandran plot was introduced in Section II.2 of Chapter 2 as a filter on the (ϕ, ψ)-values in the protein backbone (see Fig. 6.4). A more refined filter is the 'Dictionary of Protein Secondary Structures' (DSSP) program [Kabsch 83]. Here (ϕ, ψ)-values, supplemented by the geometry of the hydrogen-bonding pattern, are used to assign secondary structure type to residues.

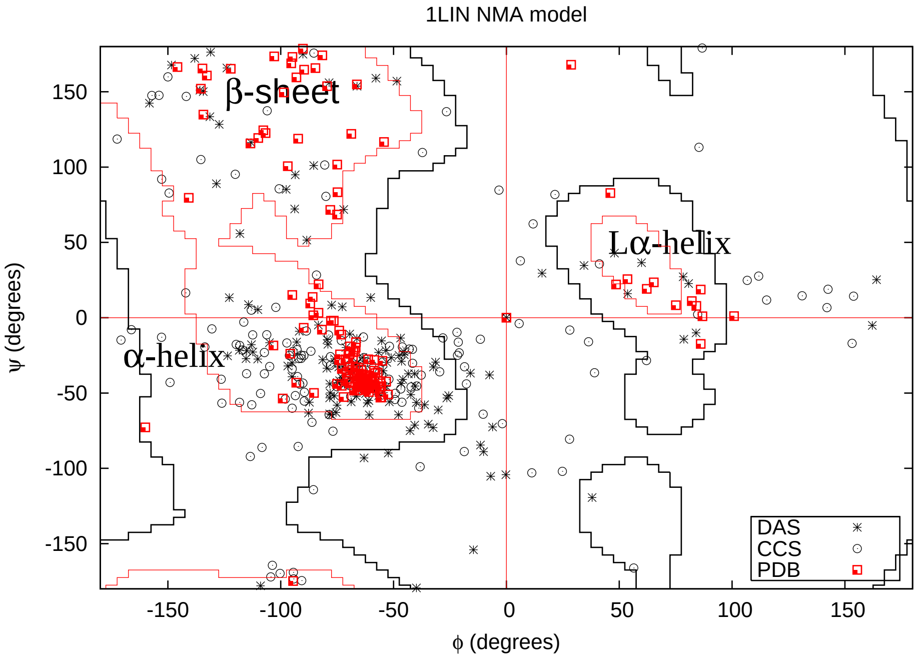


Figure 6.8: Ramachandran plot for the calmodulin target `1lin` *and model structures based on 100 DAS and CCS eigenmodes. Favored and allowed regions [Lovell 03] are delimited by stepped red and black lines respectively.*

Calmodulin Fig. 6.8 contains Ramachandran plots for the DAS and CCS models including 100 eigenmodes. Angles in the CCS model seem slightly more scattered than in the DAS model. However, there is no obvious trend, except that both models deviate significantly from the target structure's values.[11] However, using DSSP to evaluate the secondary structure, a very different picture emerges. In the CCS model there is no structure left, whereas in the DAS model even the very bent central helix is well-conserved (see Fig. 6.9).

NtrC and hemoglobin The NtrC target `1cd8` is unusual in that, a large fraction of the (ϕ, ψ)-values ($> 25\%$), lie outside the allowed regions of the Ramachandran plot. Obviously this does not change in the model structures.

[11] Ramachandran plots of the DAS and CCS model with 4 and 31 modes respectively, required to capture more than 50% of the conformational difference, are qualitatively identical to Fig. 6.8.

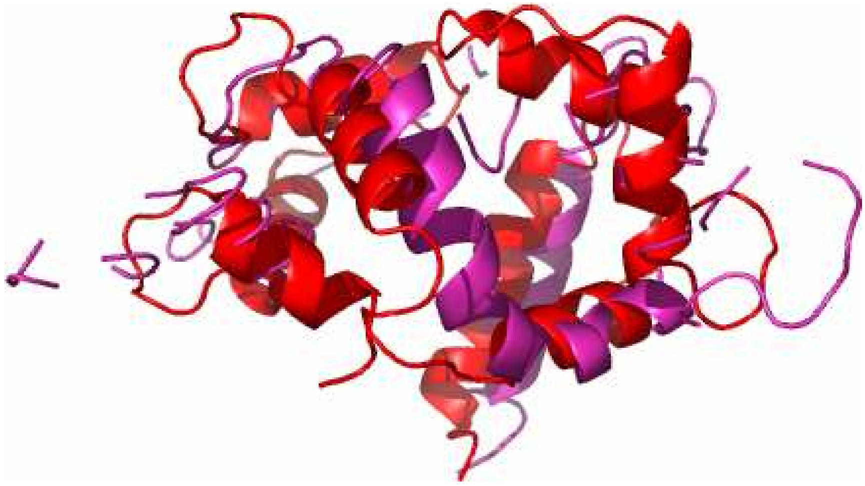


Figure 6.9: DAS model of `1lin` *including the first 4 modes. Despite a large deformation, the central helix retains the correct hydrogen bonding pattern (according to DSSP). The reason for the "floating bodies" is PyMol's strict cartoon representation. A similar representation of the CCS model displays* **no** *obvious structure.*

However, looking at the distribution of points in Table 6.3, the DAS model is seen to be closer to the target structure than the CCS model.[12]

For hemoglobin both models have Ramachandran plots almost identical to the target structure. Also DSSP finds a significant agreement between target and model structures for both NtrC and hemoglobin. This is not surprising given the small deformations observed in Fig. 6.7.

Bonded energy and normal mode deformations

We now leave the normal mode spectra and restrict our attention to the reference structures. To estimate the quality of a structural change along

[12] The distribution of points in the Ramachandran plot was determined using the web-server *Ramachandran Plot 2.0* developed by K. Gopalakrishnan, S.S. Sheik and K. Sekar.

Structure	Reference	Target	DAS model	CCS model
Fully allowed	41	45	44	38
Additionally allowed	41	27	34	47
Generously allowed	13	19	16	11
Outside	5	9	7	5

Table 6.3: Distribution of (ϕ, ψ)-angles in the Ramachandran plot for the four NtrC structures. The fully and additionally allowed regions corresponds to the allowed regions in [Lovell 03].

an eigenmode direction, we consider the variation of each term in the bonded energy

$$
\begin{aligned}
U_{\text{bonded}} = &\\
& K_{\text{bond}} \sum_{\text{bonds}} (l - l_0)^2 + K_{\text{ang}} \sum_{\text{angles}} (\phi - \phi_0)^2 + \\
& K_{\text{dih}} \sum_{\text{dihedral}} (1 + \cos(N\theta - \theta_0)^2) + K_{\text{imp}} \sum_{\text{improper}} (\tilde{\theta} - \tilde{\theta}_0)^2,
\end{aligned}
\tag{6.47}
$$

as the crystal structure is deformed along a single eigenmode. The energy was introduced in Section III of Chapter 2 and we have used the notation for the various degrees of freedom introduced in Fig. 6.2. Equilibrium values, $l_0, \phi_0, \theta_0, \tilde{\theta}_0$ and force constants, $K_{\text{bond}}, K_{\text{ang}}, K_{\text{dih}}, K_{\text{imp}}$, are taken from the CHARMM19 force-field [Reiher, III 85]. We call the bonded energy of the crystal structure the *zero-point energy.*

The behavior of U_{bonded} under deformations is a measure of the frustration in the structure and hence, of how realistic a movement along the normal mode direction is.

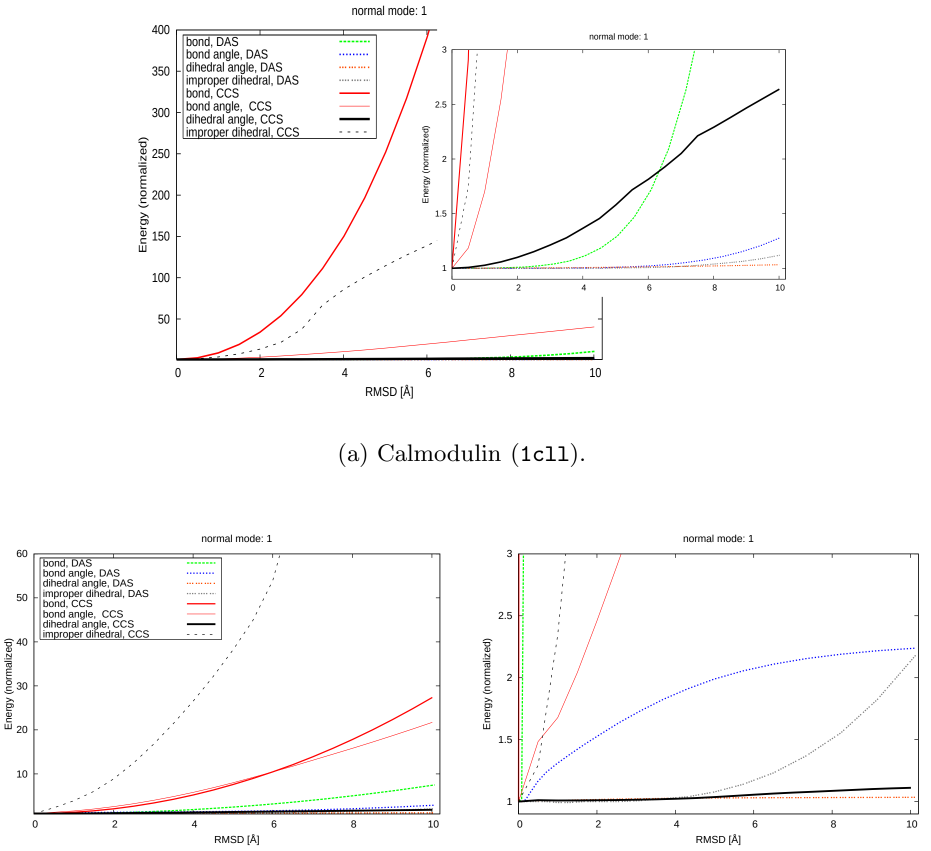
Energy: Deformation along a single normal mode We now fix the order of the mode and consider the terms in U_{bonded} as a function of RMSD from the crystal structure.

Fig. 6.10 shows the results for the first DAS and CCS eigenmode. For all protein structures and all energy terms - except NtrC above 11Å - deformation along the DAS mode gives rise to less of an increase, than a similar deformation along the CCS mode. This demonstrates, that with the same energy, a protein can move farther along a DAS mode, which makes them more plausible paths for low-energy flexibility.

Looking at other eigenmodes the qualitative picture does not change. Deformations along DAS modes exhibit a smaller increase in energy. However, as we go to higher order modes the curves become more erratic and deviate significantly from the harmonic behavior. Plots for eigenmodes 5 and 50 can be found in Appendix B.

Remark 6.15 *[Kitao 94] reports a significant overlap between the lowest-frequency DAS and CCS modes. In our case, plotting the modes together in Fig. 6.10 is merely a convenient way of visualizing representative modes. However, in calmodulin both the first modes capture* $\sim$ 19% *of the conformational difference (see Fig. 6.6).*

Energy: Fixed length deformation along each eigenmode We now take a deformed structure, at a fixed distance from the crystal structure, and consider the terms in U_{bonded} for the different eigenmodes.



(a) Calmodulin (`1cll`).

(b) Chain A and B of hemoglobin (`1a3n`).

(c) The NtrC switch (`1dc7`).

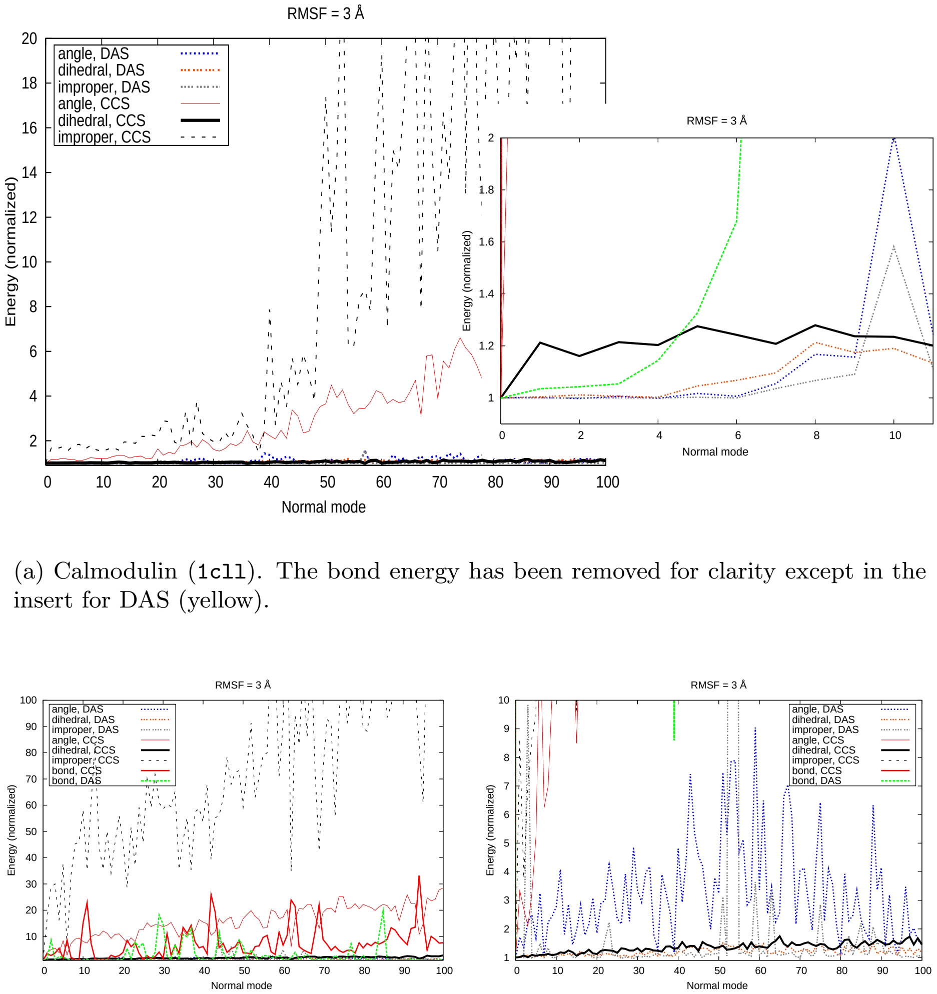
Figure 6.10: Contributions to the bonded energy, U_{bonded}, as the crystal structure is deformed, either along the first DAS eigenmode, or the first CCS eigenmode. The energy-scale is set by the zero-point energy (see text).

The result at an RMSD of 3Å is shown in Fig. 6.11. Again all structures are significantly better behaved under a deformation along DAS modes than along CCS modes. The general trends are:

- The bond energy is very erratic for both sets of eigenmodes. Though the trend seems to be lower values for DAS eigenmodes this is not too convincing.

  The exception is calmodulin. Here the first 4 modes exhibit only a slight increase in E_{bond} (Fig. 6.11(a)). In Section VI.2 it was found that the same modes captured $> 60\%$ of the conformational difference. A similar behaviour is observed for the CCS modes at 1Å (see Appendix B) but is lost at 3Å.

- The dihedral angle energy is surprisingly well preserved under CCS deformations. However, bond angle and improper dihedral angle energies deviate fast from their zero-point values. In all instances, deformations along DAS modes give well-behaved energies.



(a) Calmodulin (`1cll`). The bond energy has been removed for clarity except in the insert for DAS (yellow).

(b) Chain A and B of hemoglobin (`1a3n`).

(c) The NtrC switch (`1dc7`).

Figure 6.11: Contributions to the bonded energy, U_{bonded}, evaluated at a distance of 3Å from the crystal structure for the first 100 DAS and CCS eigenmodes. The energy-scale is set by the zero-point energy. Notice the difference in scales.

Plots at a distance of 1Å and 5Å can be found in Appendix B.

Remark 6.16 *The zero-point energy provides a natural scale but can hide significant contributions, e.g. the* $E_{bond} \sim 5.6 \cdot 10^5 kcal/mol$ *in hemoglobin. As mentioned in Section I.2 this is often observed in crystal structures and can to some extent be removed by a prior relaxation of the structure. However, this is not a part of the Tirion potential Eq. (6.2) and we do not pursue the matter.*

VII Discussion & future work

We have consider normal mode analysis (NMA) in dihedral angles using a Tirion elastic network model. Our intent has been to clarify the potential advantages of NMA in dihedral angles (DAS) as compared to the standard formulation in Cartesian coordinates (CCS).

Following [Petrone 06] we examined the ability of a few eigenmodes to represent the conformational difference between an open and a closed form of a protein. In only one of three proteins, calmodulin, did we find such a set of low-frequency modes. However, we also found, that in this case DAS eigenmodes performed significantly better than CCS modes. In the remaining two cases, the NtrC switch and hemoglobin, neither formulation provided convincing results.

Representing conformational differences between two equilibrium structures goes far beyond the regime where NMA, *a priori*, is valid. It is therefore unsurprising, that completely general statements concerning the applicability of NMA cannot be made. However, this does not mean that conformational differences can never be described, as the case of calmodulin demonstrates. The best discriminatory measure for this, seems to be the *degree of collectivity* introduced in [Brüschweiler 95]. This has been shown to correlate well with the ability of normal modes to represent a conformational difference [Tama 01]. Indeed, calmodulin has a high degree of collective, and precisely in this case, DAS shows an improvement over CCS NMA; both in terms of representing the conformational difference and in the preservation of secondary structure.

It is widely recognized, that flexibility plays a crucial role for the function of proteins [Frauenfelder 91, Ma 98, Benkovic 03]. A realistic and computationally inexpensive way of introducing flexibility into a the crystal structure of a protein is therefore of great interest. One such method is NMA.

We looked at variations in the bonded energy terms under deformations along eigenmode directions, as a way to estimate the quality of a normal mode motion. In all three cases considered, it was found, that structures obtained by deformation along DAS modes exhibit significantly better stereochemical - and hence energetic - properties, than an equivalent structure obtained from CCS modes. This makes DAS eigenmodes better paths for low-energy flexibility.

A wide range of application could benefit from introducing flexibility, e.g. noise-to-signal studies for structural measures [Røgen 05] or generating decoy structures to test for specificity in protein design [Koehl 99]. As a concrete example we mention the protein design in [Fu 07]. Here CCS eigenmodes were used to introduce flexibility into α-helical ligands. This enabled the authors to find a larger set of low-energy sequences that could accommodate the ligand. They subsequently proposed sampling ligand structures using DAS modes, in order to maintain ideal bond lengths and angles, and thereby lower the need for regularization of the structures. Our work indicates that a lot could be gained by such an approach.

A Numerical implementation

To solve the generalized eigenvalue problem Eq. (6.32) we use the routines provided in ARPACK [Lehoucq 97] specifically designed to find extremal eigenvalues and the associated vectors in large-scale eigenvalue problems.

Computational time

Initially we observed a slow convergence of the Lanczos algorithm also reported in [Yang 01]. Following [Lehoucq 97] this problem was solved by implementing the shift-invert interface of ARPACK that instead looks for the largest eigenvalues in the inverse problem.

The iterative construction of the DAS hessian described in [Abe 84] ensures much lower memory requirements. The smaller size of the matrices, compared to CCS NMA, implies a significantly faster solution of the eigenvalue problem for a given protein.[13] However, the transformation into Cartesian coordinates can take several hours, even days for larger proteins. This will of course not be problem in pure DAS applications where the transformation is left our.

Program package

The DAS normal mode analysis and the normal mode spectrum is written in Fortran77 and part of a larger package written by P. Koehl, Genome Center/Computer Science Dept., UC Davis. It consists several tools:

1. Reconstruction of side-chains and completion of the protein backbone using self-consistent mean-field theory (SCMF) [Koehl 96].
2. Gradient-based energy minimization of semi-empirical protein potentials (using the CHARMM19 force-field [Reiher, III 85] but this can be changed).
3. Computing the solvent accessible surface area and volume, together with derivatives, in the all-atom representation of a protein and using the theory of α-shapes [Edelsbrunner 05].
4. Normal mode analysis in dihedral angle and Cartesian coordinates.

The software package is written under the GNU Lesser General Public License and can be obtained by contacting P. Koehl.[14]

[13] Finding the first 100 eigenvalues and eigenvectors for myosin ($\sim$ 3600 dihedral angles) is done in the order of seconds on a standard laptop computer.

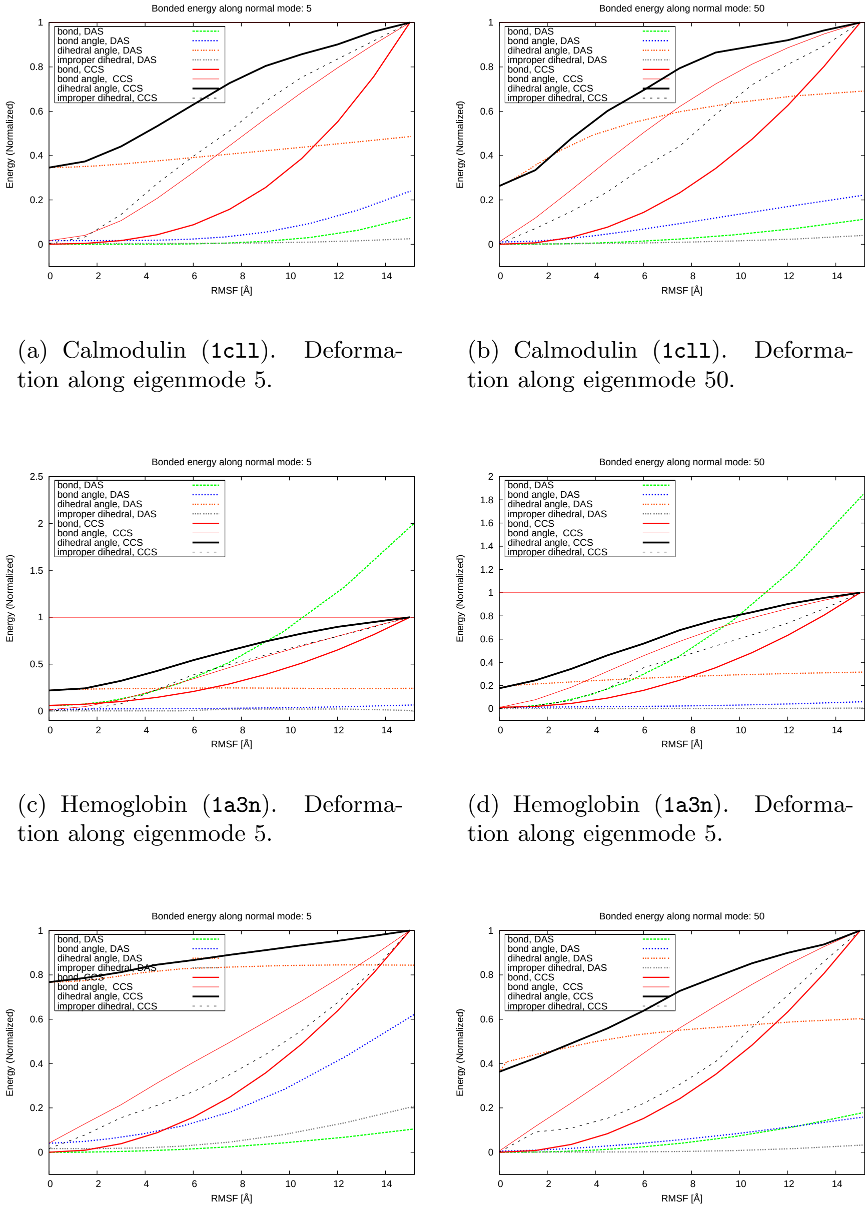
[14] `koehl@cs.ucdavis.edu`.

B Bonded energy plots

This appendix contains some further plots of the variations in the bonded energy terms

$$
\begin{aligned}
U_{\text{bonded}} = & \\
& K_{\text{bond}} \sum_{\text{bonds}} (l - l_0)^2 + K_{\text{ang}} \sum_{\text{angle}} (\phi - \phi_0)^2 + \\
& K_{\text{dih}} \sum_{\text{dihedral}} (1 + \cos(N\theta - \theta_0)^2) + K_{\text{imp}} \sum_{\text{improper}} (\tilde{\theta} - \tilde{\theta}_0)^2,
\end{aligned}
$$

as the crystal structure is deformed along single eigenmodes.



(a) Calmodulin (`1cll`). Deformation along eigenmode 5.

(b) Calmodulin (`1cll`). Deformation along eigenmode 50.

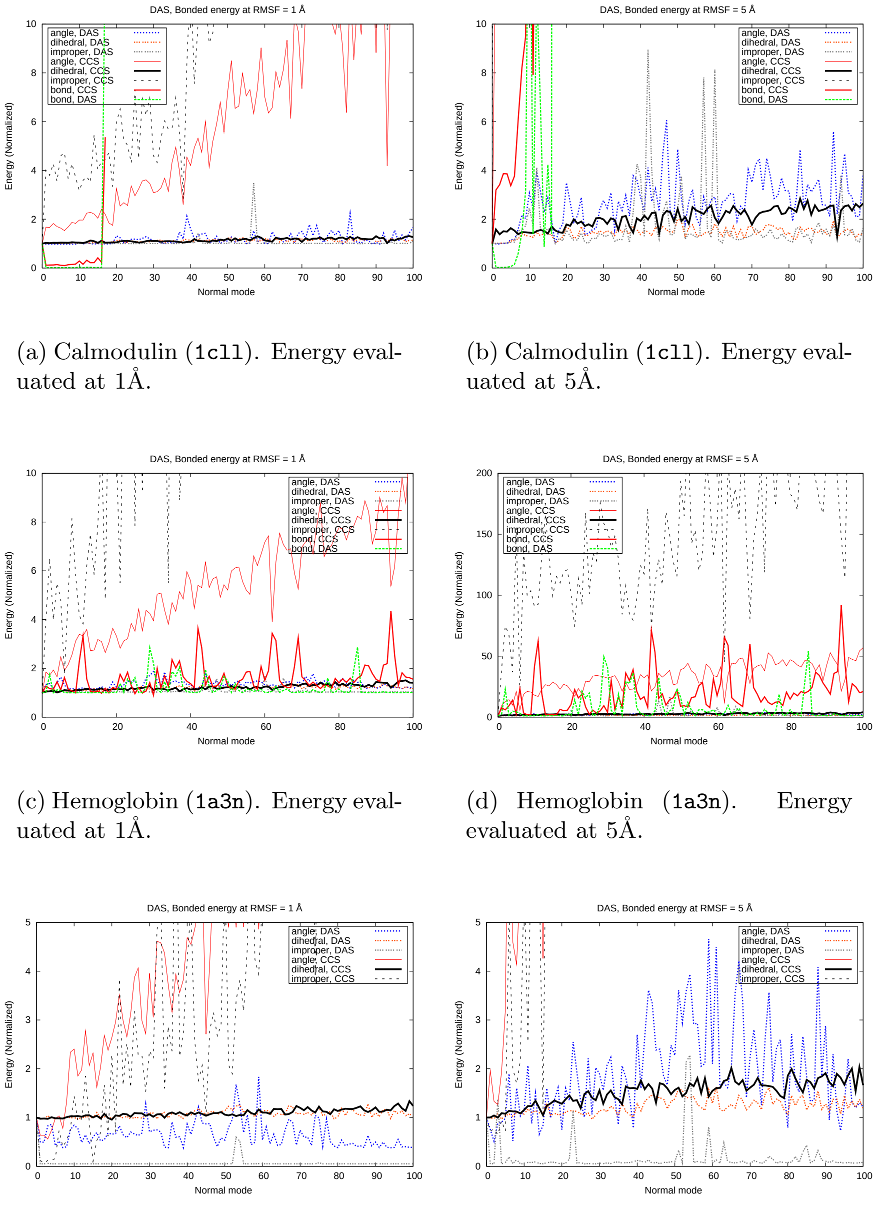
(c) Hemoglobin (`1a3n`). Deformation along eigenmode 5.

(d) Hemoglobin (`1a3n`). Deformation along eigenmode 5.

(e) NtrC (`1dc7`). Deformation along eigenmode 5.

(f) NtrC (`1dc7`). Deformation along eigenmode 50.

Figure A-1: Contributions to the bonded energy, U_{bonded}, as the crystal structure is deformed, either along a DAS or a CCS eigenmode. The energy is measured relative to the energy of the crystal structure.



(a) Calmodulin (`1cll`). Energy evaluated at 1Å.

(b) Calmodulin (`1cll`). Energy evaluated at 5Å.

(c) Hemoglobin (`1a3n`). Energy evaluated at 1Å.

(d) Hemoglobin (`1a3n`). Energy evaluated at 5Å.

(e) NtrC (`1dc7`). Energy evaluated at 1Å.

(f) NtrC (`1dc7`). Energy evaluated at 5Å.

Figure A-2: Contributions to the bonded energy, U_{bonded}, evaluated at a fixed root mean square from the crystal structure for the first 100 DAS and CCS eigenmodes. Energies are measured relative to the energy of the crystal structure. Notice the difference in scales.

Bibliography

[Abe 84] H. Abe, W. Braun, T. Noguti & N. Gō. *Rapid Calculation of First and Second Derivatives of Conformational Energy With Respect to Dihedral Angles for Proteins. General Recurrent Equations.* Comp. & Chem., vol. 8, no. 4, pages 239–247, 1984.

[Amadei 93] A. Amadei, A.B.M. Linssen & H.J.C. Berendsen. *Essential Dynamics of Proteins.* Proteins: Struct., Func. and Gen., vol. 17, pages 412–425, 1993.

[Atilgan 01] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin & I. Bahar. *Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model.* Biophys. J, vol. 80, pages 505–515, 2001.

[Bahar 97] I. Bahar, A.R. Atilgan & B. Erman. *Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential.* Fold. Des., vol. 2, no. 3, pages 173–181, 1997.

[Bahar 05] I. Bahar & A.J. Rader. *Coarse-grained normal mode analysis in structural biology.* Curr. Op. Struct. Bio., vol. 15, pages 1–7, 2005.

[Bauer 00] C.B. Bauer, H.M. Holden, J.B. Thoden, R. Smith & I. Rayment. *X-ray structures of the apo and MgATP-bound states of Dictyostelium discoideum myosin motor domain.* J. Chem. Biol., vol. 275, pages 38494–38499, 2000.

[Benkovic 03] S.J. Benkovic & S. Hammes-Schiffer. *A Perspective on Enzyme Catalysis.* Science, vol. 301, pages 1196–1202, 2003.

[Berman 03] H.M. Berman, K. Henrick & H. Nakamura. *Announcing the worlwide Protein Data Bank.* Nature Structural Biology, vol. 10, no. 12, page 980, 2003.

[Brooks 83] B. Brooks & M. Karplus. *Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor.* Proc. Nat. Acad. Sci. USA, vol. 80, pages 6571–6575, 1983.

[Brooks 85] B. Brooks & M. Karplus. *Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme.* Proc. Nat. Acad. Sci. USA, vol. 82, pages 4995–4999, 1985.

[Brüschweiler 95] R. Brüschweiler. *Collective protein dynamics and nuclear spin relaxation.* J. Chem. Phys., vol. 102, pages 3396–3403, 1995.

[Califano 76] S. Califano. Vibrational states. Wiley, London, 1976.

[Chattopadhyaya 92] R. Chattopadhyaya, W.E. Meador, A.R. Means & F.A. Quiocho. *Calmodulin structure refined at 1.7 A resolution.* J. Mol. Biol., vol. 228, pages 1177–1192, 1992.

[Cui 04] Q. Cui, G. Li, J. Ma & M. Karplus. *A Normal Mode Analysis of Structural Plasticity in the Biomolecular Motor F1-ATPase.* J. Mol. Bio., vol. 340, pages 345–372, 2004.

[Delarue 04] M. Delarue & P. Dumas. *On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models.* Proc. Nat. Acad. Sci. USA, vol. 101, no. 18, pages 6957–6962, 2004.

[Doruker 04] P. Doruker, R.L. Jernigan & I. Bahar. *Dynamics of large proteins through hierarchical levels of coarse-grained structures.* J. Comp. Chem., vol. 23, pages 119–127, 2004.

[Durand 94] P. Durand, G. Triniquier & Y-H. Sanejouand. *A new approach for determining low-frequency normal modes in macromolecules.* Biopolymers, vol. 34, no. 6, pages 759–771, 1994.

[Eckart 35] C. Eckart. *Some Studies Concerning Rotating Axes and Polyatomic Molecules.* Phys. Rew., vol. 47, pages 552–558, 1935.

[Edelsbrunner 05] H. Edelsbrunner & P. Koehl. *The Geometry of Biomolecular Solvation.* Discrete and Computational Geometry, vol. 52, pages 241–273, 2005.

[Elber 87] R. Elber & M. Karplus. *Multiple Conformational States of Proteins: A Molecular Dynamics Analysis of Myoglobin.* Science, vol. 235, no. 4786, pages 318–321, 1987.

[Flores 06] S. Flores, N. Echols, D. Milburn, B. Hespenheide, K. Keating, J. Lu, S. Wells, E.Z. Yu, M. Thorpe & M. Gerstein. *The Database of Macromolecular Motions: new features added at the decade mark.* Nucleic Acids. Res., vol. 34, pages 296–301, 2006.

[Frauenfelder 91] H. Frauenfelder, S.G. Sligar & P.G. Wolynes. *The Energy Landscapes and Motions of Proteins.* Science, vol. 254, no. 5038, pages 1598–1603, 1991.

[Fu 07] X. Fu, J.R. Apgar & A.E. Keating. *Modeling Backbone Flexibility to Achieve Sequence Diversity: The Design of Novel α-Helical Ligands for Bcl-x_L.* J. Mol. Biol., vol. 371, pages 1099–1117, 2007.

[Gō 83] N. Gō, T. Noguti & T. Nishikawa. *Dynamics of a small globular protein in terms of low-frequency vibrational modes.* Proc. Natl. Acad. Sci. USA, vol. 80, pages 3696–3700, 1983.

[Gō 89] N. Gō & T. Noguti. *Structural Basis of Hierarchical Multiple Substates of a Protein.* Chemica Scripta, vol. 29A, pages 151–164, 1989.

[Gohlke 06] H. Gohlke & M.F. Thorpe. *A Natural Coarse Graining for Simulating Large Biomolecular Motion.* Biophys. J., vol. 91, pages 2115–2120, 2006.

[Hayward 93] S. Hayward, A. Kitao, F. Hirata & N. Gō. *The Effect of Solvent on Collective Motions in Globular Protein.* J. Mol. Biol., vol. 234, pages 1207–1217, 1993.

[Hayward 94] S. Hayward, A. Kitao & N. Gō. *Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and principal component analysis.* Protein Sci., vol. 3, pages 939–943, 1994.

[Hayward 95] S. Hayward & N. Gō. *Collective variable description of native protein dynamics.* Annu. Rev. Phys. Chem., vol. 46, pages 223–250, 1995.

[Hinsen 98] K. Hinsen. *Analysis of Domain Motions by Approximate Normal Mode Calculations.* Proteins: Struct., Func. and Gen., vol. 33, page 417, 1998. Useful comments on the biological relevance of information obtainable from NMA. Also discusses various coarse-graining schemes.

[Hinsen 00] K. Hinsen, A-J. Petrescu, S. Dellerue, M-C. Bellissent-Funel & G.R. Kneller. *Harmonicity in slow protein dynamics.* Chem. Phys., vol. 261, pages 25–37, 2000.

[Hinsen 05] K. Hinsen, N. Reuter, J. Navaza D.L. Stokes & J-J. Lacapere. *Normal mode-based fitting of atomic structure into electron density maps: application to sarcoplasmic reticulum Ca-ATPase.* Biophys. J., vol. 88, no. 2, pages 818–827, 2005.

[Hollup 05] S.M. Hollup, G. Salensminde & N. Reuter. *WEBnm@: a web application for normal mode analyses of proteins.* BMC Bioinformatics, vol. 6, pages 1–8, 2005.

[Horiuchi 91] T. Horiuchi & N. Gō. *Projection of Monte Carlo and Molecular Dynamics Trajectories Onto the Normal Mode Axes: Human Lysozyme.* Proteins, vol. 10, pages 106–116, 1991.

[Horvath 05] I. Horvath, V. Harmat, A. Perczel, V. Palfi, L. Nyitrai, A. Nagy, E. Hlavanda, G. Naray-Szabo & J. Ovadi. *The structure of the complex of calmodulin with KAR-2: a novel mode of binding explains the unique pharmacology of the drug.* J. Biol. Chem., vol. 280, pages 8266–8274, 2005.

[Ikura 92] M. Ikura, G.M. Clore amd A.M. Gronenborn, G. Zhu, C.B. Klee & A. Bax. *Calmodulin-Target Peptide Complex by Multidimensional NMR.* Science, vol. 256, pages 632–638, 1992.

[Jørgensen 78] F. Jørgensen. *Orientation of the Eckart frame in a polyatomic molecule by symmetric orthonormalization.* Int. J. of Quantum Chem., vol. 14, no. 1, pages 55–63, 1978.

[Kabsch 83] W. Kabsch & C. Sander. *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.* Biopolymers, vol. 22, pages 2577–2637, 1983.

[Karplus 02] M. Karplus & J.A. McCammon. *Molecular dynamics simulations of biomolecules.* Nat. Struct. Bio., vol. 9, no. 9, pages 646–652, 2002.

[Kern 99] D. Kern, B.F. Volkman, P. Luginbuhl, M.J. Nohaile, S. Kustu & D.E. Wemmer. *Structure of a transiently phosphorylated switch in bacterial signal transduction.* Nature, vol. 402, pages 894–898, 1999.

[Kidera 90] A. Kidera & N. Gō. *Refinement of protein dynamic structure: Normal mode refinement.* Proc. Nat. Acad. Sci. USA, vol. 87, pages 3718–3722, 1990.

[Kidera 92] A. Kidera, K. Inaka, M. Matsushima & N. Gō. *Normal Mode Refinement: Crystallographic Refinement of Protein Dynamic Structure Applied to Human Lysozyme.* Biopolymers, vol. 32, pages 315–319, 1992.

[Kitao 91] A. Kitao, F. Hirata & N. Gō. *The effects of solvent on the conformation and collective motions of protein; normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum.* Chem. Phys., vol. 158, pages 447–472, 1991.

[Kitao 94] A. Kitao, S. Hayward & N. Gō. *Comparison of normal mode analysis on small globular protein in dihedral angle space and Cartesian coordinate space.* Biophys. Chem., vol. 52, pages 107–114, 1994.

[Kitao 98] A. Kitao, S. Hayward & N. Gō. *Energy Landscape of a Native Protein: Jumping-Among-Minima Model.* Proteins: Struct., Func. and Gen., vol. 33, pages 496–517, 1998.

[Kitao 99] A. Kitao & N. Gō. *Ivestigating protein dynamics in collective coordinate space.* Curr. Op. Struct. Bio., vol. 9, pages 164–169, 1999.

[Koehl 96] P. Koehl & M. Delarue. *Mean-field minimization methods for biological macromolecules.* Curr. Op. Struct. Bio., vol. 6, pages 222–226, 1996.

[Koehl 99] P. Koehl & M. Levitt. *De novo protein design. I. In search of stability and specificity.* J. Mol. Biol., vol. 293, pages 1161–1181, 1999.

[Koehl 06] P. Koehl. Protein structure classication, volume 22 of *Reviews in Computational Chemistry*, pages 1–55. Wiley and Sons, 2006.

[Lamm 86] G. Lamm & A. Szabo. *Langevin modes of macromolecules.* J. Chem. Phys., vol. 85, no. 12, pages 7334–7348, 1986.

[Landau 63] L. Landau & E. Lifchitz. Course of theoretical physics, vol 5: Statistical physics. Pergamon Press, London, first edition, 1963.

[Landau 88] L. Landau & E. Lifchitz. Physique thèorique tome i: Mécanique. Mir, Moscow, 4 edition, 1988.

[Lehoucq 97] R.B. Lehoucq, D.C. Sorensen & C. Yang. *ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.*, 1997.

[Levitt 85] M. Levitt, C. Sander & P.S. Stern. *Protein Normal-mode Dynamics: Trypsin Inhibitor, Crambin, Ribonuclease and Lysozyme.* J. Mol. Biol., vol. 181, pages 423–447, 1985.

[Levy 84] R.M. Levy, A.R. Srinivasan, W.K. Olson & J.A. McCammon. *Quasi-Harmonic Method For Studying Very Low Frequency Modes In Proteins.* Biopolymers, vol. 23, pages 1099–1112, 1984.

[Lindahl 05] E. Lindahl & M. Delarue. *Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization.* Nucleic Acids. Res., vol. 33, no. 14, pages 4496–4506, 2005.

[Lockless 99] S.W. Lockless & R. Ranganathan. *Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families.* Science, vol. 286, pages 295–299, October 1999.

[Lovell 03] S.C. Lovell, I.W. Davis, W.B. Arendall, III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson. *Structure Validation by C_α Geometry: ϕ, ψ, and C_β Deviation.* J. Mol. Biol., vol. 50, pages 437–450, 2003.

[Ma 98] J. Ma & M. Karplus. *The allosteric mechanism of the chaperonin GroEL: a dynamic analysis.* Proc. Nat. Acad. Sci. USA, vol. 95, pages 8502–8507, 1998.

[Ma 00] J. Ma, P.B. Siegler, Z. Xu & M. Karplus. *A dynamical model for the allosteric mechanism of GroEL.* J.Mol. Bio., vol. 302, pages 303–313, 2000.

[Mahan 00] G. D. Mahan. Many particle physics. Springer, third edition, 2000.

[Maragakis 05] P. Maragakis & M. Karplus. *Large Amplitude Conformational Change in Proteins Explored with a Plastic Network Model: Adenylate Kinase.* J. Mol. Biol., vol. 352, pages 807–822, 2005.

[McLachlan 79] A.D. McLachlan. *Gene Duplications in the Structural Evolution of Chrymotrypsin.* J. Mol. Biol., vol. 128, pages 49–79, 1979.

[Moritsugu 00] K. Moritsugu, O. Miyashita & A. Kidera. *Vibrational Energy Transfer in a Protein Molecule.* Phys. Rev. Lett., vol. 85, no. 18, page 3970, 2000.

[Mozzarelli 91] A. Mozzarelli, C. Rivetti, G. L. Rossi, E.R. Henry & W.A. Eaton. *Crystals of haemoglobin with the T quarternary structure bind oxygen noncooperatively with no Bohr effect.* Nature, vol. 351, pages 416–419, 1991.

[Needleman 70] S. Needleman & C. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J. Mol. Biol., vol. 48, no. 3, pages 443–453, 1970.

[Nevo 06] R. Nevo, V. Brumfeld, R. Kapon, P. Hinterdorfer & Z. Reich. *Direct measurement of the protein energy landscape.* EMBO reports, vol. 6, no. 5, pages 482–486, 2006.

[Nishikawa 87] T. Nishikawa & N. Gō. *Normal Modes of Vibration in Bovine Pancreatic Trypsin Inhibitor and Its Mechanical Property.* Proteins: Struct., Func. and Gen., vol. 2, pages 308–329, 1987.

[Noguti 82] T. Noguti & N. Gō. *Collective variable description of small-amplitude conformational fluctuations in a globular protein.* Nature, vol. 296, pages 776–778, 1982.

[Noguti 83a] T. Noguti & N. Gō. *Dynamics of native Globular Proteins in Terms of Dihedral Angles.* J. Phys. Soc. of Japan, vol. 52, no. 9, pages 3283–3288, 1983.

[Noguti 83b] T. Noguti & N. Gō. *A Method of Rapid Calculation of a Second Derivative Matrix of Conformational Energy for Large Molecules.* J. Phys. Soc. of Japan, vol. 52, no. 10, pages 3685–3690, 1983.

[Petrone 06] P. Petrone & V.S. Pande. *Can Conformational Change Be Described by Only a Few Normal Modes?* Biophysical Journal, vol. 90, pages 1583–1593, 2006.

[Petsko 83] G.A. Petsko & D. Ringe. *Fluctuations in protein structure from X-ray diffraction.* Ann. Rev. Biophys. Bioeng., vol. 13, pages 331–371, 1983.

[Press 97] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery & M. Ostermeier. Numerical recipes in fortran 77: The art of scientific computing, volume I. Cambridge Univ. Press, Cambridge, second edition, 1997.

[Qian 04] B. Qian, A.R. Ortiz, D. Baker & M. Levitt. *Improvement of Comparative Model Accuracy by Free-Energy Optimization along Principal Components of Natural Structural Variation.* Proc. Nat. Acad. Sci. USA, vol. 101, no. 43, pages 15346–15351, 2004.

[Reiher, III 85] W.H. Reiher, III. *Theoretical studies of hydrogen bonding.* PhD thesis, Harvard University, 1985.

[Røgen 05] P. Røgen. *Evaluating protein structure descriptors and tuning Gauss integral based descriptors.* J. Phys.: Cond. Mat., vol. 17, pages 1523–1538, 2005.

[Rueda 07] M. Rueda, P. Chacón & M. Orozco. *Thorough Validation of Protein Normal Mode Analysis: A comparative Study with Essential Dynamics.* Structure, vol. 15, pages 565–575, 2007.

[Sayvetz 39] A. Sayvetz. *The kinetic energy of polyatomic molecules.* J. Chem. Phys., vol. 7, no. 6, pages 383–389, 1939.

[Schlick 02] T. Schlick. Molecular modeling and simulation: An interdisciplinary guide. Springer, New York, 2002.

[Shrivastava 06] I.H. Shrivastava & I. Bahar. *Common Mechanism of Pore Opening Shared by Five Different Potassium Channels.* Biophys. J., vol. 90, pages 3929–3940, 2006.

[Silva 92] M.M. Silva, P.H. Rogers & A. Arnone. *A third quaternary structure of human hemoglobin A at 1.7-A resolution.* J. Biol. Chem., vol. 267, pages 17248–17256, 1992.

[Smith 96] C.A. Smith & I. Rayment. *X-ray structure of the magnesium(II).ADP.vanadate complex of the Dictyostelium discoideum myosin motor domain to 1.9 A resolution.* Biochemistry, vol. 35, pages 5404–5417, 1996.

[Snow 05] C.D. Snow, E.J. Sorin, Y.M. Rhee & V.S. Pande. *How Well Can Simulation Predict Protein Folding Kinetics and Thermodynamics?* Annu. Rev. Biophys. Biomol. Struct., vol. 34, pages 43–69, 2005.

[Süel 03] G.M. Süel, S.W. Lockless, M.A. Wall & R. Ranganathan. *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nature Struc. Bio., vol. 10, no. 1, pages 59–69, 2003.

[Sunada 95] S. Sunada & N. Gō. *Small-Amplitude Protein Conformational Dynamics: Second-Order Analytic Relation between Cartesian Coordinates and Dihedral Angles.* J. Comp. Chem., vol. 16, no. 3, pages 328–336, 1995.

[Sunada 96] S. Sunada, N. Gō & P. Koehl. *Calculation of nuclear magnetic resonance order parameters in proteins by normal mode analysis.* J. Chem. Phys., vol. 104, no. 12, pages 4768–4775, 1996.

[Tama 01] F. Tama & Y-H. Sanejouand. *Conformational change of proteins arising from normal mode calculations.* Protein Engineering, vol. 14, pages 1–6, 2001.

[Tama 03a] F. Tama, O. Miyashita & C.L. Brooks, III. *Flexible Multi-scale Fitting of Atomic Structures into Low-resolution Electron Density Maps with Elastic Network Normal Mode Analysis.* J. Mol. Biol., vol. 337, no. 2, pages 985–999, 2003.

[Tama 03b] F. Tama, M. Valle, J. Frank & C.L. Brooks, III. *Dynamic reorganiztion of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy.* Proc. Nat. Acad. Sci. USA, vol. 100, no. 16, pages 9319–9323, 2003.

[Tama 06] F. Tama & C.L. Brooks, III. *Symmetry, Form, and Shape: Guiding Principles for Robustness in Macromolecular Machines.* Annu. Rev. Biophys. Biomol. Struct., vol. 35, pages 115–133, 2006.

[Tame 98] J. Tame & B. Vallone. *Deoxy human hemoglobin*, 1998.

[Tirion 96] M.M. Tirion. *Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis.* Physical Review Letters, vol. 77, no. 9, pages 1905–1908, 1996.

[van Aalten 97] D.M.F. van Aalten, B.L. De Groot, J.B.C. Findlay, H.J.C. Berendsen & A. Amadei. *A Comparison of Techniques for Calculating Protein Essential Dynamics.* J. Comp. Chem, vol. 18, no. 2, pages 169–181, 1997.

[van Vlijmen 99] H.W.T. van Vlijmen & M. Karplus. *Analysis of Calculated Normal Modes of a Set of Native and Partially Unfolded Proteins.* J. Phys. Chem., vol. 103, pages 3009–3021, 1999.

[Vandonselaar 94] M. Vandonselaar, R.A. Hickie, J.W. Quail & L.T. Delbaere. *Trifluoperazine-induced conformational change in Ca(2+)-calmodulin.* Nat. Struct. Biol., vol. 1, pages 795–801, 1994.

[Wako 04] H. Wako, M. Kato & S. Endo. *ProMode: a database of normal mode analyses on protein molecules with a full-atom model.* Bioinformatics, vol. 20, no. 13, pages 2035–2043, 2004.

[Wilson Jr. 55] E.B. Wilson Jr., J.C. Decius & P.C. Cross. Molecular vibrations. McGraw-Hill, New York, 1955.

[Yang 01] C. Yang, B.W. Peyton, D.W. Noid, B.G. Sumpters & R.E. Tuzun. *Large-Scale Normal Coordinate Analysis For Molecular Structures.* SIAM J. Sci. Comput., vol. 23, no. 2, pages 563–582, 2001.

[Yang 05] L.W. Yang, X. Liu, C.J. Jursa, M. Holliman & A.J. Rader amd H.A. Karimi. *iGNM: a database of protein functional motions based on Gaussian Network Model.* Bioinformatics, vol. 21, pages 2978–2987, 2005.

[Zheng 06] W. Zheng, B.R. Brooks & D. Thirumalai. *Low-frequency modes that describe allosteric transitions in biological nanomachines are robust to sequence variations.* Proc. Nat. Acad. Sci. USA, vol. 103, no. 20, pages 7664–7669, 2006.

Chapter 7

Bubble coalescence in breathing DNA: Two vicious walkers in opposite potentials

The final chapter is a brief self-contained excursion into the world of nucleotides, more precisely that of DNA. The equilibrium structure of DNA is the double helix but sometimes, thermal fluctuations break apart the double stranded DNA, and give rise to the formation of *bubbles*. Once a bubble has been formed, a subsequent breaking and forming of bonds between nucleic acids on opposite strands can take place. This is the dynamical process known as *DNA breathing*. Depending on the temperature, salt-concentration and sequence, bubbles will have a tendency to die out or expand.

In [Novotný 07] we considered a DNA construct consisting of two soft segments, separated by a more stable barrier region. With bubbles formed in the soft segments, the question of coalescence was mapped to the previously unsolved problem of two vicious walkers in opposite linear potentials. Here, a continuum Fokker-Planck formulation was used to obtain the bubble position distribution. Below the melting temperature of the barrier region, the bubbles exhibit a barrier crossing behavior with a cross-over to mainly diffusion-drift behavior, as the temperature increases. The findings were verified by comparison with both an exact solution of the discrete master equation for the initial bubble model and a stochastic method known as the Gillespie algorithm.[1]

In Section I we introduce some aspects of DNA necessary to understand

[1] The present author's part in [Novotný 07] has mainly been in the numerical validation by the Gillespie algorithm, and the following work to determine a set of biologically reasonable parameters satisfying the assumptions underlying the Fokker-Planck approximation (to appear in [Pedersen 07]).

the breathing phenomenon. In Section II we present the discrete model, comment briefly on the transition to a Fokker-Planck formulation, and present some results on the bubble-coalescence statistics. Finally we close the chapter by the quest for biologically sensible parameters in the region where the Fokker-Planck approximation is valid. This is done in Section III.

The work is the result of a collaboration with T. Novotný, J.N. Pedersen, T. Ambjörnsson, and R. Metzler which was initiated at the *Computational Problems in Physics* workshop in Helsinki, May 2005, supported by Nord-Forsk, Nordita, and Finnish NGSMP.

I Introduction

Under a wide range of temperatures and salt-concentrations the equilibrium structure of *deoxyribonucleic acid* (or DNA) is the double helix discovered by Watson and Crick [Watson 53] based on the X-ray crystallographic data of Franklin [Franklin 53]. An example of such a structure is shown in Fig.

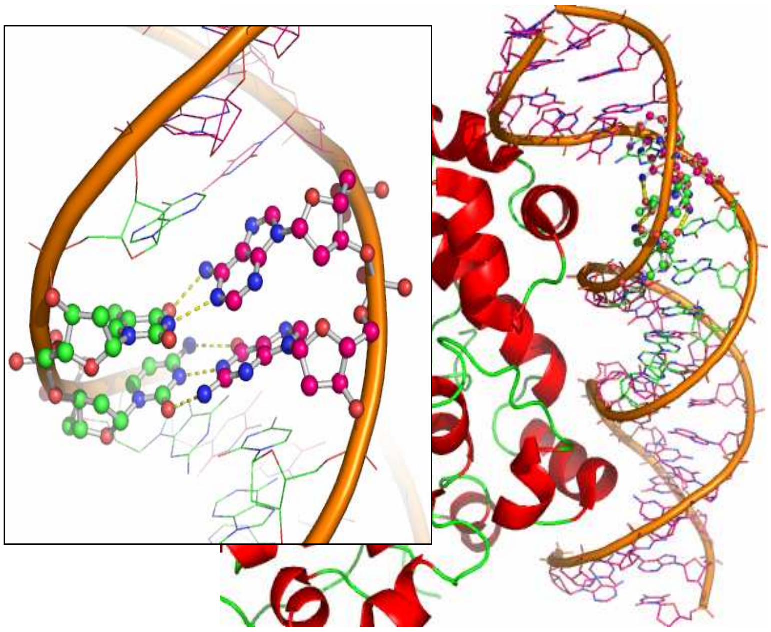


Figure 7.1: The equilibrium structure of DNA is the double helix. Here shown together with the α-subunit of RNA polymerase (`1xs9`, [Martin 04]). Part of its stability is due to hydrogen bonding between nucleic acids on opposite strands (see insert) but the largest contribution comes from a set of complex electrostatic and solvent effects known as base stacking [Kool 01]. Only two forms of base pairing can take place: Adenine-thymine *forming two hydrogen bonds and* guanine-cytosine *forming three.*

7.1 together with the α-subunit of RNA polymerase [Martin 04], the protein involved in RNA synthesis from DNA. The DNA code is written in terms of a 4-letter alphabet of nucleic acids: *adenine* (A), *cytosine* (C), *guanine* (G), and *thymine* (T). As for proteins, the DNA molecule has a regular backbone, consisting of alternating sugar and phosphate groups, and then side-chains providing the distinguishing mark between the nucleic acids.

The stability of DNA involves a pairing between nucleic acids on opposite strands. Only pairing between *adenine-thymine* (AT) and *guanine-cytosine* (GC) base pairs is possible, and leads to the formation of two and three hydrogen bonds, respectively (see Fig. 7.1). Furthermore, there is a collection of electrostatic and solvent effects known collectively as *base stacking* [Kool 01]. Stacking is the largest contributor to helix stability [Protozanova 04, Yakovchuk 06] also exhibiting a strong sequence dependence [Krueger 06]. All together, this gives rise to a difference in melting temperature for the two types of base pairs, with $T_{AT} < T_{GC}$.

Remark 7.1 *It has been found, that the sequence dependence is relevant for distinguishing coding and non-coding regions in genomes, with coding regions lying predominantly in regions exhibiting of high melting temperature [Yeramian 00a, Yeramian 00b, Carlon 05].*

Because the initiation of a bubble involves a disruption of the helical stack, an energy of 4kcal/mol is required to form a bubble [Krueger 06]. This should be compared with the characteristic energy of thermal fluctuations, $k_BT \approx 0.62$kcal/mol, which makes bubble formation a rare event under physiological conditions. However, once it has been formed, the cost of breaking additional pairs is comparable to k_BT. The subsequent breaking and forming of bonds between base pairs, giving rise to bubbles of varying sizes and positions, is the dynamical process known as *DNA breathing.*

The melting process, with base pairs flipping out of the helical stack and forming single strands, corresponds to a phase transition known as the *helix-coil transition.* The melting process has been modelled extensively, both as a (discrete) random walk in the so-called Poland-Scheraga energy landscape (see [Richard 04] and references therein) and in terms of the Peyrard-Bishop-Dauxois model [Peyrard 89]. A continuum Fokker-Planck (FP) formulation has previously been used in [Hanke 03, Bar 07, Fogedby 07].

II Discrete model and transition to Fokker-Planck formulation

In [Novotný 07] we consider a DNA construct as shown in Fig. 7.2. It consists of a (barrier) GC-region sandwiched by two (soft) AT-regions clamped at either end. The interfaces between between open and closed base pairs, denoted by the variables X and Y respectively , are called *zipper forks.*

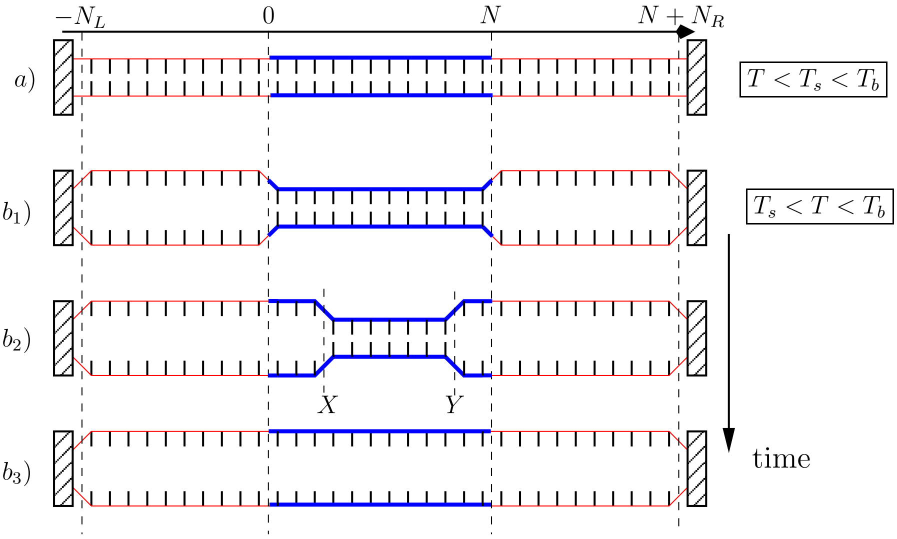


Figure 7.2: A DNA construct clamped at either end and used to model bubble coalescence in [Novotný 07]. The soft zones (thin red lines) consists of AT base pairs with melting temperature T_s. They are separated by a barrier region (thick blue line) of GC base pairs with melting temperature $T_b > T_s$. a) All base pairs are closed ($T < T_s < T_b$). $b_1 - b_3$) Soft zones open when $T > T_s$. A subsequent melting of the barrier is mainly driven by either fluctuations ($T < T_b$) or diffusion-drift ($T > T_b$). The position of the zipper forks are given by the variables X and Y.

The rate constants of base pairing

We can neglect secondary structure formation in the single strands of a bubble [Altan-Bonnet 03]. The physical properties of the DNA construct are then completely specified by a set of four rate constants, that express the probability of opening or closing a base pair at either zipper fork. The rates depend on the Boltzmann factor $\exp(\Delta G/k_B T)$, which involves the site-dependent free energy, ΔG, of breaking apart a single base pair. The free energy can be expressed in terms of temperature by

$$\Delta G = \Delta S (T_m - T), \tag{7.1}$$

where $\Delta S = -24.85\text{cal}/(\text{mol K})$ [Krueger 06] and the site dependent melting temperatures are given by the empirical relations

$$T_m^{AT} = \left[355.55 + 7.95 \ln[Na^+]\right] \text{K}, \quad T_m^{GC} = \left[391.55 + 4.98 \ln[Na^+]\right] \text{K}, \tag{7.2}$$

demonstrating the dependence of melting temperatures on the (intermediate) salt-concentration [Schildkraut 65, Frank-Kamanetskii 71].[2] The temperature relations contain contributions from both base stacking and hydrogen bonding and are the melting temperatures suitable for our model. Furthermore, the rate constants include two terms that depend on the length L of a bubble

$$s(L) = \left(\frac{L+1}{L+2}\right)^{-c}, \qquad \mathcal{K}(L) = L^{-\mu}, \tag{7.3}$$

known as the loop and the hook factor respectively. The first term represents the entropic decrease in forming a bubble from two single strands of length L.[3] The second term captures that the closing of a base pair involves moving single strands, and long strands are more difficult to move [Ambjörnsson 05]. The values for the coefficients are $c = 1.76$ and $\mu = 0.588$, the scaling coefficient for the radius of gyration of a self-avoiding loop and a self-avoiding chain in 3 dimensions respectively [Richard 04]. Finally, the time-scale of the dynamics is set by k, the rate of closing a single base pair. It is based on the diffusive encounter of two base pairs, leading to the forming of a bond. It is assumed to be identical for AT and GC pairs [Ambjörnsson 07b].

II.1 The equations of bubble coalescence dynamics

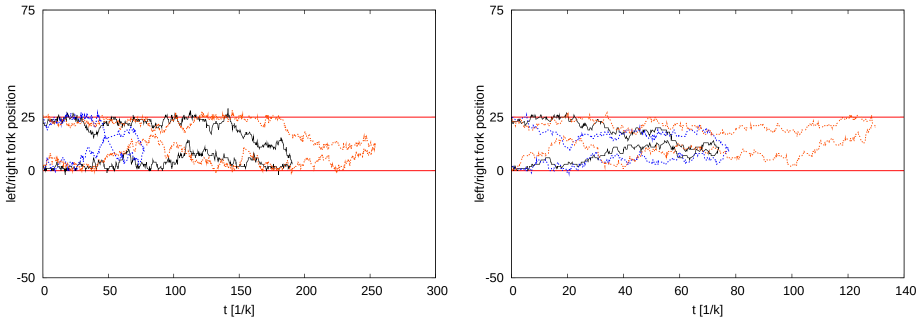
The coalescence dynamics are given by $P(X,Y;t)$, the probability distribution that the left and right zipper fork is located at X and Y, respectively, at time t. The time evolution is given by the (discrete) master equation

$$\frac{\partial}{\partial t} P(X,Y;t) = \mathbb{W} P(X,Y;t), \tag{7.4}$$

where $\mathbb{W}$ is a transfer matrix given by the rate constants, that define the random walk of zipper forks. Time and position averaged quantities of the system can then be obtained either directly, by solving the master equation [Ambjörnsson 05, Ambjörnsson 06, Ambjörnsson 07c], or by stochastic methods generating single trajectories, e.g. using the Gillespie scheme [Gillespie 76, Banik 05]. Examples demonstrating the temperature dependence of coalescence trajectories are given in Fig. 7.3. Notice the difference in time-scales.

[2] It has been shown that the dependence on salt-concentration is due to stacking term [Yakovchuk 06] and not, as previously thought, due to the hydrogen bonding [Protozanova 04]. It is known that stacking is a combination of hydrophobic, electrostatic (screening of the negatively charged phosphate groups), and dispersive interactions but there is no apparent consensus as to which term is the dominant one [Yakovchuk 06]. At high salt concentrations, ($\sim 1-5$M), the temperature dependence levels off due to a decrease in the hydrophobic effect; with most water molecules tied up in the solvation of ions, the entropy decrease involved in base stacking is small [Schildkraut 65].

[3] Due to persistence length effects in single stranded DNA, $L^{-c} \to (L+1)^{-c}$, i.e. there is a lower limit as short segments have only little freedom to wiggle [Metzler 05].



(a) $T_s < T < T_b$

(b) $T_s < T_b < T$

Figure 7.3: Trajectories of bubble-coalescence trajectories above and below the melting temperature of the barrier segment, T_b. The barrier region is delimited by horizontal lines. We see that there are effectively reflecting boundary conditions at the barrier-soft zone interfaces. This is a necessary requirement for the Fokker-Planck approximation to be valid.

A continuum Fokker-Planck formulation

In the limit where the inter-base pair distance is small, compared to the width of the barrier segment, and the temperature is such, that the soft regions are always open, the master equation Eq. (7.4) can be used to derive a bi-variate Fokker-Planck equation [van Kampen 92]

$$\frac{\partial}{\partial t}P(x,y;t) = (\underbrace{\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}}_{\text{diffusion}} \underbrace{-2f\frac{\partial}{\partial x}+2f\frac{\partial}{\partial y}}_{\substack{\text{drift in (opposite)}\\ \text{linear potentials}}})P(x,y;t), \tag{7.5}$$

where $x = X/N < y = Y/N$ and f is a dimensionless force depending only on the length N and the melting temperature of the barrier. Together with a set of reflecting boundary conditions, the viciousness condition [Fisher 84]

$$P(x,x;t) = 0, \tag{7.6}$$

saying that zipper forks cannot be at the same place, and the initial condition $P(x,y;t=0) = \delta(x-x_0)\delta(y-y_0)$, this maps the bubble coalescence problem to the problem of two vicious walkers in linear and opposite potentials with reflecting boundary conditions. Details on the solution can be found in [Novotný 07, Pedersen 07].

The Fokker-Planck equation Eq. (7.5) only determine the time evolution of zipper fork positions in the barrier region. For it to be suitable as an approximation to the discrete model in Fig. 7.2, the following assumptions must be satisfied:

1. The temperature T is sufficiently high, such that base pairs in the soft regions remain unzipped at all times. This gives effectively reflecting boundaries at the barrier-soft zone interfaces (see Fig. 7.3).

2. The length of the barrier region should be large compared to the inter-base pair distance. This allows us to take the continuum limit.

3. The soft zones should be sufficiently long, so that the loop factor in Eq. (7.3), most pronounced when bubbles are small, can be neglected.

4. The hook factor in Eq. (7.3) can be neglected.

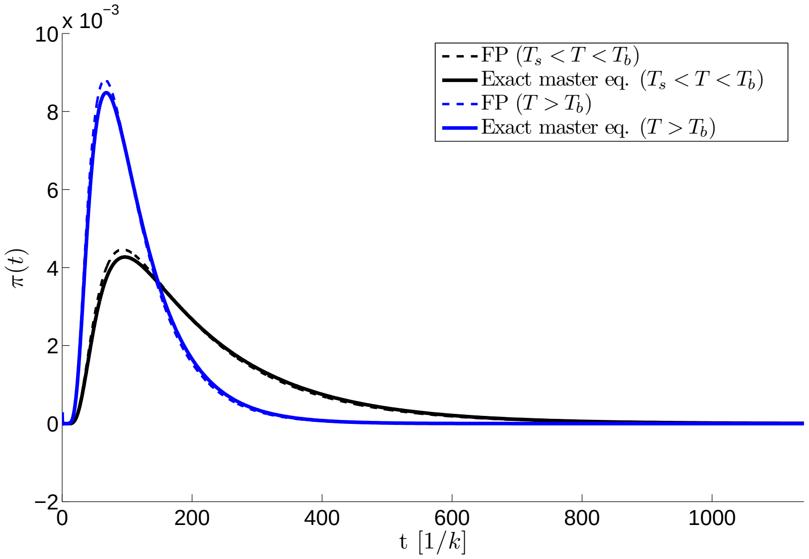


Figure 7.4: Meeting time distributions above and below the melting temperature of the barrier. The master equation solution is for a DNA construct with a 25 base pair barrier region sandwiched by soft zones each consisting of 50 base pairs. The FP approach only model dynamics in the barrier region.

When the above conditions are fulfilled, the probability distribution $P(x, y; t)$ contains the full information regarding the coalescence statistics. For example we can find the meeting time distribution (Fig. 7.4), or the mean-first passage time τ as a function of the dimensionless force f (Fig. 7.5). See [Novotný 07] for details.

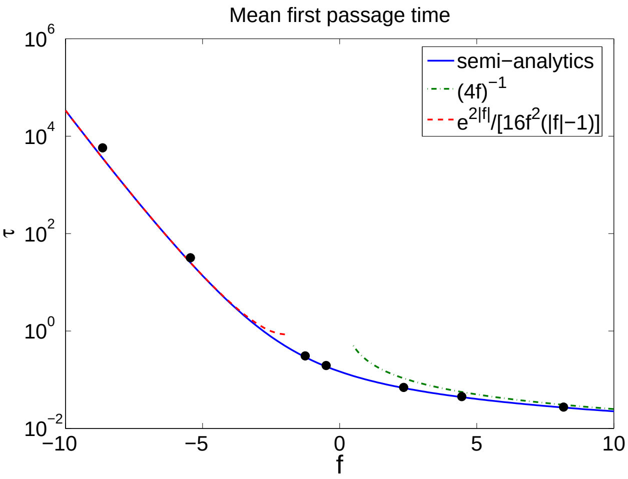


Figure 7.5: Mean first passage (bubble coalescence) time τ as a function of the dimensionless force f. f only depends on the barrier length and the temperature relative to the melting temperature of the barrier. Discrepancies for large negative values of f are due to numerical difficulties in the integration of the meeting time distribution.

III Discussion & future work: Physically based parameters

In [Novotný 07] the melting temperature of the barrier, T_b, and that of the soft zones, T_s, where fixed independently in order to validate the Fokker-Planck results. In the real world, the melting temperatures are of course not independent. We are currently investigating whether the assumptions behind the Fokker-Planck formulation, listed in Section II.1, can be satisfied under biologically realistic conditions [Pedersen 07]. We now address the assumptions one at a time:

1. From Eq. (7.2) we see, that the T_{AT}/T_{GC} ratio can be lowered by decreasing the salt-concentration. High temperatures have practical disadvantages such as the formation of air bubbles and increased evaporation of solvent molecules [Schildkraut 65]. An added benefit is therefore the general lowering of melting temperatures to below $100C^{\circ}$.

2. Even for relatively short segments (~ 20) the continuum approximation holds.

3. The loop factor can be neglected with open soft zones consisting of ~ 50 base pairs.

4. The hook factor in Eq. (7.3) gives a significant lowering of the rates for

long bubbles (required to neglect the loop factor) but is approximately constant. The effect can therefore be removed by introducing a new time-scale, in the form of a "renormalized' base pair closing rate, $\tilde{k}$.

In a setup combining fluorescence correlation spectroscopy and fluorescence quenching [Altan-Bonnet 03] the DNA is free to diffuse around in the solution. In terms of segment length, the limiting factor is the time it takes for the quencher to diffuse in and out of the confocal volume. Given this, it should be possible to work with segments of $100 - 200$ base pairs [Ambjörnsson 07a].

In short, the hunt for biologically/physically reasonable parameters appears successful. An experimental study of the two bubble setup considered here would push the boundaries of single molecule real-time measurements on DNA. A validation of the Fokker-Planck would provide information about the statistics outside the experimentally realisable situations.

Finally, this would be a further step in the fundamental understanding of the dynamics related to DNA replication, transcription and single-strand binding proteins [Ambjörnsson 05]. A model similar to ours has recently been used to demonstrate an increased bubble initiation frequency at the TATA promoter site for the RNA polymerase of the T7 phage [Ambjörnsson 07b]. The authors conjecture, that this may be connected to the initiation of transcription. Similar results, concerning the relation between melting and transcription, was obtained in [Choi 04] by molecular dynamics simulations using the Peyrard-Bishop-Dauxois model of DNA.

Bibliography

[Altan-Bonnet 03] G. Altan-Bonnet, A. Libchaber & O. Krichevsky. *Bubble Dynamics in Double-Stranded DNA*. Phys. Rev. Lett., vol. 90, page 138101, 2003.

[Ambjörnsson 05] T. Ambjörnsson & R. Metzler. *Binding dynamics of a single stranded DNA binding proteins to fluctuating bubbles in breathing DNA*. J. Phys: Cond. Matt., vol. 17, pages 1841–1869, 2005.

[Ambjörnsson 06] T. Ambjörnsson, S.K. Banik, O. Krichevsky & R. Metzler. *Sequence sensitivity of breathing dynamics in heteropolymer DNA*. Phys. Rev. Lett., vol. 97, page 128195, 2006.

[Ambjörnsson 07a] T. Ambjörnsson, 2007. Personal communication.

[Ambjörnsson 07b] T. Ambjörnsson, S.K. Banik, O. Krichevsky & R. Metzler. *Breathing Dynamics in Heteropolymer DNA*. Biophys. J., vol. 92, pages 2674–2684, 2007.

[Ambjörnsson 07c] T. Ambjörnsson, S.K. Banik, M.A. Lomholt & R. Metzler. *Master equation approach to DNA-breathing in heteropolymer DNA*. Phys. Rev. E, vol. 75, page 021908, 2007.

[Banik 05] S.K. Banik, T. Ambjörnsson & R. Metzler. *Stochastic approach to DNA breathing dynamics.* Europhys. Lett., vol. 71, page 852, 2005.

[Bar 07] A. Bar, Y. Kafri & D. Mukamel. *Loop Dynamics in DNA Denaturation.* Europhys. Lett., vol. 98, page 038103, 2007.

[Carlon 05] E. Carlon, M.L. Malki & R. Blossey. *Exons, introns, and DNA Thermodynamics.* Phys. Rev. Lett., vol. 94, page 178101, 2005.

[Choi 04] C.H. Choi, G. Kalosakas, K.Ø. Rasmussen, M. Hiromura, A.R. Bishop & A. Usheva. *DNA dynamically directs its own transcription initiation.* Nucleic Acids. Res., vol. 32, pages 1584–1590, 2004.

[Fisher 84] M.E. Fisher. *Walks, Walls, Wetting, and Melting.* J. Stat. Phys., vol. 5-6, pages 667–729, 1984.

[Fogedby 07] H.C. Fogedby & R. Metzler. *DNA Bubble Dynamics as a Quantum Coulomb Problem.* Phys. Rev. Lett., vol. 98, page 070601, 2007.

[Frank-Kamanetskii 71] M.D. Frank-Kamanetskii. *Simplification of the empirical relationship between melting temperature of DNA, its GC concentration and concentration of sodium ions in solution.* Biopolymers, vol. 10, pages 2623–2624, 1971.

[Franklin 53] R. Franklin & R.G. Gosling. *Molecular Configuration in Sodium Thymonucleate.* Nature, vol. 171, pages 740–741, 1953.

[Gillespie 76] D.T. Gillespie. *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions.* J. Comp. Phys, vol. 22, pages 403–434, 1976.

[Hanke 03] A. Hanke & R. Metzler. *Bubble dynamics in DNA.* J. Phys. A: Math. Gen., vol. 36, page 473, 2003.

[Kool 01] E.T. Kool. *Hydrogen Bonding, Base Stacking, and Steric Effects in DNA Replication.* Annu. Rev. Biophys. Biomol. Struct., vol. 30, pages 1–22, 2001.

[Krueger 06] A. Krueger, E. Protozanova & M. Frank-Kamenetskii. *Sequence-Dependent Basepair Opening in DNA Double Helix.* Biophys. J., vol. 90, pages 3091–3099, 2006.

[Martin 04] B. Dangi A.M. Gronenborn J.L. Rosner R.G. Martin. *Versatility of the carboxy-terminal domain of the alpha subunit of RNA polymerase in transcriptional activation: use of the DNA contact site as a protein contact site for MarA.* Mol. Microbiol., vol. 54, pages 45–59, 2004.

[Metzler 05] R. Metzler & A. Hanke. *Knots, Bubbles, Unwinding, and Breathing: Probing the Topology of DNA and Other Biomolecules,Bubble dynamics in DNA.* In Handbook of Theoretical and Computational Nanotechnology, pages 1–54. American Scientific Publishers, 2005.

[Novotný 07] T. Novotný, J.N. Pedersen, T. Ambjörnsson, M.S. Hansen & R. Metzler. *Bubble coalescence in breathing DNA: Two vicious walkers in opposite potentials.* Europhys. Lett., vol. 77, page 48001, 2007.

[Pedersen 07] J. N. Pedersen, T. Novotny, M. S. Hansen, T. Ambjornsson & R. Metzler. *Bubble coalescence in breathing DNA as vicious walker problem in opposite potentials*, 2007. In preparation.

[Peyrard 89] M. Peyrard & A.R. Bishop. *Statistical Mechanics of a Nonlinear Model for DNA Denaturation.* Phys. Rev. Lett., vol. 62, page 2755, 1989.

[Protozanova 04] E. Protozanova, P. Yakovchuk & M.D. Frank-Kamenetskii. *Stacked-Unstacked Equilibrium at the Nick Site of DNA.* J. Mol. Biol., vol. 342, pages 775–785, 2004.

[Richard 04] C. Richard & A.J. Guttmann. *PolandScheraga Models and the DNA Denaturation Transition.* J. Stat. Phys., vol. 115, no. 3/4, pages 925–947, 2004.

[Schildkraut 65] C. Schildkraut & S. Lifson. *Dependence of the Melting Temperature of DNA on Salt Concentration.* Biopolymers, vol. 3, pages 195–208, 1965.

[van Kampen 92] N.G. van Kampen. Stochastic processes in physics and chemistry. North-Holland, 2nd edition, 1992.

[Watson 53] J.D. Watson & F.H.C. Crick. *A Structure for Deoxyribose Nucleic Acid.* Nature, vol. 171, pages 737–738, 1953.

[Yakovchuk 06] P. Yakovchuk, E. Protozanova & M.D. Frank-Kamenetskii. *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.* Nuc. Acid. Res., vol. 34, pages 564–574, 2006.

[Yeramian 00a] E. Yeramian. *Genes and the physics of the DNA double-helix.* Gene, vol. 255, pages 139–150, 2000.

[Yeramian 00b] E. Yeramian. *The physics of DNA and the annotation of the Plasmodium falciparum genome.* Gene, vol. 255, pages 151–168, 2000.