Biophysical and Computational Analysis of Protein Structure and Stability

Protein H and the M1 protein from

Streptococcus pyogenes

- A Case Study

Bo H.K. Nilson

Masters Thesis

Department of Atomic Physics
Lund Institute of Technology
1995

Contents

1. Abstract

M proteins and other members of the M protein family, expressed on the surface of *Streptococcus pyogenes*, bind host proteins such as immunoglobulins, albumin, and fibrinogen. Protein H and the M1 protein are expressed by adjacent genes and both belong to the M protein family. In this Case Study, the structure and stability of these two proteins have been investigated. As judged from computational sequence analysis and circular dichroism spectroscopy, the proteins are almost entirely in an α-helical conformation. The amino acids are arranged in a seven-residue (heptad) repeat pattern along the greater part of the proteins. These observations support the previously accepted model of M proteins as coiled-coil dimers. However, it was also found that the structures of both proteins were thermally unstable, i.e. the content of helix conformation was greatly reduced at 37°C as compared to 25°C or below. Together with previous findings that these proteins appear as monomers at 37°C and dimers at low temperatures the results suggest that the coiled-coil dimers are unfolded at 37°C. The heptad patterns of protein H and the M1 protein showed a non-optimal distribution of residues expected for a coiled-coil conformation. This is a possible explanation for the low thermal stability of the proteins. It was also demonstrated that the proteins were stabilised in the presence of the ligands IgG and/or albumin. Protein H and the M1 protein show a high degree of sequence similarity in their C-terminal regions, and a fragment from this region displayed a high content of helix conformation, whereas fragments from the dissimilar N-terminal parts did not adopt any stable folded structure. Thus, the C-terminal parts, which are conserved within the M protein family, may constitute a framework for the formation of the parallel helical coiled-coil structure, and we propose that the less stable N-terminal part may also participate in antiparallel interaction with M protein on adjacent bacteria. The results suggest that temperature fluctuations in the environment could change the properties of bacterial surface proteins, thereby affecting the molecular interactions between the bacterium and its host. Finally, this Case Study illustrates the great benefit of combining several different analytical techniques, biophysical, computational and biochemical, when sudying important biological processes.

2. Introduction

The ultimate goal of molecular biology is to understand biological processes in term of the chemistry and physics of the macromolecules that participate in them. One of the essential differences between the chemistry of living systems and that of nonliving is the great structural complexity of biological macromolecules. To be able to understand the chemistry of life in molecular detail it is essential to unravel the structure of these biological macromolecules, especially the proteins.

The pace of discovery in biochemistry has been exceptionally rapid during the past few years. This process has greatly enriched our understanding of the molecular basis of life and has opened many new areas of inquiry. Recombinant DNA technology have provided tools for the rapid determination of DNA sequences and, by inference, the amino acid sequences of proteins from structural genes. The number of such genes is now increasing almost exponentially, but by themselves these sequences tell little about the biology of the system.

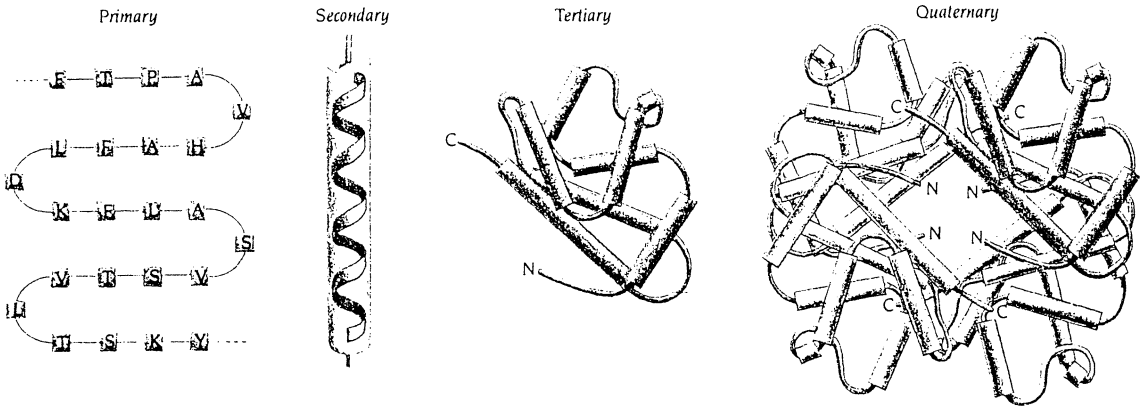


Fig. 2.1 The amino acid sequence of a protein's polypeptide chain is called its primary structure. Different regions of the sequence form local regular secondary structure, such as alpha (α) helices or beta (β) strands. The tertiary structure is formed by packing such structural elements into one or several compact globular units called domains. The final protein may contain several polypeptide chains arranged in a quaternary structure. By formation of such tertiary and quaternary structure amino acids far apart in the sequence are brought close together in three dimensions to form a functional region, an active site.

The proteins we observe in nature have evolved, through selection pressure, to perform specific functions. They play crucial roles in virtually all biological processes and the functional properties of proteins depends on their three-dimensional structures. The three-dimensional structure arises because particular sequences of amino acids in polypeptide chains fold to generate from linear chains, compact domains with specific three-dimensional structures (Figure 2.1). The folded domains can serve as virus particles or

muscle fibers or can provide specific catalytic or binding sites as found in enzymes or proteins that carry oxygen or that regulate the function of DNA, or as bacterial surface molecules which interact with human macromolecules, which have been studied in great detail in this Case Study (see Chapter 4).

3. Biophysical and Computational Methods for Study of Protein Structure and Stability

To understand the biological function of proteins we would like to be able to deduce or predict the three-dimensional structure from the amino acid sequence. This is very difficult. In spite of considerable efforts over the last 25 years, this folding problem is still unsolved and remains one of the most basic intellectual challenges in molecular biology. The fundamental reason why the folding problem remains unsolved lies in the fact that there are 20 different amino acids and therefore a vast number of ways in which similar structural domains can be generated in proteins by different amino acid sequences. By contrast, the structure of DNA, made up of only four different nucleotide building blocks that occur in two pairs, is relatively simple, regular, and predictable.

Since the three-dimensional structure of individual proteins cannot be predicted with full accuracy, they have instead to be determined experimentally by physical methods by X-ray crystallography or Nuclear magnetic resonance- (NMR) techniques. Over the past 30 years the structures of around 500 proteins have been solved by X-ray methods and over the past 15 years approximately 100 different structures have been determined by NMR methods. Crystallising proteins is a hit-and-miss business with little theory. Some proteins crystallise readily, others not at all; some investigators seem to have "green thumbs", like good gardeners, and can grow crystals where others fail. It is laborious and has a number of limitations. NMR methods are restricted by the size of the protein molecules whose structures can be determined. Currently the upper limit is molecules with molecular weight of around 20 kDa. Even though the methods are rapidly being improved, there seems little hope today, that the size limit can be changed by more than a factor 2 to 40 kDa. Furthermore, the method requires highly concentrated protein solutions, on the order of 1-2 mM, with the additional requirement that the protein molecules must not aggregate at these concentrations.

Many interesting questions can be explored and even answered without the knowledge of the complete three-dimensional structure. In order to proceed one has to simplify the problem. Two general approaches can be used. The first is to consider the system at a level much less precise than atomic resolution. Simplest are techniques that seek information about rough size and shape. In many of these, all molecular detail is put aside. The macromolecule is modelled as an ellipsoid of revolution, a coil with uniform stiffness or lack of it, a rod, or other simple shapes. Most hydrodynamic and certain scattering techniques fall into this category (see Chapter 4.3.4).

At the other extreme are techniques that look at only one small part of a macromolecule. These can sometimes attain the precision of X-ray diffraction (or even better). What is sacrificed is any overview of the entire structure. Usually, these types of techniques make use of probe molecules, whose properties allow them to be singled out from the rest of the system. Sometimes they are an intrinsic part of the macromolecule; in other cases they are added extrinsically by the investigator. In principle, by use of numerous probes, most of the structure can be mapped out. In practice, it is usually feasible to concentrate on only a few selected regions. These are chosen either to be of special biological interest or else for expediency (that is, a probe naturally exists or can be easily be introduced). Examples of successful and widely used probe techniques are fluorescence, electron paramagnetic resonance (EPR), some aspects of NMR, and chemical modification.

In between these two extremes are techniques that can selectively examine certain general aspects of a structure while ignoring others. These provide an average picture that can offer a lead to constraints on the possible three-dimensional structure. Very often such information is accessible for regions of regularly ordered secondary structure. Experimental techniques such as circular dichroism (CD), ultraviolet (UV), infrared (IR), and Raman spectroscopy, tritium exchange and computational secondary structure prediction afford a reasonable overview of the amount of helix in a protein. Sometimes certain details about the nature, size and composition of the helices are accessible. None of these techniques can permit the definition of an entire three-dimensional structure, but they often yield a convenient and accurate approximate picture for the expenditure of relatively little effort.

Virtually all the techniques in the biophysical chemist's arsenal gain strength when used in concert. In many cases, parallel studies from a variety of approaches on a single macromolecule have led to the wealth of detailed information. In no case yet has a sufficient body of data ever been accumulated to match the total static structural picture that X-ray techniques can provide. However, many insights have been obtained that would be available from pure X-ray studies only with a great deal of luck and extrapolation. Table 2.1 compares some of the basic characteristics of many of the techniques commonly used. It should be readily apparent that a detailed discussion of the principles and practice of each of these is beyond the scope of this introduction. In Chapter 3 four techniques will be described, the two major methods that have been used in this Case Study (in Chapter 4), circular dichroism spectroscopy and computational structure prediction, and in addition the two major methods that are used to determine high resolution three-dimensional structures of macromolecules, X-ray crystallography and NMR. Even within this limited set, the coverage is far from comprehensive.

Table 2.1 Techniques for structural analysis of proteins

	Information available											Experimental constraints					
Technique	Complete 3° structure	Detailed subunit arrangement	Specific site structure, bonding	Proximity between specific sites	Orientation of 2°-structure units	Type and extent of 2° structure	Molecular weight	Shape, topology	Flexibility	Ionization of individual residues	Amounts & sites of small-molecule binding	Side-chain exposure, environment	Purity of sample needed	Size of sample needed	Ease of environmental variation	Effort involved in experiment	Need for 1°-structure knowledge	Possibilities for kinetic measurements
X-ray diffraction	3	3	2	2	3	3	3	3	–	1	2	2	–	–	–	–	–	–
Electron diffraction	3	3	2	2	3	3	3	3	–	1	2	2	–	–	–	–	–	–
Electron microscopy	1	2	–	1	–	–	1	3	1	–	–	–	3	3	–	2	3	–
Autoradiography	1	–	–	1	–	–	1	2	–	–	–	–	3	3	–	1	3	–
Neutron scattering	–	2	–	–	–	–	1	2	–	–	–	–	1	–	1	–	2	–
X-ray scattering	–	1	–	–	–	–	1	2	1	–	–	–	1	–	1	–	2	–
EXAFS	–	–	3	1	–	–	–	–	–	–	2	–	–	–	1	2	–	–
Rayleigh scattering	–	1	–	–	–	–	2	1	1	–	–	–	1	2	2	1	2	3
Inelastic light scattering	–	1	–	–	–	–	2	–	–	–	–	–	–	–	–	–	–	–
Sedimentation velocity	–	–	–	–	–	–	2	1	1	–	–	–	1	1	2	2	2	–
Sedimentation equilibrium	–	–	–	–	–	–	2	1	1	1	–	–	1	3	2	2	2	–
Diffusion	–	–	–	–	–	–	2	1	–	–	–	–	1	2	–	–	2	–
Viscosity	–	–	–	–	–	–	2	2	2	–	–	–	1	3	3	3	2	1
Gel filtration	–	–	–	–	–	–	2	1	–	–	1	–	3	3	3	3	2	1
Absorption: UV/vis	–	–	–	–	–	1	–	–	–	1	1	–	1	3	3	3	1	3
Linear dichroism, birefringence	2	–	–	–	2	1	1	2	2	–	–	–	1	1	1	1	1	1

Table 2.1 (cont.) Techniques for structural analysis of proteins

Technique	Information available										Experimental constraints					
	Complete 3° structure	Detailed subunit arrangement	Specific site structure, bonding	Proximity between specific sites	Orientation of 2°-structure units	Type and extent of 2° structure	Molecular weight	Shape, topology	Flexibility	Ionization of individual residues	Amounts & sites of small-molecule binding	Side-chain exposure, environment	Purity of sample needed	Size of sample needed	Ease of environmental variation	Effort involved in experiment	Need for 1°-structure knowledge	Possibilities for kinetic measurements
CD/ORD	–	–	1	–	–	2	–	–	–	1	2	1	1	3	3	3	1	1
Absorption: IR	–	–	–	–	–	1	–	–	–	–	–	1	–	1	1	2	1	1
Dichroism: IR	–	–	–	–	2	1	–	–	–	–	–	–	–	1	1	1	1	–
MCD/MORD	–	–	2	–	–	–	–	–	–	1	–	–	1	3	3	3	1	1
Fluorescence	–	–	–	2	–	–	–	–	–	1	1	1	3	3	3	3	1	3
Polarized fluorescence	–	1	–	–	1	–	–	–	–	–	–	–	3	3	3	2	1	2
Phosphorescence	–	–	1	1	–	–	–	–	2	2	–	–	3	3	–	2	1	–
Raman scattering	–	–	1	–	–	1	–	–	–	–	–	1	1	1	2	2	1	1
Resonance Raman	–	–	2	–	–	–	–	–	–	–	–	–	3	3	3	2	1	1
EPR	–	–	2	2	1	–	–	–	2	2	2	1	3	3	2	2	1	1
NMR–^{1}H	1	–	3	2	–	–	–	–	1	3	3	1	–	1	1	1	1	1
NMR–^{13}C	–	–	2	1	–	–	–	–	2	2	2	–	–	–	2	–	2	–
^{3}H–Exchange	–	–	1	–	–	2	–	–	1	–	–	1	1	2	2	1	1	1
Potentiometric titration	–	–	1	–	–	–	–	–	2	1	2	–	1	3	1	1	1	1
Electrophoresis	–	–	–	–	–	–	1	–	–	1	–	–	3	3	1	–	2	–
Photoelectron spectroscopy	–	–	2	–	–	–	–	–	–	1	1	–	2	1	–	2	1	–
Mössbauer	–	–	2	–	–	–	–	–	–	1	1	–	2	1	–	2	1	–

NOTE: Under "Information available," larger numbers indicate more powerful techniques, and a dash indicates not usually applicable; under "Experimental constraints," larger numbers indicate easier measurements or interpretations, and a dash indicates difficult experimental hurdles.

3.1 Computational Sequence Analysis

3.1.1 Secondary Structure Prediction

A statistical analysis of conformations of known proteins reveals that certain amino acids seem to prefer certain conformations. There is a relationship between these statistical tendencies and the measured tendency to stabilise a helical conformation observed in experiments in which each residue is a guest in a block polymer (Suzuki and Robson, 1976). Although these preferences are not strict, the hope is that by analysing the pattern of residues in a region one could make a correct prediction.

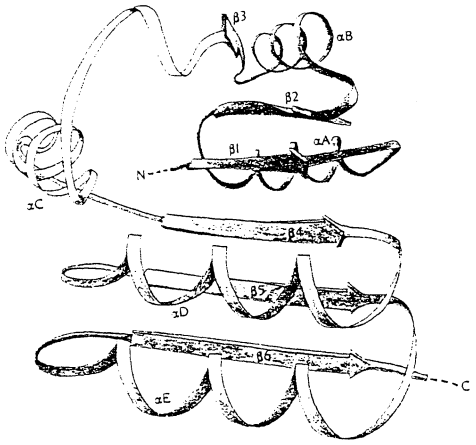


Fig. 3.1 A ribbon diagram of the three dimensional structure of a protein showing the secondary structure elements, αA to αE for α-helices and β1 to β6 for β-strands.

Most attempts to predict protein structure have concentrated on predicting the elements of secondary structure, because 90% of the residues in most proteins are involved either in α-helices (38%), β-strands (20%), or reverse turns (32%) (Figure 3.1). If the secondary-structure elements could be predicted accurately, it might be feasible to pack them together to generate the correct folded conformation. The various amino acid residues do demonstrate conformational preferences. The relative tendencies of the various residues to be involved in α-helices, β-strands, and reverse turns are given by the conformational preferences P_α, P_β, and Pt, respectively. (Table 3.2). These preferences, however, are only marginal. For example, the most helix-preferring amino acid, glutamine, occurs in helices only 59% more frequently than random, and even glycin and proline residues are found in helices about 40% as often as random, even though they are not stereochemically compatible with helical conformation. Fortunately, helices, β-sheet, and reverse turns are not determined by a single residue but by a number of them adjacent in the sequence. A segment of a particular secondary structure is much more probable when several adjacent residues prefer that structure. A number of

prediction schemes based on such empirical observations have been proposed. The classical method, and the best known, is that of Chou and Fasman, which classifies the amino acids as favouring, breaking, or being indifferent to each type of conformation. An α-helix is predicted if four out of six adjacent residues are helix-favouring and if the average value of Pα is greater than 1.0 and greater than Pβ; the helix is extended along the sequence until either proline or a run of four sequential residues with average value of Pα < 1.0 is reached. A β-strand if three out of five residues are sheet-favouring and if the average value of Pβ is greater than 1.04 and greater than P_α; the strand is extended along the sequence until a run of four residues with an average value of Pβ < 1.0 is reached. A reverse turn is the likely conformation when sequences of four residues characteristic of reverse turns are found (Table 3.2). Random assignment of these three states to residues in a polypeptide chain will give an average score of 33% correctly predicted states. The accuracy of Chou and Fasman's method is only 50%, but fortunately, this method has been further developed by using steriochemical and evolutionary information in combination with neural networks and today the accuracy for single amino acid residues has reached above 70%, and above 85% for segments (Rost *et al.*, 1994). Furthermore, it is possible to predict transmembrane helices at 95% accuracy using a related neural network method by Rost *et al.* (1995).

Table 3.2 Conformational preferences of the amino acids in the secondary structures of proteins.

	Preference[a]		
Amino acid residue	α-helix (P_α)	β-strand (P_β)	Reverse turn (P_t)
Glu	**1.59**	0.52	1.01
Ala	**1.41**	0.72	0.82
Leu	**1.34**	1.22	0.57
Met	**1.30**	1.14	0.52
Gln	**1.27**	0.98	0.84
Lys	**1.23**	0.69	1.07
Arg	**1.21**	0.84	0.90
His	**1.05**	0.80	0.81
Val	0.90	**1.87**	0.41
Ile	1.09	**1.67**	0.47
Tyr	0.74	**1.45**	0.76
Cys	0.66	**1.40**	0.54
Trp	1.02	**1.35**	0.65
Phe	1.16	**1.33**	0.59
Thr	0.76	**1.17**	0.90
Gly	0.43	0.58	**1.77**
Asn	0.76	0.48	**1.34**
Pro	0.34	0.31	**1.32**
Ser	0.57	0.96	**1.22**
Asp	0.99	0.39	**1.24**

[a] The normalized frequences for each conformation (e.g., P_α, P_β, P_t) were calculated from the fraction of residues of each amino acid that occurred in that conformation, divided by this fraction for all residues. Random occurrence of a partricular amino acid in a conformation would give a value of unity.

3.1.2 Coiled-Coil Prediction

A special case of α-helices is the coiled coil structures which are formed by two or more parallel or anti-parallel α-helices and they display a typical pattern of hydrophilic and hydrophobic residues that is repeated every seventh residue. The seven positions of the heptad repeat are designated *a* to *g*, *a* and *d* being generally hydrophobic. The sequences of coiled-coil proteins exhibit common patterns of amino acid distribution that appear to be distinct from those of other proteins. Lupas *et al.* (1991) showed that by comparing the sequence of a given protein with that of known coiled-coil proteins it is possible to predict regions of coiled-coil structures. A data base of known coiled-coil sequences was used to analyse the frequency of occurrence of each amino acid at each position of the coiled-coil heptad repeat and then divided by the overall frequency of occurrence of each amino acid in the complete GenBank to establish the relative frequencies in coiled coils. The relative frequency is then used for the evaluation of protein sequences by means of a gliding window of 28 residues. For each position of the gliding window a score is calculated that allowed estimation of the probability that an amino acid would be in a coiled-coil structure.

3.2 Circular Dichroism (CD) spectroscopy.

Circular dichroism is a spectroscopic technique which is sensitive to the secondary structure of proteins (Johnson, 1988). As shown in Figure 3.2, the different types of secondary structures show distinct CD spectra.

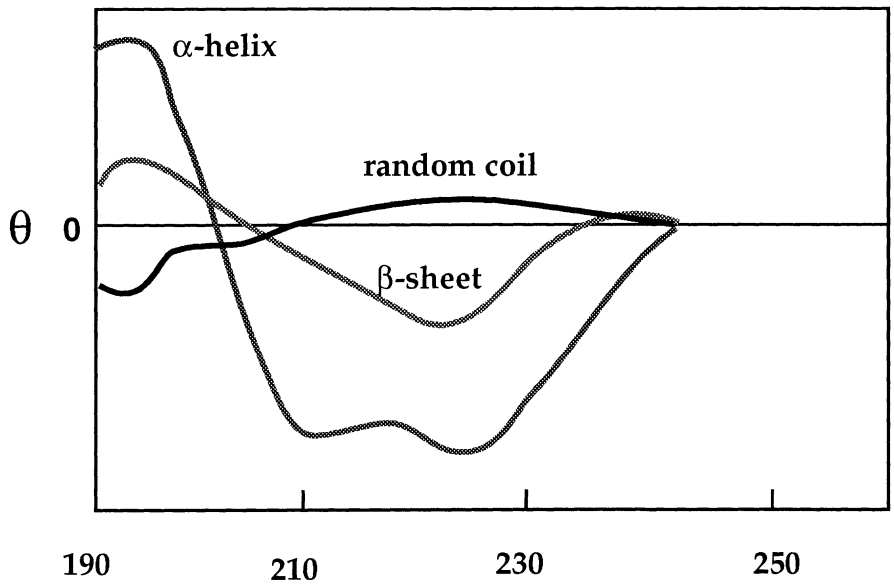


Fig.3.2 Characteristic CD-signal for the secondary structures α-helix and β-sheet, and for an unfolded peptide chain.

CD spectroscopy is a variant of electronic absorption spectroscopy using isotropic light. In CD, plane polarised light is used and one detects the difference in absorption of the left-and right circularly polarised components. CD offers the possibility to relatively easily study the effects of solvent, temperature and ligand binding on a protein's structure. CD is ideally suited for the initial characterisation of mutant proteins to investigate if the changes in the primary structure will still result in a folded conformation. In absorption spectroscopy, the amount of light absorbed by the sample is measured. The absorbance A, depends on the concentration of the sample, C, and the path length of the sample cell, l. This can be expressed in Beer-Lamberts law:

$$A(\lambda)=\varepsilon(\lambda)lC \tag{3.1}$$

where the proportionality constant, $\varepsilon(\lambda)$ is called the absorption coefficient. The CD of a sample is, as mentioned above, the difference in absorption of left and right circularly polarised light. Using the equation above this can be written:

$$A_L(\lambda)-A_R(\lambda)=[\varepsilon_L(\lambda)-\varepsilon_R(\lambda)]lC=\Delta\varepsilon(\lambda)lc \tag{3.2}$$

From the above expression can be seen that the CD of a molecule can be either positive or negative (see Figure 1), and that the molecule under study must be optically active to give a CD effect. This present no problem since most biological molecules are asymmetric. The data from the CD spectrometer is, for historical reasons, usually reported as the ellipticity, θ:

$$\theta = 32.98\ x[A_L(\lambda)-A_R(\lambda)] \tag{3.3}$$

Instead of ellipticity the mean residue ellipticity $[\theta]_{MRW}$ is normally the unit used when reporting CD-data, since this unit is independent of the sample concentration, the path length, and the molecular weight of the studied molecule.

$$[\theta]_{MRW} = \frac{\theta \times 100 \times MRW}{C \times l} = \frac{\theta \times 100 \times Mr}{C \times l \times N_A} \tag{3.4}$$

where MRW is the mean residue weight, M_r is the molecular weight, and N_A is the number of amino acid residues in the studied protein molecule.

A number of different methods have been proposed in the literature for the estimation of the secondary structure content of proteins through the use of CD-spectroscopy

(Greenfield & Fasman, 1969; Chen *et al.*, 1974; Provencher & Glöckner, 1981; Manvalan & Johnson, 1987; Sreerama & Woody, 1993). Usually, the CD spectrum of the studied peptide is fit to a database containing a number of CD-spectra of proteins with known three-dimensional structure. The concentration of the sample must be accurately determined to allow a reasonable estimation of the secondary structure content. This can be done, for example, by quantitative amino acid analysis. It is also important to measure CD to as a low wavelength as possible, since data measured to 200 nm are sufficient for determining two structure elements, whereas data measured to 190 nm are adequate to estimate nearly four structure elements (Johnson, 1990).

3.3 X-ray Crystallography

In diffraction experiments a narrow and parallel X-ray beam is directed onto a protein crystal to produce a diffraction pattern (Figure 3.3). The primary beam must strike the crystal from many different directions to produce all possible diffraction spots, and so the crystal is rotated in the beam during the experiment. To determine the structure of a protein it is necessary to compare X-ray data from native crystals of the protein with those from crystals in which different atoms of the protein are complexed with heavy metals.

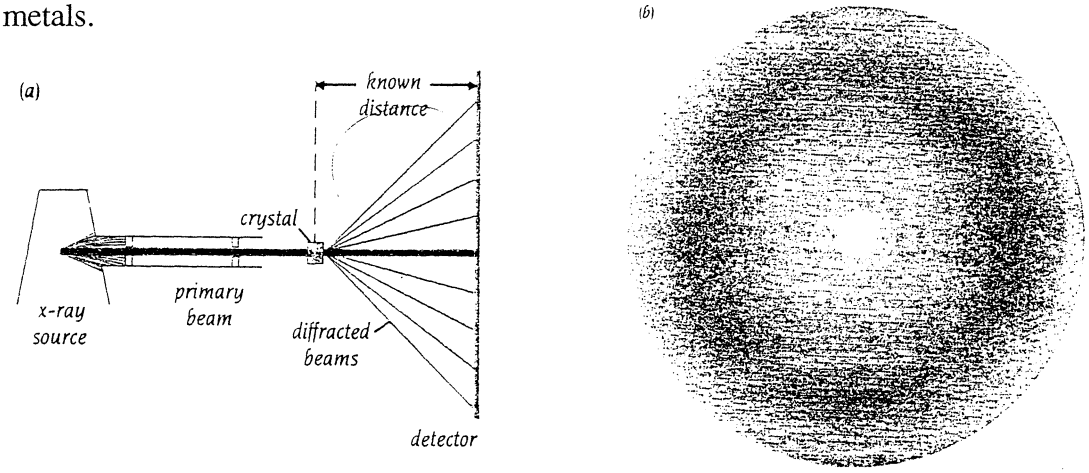


Fig. 3.3 Schematic representation of a diffraction experiment. (a) A narrow beam of X-rays is emitted from the X-ray source. When then the primary beam hits the crystal, most of it passes straight through, but some is diffracted by the crystal. These diffracted beams, which leave the crystal in many different directions, are recorded on a detector. (b) An example of a diffraction pattern from a crystal of protein.

In total several hundred thousand diffraction spots are usually collected and measured for each protein. When the primary beam from a X-ray source strikes the crystal, most of the X-rays travel straight through it. Some, however, interact with electrons on each atom and cause them to oscillate. The oscillating electrons serve as a new source of X-rays, which are emitted in almost all directions. We refer to this rather loosely as scattering.

When atoms and hence their electrons are arranged in a regular three-dimensional array, as in a crystal, the X-rays emitted from the oscillating electrons interfere with one another. In most cases, these X-rays, colliding from different directions, cancel each other; those from certain directions, however, will add together to produce diffracted beams of radiation that can be recorded as a pattern on a photographic plate or detector. Only those X-rays that positively interfere with one another, according to Bragg's law, give rise to diffracted beams that can be recorded as a distinct diffraction spot above background. Each diffraction spot is the result of interference of all X-rays with the same diffraction angle emerging from all atoms. For a typical protein crystal, myoglobin, each of about 20,000 diffracted beams that have been measured contains scattered X-rays from each of the around 1500 atoms in the molecule. The diffraction analysis determines the structure of not just of one molecule but of the approximately 10^{15} molecules that makes up the crystal.

Each diffracted beam, which is recorded as a spot on the film, is defined by three properties: the amplitude, which can be measured from the intensity of the spot; the wavelength, which is set by the X-ray source; and the phase, which is lost in the X-ray experiment. We need to know all three properties for all of the diffracted beams to determine the position of the atoms giving rise to the diffracted beams. How do we find the phases of the diffracted beam? This is the so-called phase problem in x-ray crystallography. The phase problem in the case of proteins was initially solved by the method of multiple isomorphous replacement (MIR) described by Max Perutz and John Kendrew and their co-workers (Perutz, 1956). This technique depends on the preparation of protein crystals into which heavy atoms that scatter X-rays very strongly (e.g., uranium, platinum, or mercury) have been introduced at a few specific positions without otherwise affecting the crystal structure. The heavy atoms must be in only one or a few positions of each asymmetric unit so that their positions in the unit cell can be deduced from the way they alter the protein diffraction pattern. After the positions of the heavy atoms have been determined, both the intensities and the phases of their diffraction pattern can be calculated. And when the phase of the heavy-atom contribution is known, the phase of the protein contribution can be determined. The crystal structure of a protein can often be determined without measuring the phases of the reflections in the three-dimensional structure if the structure of a similar molecule is known, by means of the molecular replacement technique. The known structure is used as an initial model for the new structure, and the information in the amplitudes of the reflections of the new crystal is used to find the position and orientation within the unit cell of the model structure. This initial model is then refined using the measured amplitudes and the phases calculated from the model.

The structure of the unit cell of the crystal is reconstructed by recombining mathematically the individual reflections of the diffraction pattern, a computation know as a Fourier synthesis. The electron density ρ at a point (x, y, z) in the unit cell, where x, y, and z are expressed as fractions of the unit cell dimension a, b, and c, is given by

$$\rho(x, y, z) = 1/V \left(\sum_h \sum_k \sum_l F(h, k, l)\, e^{i\alpha(h, k, l)}\, e^{-2\pi i(hx + ky + lz)} \right)$$

where V is the volume of the unit cell; F(h, k, l) is the amplitude (the square root of the intensity) of the reflection with indices h, k, and l; and α(h, k, l) is the phase. To extract information about individual atoms from such a system requires considerable computation.

The amplitudes and the phases of the diffraction data from the protein crystal are used to calculate an electron-density map of the repeating unit of the crystal. Building the initial model is a trial-and error process. First, one has to decide how the polypeptide chain weaves its way through the electron-density map. The resulting chain trace constitutes a hypothesis, by which one tries to match the density of the amino acid side chains to the known sequence of the polypeptide. This sounds easy, but it is not; a map showing continuous density from N terminus to C terminus is rare. More usually one produces a number of matches between the electron density and discontinuous regions of the sequence that may initially account for only a small fraction of the molecule and may be internally inconsistent. When a reasonable chain trace has finally been obtained, an initial model is built to give the best fit of the atoms to the electron density. Today computer graphics are exploited both for chain tracing and for model building to present the data and manipulate the models.

3.4 Nuclear Magnetic Resonance (NMR) spectroscopy.

Certain atomic nuclei, such as 1H, 13C, 15N, and 31P have a magnetic moment or spin. The chemical environment of such a nuclei can be probed by nuclear magnetic resonance, NMR, and this technique can be exploited to give information on distances between atoms in a molecule. These distances can then be used to derive a three-dimensional model of the molecule. Most structure determinations of protein molecules by NMR have used the spin of 1H, since hydrogen atoms are abundant in proteins.

When protein molecules are placed in a strong magnetic field, the spin of their hydrogen atoms align along the field. This equilibrium alignment can be changed to an excited state by applying radio frequency (RF) pulses to the sample. When the nuclei of the protein molecule revert to their equilibrium state, they emit RF radiation that can be measured. The exact frequency of the emitted radiation from each nucleus depend on the molecular environment of the nucleus and is different for each atom, unless they are chemically equivalent and have the same molecular environment (Figure 3.4a). These different frequencies are obtained relative to a reference signal and are called chemical shifts. The nature, duration, and combination of applied RF pulses can be varied enormously, and different molecular properties can be probed by selecting the appropriate combination of pulses.

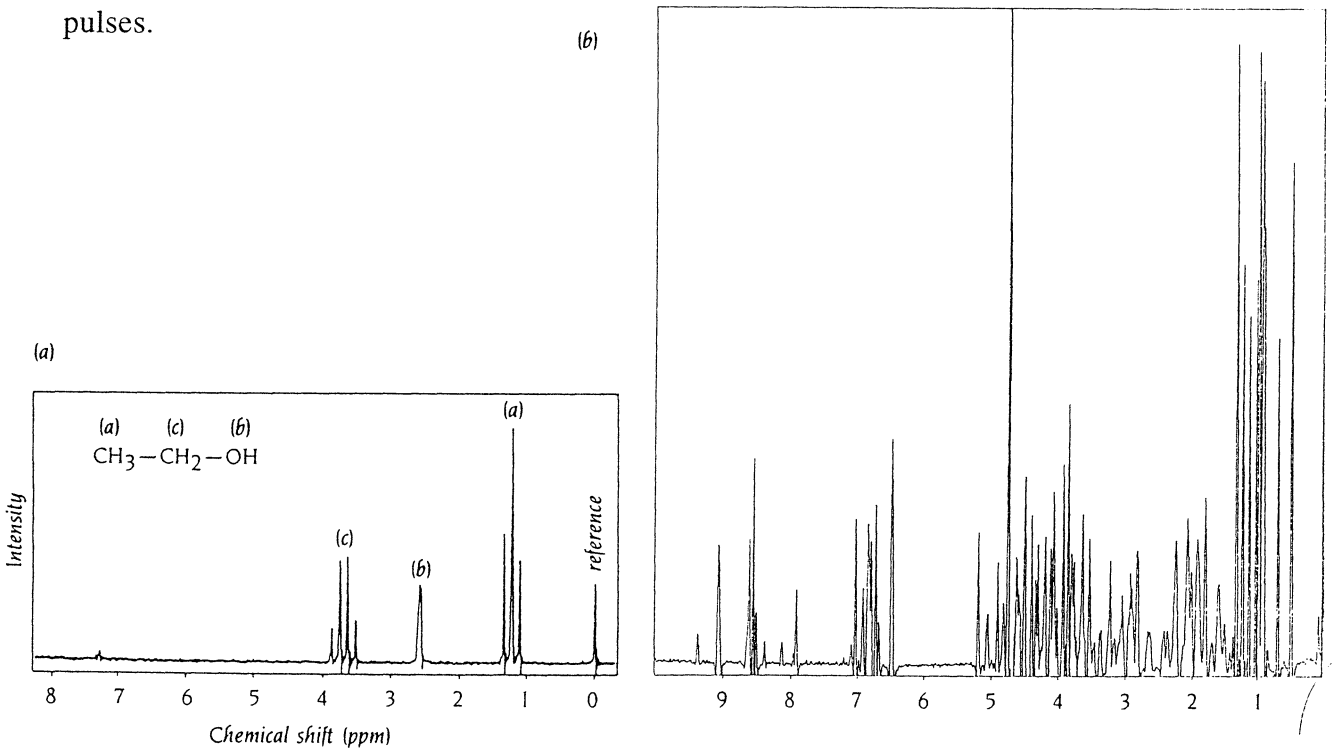


Fig 3.4 One-dimensional NMR spectra. (a) 1H-NMR spectrum of ethanol. The NMR signals (chemical shifts) for all the hydrogen atoms in this small molecule are clearly separated from each other. In this spectrum the signal from the CH_3 protons is split into three peaks and that of the CH_2 protons into four peaks close to each other, due to the experimental conditions. (b) 1H-NMR of a small protein comprising 36 amino acid residues. The NMR signals from many individual hydrogen atoms overlap and peaks are obtained that comprise signals from many hydrogen atoms.

In principle, it is possible to obtain a unique signal (chemical shift) for each hydrogen atom in a protein molecule, except those that are chemically equivalent, for example, the protons on the CH_3 side chain of an alanine residue. In practice, however, such one dimensional NMR spectra of protein molecules (Figure 3.4b) contain overlapping signals from many hydrogen atoms because the differences in chemical shift are often smaller than the resolving power of the experiment. In resent years this problem has been

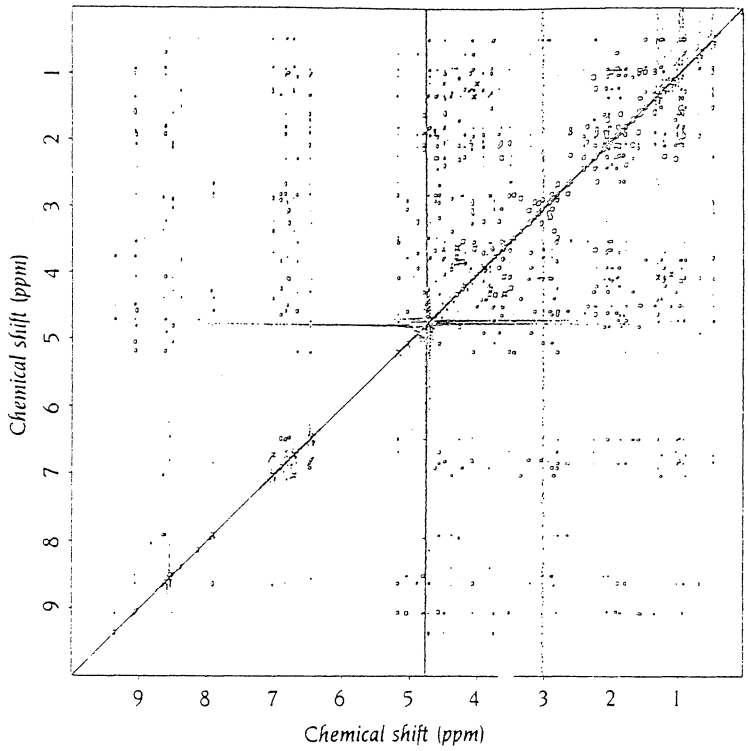


Fig. 3.5 Two dimensional NMR spectrum of the C-terminal of a small protein. The peaks along the diagonal correspond to the spectrum shown in Figure 3.4. The off-diagonal peaks in this NOE spectrum represent interactions between hydrogen atoms that are closer than 5Å to each other in space. From such a spectrum one can obtain information of both the secondary and tertiary structures of the protein.

overcome by designing experimental conditions that yield a two-dimensional NMR spectrum, the result of which is usually plotted in a diagram as shown in Figure 3.5. The diagonal in such a diagram corresponds to a normal one-dimensional NMR spectrum. The peaks off the diagonal result from interactions between hydrogen atoms that are close to each other in space. By varying the nature of the applied RF pulses these off-diagonal peaks can reveal different types of interactions. A COSY (correlation spectroscopy) experiment gives peaks between hydrogen atoms that are covalently connected through one or two other atoms, for example, the hydrogen atoms attached to the nitrogen and Cα

atoms within the same amino acid residue (Figure 3.6a). An NOE (nuclear Overhauser effect) spectrum, on the other hand gives peaks between pairs of hydrogen atoms that are close together in space even if they are from amino acids residues that are quite distant in the primary sequence (Figure 3.6b).

It is far from trivial, to assign the observed peaks in the spectra to hydrogen atoms in specific residues along the polypeptide chain because the order of peaks along the diagonal has no simple relation to the order of amino acids along the polypeptide chain. This problem has in principle been solved by Kurt Wüthrich and co-workers in Zürich, where the method of sequential assignment was invented (Wüthrich, 1986). Each type of amino acid has a specific combination of cross-peaks, a "fingerprint", in a COSY spectrum. From the COSY spectrum it is therefore possible to identify the H atoms that belong to each amino acid residue and, in addition, determine the nature of the side chains of that residue. However, the order of these fingerprints along the diagonal has no relation to the amino acid sequence of the protein. The sequence assignment, however, can be made from NOE spectra that record signals from H atoms that are close together in space. The signals in the NOE spectra in principle make it possible to determine which fingerprint in the COSY spectra comes from a residue adjacent to the one previously identified.

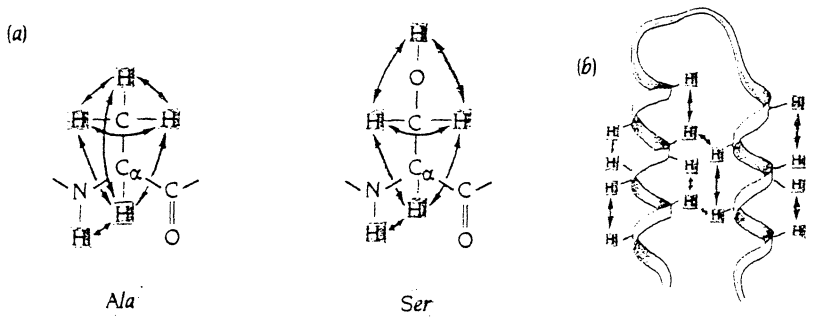


Fig. 3.6 (a) COSY NMR experiments give signals that correspond to hydrogen atoms that are covalently connected through one or two other atoms. Since hydrogen atoms in two adjacent residues are covalently connected through at least three other atoms (for instance, HCa-C'-NH), all COSY signals reveal interactions within the same amino acid residue. These interactions are different for different types of side chains. The NMR signals therefore give a "finger print" of each amino acid. The diagram illustrate the fingerprints of residues alanine and serine. (b) NOE NMR experiments give signals that correspond to hydrogen atoms that are close together in space (less than 5Å), even they may be far apart in the amino acid sequence. Both secondary structure and tertiary structures of small protein molecules can be derived from a collection of such signals, which define distance constraints between a number of hydrogen atoms along the polypeptide chain.

The final result of the sequence-specific assignment of NMR signals is a list of distance constraints from specific hydrogen atoms in one residue to hydrogen atoms in an second amino acid residue. The list immediately identifies the secondary structure elements of the protein molecules because both α helices and β sheets have very specific sets of interactions of less than 5 Å between their hydrogen atoms. Furthermore, this list of distance constraints between atoms in the molecule, can also be used to calculate the three dimensional structure of the protein molecule.

NMR and X-ray crystallography are in many respects complementary. X-ray crystallography deals with the structure of the proteins in the crystalline state, while NMR determines the structure in solution. The time scales of the measurements are different: NMR is more suitable for investigation of various dynamic processes such as those during folding, while X-ray crystallography is more suitable for characterisation of protein surfaces and the water structure around the protein. X-ray crystallography remains the only method to determine the structure of large protein molecules, whereas NMR is the method of choice for small protein molecules that might be difficult to crystallise.

3.5 References

Chou, P.Y. and Fasman, U.D. (1974) *Biochemistry* **13**, 211-215.

Chen, Y.-H., Yang, J. T. & Chay, K. H. (1974) *Biochemistry* **13**, 3350-3359.

Greenfield, N. & Fasman, G. D. (1969) *Biochemistry* **8**, 4108-4116.

Johnson, W, C. Jr. (1988) Ann. Rev. Biophys. *Biophys. Chem.* **17**, 145-166.

Lupas, A., van Dyke, M., and Stock, J. (1991) *Nature* **252**, 1162-1164.

Manavalan, P. & Johnson, W. C., Jr. (1987) *Anal. Biochem.* **167**, 76-85.

Perutz, M.F. (1956) *Acta Crystallogr.* **9**, 867-873.

Provencher, S. W. & Glöckner, J. (1981) *Biochemistry* **20**, 33-37.

Rost, B. and Sander, C. (1994) *Proteins Struct Funct Genet* **19**, 55-72.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995) *Protein Science* **4**, 521-533.

Sreerama, N. & Woody, R. W. (1993) *Anal. Biochem.* **209**, 32-44.

Suzuki, E. and Robson, B. (1976) *J. Mol. Biol.* **107**, 357-367.

Wüthrich, K. (1986) NMR of proteins and nucleic acids, Wiley & Sons Inc.

4. Structure and Stability Determination of Protein H and the M1 Protein from *Streptococcus pyogenes* - A Case Study

4.1 Introduction

The M protein family is a large group of fibrous streptococcal surface molecules that share similar C-terminal halves proximal to the cell surface, and have highly variable N-terminal halves distal to the cell surface. M proteins constitute one of the key virulence factors of *Streptococcus pyogenes* due to their anti-phagocytosis properties and their capacity to induce host-cross reactive antibodies (for references see Fischetti, 1989 and Kehoe, 1994). Ig-binding proteins from *S. pyogenes* were originally regarded as a separate group of molecules, but sequence analysis showed that these proteins are structurally closely related to M proteins (Heath and Cleary, 1989; Frithz *et al.*, 1989), and they are now regarded as members of the M protein family. In some cases, M proteins have also been shown to bind Ig (Schmidt & Wadström, 1990; Retnoningrum *et al.*, 1993; Retnoningrum and Cleary, 1994; Åkesson *et al.*, 1994). The role for Ig-binding in virulence and pathogenesis is yet not understood, but most clinical isolates of *S. pyogenes* do express these surface molecules (Lindahl and Stenberg, 1990).

Sequence analysis of M proteins has revealed similarities to tropomyosin, myosin, laminin, and/or keratin, with a characteristic seven-residue (heptad) repeat structure (Hosein et a, 1979; Fischetti, 1989). Furthermore, antibodies against M proteins can cross-react with these proteins (Kaplan *et al.*, 1962; Dell *et al.* 1991; Vashishtha and Fischetti, 1993). Electron microscopy and ultracentrifugation studies reveals that M6 protein is a dimeric 50-60 nm long fibrillar molecule, which contains approximately 70% α-helix as estimated from circular dichroism spectra (Phillips *et al.*, 1981). In analogy with the mammalian fibrous proteins, a two-chain coiled-coil structure has therefore been proposed for the M proteins (Phillips *et al.*, 1981).

Coiled-coil proteins have a characteristic seven-residue (heptad) repeat, $(a{\cdot}b{\cdot}c{\cdot}d{\cdot}e{\cdot}f{\cdot}g)_n$, where positions *a* and *d* are normally occupied by hydrophobic residues (Hodges *et al.*, 1972; Cohen and Parry, 1990) and positions *e* and *g* by oppositely charged residues (McLachlan and Stewart, 1975; Stone *et al.*, 1975). Supercoiled α-helices have approximately 3.5 residues per turn which places the hydrophobic residues on one side of the helix (Figure 4.1)(Crick, 1953). Interactions between residues in these positions provide the major force in the formation of a two-chain coiled-coil structure. Ion pairs between the *e* and *g* positions in the interacting chains are, however, also important for the stability as well as the orientattion of the coiled-coil conformation (Zhou *et al.*, 1994; Monera *et al.*, 1994).

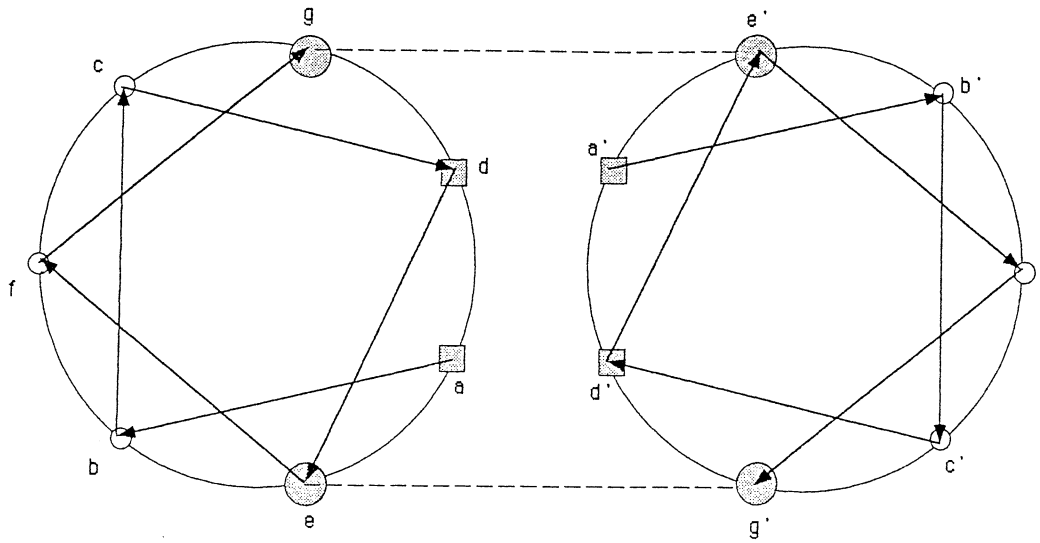


Fig. 4.1. Helical wheel projection of an α-helical coiled-coil structure as viewed from the N-terminus. The amino acids in the heptad are labeled *a* through *g*. The two wheels represent the two chains of the coiled-coil dimer. Positions *a* and *d* interact with positions a′ and *d′* on the adjacent chain. Interhelical ion-pairs can be formed between positions *e-g′* and *e′-g*, and are represented with dashed lines.

Protein H (Åkesson *et al.*, 1990) and the M1 protein (Åkesson *et al.*, 1994) are two IgG-binding proteins of the M protein family that are coexpressed on the surface of the AP1 strain of *S. pyogenes*. Besides its affinity for IgG, protein H also has affinity for albumin, fibronectin and N-CAM receptors (Åkesson *et al.*, 1994; Frick *et al.*, 1994; Frick *et al.*, 1995), whereas the M1 protein also binds albumin and fibrinogen. The IgG, albumin, N-CAM, and fibrinogen binding sites have been localized to different regions of the two bacterial molecules. It has recently been shown that the binding of IgG to protein H is temperature dependent, with high affinity for IgG at 4°C and 22°C, but surprisingly no affinity at physiological temperature (37°C) (Åkerström *et al.*, 1992). It was also shown that protein H exists as a dimer at these lower temperatures, while at 37°C it occurs as monomers, indicating that dimerisation of protein H is a prerequisite for the binding of IgG. A similar behavior was also observed for another Ig-binding protein, protein Arp4, and more recently also for protein Sir and the M1 protein (Cedervall *et al.*, 1995).

In this Case Study the structure and stability of protein H and the M1 protein have been further characterized. The results obtained are discussed and related to the structural and functional properties of other cell surface proteins of Gram-positive bacteria.

4.2 Material and Methods

4.2.1 Proteins.

Protein H was prepared by expression of the corresponding gene in *Escherichia coli* as previously described (Åkesson *et al.*, 1990; Gomi *et al.*, 1990). The A fragment of protein H was expressed (Frick *et al.*, 1994) and purified using size exclusion chromatography (Frick & Wikström, unpublished). The M1 protein and the M1 protein fragments were prepared as described by Åkesson *et al.* (1994). Polyclonal human IgG and human serum albumin were purchased from Sigma Chemical Co. These proteins were further purified using size exclusion chromatography. The parvalbumin sample was a generous gift from Eva Thulin, Lund University.

4.2.2 Physicochemical characterization.

The Stokes radius of protein H was measured using gel chromatography as described by Laurent and Killander (1964) at 20°C and 37°C. Bovine serum albumin (BSA) was used as a standard with known Stokes radius. The M_r of the denatured protein H was measured by gel chromatography in 6 M guanine HCl using BSA and ovalbumin as M_r-standards. Frictional ratios were calculated as described previously for protein L (Åkerström & Björck, 1989).

4.2.3 Circular Dichroism Spectroscopy.

Circular dichroism (CD) spectra were recorded on a Jasco J-720 Spectropolarimeter equipped with a thermostated cell holder. The spectra were recorded in the far ultraviolet (u.v.) region (260-190 nm) in cells with path lengths of 0.1 and 1.0 cm. The experiments were recorded in PBS buffer (8 mM sodium phosphate, 1.5 mM potassium phosphate, 0.12 M sodium chloride, 2.7 mM potassium chloride), pH 7.4. The concentration of the protein samples was determined by quantitative amino acid analysis. Spectra were acquired at a scan speed of 10 nm/min. and a 4 s response time. The solvent dichroic absorbance was subtracted using the Jasco software. The thermal unfolding curves were run both for the whole spectrum region and at a single wavelength (222 nm) characteristic for α-helical structures. The temperature was increased from 4°C to 90°C at a scan rate of 50 °C/h. A slower scan rate (20 °C/h) was also tested, but was shown to give the equivalent unfolding behavior. The higher scan rate was therefore used for all experiments.

4.2.4 Secondary structure estimation from CD spectra.

The program SELCON (Sreerama and Woody, 1993) was kindly made available by the authors. This program estimates the amount of secondary structure using a database of 17 proteins with well-characterized three dimensional structures.

4.2.5 Equilibrium urea denaturation.

The equilibrium denaturation experiments on protein H and the M1 protein in urea were performed in PBS buffer, pH 7.4 from 0 M to 5 M urea at a protein concentration of 4.2 μM. The samples were monitored at 2°C to prevent any effect of thermal denaturation of the sample. The unfolding was followed at a single wavelength (222 nm). The values for $[urea]_{50\%}$, the concentration of urea of which 50% of the protein is unfolded, and the free energy of unfolding were calculated from (Clarke and Fersht, 1993)

$$\Delta G_{U\text{-}F}^{D} = \Delta G_{U\text{-}F}^{H2O} - m[urea] \quad (4.1)$$

$$K_{U\text{-}F} = (F_F\text{-}F)/(F\text{-}F_U) \quad (4.2)$$

where m is the slope of the plot, F_F is the intensity of the CD of the folded state, F_U is the intensity of the CD of the unfolded form, and F is the intensity of the CD at a given urea concentration.

4.2.5 Computational sequence analysis.

The secondary structure analysis of protein H and the M1 protein was performed with the method of Rost *et al.* (1993, 1994 & 1995). The sequence Fourier transform and probabilities were determined with the method of McLachlan and Stewart (1976). An algorithm described by Parry (1982) was used to predict coiled-coil propensities based on the statistical preference of different amino acids for each position of the heptad repeat. The sequences were analyzed using a computer program ("coiledcoil", Lupas *et al.*, 1991) based on this algorithm.

4.3 Results

4.3.1 Secondary structure analysis.

Secondary structure analysis of protein H indicated that the extracellular part (sS, A, B, C regions) and part of the D region comprising amino acids 1-296 is predominantly (64%) α-helical (Figure 4.2). The proline rich sequence further towards the C-terminus (amino acids 299-350), a region likely to be located within the peptidoglycan of the cell wall, has a high content of non-helix structure. This region ends with the conserved bacterial cell wall LPSTGE motif and is followed by a helical hydrophobic region (amino acids 351-370) typical of a transmembrane sequence (Rost *et al.*, 1995). The analysis of the M1 protein reveals a similar secondary structure profile, with the extracellular part of the molecule (amino acids 1-404) consisting of 76% helical structure (Figure 4.2).

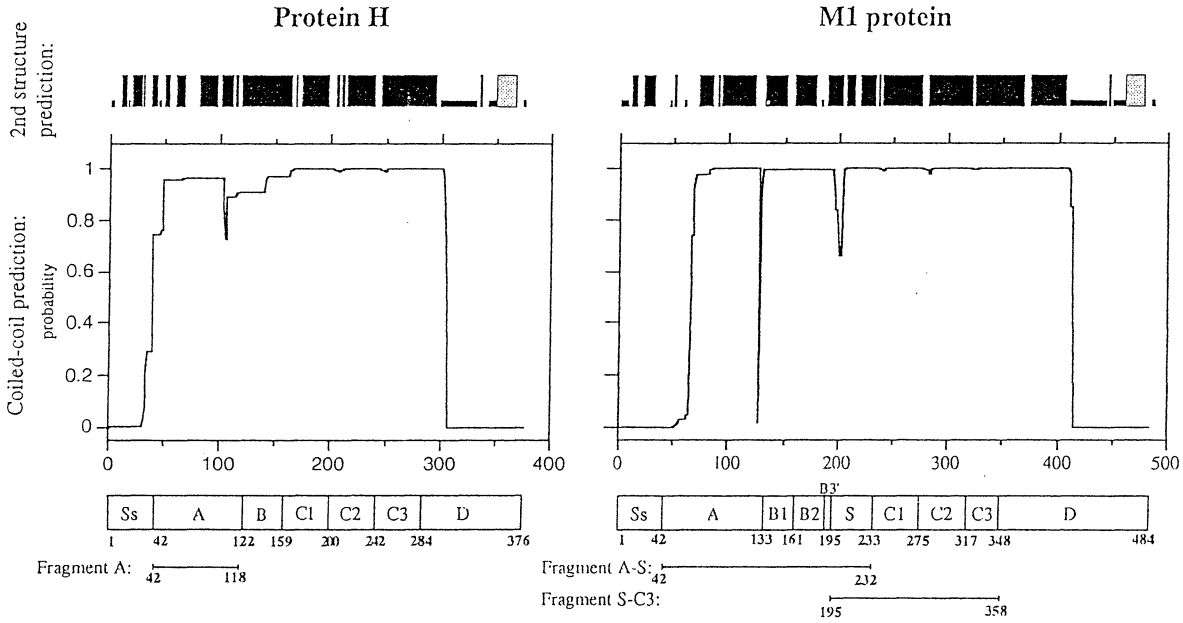


Fig. 4.2. Schematic representations of the primary structure of protein H and the M1 protein. Peptide fragments of the two proteins used in this work are depicted (lower). Prediction of secondary structure content are shown at the top with α-helix filled high bars, loop structures filled low bars, and transmembrane regions shaded. The probability of coiled-coil formation is shown in the middle of the Fig..

4.3.2 Prediction of coiled-coil structure.

The sequence of protein H (Gomi *et al.*, 1990) was analyzed by the statistical method described by Parry (1982). As depicted in Figure 4.2, a high probability for α-helical coiled-coil structure was predicted for almost the entire extracellular part (amino acids 49-302). The probability curve indicates that the A and B regions of protein H do not fit a coiled-coil structure as well as the C repeats. For the M1 protein, the major extracellular part of the molecule is also predicted to adopt a coiled-coil structure (amino acids 63-405). The signal peptide, and also a short part of the N-terminus of the mature protein,

have zero coiled-coil probability. Several other sequences from bacterial cell wall proteins were analyzed for their contents of α-helical coiled-coil structure. Some are shown in Figure 4.3. The result of the analysis show that the extracellular part of the

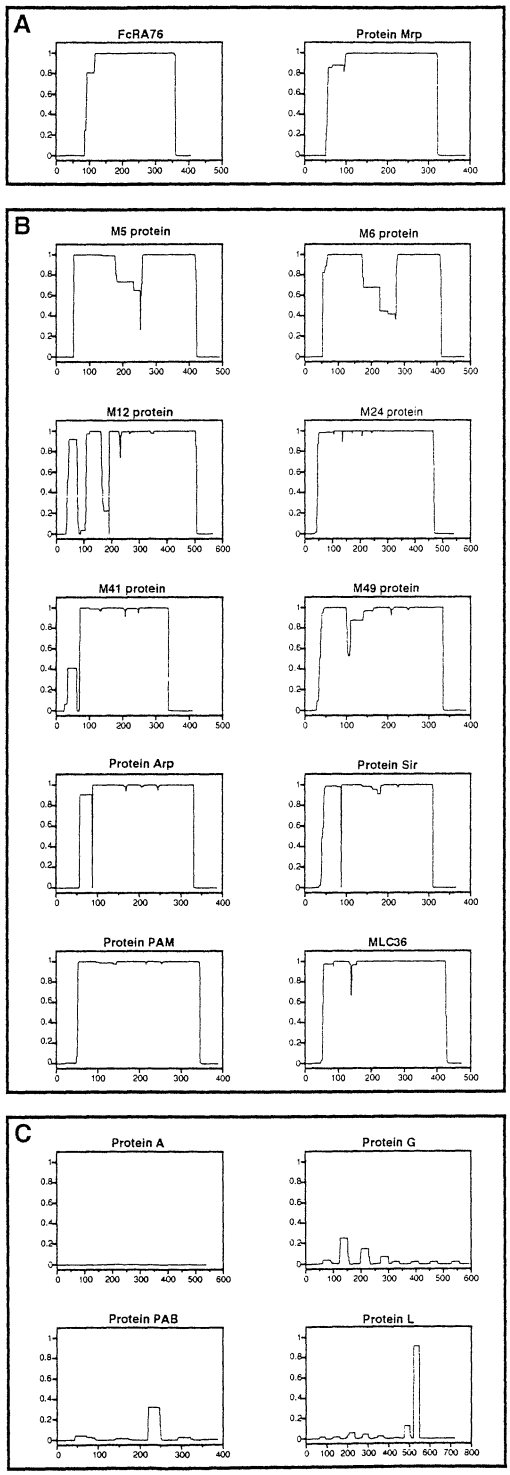


Fig. 4.3. Probability for the formation of coiled-coil structure for a selected set of bacterial surface proteins. (A) Proteins with A repeats: FcRA76 (Heath & Cleary, 1987) and Mrp (O'Toole *et al.*, 1992), (B) proteins with C repeats: M5 (Miller *et al.*, 1988), M6 (Hollingshead *et al.*, 1986), M12 (Robbin *et al.*, 1987), M24 (Mouw *et al.*, 1988), M41 (Podbielski, 1993), M49 (Khandke *et al.*, 1988), proteins Arp (Fritz *et al.*, 1989; Lindahl and Åkerström, 1989), Sir (Stenberg *et al.*, 1994), and PAM (Berge and Sjöbring, 1993), and MLC36 (Ben Nasr *et al.*, 1994). (C) Bacterial surface proteins not considered as members of the M protein family: protein A (Uhlén *et al.*, 1984), G (Guss *et al.*, 1986), PAB (de Chateau & Björck, 1994), and L (Karsten *et al.*, 1992).

various members of the M protein family show a high probability for coiled-coil structure, whereas the cell wall- and the transmembrane-spanning regions do not. Thus, besides the M proteins, the prediction analysis suggests that several of the M protein-related Ig-binding proteins, FcRA76 (Heath & Cleary, 1989) (amino acids 90-358), protein Mrp (O'Toole *et al.*, 1992) (amino acids 55-425), protein Arp (Fritz *et al.*, 1989; Lindahl & Åkerström, 1989) (amino acids 57-330), and protein Sir (Stenberg *et al.*, 1994) (amino acids 39-308) as well as the plasminogen-binding protein PAM (Berge & Sjöbring, 1993) (amino acids 50-334) and MLC36 (Ben Nasr *et al.*, 1994) (amino acids 53-425) adopt a coiled-coil structure. The length of the non-coiled-coil sequence of the N-terminal part of the mature protein seems to vary between different M proteins, from zero amino acids for protein H to approximately 50 residues for FcRA76. The M protein family can be divided into two major classes depending on the type of repeats found N-terminally of the wall spanning region. These two types of repeats are designated A and C (O'Toole *et al.*, 1992), and are shown in Figure 4.3A and 4.3B, respectively. As shown, the conserved C-terminal part of the extracellular region for both classes of the M protein family fit well in a coiled-coil structure. No general pattern can be seen for the less conserved N-terminal part of these molecules. M24 protein and protein PAM are, for example, given very high probability scores for the whole region, whereas for M5, M6 and M12 only segments receive the highest probability score, indicating that the heptad structure of the other regions are less optimal. This may be related to the fact that these M proteins have insertions (or deletions) resulting in distortions of their heptad structures (Khandke *et al.*, 1987, 1988; Fischetti *et al.*, 1988; Manjula *et al.*, 1991).

Proteins A, G, and L are Ig binding bacterial proteins which do not belong to the M protein family. The three-dimensional structure of the immunoglobulin-binding repeats of proteins A, G and L have been determined and found to be folded into well-defined globular structures (Gouda *et al.*, 1992; Gronenborn *et al.*, 1991; Lian *et al.*, 1992; Wikström *et al.*, 1993, 1994). As expected, no coiled-coil structure is predicted for these parts of the proteins, but the C2 repeat (amino acids 520-548) of protein L have a coiled-coil probability >0.8 (Figure 4.3C). The analysis also included the albumin-binding protein PAB (de Chateau & Björck, 1994), the IgD-binding protein D (Janson *et al.*, 1991), and the fibronectin-binding protein F (Hanski & Caparon, 1992). None of these bacterial surface proteins, which are not structurally related to the M protein family, have significant probabilities for coiled-coil structure (not shown).

4.3.3 The heptad structure of protein H and the M1 protein.

The amino acid residues in the region 40-303 of protein H and region 63-405 of the M1 protein were analyzed for heptad sequence periodicities by Fourier analysis (MacLachlan and Stewart, 1976). For the hydrophobic residues (L, I, V, Y) of protein H, significant

intensities were detected for the periodicity of seven and for harmonics of heptad repeats: 3.51 and 3.48~7/2, and 2.33~7/3 and the periodicity of 2 (Figure 4.4A). The hydrophobic amino acids of the M1 protein also showed significant intensities for heptad harmonics, especially at 3.5~7/2 (Figure 4.4B). Further examination of the amino acid sequences of protein H and the M1 protein revealed patterns specific to α-helical coiled-coil proteins. These are characterized by the presence of the form $(a{\cdot}b{\cdot}c{\cdot}d{\cdot}e{\cdot}f{\cdot}g)_n$ (McLachlan and Stewart, 1975) with the *a* and *d* positions being preferentially occupied by apolar amino acids (Figure 4.1). The periodicity, which extends from amino acid 42 through 297 of protein H (Figure 4.5A), is disrupted in two positions, after amino acids 103 and 139. For the M1 protein the periodicity starts at position 63 and extends to amino acid 405, and is disrupted in three locations, after amino acids 133, 196 and 392 (Figure 4.5B).

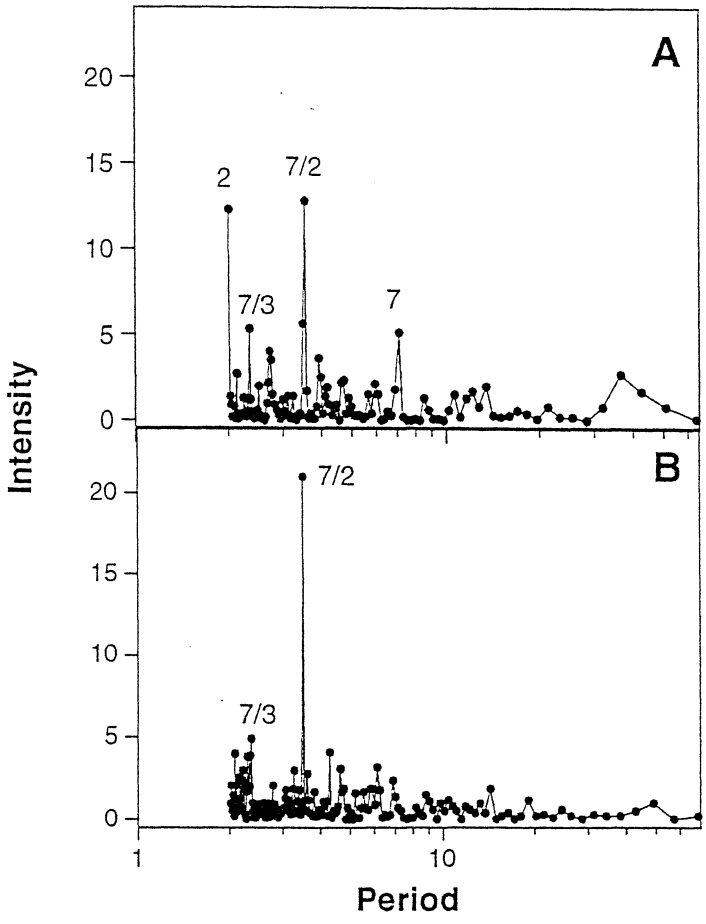
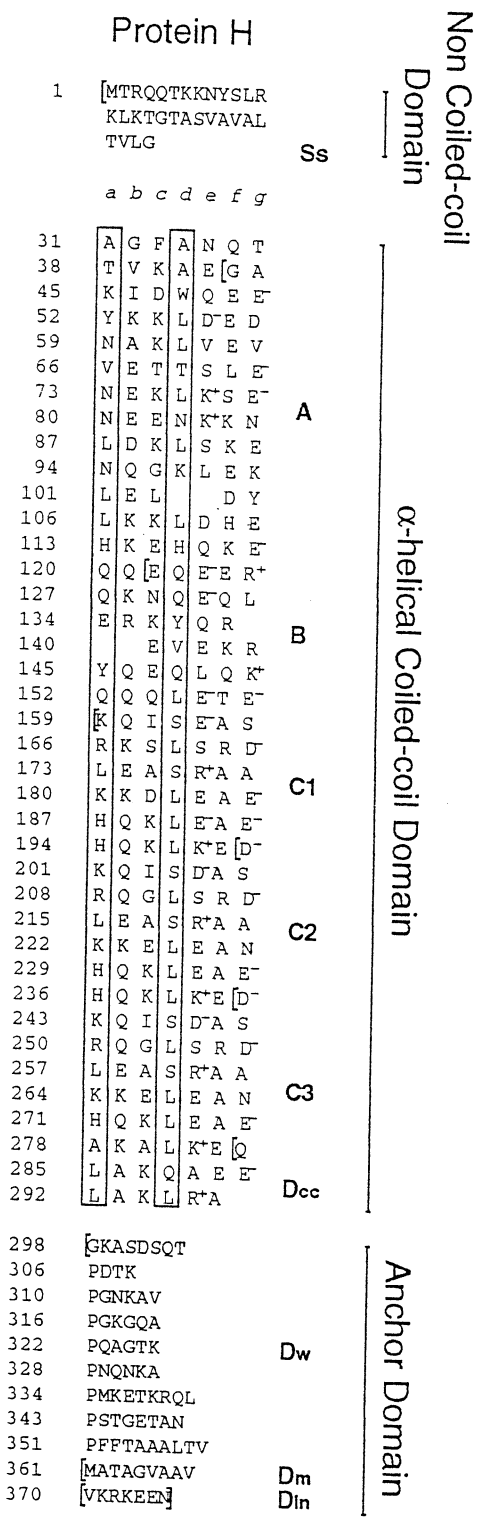


Fig. 4.4. Fourier analysis of the periodicity for the hydrophobic residues L, V, Y, and I in protein H (A) and M1 protein (B). Period values are presented in a logarithmic scale with significant peaks (i.e. a probability of random occurance that are less than 0.01) labeled by the period values.

For protein H and the M1 protein respectively, approximately 62% and 50% of the residues found in the *d* position are apolar (predominantly leucine), whereas only 34% and 35% apolar residues are found in the *a* position. In this position a high distribution of basic residues is seen (26% and 35%, respectively), which is a common property for all described M proteins that differentiate them from other described α-fibrous proteins (Conway and Parry, 1990). When the distribution of amino acids in the A, B, and C domains of protein H and the M1 protein is analyzed, similarities with other M proteins are evident. Thus, the A domain of protein H has a high content of asparagine (40%) in the *a* position and of apolar residues (50%) in *d* position. A similar pattern has previously been found in the corresponding domains (subdomain I) of proteins M5, M6, and M24 (Manjula *et al.*, 1991). The high occurrence of glutamine found in the *a* and *d* position of the B domain of protein H can be seen in subdomain II of M49 as well. The B segment is also a highly charged region which result in a lower probability for the formation of a coiled-coil in this part of protein H (Figure 4.2). Figure 5A also shows that the B-portion does not contain the expected Leu residues in either the *a* or the *d* position. Interestingly, when the sequence representing the B segment was aligned with other sequences in the database a significant identity score (41%) was found for the protein trichohyalin (Fietz *et al.*, 1993). Biophysical studies of this protein, indicate that it exists as a single extended α-helical structure in solution (Lee *et al.*, 1993). The A domain of the M1 protein has a high content of apolar residues in *a* (50%) and *d* (80%) positions which has previously been observed in M12, M49 and M57, and the high frequency of apolar residues in *a* and *d* position in the BS region has earlier been reported for the M12 and M57 proteins (Manjula *et al.*, 1991). Finally, the C repeats of proteins H and M1, show have a high sequence similarity with C repeats of other M proteins of the C class (Manjula *et al.*, 1991). This part of the molecule has a high proportion of basic residues (50%) in the *a* position in both protein H and the M1 protein, whereas a marked prevalence of histidine (27%) is found in this position only in protein H. Both proteins have a high portion of apolar residues (67% and 60%, respectively) and serine (33% and 40%, respectively) in the *d* position.

In thirty-four of the heptades identified in protein H, interhelical ionic interactions (Figure 4.1) are likely to occur between residues in the *g* and *e* positions of one chain, with the *e'* and *g'* positions of the other chain. Twenty-four of them could result in electrostatic attractions while twelve have the potential to induce electrostatic repulsions (Figure 4.5A). Similar results are found for the heptads of the M1 protein which contain twenty-six favorable ion pairs and eight unfavorable (Figure 4.5B). Analysis of the described heptad structures of M6, M12, M24 and M57 (Fischetti *et al.*, 1988; Manjula *et al.*, 1991) showed similar ratios between repulsive and attractive interactions in the *e* and *g* positions, 2/20, 6/28, 12/42, 12/18, respectively.



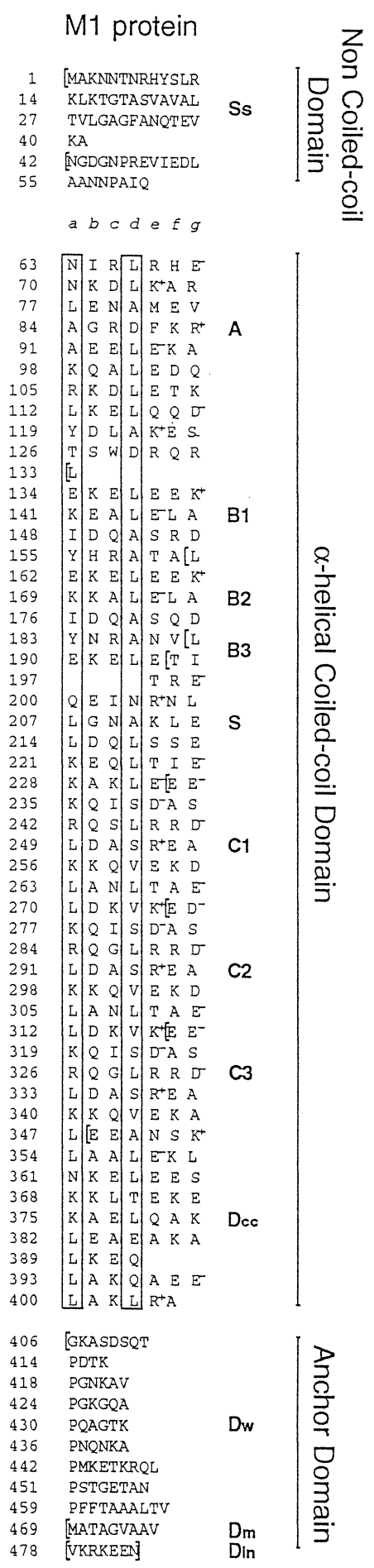


Fig. 4.5. Heptad structures of protein H (A), and the M1 protein (B). The domains are named as in Fig. 4.2 except for the D-domain which is divided into D_{cc} (coiled-coil), D_W (wall spanning), D_m (membrane associated), and D_{in} (intracellular) regions. Brackets, [, in the sequences indicate the beginning of each domain. The sign of the charged residues involved in potential electrostatic interactions between the e and g positions are denoted.

4.3.4 Physicochemical properties of protein H.

The Stokes radius of protein H was determined from gel filtration experiments at 20°C and 37°C. From these values, and the molecular mass from the amino acid sequence, the frictional ratio was calculated at the two temperatures (Table 4.1). The frictional ratio was calculated with the assumption that at the lower temperature, protein H is in a predominantly dimeric state, whereas at the higher temperature the molecule is found in a monomeric state. The frictional ratio is in both cases indicative of a highly elongated structure (Cantor & Schimmel, 1980). Interestingly, the frictional ratio representing the dimeric form is higher than the monomer form suggesting that the dimeric coiled-coil state, observed at lower temperatures, is more elongated than the monomeric form. Since the dimeric state is expected to be represented by an elongated α-helical coiled-coil, while the high temperature conformation is considered to adopt a flexible conformation, this is not surprising.

Table 4.1. Physicochemical properties of protein H.

Property	Value
Molecular mass	
from amino acid composition[a]	38,162
from SDS-polyacrylamide gel electrophoresis	42,000
from gel cromatography in guanidine-HCl	34,700
Stokes radius:	
r_S (T=20°C)	51.4 Å
r_S (T=37°C)	36.1 Å
Frictional ratio:	
f/f_O (T=20°C)	1.84
f/f_O (T=37°C)	1.63
Disulfide bonds	none
Absorption coefficient at 280 nm (ε)	$3.9 \times 10^4 cm^{-1}M^{-1}$ [b]

[a]The protein sequence minus signal peptide was deduced from the gene sequence (Gomi *et al.*, 1990).

[b]The M_r value of 38,162 was used.

4.3.5 Urea denaturation.

The stability of protein H and the M1 protein upon titration with urea was also measured by CD at 222 nm. Data were collected at 2°C to prevent effects due to thermal denaturation of the sample. The data fit, in both cases, a two state transition (Figure 4.6) and the entire set of data was fit to an unfolding transition curve based on equation 1 & 2 (see Material and Methods). The concentration midpoint for unfolding, $[urea]_{50\%}$, is 1.79 ± 0.005 M urea for protein H and 2.39±0.005 for the M1 protein, confirming the low stability of the proteins. The free energy of unfolding in water, $\Delta G_{U\text{-}F}^{H2O}$, was 25.1±0.1 kJ mol^{-1} and 31.6±0.1 kJ mol^{-1} for protein H and the M1 protein, respectively.

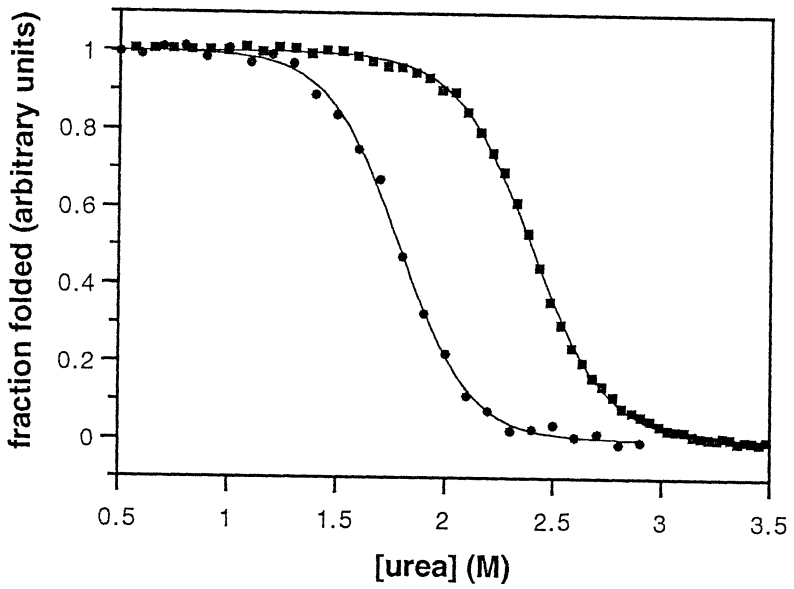


Fig. 4.6. The normalized unfolding curves for protein H ( ) and the M1 protein ( ) as a function of the concentration of urea. Data points representing CD intensities at 222 nm and solid curves represent theoretical two-state transitions fitted to the data (based on eqs. 4.1 & 4.2). The concentration of urea at which 50 % of the proteins are unfolded, $[urea]_{50\%}$, were 1.79±0.01 M and 2.40±0.01 M for protein H and the M1 protein, respectively. The ΔG_{H2O} and m values obtained from the analysis was for protein H 25.1±0.01 kJ/mol and 14.1±0.40 kJ/mol^2, respectively, and for the M1 protein 31.7±0.01 kJ/mol and 13.2±0.14 kJ/mol^2, respektivly. Shown are one selected set of data from several repeated analysis.

4.3.6 Secondary structure analysis of protein H using CD spectroscopy.

Far-u.v. CD spectra of protein H were recorded at various temperatures (Figure 4.7A). They reveal that protein H is a predominantly α-helical protein at 4°C, as indicated by the

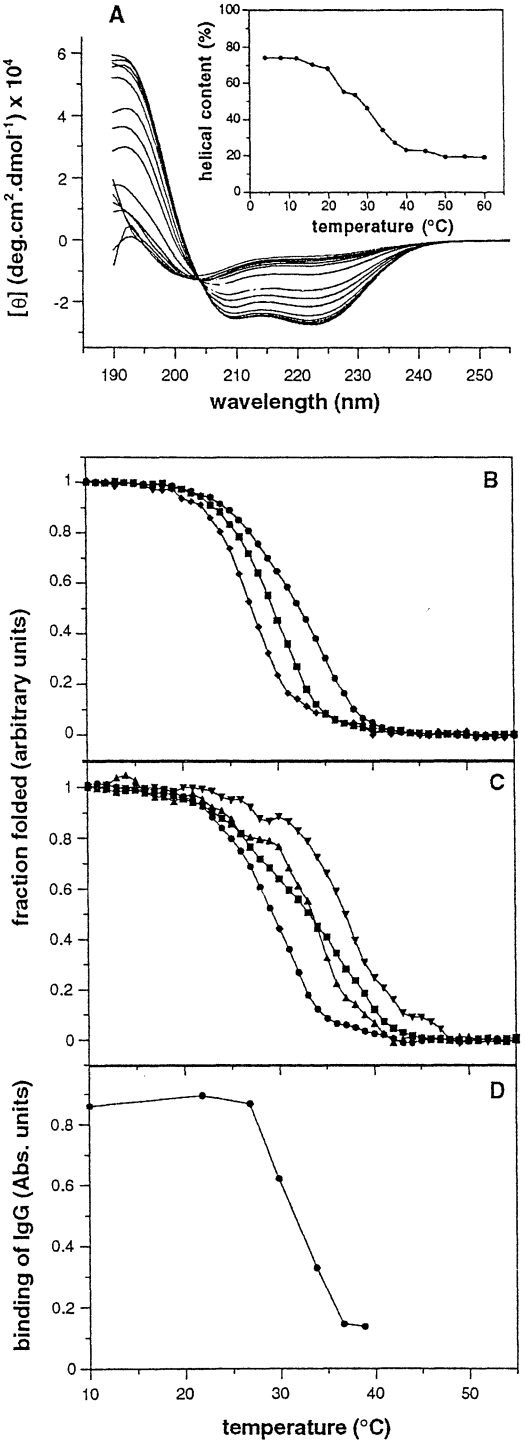


Fig. 4.7. CD analysis of protein H. (A) shows the far-u.v. spectral region of the intact protein H at increasing temperatures with the estimated helical content in the upper right corner. (B) The normalized CD-signal at 222 nm for protein H at 0.42 μM ( ), 4.2 μM ( ), and 65 μM ( ) as a function of temperature. (C) The CD at 222 nm for protein H free ( ), together with IgG ( ) or HSA ( ), and with both IgG and HSA ( ), and (D) the IgG-binding activity for protein H adapted from Åkerström *et al.* (1992), as a function of temperature .

double minima at 207 nm and 222 nm, as well as the maximum at 192 nm. The α-helical content was estimated from the CD spectra to be 74% using the method described by Sreerama & Woody (1993). The ratio of molar ellipticity ($[\theta]_{222}/[\theta]_{207}$=1.18 is similar to that observed previously for coiled-coils (Lau *et al.* 1984; Hodges *et al.* 1988, 1990; Zhou *et al.* 1992b,c) and different from structures with α-helices in non coiled-coil conformations in which the ellipticity at 207 nm is generally more negative than that at 222 nm (Cooper & Woody, 1990; Zhou *et al.*, 1993). The conformation of protein H is dramatically affected by an increase in temperature. The estimated α-helical content of protein H drops from approximately 55% to 27% when the temperature is raised from 24°C to 37°C. When the thermal denaturation of protein H is followed at 222 nm a smooth transition curve can be seen (Figure 4.7B) which correlates with the binding activity of IgG to protein H at various temperatures (Figure 4.7D). The denaturing process could be completely reversed by lowering the temperature back to 4°C. When three different concentrations of protein H were analyzed (0.42 μM, 4.2 μM and 65 μM), different melting temperatures (T_m values) could be observed, 27°C, 30°C and 32 °C, respectively (Figure 4.7B). Since the melting curve was concentration dependent, all further CD study of protein H were performed at a fixed concentration, 4.2 μM. CD spectra were also measured on a fragment of protein H representing the A domain (Figure 4.2). This domain showed a weak CD signal indicating no appreciable amount of secondary structure, and when heated to 90°C no change in the ellipticity was observed (not shown). This demonstrates that the isolated A domain does not adopt to a stable folded structure in solution.

4.3.7 Secondary structure analysis of the M1 protein using CD spectroscopy.

To further investigate its structural properties, the M1 protein was subjected to analysis by CD methods. The far u.v. region was, as in the case of protein H, indicative of an α-helical structure (~70% α-helix at 4°C) (Figure 4.8A). When heated, the M1 protein starts to unfold above 30°C and at 37°C only 52% are in a folded state. Analogous to protein H, this correlates with the drop in affinity for IgG to the M1 protein at this temperature (Cedervall *et al.*, 1995). The unfolding behavior was, however, different as compared to protein H, since the CD-intensity continued to decrease with increasing temperature (Figure 4.8B). This behavior suggests that the unfolding of the intact M1 protein does not undergo a simple two-state transition. A fragment from the N-terminal portion (fragment A-S) had a weak CD-signal with no indication of stable secondary structure and, as for the A-domain from protein H, no change in the CD intensity was observed when heating the sample (not shown). However, the isolated fragment of the M1 protein representing the S-C3 region was, according to the CD analysis, in an α-

helical conformation with an α-helix content of approximately 85% (Figure 4.8C). The temperature denaturation experiment is indicative of a two-state transition with a T_m~17°C (Figure 4.8D).

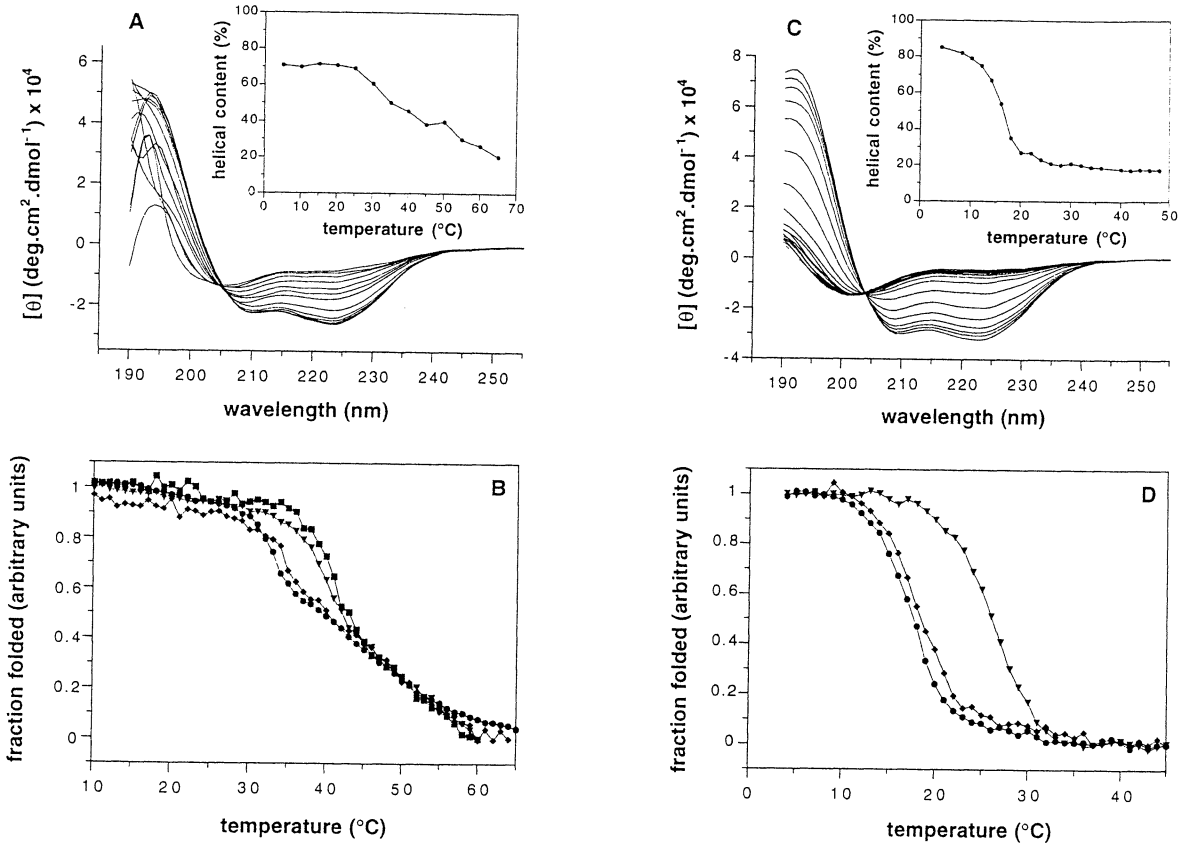


Fig. 4.8. CD analysis of the M1 protein. (A) The far-u.v. spectral region of the intact M1 protein at increasing temperatures with the estimated helical content in the upper right corner. (B) The unfolding curves (CD at 222 nm) for the M1 protein free ( ), together with IgG ( ) or HSA ( ), and both IgG and HSA ( ). (C) The far-u.v. spectral region of fragment S-C3 at increasing temperatures with the estimated helical content in the upper right corner. (D) Unfolding curves (CD 222 nm) for the S-C3 fragment free ( ), together with IgG ( ) or HSA ( ).

4.3.8 Effect on the thermal stability of protein H in the presence of ligands.

The far-u.v. CD spectra at 4°C were recorded for protein H alone and in the presence of equimolar concentration of different ligands: IgG, albumin, or IgG plus albumin. The concentrations were chosen such that >90% of protein H was calculated to be in complex with the ligands at low temperature. The final spectra were derived by subtracting the spectrum of the ligand(s) alone from the spectrum of the mixture of protein H and ligand(s). The final spectra at 4°C were not changed when IgG or albumin were added, indicating that no conformational change occurs when protein H binds these ligands (data not shown). Figure 4.7C shows the CD thermal transition curves at 222 nm of protein H alone and with the different ligands. When IgG is added to protein H, the low temperature conformation is stabilized, as seen by the increase of the melting temperature with 3°C (from 30 to 33°C). A similar effect is observed by the addition of albumin, an increase of 2°C (from 30 to 32°C). A synergistic effect is seen when both IgG and albumin are added to protein H, resulting in an even higher melting temperature (37°C), i.e. an increase of 7°C. In a negative control experiment parvalbumin was added to the protein H sample with no effect on the melting temperature (data not shown).

4.3.9 Efect on the thermal stability of the M1 protein and the S-C3 fragment in the presence of ligands.

As in the case of protein H, the thermal stability of the M1 protein was investigated without and together with the ligands HSA and IgG. The change in unfolding behavior of the M1 protein when adding IgG and HSA is shown in Figure 4.8B. HSA was shown to have a much larger effect on the unfolding temperature than IgG. Also the C-terminal fragment of the M1 protein, S-C3, was investigated together with ligands (Figure 4.8D). The addition of HSA had a dramatic effect on the unfolding behavior of the S-C3 fragment. The transition temperature, T_m~17°C observed for the free peptide increases to T_m~27°C with HSA present.

4.3.10 Antiparallel alignment of the heptad structures of protein H and the M1 protein.

Electron microscopy studies on streptococci reveals that M proteins of neighboring bacteria can interact with each other (Swanson *et al.*, 1969; Philips *et al.*, 1981; Fischetti, 1989). Our studies show that protein H and the M1 protein can fold and unfold their coiled-coil structures closes to 37°C. The low thermal stability may therefore allow these proteins to refold in an antiparallel fashion with M proteins on adjacent bacteria. Since intermolecular electrostatic interactions have been shown to be important for determining the orientation of two-stranded coiled-coils (Monera *et al.*, 1994), we decided to examine these interactions for aligned antiparallel sequences of protein H and the M1 protein with

themselves or with each other. All possible overlaps that fitted with the heptad structure of an antiparallel coiled-coil of protein H and the M1 protein were analyzed in respect to the distribution of different potential interhelical electrostatic interactions in the *g* and *e* positions. The electrostatic interactions of the antiparallel paring of protein H with itself are in a few cases more favorable than the electrostatic interaction of the equivalent parallel alignment. The best antiparallel alignments of protein H was found to have an overlap of 120 and 141 amino acid residues, respectively (Figure 4.9A). This would thus

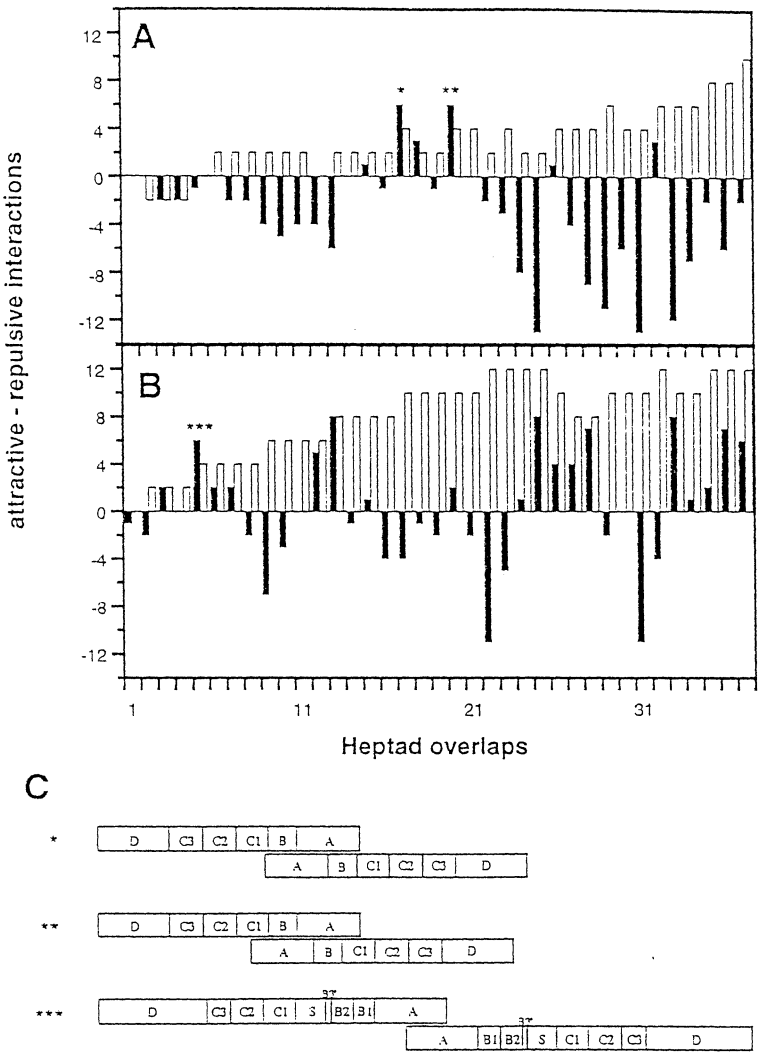


Fig. 4.9. Antiparallel and parallel alignment of the hepta structure of protein H and the M1 protein. The hepta structure of protein H (A) and the M1 protein (B) were aligned in an antiparallel (filled bars) orientation (with *a* paring with *d'* and *d* paring with *a'* in the hepta structure) or in the parallel (open bars) orientation (with *a* pairing with *a'* and *d* pairing with *d'*). An alignment of one hepta repeat overlap, is interpreted as an alignment of the first N-terminal hepta repeat, a two hepta repeat overlap is an alignment of the two first hepta repeats in the N-terminal of the protein, etc. The antiparallel and parallel alignments were analyzed for intermolecular electrostatic interactions in the *e* and *g* positions (i.e. for the antiparallel alignment *g* pairs with *g'* and *e* with *e'*, and for the parallel alignment *g* pairs with *e'* and *e* with *g'*) and the difference between attractive and repulsive interactions were calculated. Favorably antiparallel aligned sequenses are labled with stars and represented schematically in C.

involve the A and B regions, and in the latter case also the beginning of the C1 region (Figure 4.9C). This latter alignment results in eight attractive and two repulsive interactions. The repulsive interactions are, however, located in the N-terminus of each chain and can therefore be expected to be of less importance than the interior interactions. The same region of the parallel structure has ten attractive bonds and six repulsive interactions. The alignment of the M1 protein sequence with itself suggest that one position (an overlap of 38 residues), has potentially favorable ion-interactions for the antiparallel arrangement (Figure 4.9B). This alignment involves the N-terminal portion of the A domain (Figure 4.9C). Finally, the alignment of protein H with the M1 protein indicates also an overlap of 68 amino acids with potentially favorable (eight attractive- and no repulsive-) ion-interactions (not shown). These results suggest that an antiparallel coiled-coil structure could occur, and may be more favorable than the parallel organization.

4.4 Discussion

The Ig-binding proteins H, Arp, Sir, FcRA76, and Mrp as well as two plasminogen-binding proteins, protein PAM and PBP1 were shown to share the common overall coiled-coil structure known for M proteins, and they should therefore be regarded as members of the M protein family. This prediction method distinguish coiled-coil proteins from other α-helix rich proteins, like protein A, and is therefore a valuable tool in the evaluation of the structural relationship between various bacterial cell wall-associated proteins.

The sequence in the IgGFc-binding regions are different for proteins H, A, and G although it has been shown that they all bind to the Cγ2-Cγ3 interface area of IgG (Deisenhofer, 1981; Stone *et al.*, 1989; Frick *et al.*, 1992; Sauer-Eriksson *et al.*, 1995). Furthermore, the structure of these IgG-binding domains also exhibit considerable differences; the protein A domains are constructed as a three helix bundle, the protein G domains consist of a four stranded β-sheet with a single α-helix on top, and the protein H domains, as further evidenced in this report, have a two-stranded parallel α-helical coiled-coil structure. This indicates strongly that the three IgGFc-binding bacterial proteins have evolved their binding capacities through convergent evolution, suggesting that these surface proteins are connected with essential microbial functions adding selective advantages to the bacteria.

Previous findings have shown that the binding of Ig to protein H, M1 protein and proteins Arp and Sir is temperature dependent, i.e. they bind Ig with high affinity only at temperatures < 37°C (Åkerström *et al.*, 1992; Cedervall *et al.*, 1995). The temperature-

dependent binding was also seen at the bacterial surface. All four molecules belong to the M protein family and it is therefore possible that temperature dependent ligand binding is a common feature of this protein family. Interestingly, our results demonstrate that the drop in binding affinity can be directly correlated to the unfolding of the coiled-coil dimer of protein H (Figure 4.7B and 7D). The temperature unfolding curve of intact protein H suggests that the protein undergoes a two-state transition from a folded dimeric coiled-coil to an unfolded monomeric state at higher temperatures with a T_m of 30°C. Such behavior is in line with what has been observed in gel filtration experiments (Åkerström *et al.*, 1992), where the dimeric state is observed at low temperature (4°C and 10°C), whereas monomers occur at high temperature (37°C). As a comparison, the average T_m of unfolding, summarizing thermodynamic data of a large number of proteins (Pheil, 1986), is ~63°C. The M1 protein was also found to be thermally unstable, but had a more complex unfolding behavior. The unfolding started just above 30°C but did not confirm to a simple two-state process (Figure 4.8B). This behavior of the M1 protein could be due to different unfolding temperatures in different regions of the coiled-coil structure. The regions of protein H and the M1 protein with the lowest degree of similarity are found in the N-terminal regions, and M1 protein has an additional block of 44 amino acids in the D domain as compared to protein H. It is possible that the degree of stability in different fragments and the intact M1 protein is: A-S < S-C3 < S-D < intact M1 protein. These differences could explain the different unfolding behavior of the two proteins. Furthermore, the melting temperatures were modulated by the binding of ligands to the coiled-coil structure of protein H and the M1 protein. Thus, free protein H was less stable than protein H in complex with either IgG or albumin and further stabilization was observed in the presence of both ligands, which result in an increase from <10% to ~50% (with both ligands) of folded protein H at 37°C (Figure 4.7C).

As judged by CD, the N-terminal portions of protein H and the M1 protein were not able to adopt stable structures as isolated peptides. This observation supports gel chromatography experiments (Cedervall *et al.*, 1995), which suggest that these peptides exist as monomers in solution. Interestingly, the peptide fragment containing the S domain followed by the three C repeats (S-C3) from the M1 protein was shown to be in a folded state at 4°C (Figure 4.8C). The unfolding behavior of this peptide suggest a two-state transition going from a folded helical state at low temperature to an unfolded state at higher temperatures. This fragment was also markedly stabilized by the addition of HSA, supporting previous mapping of the albumin-binding to the C repeats (Frick *et al.*, 1994; Åkesson *et al.*, 1994). A majority of the sequenced genes of M proteins contain C repeats. This region is in fact, together with the signal sequence and the membrane associated part, the most well-conserved region within the M protein family. The N-terminal region, on the other hand, shows a much lower degree of sequence similarity

among various M proteins. The C repeats may represent a framework for the formation of the α-helical coiled-coil structure, and this could be one of the explanations why the C repeats are highly conserved among M proteins of class C.

Electron microscopy has indicated that M proteins of adjacent bacteria may form specific end-to-end interactions involving the distal N-terminal regions (Swanson *et al.*, 1969; Phillips *et al.*, 1981). The present work show that protein H and the M1 protein both have the ability to fold and unfold their coiled-coil structure within a small temperature range close to 37°C. It therefore seems possible that these surface molecules could unfold and refold in an antiparallel fashion together with M proteins on adjacent bacteria. To test this hypothesis, sequences of both protein H and the M1 protein were positioned in all possible antiparallel heptad alignments. It was found that the antiparallel arrangement can indeed be more favorable than the parallel structure. Interestingly, these favorable antiparallel alignments exclusively involves residues in the more flexible N-terminal portion of the two molecules. This kind of interaction could therefore provide a mechanism for cell-cell interactions among *S. pyogenes* and may also allow the bacteria to bind similar coiled-coil structures on human cells, e.g. receptors of the C-type lectin superfamily (Beavil *et al.*, 1992).

The M protein family of *S. pyogenes* provides these bacteria with the ability to bind host proteins like IgG, albumin, and fibrinogen. The present data suggest that members of this family can switch between two temperature-dependent states - an active folded state and an inactive unfolded state, and that the switch *in vivo* is triggered by temperature changes close to 37°C. Thus, temperature fluctuations in the bacterial environment could change the physicochemical surface properties of the bacterium, and thereby influence the host-parasite relationship during *S. pyogenes* infections.

4.5 References

Åkesson, P., Conney, J., Kishimoyo, F. & Björck, L. (1990) *Mol. Immunol.* **27**, 523-531.

Åkesson, P., Cooney, J., Kishimoto, F. & Björck, L. (1994). *Biochem. J.* **300**, 877-886.

Åkerström, B. & Björck, L. (1989) *J. Biol. Chem.* **264**, 19740-19746.

Åkerström, B., Lindahl, G., Björck, L. & Lindqvist, A. (1992) *J. Immunol.* **148**, 3238-3242.

Beavil, A.J., Edmeades, R.L., Gould, H.J. & Sutton, B.J. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 753-757.

Ben Nasr A., Wistedt A., Ringdahl U. & Sjöbring U. (1994) Eur. J. Biochem. 222:267-276.

Berge, A. & Sjöbring, U. (1993) *J. Biol. Chem.* **268**, 25417-25424.

Cantor & Schimmel (1980) *Biophysical chemistry, Part II: Techniques for the study of biological structure and function*, W. H. Freeman and Company, New York.

Cedervall, T., Åkesson, P., Björck, L., Lindahl, G. & Åkerström, B. (1995) submitted.

Clarke, J. & Fersht, A. R. (1993) *Biochemistry* **32**, 4322-4329.

Cohen, C. & Parry, D.A.D. (1990) Proteins: Struct. Funct. Genet. 7, 1-15.

Conway, J.F. & Parry, D.A.D. (1990) *Int. J. Biol. Macromol.* **12**, 328-334.

Cooper, T. & Woody, R. W. (1990) *Biopolymers* **30**, 657-676.

Crick, F.H.C. (1953). *Acta Crystallogr.* **6**, 689-697.

de Château, M. & Björck, L. (1994) *J. Biol. Chem.* **269**, 12147-12151.

Deisenhofer, J. (1981) *Biochemistry* **20**, 2361-2370.

Dell, A., Antone, S.M., Gauntt, C.J., Crossley, C.A., Clark, W.A. & Cunningham, M.W. (1991) *Eur. Heart J.* **12**, 158-162.

Fietz, M. J., McLauglan, C. J., Campbell, M. T. & Rogers, G. E. (1993) *J. Cell. Biol.* **121**, 855-865.

Fischetti, V. A., Parry, D. A. D., Trus, B. L., Hollingshead, S. K., Scott, J. R. & Manjula, B. N. (1988) *Proteins Struct. Funct. Genet.* **3**, 60-69.

Fischetti. V.A. (1989). *Clin. Microbiol. Rev.* **2**, 285-314.

Frick, I-M., Åkesson, P., Cooney, J., Sjöbring, U., Schmidt, K., Gomi, H., Hattori, S., Tagawa, C., Kishimoto, F. & Björck, L. (1994) *Mol. Microbiol.* **12**, 143-151.

Frick, I-M., Wikström, M., Forsén, M., Forsén, S., Drakenberg, T., Gomi, H., Sjöbring, U. & Björck, L. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 8532-8536.

Frick, I.-M., Crossin, K.L., Edelman, G.L. & Björck, L. (1995). *EMBO J.* in press

Frithz, E. Hedén, L.-O. & Lindahl, G. (1989). *Mol. Microbiol.* **3**, 1111-1119.

Gomi, H., Hozumi, T., Hattori, S., Tagawa, C., Kishimoto, F. & Björck, L. (1994) *J. Immunol.* **144**, 4046-4052.

Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y. & Shimada, I. (1992) *Biochemistry* **31**, 9665-9672.

Gronenborn, A. M., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T. & Clore, G. M. (1991) *Science* **253**, 657-661.

Guss, B., Eliasson, M., Olsson, A., Uhlén, M., Frej, A.-K., Jörnvall, H., Flock, J.-I., & Lindberg, M. (1986) *EMBO J.* **5**, 1567-1575.

Heath D.G. & Cleary, P.P. (1987) *Infect. Immun.* **55**, 1233-1238.

Heath D.G. & Cleary, P.P. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4741-4745.

Hodges, R.S., Sodek, J., Smillie, L.B. & Jurasek, J. (1972) Cold Spring Harbor Symp. Quant. Biol. 37, 299-310.

Hodges, R. S, Semchuk, P.D., Taneja, A.K., Kay, C.M., Parker, J.M.R & Mant, C.T. (1988) *Peptide Res.* **1**, 19-30.

Hodges, R. S., Zhou, N. E., Kay, C. M. & Semchuk, P. D. (1990) *Peptide res.* **3**, 123-137.

Hollingshead, S.K., Fischetti, V.A. & Scott, J.R. (1987) *Infect. Immun.* **55**, 3237-3239.

Hosein, B., McCarte, M. & Fischetti, V.A. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3765-3768.

Janson, H., Hedén, L.-O., Grubb, A., Ruan, M. & Rorsgren, A. (1991) *Infect. Immun.* **59**, 119-125.

Kastern, W., Sjöbring, U., and Björck, L. (1992) *J. Biol. Chem.* **267**, 12820-12825.

Kehoe, M.A. (1994) *New Comp. Biochem.* **27**, 217-261.

Khandke, K. M., Fairwell, T. & Manjula, B. N. (1987) *J. Exp. Med.* **166**, 151-162.

Khandke, K. M., Fairwell, T., Acharya, A. S., Trus, B. L. & Manjula, B. N. (1988) *J. Biol. Chem.* **263**, 5075-5082.

Kraus, W., Ohyama, K., Snyder, D.S. & Beachey, E.H. (1989). *J. Exp. Med.* **169**, 481-492.

Kraus, W., Dale, J.B. & Beachey, E.H. (1990). *J. Immunol.* **145**, 4089-4093.

Lau, S.Y.M., Taneja, A.K. & Hodges, R.S. (1984) *J. Biol. Chem.* **259**, 13253-13261.

Laurent, T. C., & Killander, J. (1964) *J. Cromatogr.* **14**, 317-330.

Lee, S-C., Kim, I-G., Marekov, L. N., O'Keefe, E. J., Parry, D. A. D. & Steinert, P. M. (1993) *J. Biol. Chem.* **268**, 12164-12176.

Lupas, A., Van Dyke, M., and Stock , J. (1991). *Science* **252**, 1162-1164.

Lian, L.-Y., Derrick, J. P., Sutcliffe, M. J., Yang, J. C ., & Roberts, G. C. K. (1992) *J. Mol. Biol.* **228**, 1219-1234.

Lindahl, G. & Åkerström, B. (1989) *Mol. Microbiol.* **3**, 239-247.

Lindahl, G. & Stenberg, L. (1990). *Epidemiol. Infect.* **105**, 87-93.

McLachlan, A.D. & Stewart, M. (1975) *J. Mol. Biol.* **98**, 293-304.

Manjula, B. N., Khandke, K. M., Fairwell, T., Relf, W. A. & Sriprakash, K. S. (1991) *J. Protein Chem.* **10**, 369-384.

McLachlan, A.D., and Stewart, M. (1976). *J. Mol. Biol.* **103**, 271-298.

Miller L., Gray L., Beachey E. & Kehoe M. (1988) *J. Biol. Chem.* **263**, 5668-5673.

Monera, O., Kay, C.M. & Hodges, R. (1994). *Biochemistry* **33**, 3862-3871.

Mouw, A.R., Beachey, E.H. & Burdett, V. (1988) *J. Bacteriol.* **170**, 676-684.

O'Toole, P., Stenberg, L., Rissler, M. & Lindahl, G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 8661-8665.

Parry, D.A.D. (1982) *Biosci. Rep.* **2**, 1017-1024.

Pheil, W. (Hinz, H.-J. Ed.) (1986) *Thermodynamic data for Biochemistry and Biotechnology*, Springer., 349-376.

Phillips, G.N., Flicker, P.F., Cohen, C., Manjula, B.N. & Fischetti, V.A. (1981). *Proc. Natl. Acad. Sci. USA* **78**, 4689-4693.

Podbielski A. (1993) *Mol. Gen. Genet.* **237**, 287-300.

Retnoningrum, D.S. & Cleary, P.P. (1994). *Infect. Immun.* **62**, 2387-2394.

Retnoningrum, D.S. Podbielski, A. & Cleary, P.P. (1993). *J. Immmunol.* **150**, 2332-2340.

Robbins, J.C., Spanier, J.G., Jones, S.J., Simpson, W.J. & Cleary, P.P. (1987) *J. Bacteriol.* **169**, 5633-5640.

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

Rost, B. & Sander, C. (1994) *Proteins* **19**, 55-72.

Rost, B., Casadio, R., Fariselli, P. & Sander, C. (1995) submitted.

Sauer-Eriksson, A, E., Kleywegt, G. J., Uhlén, M. & Jones, T. A. (1995) *Structure* **3**, 265-278.

Sreerama, N. & Woody, R.W. (1993) *Anal. Biochem.* **209**, 32-44.

Schmidt, K.-H. & Wadström, T. (1990) *Zbl. Bakt.* **273**, 216-228.

Stenberg, L., O'Toole, P. W., Mestecky, J. & Lindahl, G. (1994) *J. Biol. Chem.* **269**, 13458-13464.

Stone, D., Sodek, J., Johnson, P. & Smillie, L.B. (1975) in Proteins of Contractile Systems, Preceedings of the IX Federation of European Biochemical Societies Meeting (Biro, E.N.A., Ed) Vol. 31, pp 125-136, North Holland Publishing, Amserdam.

Stone, G., Sjöbring, U., Björck, L., Sjöquist, J., Barber, C. & Nardella, F.(1989) *J. Immunol.* **143**, 565-570.

Swanson, J., Hsu, K. C. & Gotschlich, E. C. (1969) *J. Exp. Med.* **130**, 1063-1091.

Uhlén, M., Guss, B., Nilsson, B., Gatenbeck, S., Philipson, L. & Lindberg, M. (1984) *J. Biol. Chem.* **259**, 1695-1702.

Vashishtha, A. & Fischetti, V.A. (1993) *J. Immunol.* **150**, 4693-4701.

Wikström, M., Sjöbring, U., Kastern, W., Björck, L., Drakenberg, T. & Forsén, S. (1993). *Biochemistry* **32**, 3381-3386.

Wikström, M., Drakenberg, T., Forsén, S., Sjöbring, U. & Björck, L. (1994)*Biochemistry* **33**, 14011-14017.

Zhou, N. E., Kay, C. M. & Hodges, R. S. (1992a) *Biochemistry* **31**, 5739-5746.

Zhou, N. E., Kay, C. M. & Hodges, R. S. (1992b) *J. Biol. Chem.* **267**, 2664-2670.

Zhou, N. E., Kay, C. M., Sykes, B. D. & Hodges, R. S. (1993) *Biochemistry* **32**, 6190-6197.

Zhou, N. E., Kay, C. M. & Hodges, R. S. (1994) *J. Mol. Biol.* **237**, 500-512.

5. Summary

The computational and biophysical analysis described in this Case Study have given us a picture of what the two bacterial surface molecules, protein H and the M1 protein, look like (Figure 5.1): dimeric parallel coiled-coil proteins.

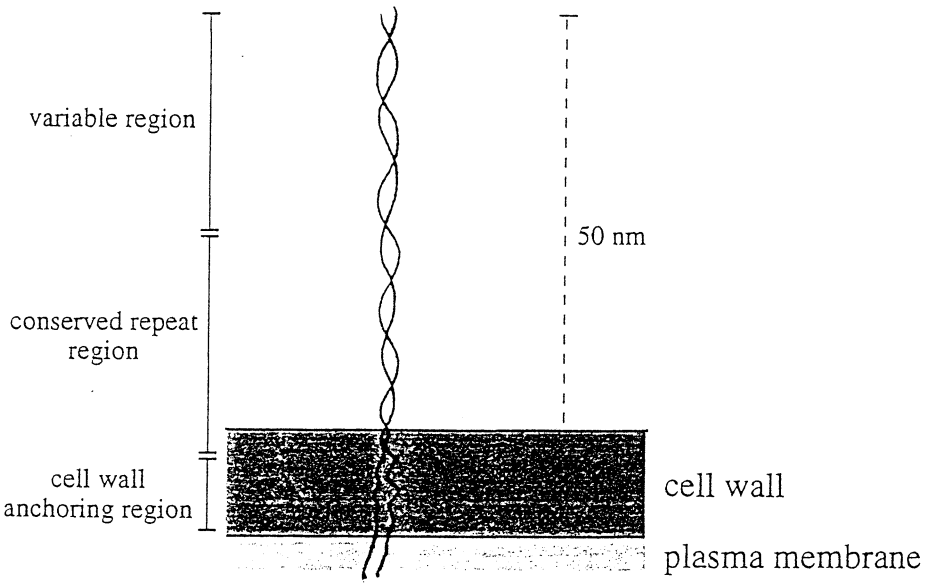


Fig 5.1 Schematic representation of a dimeric coiled-coil bacterial surface molecule.

How did we come to this conclusion? Let us summarise the data that have helped us to determine the structure and the stability of one of these molecules, protein H. The molecular weight of a single chain of protein H was calculated from the DNA sequence of the protein H gene and the size of the protein was confirmed by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) analysis (Table 4.1). Gel chromatography experiments determined that the expressed protein exists either as a dimer or as a monomer depending on the surrounding temperature. Furthermore,the computational secondary structure analysis of the proteins amino acid sequence indicated a high content of α-helical structure (Figure 4.2), and this was confirmed by circular dichroism (CD) spectroscopy showing that approximately 70% of the protein had α-helical structure (Figure 4.7A). The computational secondary structure prediction analysis also clearly pointed out that a region in the very C-terminal region of protein H is anchored in the bacterial cell membrane (Figure 4.2). Finally, a heptad pattern of hydrophobic amino acids typical for coiled-coil structures was identify by manually analysing the protein sequence in detailed (Figure 4.5A), this heptad periodicity was also confirmed by Fourier analysis (Figure 4.4A), and furthermore by the coiled-coil

prediction program (Figure 4.2). This program predicted a coiled-coil structure for a major part of the extracellular region of protein H with very high probability . The coiled-coil structure model fit very well with the fibrous hair-like structure that one can see by electron microscopy (Figure 5.2). In the literature coiled-coil proteins are described to be built up by two, three or four helixes coiled together and they can either be in a parallel or antiparallel orientation to each other. In our case we could conclude that the structure of protein H must be a coiled-coil built up by two parallel chains (Figure 5.1). This because the protein has the size of a dimer and it is anchored to the bacterial cell membrane in one orientation with the C-terminal end located in the bacterium.

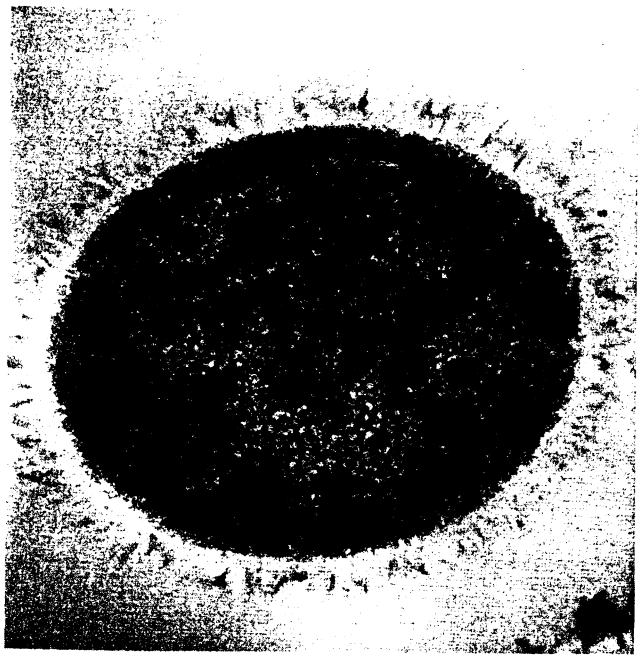


Fig 5.2 Electon microscope image of a group A streptococcus.

The next obvious goal would be to determine the three dimensional structure of these proteins to high resolution. Unfortunately, no one has been able to crystallise any of these proteins, because they precipitate (form aggregates) at high concentrations instead of forming crystals suitable for X-ray diffraction studies. For the same reason, these proteins are not suitable for NMR studies, since NMR requires highly concentrated protein solutions. Secondly, the size of these proteins is far too large for NMR studies. Furthermore, it would not be possible to study smaller fragments (< 20 kDa) of these proteins with NMR because it is likely that they are unable to fold into stable dimeric complexes, as the full length protein does. In the future when methods for computer

modelling of proteins have been improved it might be possible to model the three dimensional structure of protein H and the M1 protein. The X-ray crystal structure of tropomyosin, another related dimeric parallel coiled-coil protein isolated from muscle has been described (Figure 5.3). This structure could probably be used as an initial three-dimensional template when modelling the coiled-coil region of protein H and the M1 protein.

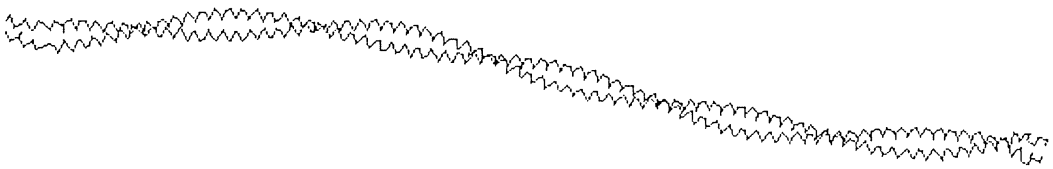


Fig 5.3 The parallel coiled-coil crystal structure of rabbit tropomyosin.

The results described in this Case Study have not only given us a picture of a parallel dimeric coiled-coil protein, but also shown how sensitive the structure of protein H is to changes in the temperature. To our surprise the circular dichroism analysis showed that the structure of protein H was dramatically affected by a small increase in the surrounding temperature and resulted in a random conformation of the protein at physiological temperature (Figure 4.7). These results explained why we saw only monomeric forms of protein H at physiological temperature, this is because when the dimeric coiled-coil unfolds the chains separate and become monomeric. The loss of structure also explained why protein H does not bind its ligands (immunoglobulin and albumin) at this increased temperature (Figure 4.7D). In comparison, most other proteins are known to have a stable structure in this temperature interval. Biologically these results were puzzling. These bacteria live in physiological environments (at 37°C) and it is under these conditions it encounter its ligands. Why has protein H evolved to be so unstable? Does this characteristic have any physiological importance? The answer to this question is still not proven, but it might be linked the fact that these bacteria can aggregate.

When these bacteria are studied under electron microscopy one can see that one fibrous surface molecule (e.g. protein H) on one bacteria interacts with another one on an other bacteria (Figure 5.4). It is this binding that allows the bacteria to aggregate. It is tempting

to speculate that the temperature sensitive unfolding of protein H can regulate the degree of bacterial aggregation. This is because a temperature increase does results in the unfolding of the helical structure of protein H which would disrupt the interaction between the two protein H molecules. In Chapter 4.3.10 it was suggested that this temperature sensitive binding occured through an anti-parallel coiled-coil pairing (folding/unfolding) of the N-terminal region of protein H. Such an increase in temperature occurs naturally when streptococci enter the human host. For instance, it might be more favourable for these bacteria to live in colonies when being located on the skin or outside the body, but not when they are living in it. Thus, one could speculate on the following senario: when a colony of streptococci enter the interior environment of its host, their surrounding temperature increases to physiological temperature; the protein H molecules unfolds and thus disrupt the interaction between adjacent bacteria; this will allow single bacteria to spread and infect the host. One might further speculate that when these bacteria reach the circulation, the presence of the ligands immunoglobulin and albumin might stabilise the protein H structure (as shown in Figure 4.7C) and allow this stabilised structure to bind its other two ligands, fibronectin and the N-CAM receptors located on surface of human cell. Of course, this is still just a theory and many more experiments are needed before we know if this will explain why protein H molecule is so unstable.

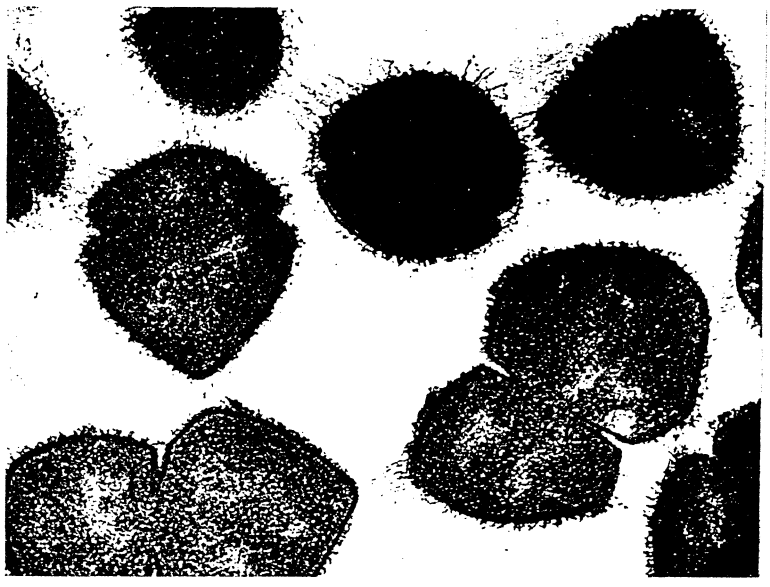


Fig 5.4 Electon microscope image of group A streptococci.

In summary, the biophysical method circular dichroism spectroscopy has been an excellent tool for analysing the secondary structure content and the stability of the two bacterial surface molecules, protein H and the M1 protein. By combining this information with computer structure analysis and biochemical analysis a model for the proteins H and the M1 protein structure could be proposed. The stability measurements of these proteins with circular dichroism spectroscopy have suggested that temperature fluctuations in the bacterial environment could change the properties of bacterial surface proteins, thereby influence the molecular interactions between the bacterium and its host. Thus, temperature might regulate molecular processes during bacterial infection. Finally, this Case Study illustrates how one can gain a stronger position by combining several different techniques, biophysical, computational and biochemical, when analysing interesting biological macromolecules.

6. Acknowledgements

Sune Svanberg for support and giving me the opportunity to plan and execute this Case Study for my Masters Thesis. I like to thank Mats Wikström for experimental help with this project. Mats Wikström, Bo Åkerström, Lars Björck, Sture Forsén, and Sune Svanberg are gratefully acknowledged for critically reading this manuscript. Inga-Maria Frick and Per Åkesson for preparing protein H and the M1 protein and fragments thereof.

7. Abbreviations

CD	Circular dichroism
EPR	Electron paramagnetic resonance
Fc	Constant region of the heavy Ig chain
HSA	Human serum albumin
Ig	Immunoglobulin
nm	Nanometer
NMR	Nuclear magnetic resonace
T_m	Temperature of thermal melting
u.v.	Ultraviolet