Calmodulin: Protein - Target Interactions

Twelve Month Report

1.0 Introduction

Calmodulin is a ubiquitous intracellular protein in higher eukaryotes. It is a member of a large family of structurally homologous proteins known as the EF-Hand family of proteins. Each member of the family has multiple EF-hand motifs; a helix - loop - helix conformation effectively pre-organised for Calcium ion co-ordination (Ca^{2+}).

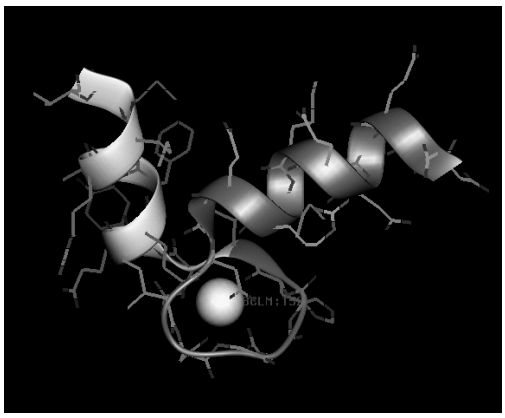

Figure 1.1.
Ribbon representation of the EF-Hand with a bound Calcium ion

The EF-hand nomenclature originates from the Parvalbumin structure, the first of the EF-hand calcium binding protein structure to be determined[1] in which the helices adjacent to the Calcium binding loop were labelled E and F.

Calmodulin is one of the more interesting members of this family due to its ability to bind a variety of different target sequences with nanomolar affinity. Nature can normally only attain this type of affinity through a highly specific interaction. It is this high-affinity, low-specificity phenomenon which is the subject of this project. How is it possible for a protein to form a binding surface appropriate for high affinity interaction with such a large variety of different targets?

[1] Kretsinger et al.

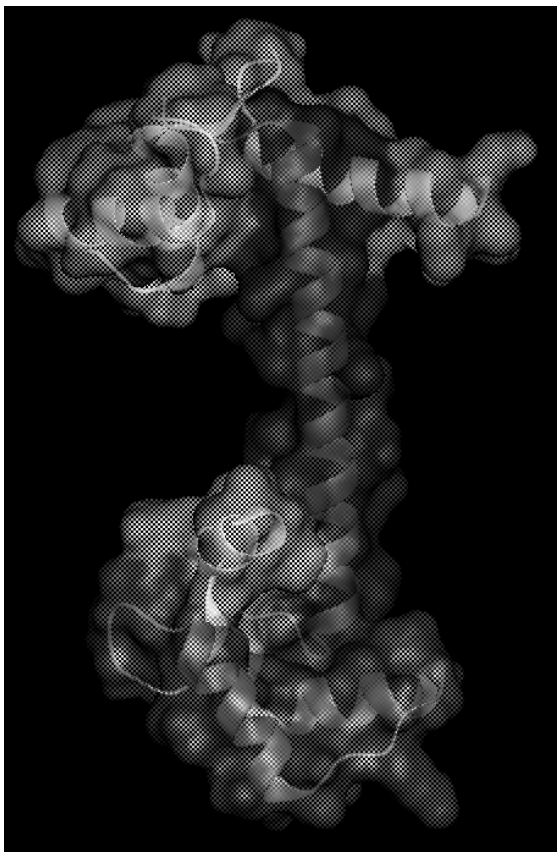

Figure 1.2.
"Surfaced" ribbon model of Calmodulin, coloured to highlight the individual domains.

The above figure shows a model of Calmodulin displaying a solvent accessible surface coloured by domain; the four EF-Hands are coloured in blue, white, green and yellow and the central linker helix (often referred to as the *tether*) is coloured in red. This central tether is another key to the unique action of Calmodulin as it enables the two lobes to move relatively independently of one another and interact with the target in a variety of ways.

Calmodulin generally has a much greater affinity for its targets in the presence of Calcium and interestingly it has a much higher affinity for calcium in the presence of target peptides. The associated process of conformational change in Calmodulin upon Calcium binding has proved to be another area of considerable interest. The results of NMR studies of the apo (Ca^{2+}-free) protein clearly illustrate a tightening of the whole structure (figure 1.3.) as a result of calcium binding in the EF-hands and a study by Finn et al.[2] (figure 1.4) was able to show in some detail some of the changes that take place in the C-terminal domain.

[2] Finn et al. *Na. Str. Biol.* **2 No. 9** 777

Figure 1.3.
A comparison of the apo (blue) and Ca^{2+}(green) loaded states of Calmodulin

Figure 1.4a.
C-Terminal lobe of Ca Free Calmodulin with individual amino acids coloured by polarity, red (negative), blue (positive), yellow (polar) and white (hydrophobic).

Figure 1.4b.
C-Terminal lobe of Ca^{2+} Calmodulin in the same orientation as figure1.4a. Note the extended area of hydrophobic residues and large hydrophobic pocket.

It is this binding surface which is of particular interest and how it can be adapted to bind a variety of targets with such high affinity.

This project falls into two sections;

I. A largely experimentally based project centring around the characterisation of a novel Calmodulin-target system thought to be of a lower affinity:
II. A more theoretically based project to look in detail at the variability of the Calmodulin target binding surfaces.

- Experimental Project:

  By performing a detailed analysis of the interaction between Calmodulin and a novel target peptide, it is hoped that a characterisation of the mode of binding of the protein can developed for this particular system. By using a wide variety of techniques to study the interaction, different features can be explored and it is hoped that crystallisation trials will be successful and a structure for the complex may be solved.

- Computational Project:

  Whilst the experimental side of the project will explore the details of a single Calmodulin target system, the computational part will deal initially with a statistical analysis of all the information available for the different Calmodulin studies for which structures have been solved. Once this has been carried out it is hoped that patterns will become apparent which will help to explain its functional characteristics.

2. Experimental Work

2.1 Background

A novel target has been identified for Calmodulin in the non-muscle myosin II family. Myosins are generally involved in motor action and have long helical neck regions which are normally bound by light chains which have a very high degree of structural homology with Calmodulin. A notable difference however between Calmodulin and the myosin light chains is the global conformation in which they bind their respective targets. Calmodulin is known to generally exhibit the locally opposite mode of interaction whereas the light chains tend to adopt a slightly off-set conformation (see figure 2.1).

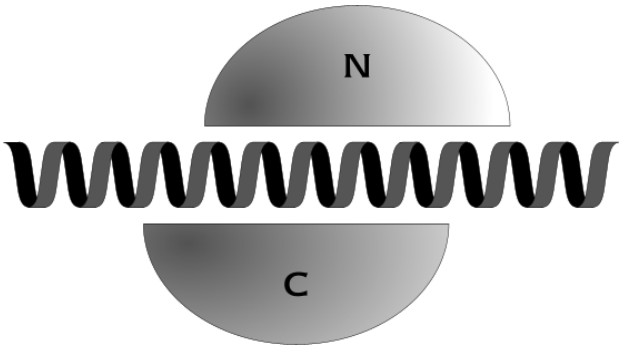

Figure 2.1a
Schematic representation of the two lobes of Calmodulin binding a helical target in a globally opposite conformation.

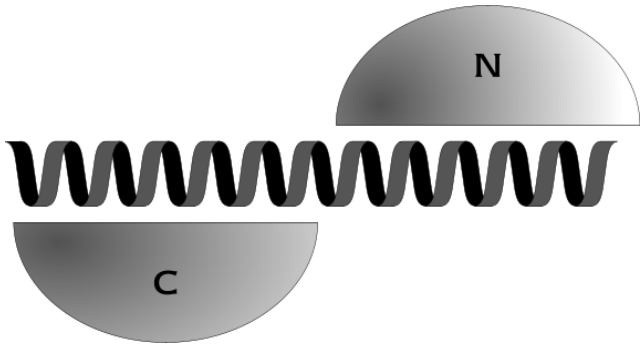

Figure 2.1b
Schematic representation of a globally off-set binding conformation as found in Myosin light chain – heavy chain interactions where the two lobes bind different parts of the target motif.

The interaction between Calmodulin and the appropriate domain of this non-muscle myosin was identified in this laboratory as part of work on another of the EF-hand calcium binding proteins, S100 A4 which has a single globular domain containing two calcium binding units but is found in nature as a dimer and involved in cell differentiation and motility.

It is well established that Calmodulin has an affinity for amphiphilic helices (i.e. with a hydrophobic and a hydrophilic side) essentially as a complement to the distinct characteristics of the two lobes of Calmodulin. Another common motif found in Calmodulin binding domains is the B-12-B; two bulky hydrophobic residues separated by 12 other residues. This myosin peptide can be seen to exhibit a B-10-B motif though either of these identifed hydrophobic residues could prove to be the key locator pin and a shorter hydrophobic residue act at the other locator pin. The sequence identified from Myosin has the potential to fit this description in various

places (see figure below) and in different orientations which gives the potential for interesting results when it is finally established exactly where the Calmodulin binds. It is this "multi-motif" nature and the unusual binding activities that have caused the interest to date and require the most in-depth investigation. In addition to these features it was also noted above that the target sequence was found where the myosin tail splits from a coiled coil to two individual helices of two separate myosin head units.

Figure 2.2
A surfaced representation of a model of the MyosinII peptide. Note the sequence was modelled to an α-helical conformation and this is not an observed structure. Both sequence and surface are coloured according to amino acid polarity; yellow (polar) red (acidic) blue (basic) white (hydrophobic).

Another peptide of interest comes from the cytoplasmic domain of L-Selectin. This interaction has been reported elsewhere[3] but not studied in detail. L-Selectin is a trans-membrane protein which is susceptible to attack from various proteases in the cytoplasm. Calmodulin binding to the cytoplasmic domain however appears to cause some sort of conformational change throughout the protein so as to protect it from attack in the cytoplasm. As this is a trans-membrane protein there is little chance of a crystallographic study of the interaction with the entire protein. The potential still exists however to work with peptides from the protein as analogues of the sequence of interest.

Figure 2.3 shows an alignment of these two sequences along with the sequence from the calmodulin binding domain of smooth muscle myosin light chain kinase (MLCK) which is often

[3] Kahn, J. *Cell* **92** 809 - 918

used as an example of a classic calmodulin target. Although the sequences are somewhat dissimilar, it is clear that they have some common features in for example the period distribution of basic residues and matched hydrophobic residues at positions 11 and 25.

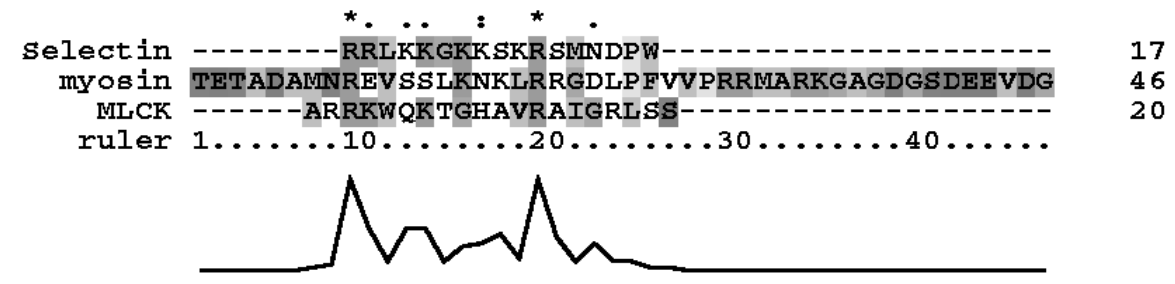

Figure 2.3
Output from the sequence alignment program ClustalX for the two target sequences of interest and the calmodulin binding domain from MLCK.

2.2 Areas of Interest

Our prime concern here is to establish that there is indeed an interaction taking place between Calmodulin and these proposed novel targets. Once that has been established it will then be of interest to characterise the nature of this interaction in order to answer questions such as how strong is the interaction and how calcium dependent is it. Ideally, structures of the complexes would be obtained which would give us highly detailed information about the interaction and the topography of the binding surface. This information could then supplement the database of structural information used in the survey of CaM-Target interaction or test the validity of any rules proposed.

In addition to manipulating the actual binding surfaces of each lobe, Calmodulin had great potential for global rearrangement of its two lobes through use of its central tether. It is our belief that if the interaction between Calmodulin and this new target from myosin does indeed prove to be a novel low affinity interaction, that this may well be something to do with a different global arrangement of the Calmodulin lobes. This is implicit from the location of the target sequence in the Myosin protein and the type of interactions usually observed between the myosin neck and the two light chains shown previously (see figure 2.1).

2.3 Production of material

A tried and tested system[4] for the expression of Calmodulin in E.Coli already exists and works very well. However for most of the proposed work to be carried out, large amounts of the desired peptides will be required at a high level of purity. Companies exist which synthesise peptides to order at various levels of purity and whilst this has proved successful in the past it can be expensive. Selectin and Myosin peptides have been synthesised with Tryptophan markers for fluorescence experiments in other projects[5] and have proved to be a useful starting point. For detailed analysis of the interaction with Calmodulin however, the effect of the mutations is unknown and unpredictable and it would be preferable to work with wild type systems. This being the case, wild type synthetic peptide was ordered but the company later informed us that the synthesis had proved inexplicably unsuccessful and the attempt was abandoned.

Once the failure of synthetic sources of peptide had been acknowledged, a route for expression was explored. Current investigations centre on a bacterial expression system of the myosin peptide using the pQE60 expression vector in E.Coli. Primers have been designed for a PCR reaction to make sufficient DNA to clone into the vector using appropriate restriction enzymes. It has been decided that the peptide should be expressed as a construct with an S-Tag to enable a technique known as *Affinity Purification* to be used as it was feared that the peptide alone may prove too small to purify. The S-peptide tag, which is separated from the myosin peptide by a cleavage site, adheres to S-Protein on a column support and depending on the strength of interaction between Calmodulin and the target peptide, it should be possible to co-purify Calmodulin along with the S-tagged target peptide.

2.4 Binding Assays

Of the many various techniques that are to be explored, the only data available at the moment is the Fluorimetric data.

Fluorimetric analysis has shown a change in the environment around the tryptophan markers of both the myosin and selectin peptides upon addition of Calmodulin to the sample. As the Calmodulin solution was added, there was a shift in the emission wavelength and a gradual increase in its intensity. Saturation appeared to have been reached at a ratio of 0.5 : 1 Peptide:CaM which would imply that each molecule of Calmodulin is binding two peptide

[4] Work of Ken Johnson at YSBL

molecules. Whilst such a stoichiometry has been observed before[6] it remains a somewhat unexpected result.

A control experiment was performed using free tryptophan ester which confirmed that the observations were not due to any anomolous effects of buffer or solvent and were indeed due to the Calmodulin binding the peptide. Similar effects have also been observed in a variety of different buffers. Although these experiments were performed with calcium added to the sample, control experiments performed with no added calcium did not seem to inhibit the binding action but this will need to be repeated in the presence of EDTA to ensure that there is no residual Calcium in the sample before it can be confirmed whether this interaction is Calcium dependant or independant. This technique has been shown to work for similar systems in the past[7] and there is no apparent source of obvious error although we will continue to establish a reason for these results.

In addition to the fluorimetry work, once both protein and target peptide have been expressed at suitable levels a number of other types of binding assays will be attempted in order to confirm the data obtained from fluorimetry. These techniques should include:

- Micro-Calorimetry
- Analytical Ultra-centrifugation
- Surface Plasmon Resonance
- Protein affinity Chromatography
- In vivo FRET analysis

2.5 Crystallisation Studies

Various complexes of Calmodulin have already been studied by both X-Ray crystallography and NMR but one of the basic aims of this study will be to obtain a structures of these new complexes about which little so is known. Calmodulin complexes are notoriously difficult to crystallise although precedents exist in the lab[8].

---

[5] Afshar, M., Bronstein, I.

[6]

[7] Chapman et al. *J.Biol.Chem.* **266** (1) 207 - 213

[8] Johnson K. unpublished work

Initial attempts following the protocols of Ken Johnson developed in this laboratory have as of yet proved unsuccessful. However, crystals were obtained from the Hampton Screen[9]. These crystals were very small and there was no observable diffraction (see below), which is at least indicative that they were not samples of salt crystals. This result provides a useful starting point for further trials once sufficient amounts of material have been produced.

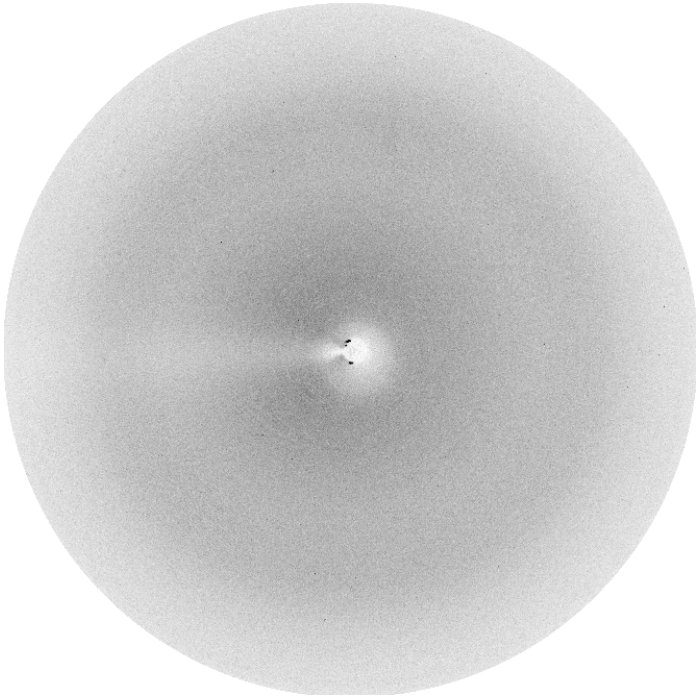

Figure 2.2
Sample image from marplate of diffraction pattern from CaM-Myosin crystal

2.6 Future Work

As outlined in the sections above, there remains a considerable amount of work still to be carried out on this project. However various processes have been set in motion and some of the preliminary results are quite promising. There is evidence that Calmodulin is interacting with both of these peptides further analysis will enable a more detailed characterisation of the binding taking place. This characterisation, in particular crystal structures, will reveal in great detail what interactions are taking place in these two new systems.

[9] Commercially available set of solutions for setting up a screen of general crystallisation conditions.

3.0 Computational Work

3.1 Background

One of the main keys to understanding Calmodulin-target interactions is to perform a detailed analysis of the structural information available as published in the Protein Data Bank (PDB) currently hosted by the Research Collaboratory for Structural Bioinformatics (RCSB). This is the central database for all protein structures as solved by X-ray crystallography, NMR and molecular modelling. The PDB currently contains some 8551 protein structures along with nucleic acid and carbohydrate structures. This represents the protein modellers basic library and the origin of nearly all the structural information available to people working in this large and expanding field.

3.2 Structural Database Analysis

Before any analysis on a family of structures can be performed it is first necessary to sort through the database in an attempt to gain as much information as possible about the structures available. In addition to all the available structural information on Calmodulin itself it was thought appropriate to extract co-ordinate files for those related proteins with highly homologous structures such as Troponin C, Parvalbumin and other EF-hand proteins.

In order to do this a search was performed on the PDB for calcium binding proteins. The hit list generated was then searched by eye to extract those structures which appeared to feature the helix - turn - helix conformation of the EF-Hand calcium binding site. The quickest way to look at these structures proved to be by using MDL Chime on a PC which, once downloaded as a plug-in, can be activated by clicking on a link in the PDB browser.

It should be noted at this point that this is a laborious and somewhat subjective method of searching the structural database. Methods are currently being explored to make a better assessment of the informational available. This would involve a search tool that looks for structures with homolgy to a given template and would provide a far more rigorous search criterion for selecting structures (see further work).

This procedure facilitated a detailed examination of the features of the 3DB browser of the Brookhaven structural database[10] and useful exposure to the huge range of calcium binding proteins contained within it. The file names of these structures were then entered into an ASCII file with the name `#ef_like.lst` with each filename on a new line. The complete list is included as an Appendix.

A script was used to copy the selected files from the Brookhaven database mirror to the working directory and to extract parts of the PDB header and write them into a log file called `details`. It was hoped that this log file would contain enough information to be able to appropriately categorise each structure e.g. X-ray structures, NMR structures, structures with bound ligands, structures with Calcium etc. Unfortunately, although all this information is contained with each structural data file, there is no consistency to the format which can be utilised in this way. Whilst the RCSB (and previously Brookhaven) database has most of this information in a readily accessible and searchable format, it would appear that this information has been formatted manually and there is no access for the public user. This means that in order to perform any large scale searches on the basis of such categories, it will be necessary to set up a new database with all the information in a suitable format. This process is also underway but in the early stages, the exact details required will become more apparent as the structural information becomes available.

For the present however, from this complete set of EF-Hand containing proteins, a sub-category of Calmodulin structures was assigned, which was then split into two further sub-groups: structures with a ligand or inhibitor bound to Calmodulin and those without. The members of these groups are also included in the Appendix.

3.3 Data Preparation

In order to carry out any analysis on the structural data contained within these PDB files it is first necessary to format the raw data into comparable formats. At this stage in the process it is necessary to clean up the data in a manner of different ways prior to any statistical analysis. At this point in time two distinct aims have been identified but more are certain to become apparent as the process develops.

[10] NB this database is now hosted by the RCSB as mentioned above and has a slightly different interface. However the same system can still be used as described above.

1) Selecting Representative Structures from Multi-model files

The nature of NMR protein structure determination experiments is such that rather than a single refined set of co-ordinates being published, it is far more appropriate to publish several sets of co-ordinates or *Models*. This is primarily due to the fact that the NMR experiment observes proteins in a highly dynamic manner as it moves in the solvent. As it is impossible to define a correct structure from the ensemble of models produced, it is normal for some twenty or so models to be published from which the user can select or determine a defining structure as required. Sometimes the NMR PDB files are published as a pair where the first file will contain the ensemble of structures and a second file will contain an averaged structure. However this is not always the case and even when it is, an averaged structure may not be the best representation of the results of the experiment.

It has been mentioned above that another member of the EF-Hand family of proteins is, or are, the Myosin Light Chains. This presents another such problem as here we have various files containing co-ordinates for Myosin structures, several of which include the atomic positions of both light chains which need to be treated independently as they are different molecules but are contained within the same file.

In other cases, the PDB files may contain multiple chains. Often different chain ID's are used to define different molecules, for example a structure of Calmodulin with a target peptide would define the Calmodulin molecule as Chain A and the target peptide as Chain B. In some cases however two chains are given for the same molecule with no apparent rationale as is the case for 1trc.pdb. In such cases it appears to be common[11] to use only the A chain and ignore all subsequent structural information.

This problem of multi-model files from the NMR experiment has been made the subject of at least one Software development project[12] which devises a system for selecting a representative structure from an ensemble. Development of such a technique may be an appropriate way of dealing with this problem but there are other potential solutions. An alternative solution might be to consider **all** the structural information contained within these files and assign a weighting: If for example an ensemble contained 20 models, include all the structures in a

[11] e.g. *Avgcore* software by Altman et al.
[12] See section on *NMRClust*

statistical analysis but assign each structure only a 20th of unit significance. Similarly for multi-chain files.

2) Structural Super-position

Before any statistical analysis can be performed on such a large set of structures it is also necessary to have all the structures in a consistent spatial reference frame. A development of this principle would be that some sort of core structure would be identified as a defining model to which all other structures could be fitted. This problem presents similar aims to those of selecting representative models from an ensemble of structures as outlined above and for that reason, there is considerable potential for similar tools to be used for carrying out both of these aims.

This Structural template could then be used for all manner of comparisons for different structures but a system would probably have to be developed for *improvement* of the model as more structural information was included in the dataset.

Various tools have been identified which offer solutions to either - or potentially both of these problems and they are looked at below. It will become apparent as they are discussed which tools are favoured for which task but a rational decision has yet to be made as to which tool(s) should be applied to which situation(s).

NMRCore/NMRClust

This suite of programs was developed by Michael Sutcliffe and co-workers[13] at Leicester to provide a process for the selection of a representative structure from an ensemble of structures. According to the associated literature the software was developed in two stages;

1. NMRClust to perform a cluster analysis of a set of structures
2. NMRCore to define structurally conserved regions from an ensemble of structures upon which the cluster analysis should be performed

The version discussed below was that found on the internet[14] although it is not clear whether this version uses the entire structures to perform the cluster analysis or just the regions of conserved structure as is outlined in the literature.

[13] Kelly et al. *Prot. Eng.* **10** 737-731
[14] http://neon.chem.le.ac.uk/nmrcore

- **NMR Core**

Once connected to the site the user is prompted to enter the PDB ID of the structure of interest and the program then automatically performs the required calculations. The output generated can be seen below.

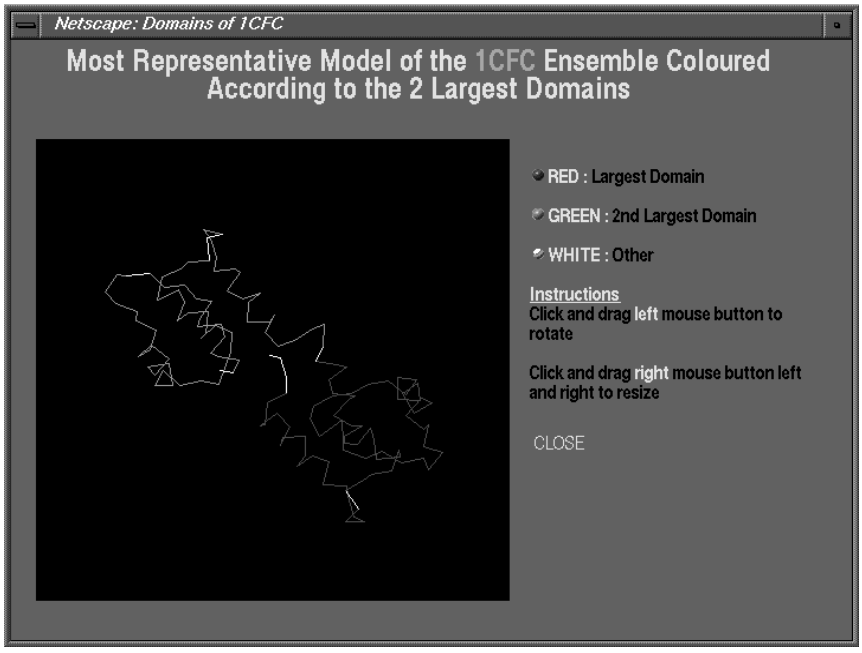

Figure 3.1
A picture of the output window generated from the *NMRCore* analysis of the file 1cfc.pdb. LSD's are defined and illustrated in different colours.

This program defines "Local Structural Domains" (LSD's) for the structure of interest and on the basis of only Cα atoms it defines the parts of the structure where there is least variability across the ensemble of structures and then uses the program *NMRClust* to perform a cluster analysis of these LSD's across the ensemble and choose the most representative model for the structure.

- **NMR Clust**

Whilst it is clear that the program *NMRCore* uses the clustering method of *NMRClust*, it is unclear whether the clustering is performed only on these LSD's or on the entire structures. However the output format of *NMRClust* for the same structure is shown below.

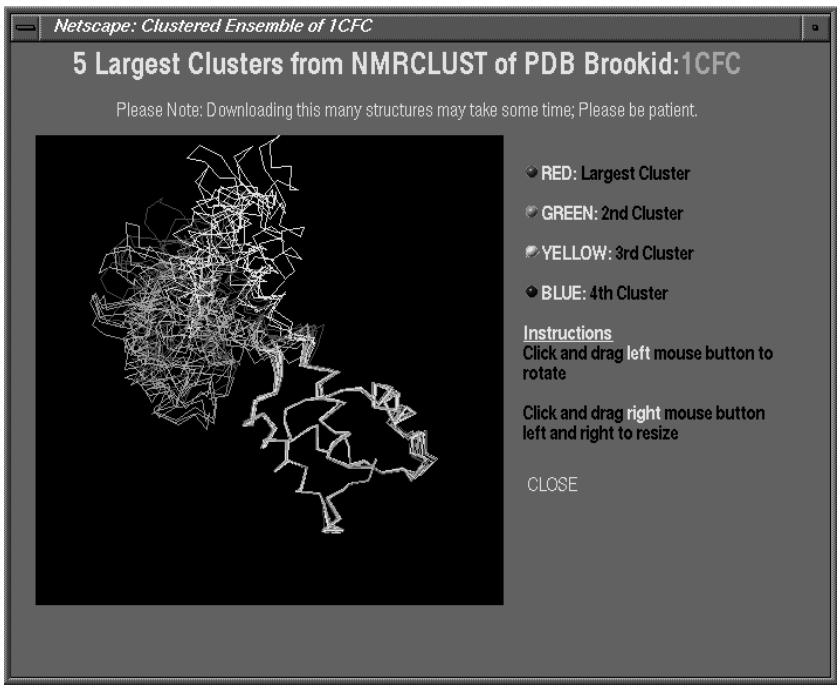

Figure 3.2

This is the output window generated by *NMRClust*. All the different model structures are displayed and superposed. The structures are coloured according to the clusters assignment of each model, in this case four different clusters with the cluster in red the largest. Representative models are chosen for each cluster and the most representative comes from the largest cluster.

All these multi - model PDB files were then read into *insightII* and the representative model as chosen by the Sutcliffe software was pulled out and saved separately for use in the further analysis.

In addition to the multi - model NMR structures from the database there are also examples of multi - chain structures which also need to be dealt with. In most cases this multi - chain nature is due simply to a ligand or inhibitor which is solved as part of the structure. For the moment these files have been edited with a script which takes only the A chain from each file and uses that for the further analysis.

As the project develops, it seems likely that this suite of programs has the potential to be adapted to take the structural information from all the files of interest and perform a similar analysis on a much larger data-set to generate these LSD's and a then perform a cluster analysis. It would then be of interest to look at the members of the individual clusters and examine their membership with respect to the different classes of structure; apo, target-bound, Ca^{2+} loaded etc.

Avgcore

This is a program suite developed by Russ Altman and Mark Gerstein[15] from the Scripps Institute to calculate a core of homologous atoms from a family of structures. In principle the aim is much the same as *NMRCore* although the analysis is carried out in a rather different way and the software can deal with a set of homologous proteins whereas *NMRCore* is set up to deal with structures of identical sequence.

In addition to the files containing the structural information, an alignment file is required (the format of which is the as that used in *Modeller*, the homology modelling program) and only uses the Cα atoms for residues aligned across all the structures. This alignment is then adapted and cross-checked against the files containing the structural information to check for missing atoms. A sample alignment is shown below where all the matched atoms are indicated by a #. Since this dataset contains only information from Calmodulin structures, the match here is very good apart from at the termini where atomic co-ordinates are often difficult to resolve. Note that it is only once this sequence alignment has been established that the software is able to perform any structural comparisons.

[15] Gerstein et al. *JMB* **251** 161 - 175

```
012345678----#################################################################*
1a29   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMT*
1ahr   . -SEEEIREAFRVFDKDGNGFISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVTMMTSK*
1bbn   . DSEEEIREAFRVFDKDGNGFISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVTMMTSK*
1cdl   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMT*
1cdm   . ----EIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMT*
1cfc   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK*
1cfd   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK*
1cll   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTA*
1clm   . DSEEELIEAFKVFDRDGNGLISAAELRHVMTNLGEKLTDDEVDEMIREADIDGDGHINYEEFVRMMVS*
1cm1   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMT*
1cm4   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMT*
1cmf   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK*
1cmg   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK*
1ctr   . -SEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTA*
1deg   . -DSEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREANIDGDGQVNYEEFVQMMTA*
1dmo   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREANIDGDGQVNYEEFVQMMTAK*
1lin   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK*
1mux   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK*
1osa   . DSEEELIEAFKVFDRDGNGLISAAELRHVMTNLGEKLTDDEVDEMIREADIDGDGHINYEEFVRMMVSK*
1trc   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTA*
3cln   . DSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREANIDGDGQVNYEEFVQMMTA*
4cln   . DSEEEIREAFRVFDKDGNGFISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVTMMTSK*
```

Various formatting problems were encountered at this stage due to the usage of altered files as the program prefers to collect a fresh copy of the PDB file from the Brookhaven database and work with the native file. This causes all sorts of errors when the sequence alignment does not match up with the atom list in the PDB files because the user has decided to work, for example, with only one lobe of Calmodulin. These problems can be overcome by manually editing various files but this can become very tedious and tricky in order to get everything correct.

The Cα atoms for each matched residue are then considered as the *putative core* from which an average structure is calculated. All the other structures are then fitted to this averaged structure. The variability of the atoms about the average is then calculated for each position (i.e. each residue) and error ellipsoids generated. The position with the largest error ellipsoid is then thrown out of the core, a new average structure is calculated and the new error ellipsoids generated. This process is then repeated until all the atoms have been thrown out of the core and the final average structure is calculated about a single atom across the ensemble.

A cut-off can then be defined by a variety of different methods. This will be a structure which has been averaged over the *core* at this point, a *core* which contains a percentage of less variable

atoms. A special viewer is used to look at the averaged structures with the atoms represented by their error ellipsoids and coloured according to the membership of the core. A script has also been written[16] to convert the output to a PDB format so that the averaged structure can be used in molecular visualisation packages.

Other Structural Superposition Tools

In addition to the structural alignment program outlined above there are various other tools which may hold some potential for the future. Tom Oldfield has spent some time developing various pieces of data mining software to work quickly through the large amounts of structural information in the PDB and give an RMSD fit to the compound of interest. A structural template can be defined and the structural database can then be searched for all similar fragments. It was noted that this may have potential applications for the future but further development is still necessary.

Once a sequence alignment has been obtained within QUANTA it is then possible to do a structural superposition using a least squares fitting algorithm developed by Sutcliffe[17] which provides a useful starting point. The new molecular positions can then be exported in a PDB file format for structural comparison purposes. Structural superpositioning is also a facility of *InsightII* which uses another relatively standard algorithm[18] but applies it only to residues that have been matched in a sequence alignment. The Levenburg-Marquardt algorithm helps to minimise the RMS deviation over all the structures concerned simultaneously for the residues that have been matched to give a globally optimal alignment - but only with respect to *matched* residues.

3.4 Further Work

It is very clear now that there exists a multitude of different methods for performing a structural superposition of a set of structures from which some kind of average can be derived. Any one of these methods *could* be used to generate some sort of average structure for use in any further analysis. Obviously the programs *NMRCore* and *Avgcore* both have the added subtlety of developing some sort of sophisticated weighting system whereby a core can be set aside as the defining part of the structure and be more important in the fitting process. Notably of course, in

[16] Acknowledged to Ryan Smith, YSBL
[17] Sutcliffe et al. *Prot. Eng.* **1** 377 - 384
[18] Press et al. *Recipes in C* (C.U.P. 1988) p. 683

contrast to *NMRCore* which uses a representative *observed* model, *Avgcore* generates a new averaged model as the representative structure.

The next step of this procedure is to assess the methods of structural superposition. Once a procedure has been selected, it will then be possible to superpose structures, or parts of structures, as a whole or in families or groups which can then be compared statistically for specific features or characteristics. For Calmodulin, this focus will be a detailed comparison of the target binding surfaces for each of the two lobes.

4.0 Summary and Conclusions

The details and progress of both major arms of this project have been discussed in some detail and their progress evaluated. It is clear that there are areas of interest in both fields. Both projects could easily be developed as the sole area of study but it is the combination of the two that makes the project so interesting.

There is a great deal of work to be done here and so priorities have to be set and adhered to in order to maintain a balance between the two disciplines. This will require frequent evaluation of progress throughout the project and re-prioritisation at regular intervals in order to keep on top of all the developments.

The unique feature of Calmodulin that enables it to bind highly diverse target sequences with high affinity presents an interesting paradox to the scientist. One which can be better understood through the analysis of available information on observed interaction and investigative characterisation of a new interaction. Whilst the whole phenomenon cannot be completely explained through such an analysis, interesting new results will be observed that will significantly advance our understanding of this system.