Prediction of calcium-binding DxDxDG motifs in protein sequences.

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by *Duncan David Woodhead*

January 2013

Contents

D.D. Woodhead - Prediction of calcium-binding DxDxDG motifs in protein sequences.

The binding of calcium is important in many biological processes, from signal transduction in prokaryotes to muscle movement in higher eukaryotes. As such, studies of protein function based on sequence would be greatly enhanced by a simple method of identification of calcium binding sites.

Many calcium binding proteins contain the DxDxDG motif. This motif is most often associated with EF-hands, however it is also found in a number of other, seemingly independently evolved, structural environments. Although the DxDxDG motif alone is not enough to confer the ability to bind calcium, it has been suggested that it may be possible to predict this ability using the sequence surrounding the motif.

This project is roughly split into four main parts. In the first, a search for examples of the DxDxDG motif in all of its varying structural contexts was performed. This was done using software that is able to identify motifs that match a given three-dimensional arrangement. These examples were further filtered to ensure a metal-binding function, and classified into structural SCOP super-families.

Once as many examples of the DxDxDG motif in its various structural contexts were identified and confirmed, the properties of these proteins were analysed. In particular, the characteristics of the amino acids surrounding the DxDxDG motif from these proteins were looked at. An alignment was produced using a representative example from each super-family. The predicted secondary structure, peptide chain solvent accessibility, and the physiochemical characteristics of the surrounding amino acids were obtained, then contrasted with a set of proteins that contained the DxDxDG motif but were known not to bind any metals.

Artificial intelligence methods, such as decision trees and support vector machines, were then used with the characteristics identified, to discriminate between binding and non-binding examples. This allowed rules to be established so other potential calcium-binding proteins could be classified, and a tool developed that is able to determine if it is likely that a protein will bind calcium, based on protein sequence data alone.

Finally, the resulting rule-sets were used on two full genome sequences. *Escherichia coli* is well known and has been extensively studied. This proved useful as a method of verification that the tool was effective at identification of proteins that bind calcium through the DxDxDG motif. *Bacillus.coahuilensis* is of interest due to its native environment in the desiccation lagoons of Cuatro Cie´negas Valley, Mexico. These lagoons show high calcium ion concentrations and *B. coahuilensis* may display interesting calcium proteins as a result.

The tools developed were successful in the classification of binding and non-binding DxDxDG-containing proteins. Additionally, a number of examples of proteins from *E. coli* and *B. coahuilensis* that are likely to bind calcium were identified.

Section 1:- Background

1.1 Section Introduction

1.1.1 Section Overview

In this initial section, the context of this project will be laid out. Why is calcium and its binding important, biologically? What does a typical calcium-binding protein look like at the structural level? What, specifically, is interesting about the DxDxDG motif? Here, these questions will be addressed, as well as summarise the research that has been carried out previously in relation to the DxDxDG motif and the work that led to this project.

This will help to set the scene for the project and show how its aims developed. It will also help lead into the next section which concerns the initial search for the DxDxDG motif and why this was more complicated than it initially might appear.

1.1.2 Glossary and abbreviations

Adenosine-diphosphate (ADP) – the product of the reversible dephosphorylation of ATP, an important molecule in energy transfer in cells.

Angstrom (Å) – unit of length equal to one ten billionth of a metre or 100 picometers.

Asparagine (ASN, N) – hydrophilic amino acid, (H_3NHCOO)- CH_2CONH_2

Aspartic Acid (ASP, D) – hydrophilic amino acid, (H_3NHCOO)-CH_2COOH

Calmodulin (CaM) – common eukaryotic calcium-binding protein that regulates many cellular processes.

Convergent evolution - describes the acquisition of the same biological trait in unrelated lineages.

Gap phase (G_1, G_2, G-phase) – stage of the cell cycle where majority of growth takes place.

Glutamic Acid (GLU, E) – hydrophilic amino acid, (H_3NHCOO)-CH_2CH_2COOH

Inositol 1, 4, 5-trisphosphate receptor (InsP3R) – receptor on the endoplasmic reticulum and sarcoplasmic reticulum; its activation triggers Ca^{2+} release.

Meiosis phase (M, M-phase) – stage of the cell cycle where actual cellular division takes place.

Migration inhibitory factor-related protein (MRP8/MRP14) – acts on macrophages to inhibit migration and regulate differentiation.

Protein Data Bank (PDB) –repository of macromolecular structures.

Quiescence (G_0) – a resting stage; the cell cycle has been exited.

Root mean squared deviation (RMSD) - measure of how different the arrangement of two sets of atoms is.

Synthesis phase (S, S-phase) – stage of the cell cycle where DNA is replicated.

Serine (SER, S) – hydrophilic amino acid, R-CHOH

Three dimensional (3D)

1.1.3 List of Figures

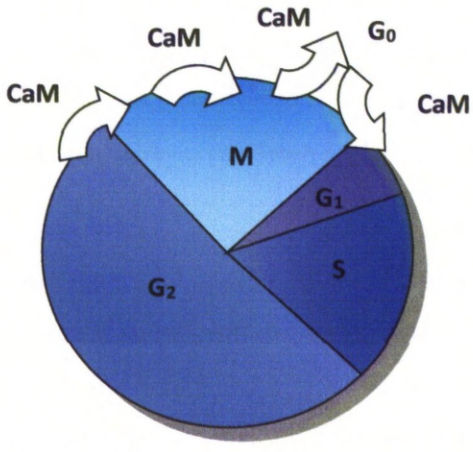
1.2 Calcium-binding

It is widely known that calcium is important biologically, and it is often associated with the development of bones and teeth; this, however, is only one of many biological processes that calcium is involved in. The majority of these actions are mediated by binding with proteins. The proteins that bind calcium are as varied as its functions; these include everything from basic cell chemistry to signal transduction to structural roles within proteins.

Its messenger function, in particular, is important in multi-cellular organisms, where it can be seen taking a role in processes as diverse as the cell cycle (Kahl and Means, 2003), differentiation (Lagasse and Clerc, 1998) and motor functions (Matthews, 2001; Pitta et al., 1997; Tassinary and Cacioppo, 2000). It is becoming increasingly apparent that calcium-binding proteins are also vital to the biology of prokaryotic organisms, where they have been associated with chemotaxis (Tisa and Adler, 1992) and structural roles (Torrance et al., 2007).

Calcium is involved in every part of a cell's life, from its creation to its differentiation and its eventual demise. A new cell is created by division during the normal cell cycle. This can be divided into four distinct phases (see figure 1.2-1): synthesis phase (S) and mitosis phase (M), separated by two gap phases (G1 and G2). Synthesis phase is where the cell's DNA is replicated; mitosis is where the cell actually divides; the

Figure 1.2-1, The Cell Cycle and the role of Calmodulin



The cell cycle consists of 4 phases G1, M, S and G2. G_0 is senescence the resting phase of the cell cycle. Calmodulin (CaM) has been shown to be important in progression from G_0 to G_1, and the initiation (G_2 to M), progression and exit from mitosis phase (M to G_0).

gap phases are periods of growth and preparation for the other two. There is also a non-dividing state called quiescence (G0) (Forsburg and Nurse, 1991). Calcium is important in the regulation of the cell cycle: several checkpoints are regulated by increases in the intracellular concentration of Ca^{2+} ions (Kahl and Means, 2003). Calmodulin (CaM) is the calcium-binding protein that detects these fluctuations in the cytoplasm; its name is, in fact, an abbreviation of Calcium Modulated Protein (Stevens, 1983). Ca2+/CaM is required in two steps in the initiation of the cell cycle (transition from G0 to G1). The Ca2+/CaM is also associated with the initiation, progression and exit from mitosis (Kahl and Means, 2003).

Differentiation is the process that transforms a cell from a stem cell to a specialised cell. Calcium-binding proteins have been implicated in the differentiation of a number of different cell lines (Schafer and Heizmann, 1996). An example of this is the differentiation of myeloid cell lineages, where macrophage migration inhibitory factor is thought to play an important role (Lagasse and Clerc, 1998). Specifically, MRP8 and MRP14, part of the macrophage migration inhibitory factor complex, have been shown to be activated during the mononuclear phagocyte differentiation pathway (Lagasse and Clerc, 1998). These proteins show a high degree of similarity to calcium-binding protein sequences and show the characteristics of a subfamily of calcium-binding proteins (Lagasse and Clerc, 1998).

Apoptosis is the process of programmed cell death, essential in normal development and maintenance. Calcium signals are involved in several of the pathways involved in this process (Mattson and Chan, 2003). The programmed death of a cell is a complex process. First, a cell death stimulus is received, leading to mitochondrial calcium overload, which activates the permeability transition pore (PTP); this is thought to cause the release of mitochondrial proteins such as cytochrome c (Boehning et al., 2003). Cytochrome c then diffuses into the endoplasmic reticulum (ER) and binds the inositol 1, 4, 5-trisphosphate receptor (InsP3R), resulting in sustained release of calcium into

the cytoplasm (Boehning et al., 2003). This increase in calcium in the cytoplasm can either lead to further release of cytochrome c from all the mitochondria, or activation of cell survival pathways. An efflux of cytochrome c from the mitochondria leads to the formation of the apoptosome, cleavage of the DNA and various proteins and eventual cell death (Mattson and Chan, 2003). Calmodulin, in addition to its role in the cell cycle, is an important mediator of the apoptotic response. Calmodulin's action is late on in the apoptotic response, several hours after initial stimuli. Calmodulin is itself regulated by calmodulin kinase II (CaMKII), inhibition of which results in inhibition of apoptosis (Olofsson et al., 2008).

Another important function of calcium ions, in animals, is as a messenger for muscle contraction. When movement is initiated by the brain, an action potential travels through the motor neuron network to the muscle tissue. When this action potential reaches the end of the motor neuron near the muscle, vesicles containing the neurotransmitter acetylcholine fuse with the plasma membrane. The acetylcholine diffuses across the synapse and activates nicotinic acetylcholine receptors on the end plate of the muscle cell. This results in an action potential that spreads through the muscle cell (Tassinary and Cacioppo, 2000).

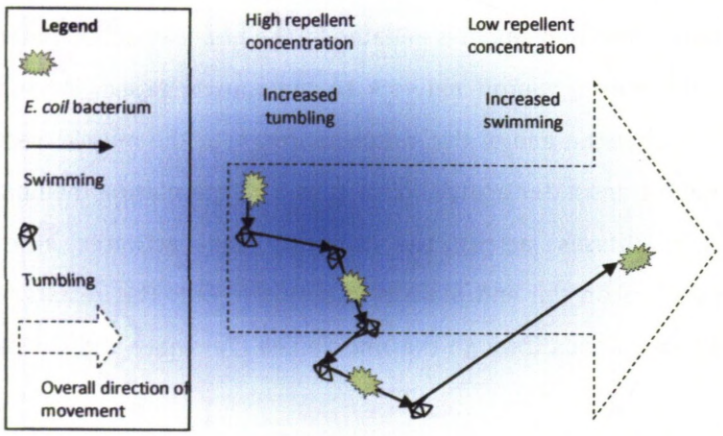
Calcium is released from the sarcoplasmic reticulum and interacts with troponin via calcium-binding motifs known as EF-hands. The bound calcium causes an allosteric change, which reveals the binding sites for myosin normally obscured by tropomyosin. This allows the thick myosin filament to anchor to the thin filament, allowing the filaments to pull past each other as the myosin releases ADP (Matthews, 2001).

Calcium-binding proteins are not only important in complex multicellular eukaryotes. It is also becoming increasingly evident that they are important to the biology of prokaryotes and single celled eukaryotes.

Calcium has been shown to be a messenger in chemotaxis in *E. coli* (Tisa and Adler, 1992). Chemotaxis is the response to a chemical gradient, either towards a higher concentration of an attractant or away from a repellent.

In *Escherichia coli*, flagella drive movement of the bacteria. These flagella can act in two ways: swimming and tumbling (see Figure 1.2-2). Swimming propels the bacterium forward, whereas tumbling reorients the bacteria in a random direction. The bacteria tend to tumble more frequently and swim less as the concentration of a repellent chemical increases, causing a general movement away from a repellent (Ordal, 1977). (Kahl and Means, 2003).

Figure 1.2-2, Calcium signalling in the Chemotaxis of E.Coli



An *E. coli* bacterium has two modes of movement; swimming in a straight line and tumbling. Tumbling results in a random change of direction. Detection of a repellent by receptors on its surface triggers the release of calcium into the cytoplasm. The increase in calcium concentration causes an increase in tumbling. Therefore, as an *E. coli* bacterium moves from a high concentration to a low concentration of repellent, its tendency to tumble decreases. This results in an average movement away from the repellent.

This behaviour is modulated by calcium; detection of a repellent causes the intracellular concentration of calcium to increase, which, in turn, increases the likelihood of the flagella to initiate tumbling (Tisa and Adler, 1992).

The single-celled eukaryotes, *Trypanosoma cruzi* and *Trypanosoma brucei,* have also been shown to contain calcium-binding proteins, such as calmodulins and calflagins, at high concentrations in their flagella. It has been suggested, owing to their localisation, that these proteins may be involved in the molecular processes that confer a high level of motility to these parasites (Engman et al., 1989; Wu et al., 1994). Calcium has also been shown to be required for swimming in non-flagellated bacteria, such as the *Cyanobacterium synechococcus* (Pitta et al., 1997).

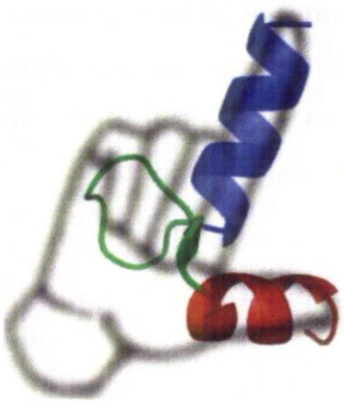
Calcium and the proteins that bind it have roles in a variety of biological processes in many diverse organisms, so it follows that there are a number of groups of calcium-binding proteins that mediate the biological functions calcium performs. Given the variety of calcium-binding proteins, it is interesting to find that many employ the same helix-loop-helix structure in their interaction with the Ca^{2+} ions, often referred to as the EF hand. Therefore, the understanding of calcium and how it interacts with EF hands has implications for our understanding of many important biological functions.

1.3 EF hands

The EF hand structure is common to many Ca^{2+}-binding proteins. Its helix-loop-helix structure utilises three aspartic acid residues to bind calcium ions, while the surrounding helices help stabilise their conformation.

EF hands were originally observed in carp muscle tissue protein parvalbumin B (Kretsinger and Nockolds, 1973), and have been identified in a variety of calcium-sensitive proteins since, appearing in, for example, calmodulins, calbindin and troponin. These proteins are involved in processes such as the cell cycle (Kahl and Means, 2003), apoptosis (Olofsson et al., 2008) and muscle contraction (Matthews, 2001).

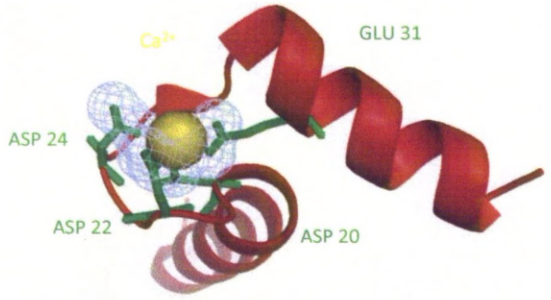
Figure 1.3 -1, Diagram showing a typical EF hand



The typical EF hand consists of a helix-loop-helix configuration (calmodulin PDB code: 1exr, residues 4-40 shown here (Wilson and Brunger, 2000). The first alpha helix is a little longer than the other, and the 2 helices are approximately perpendicular, resembling an extended finger and thumb, with the binding loop forming the 3 clenched fingers. The residue-binding loop is typically around 12 residues long and contains a DxDxDG motif that confers calcium-binding properties.

The typical EF hand is made up of an 'E' alpha helix (the fifth of six labelled A to F), of approximately ten residues, followed by a binding loop of around twelfth residues, followed by a shorter 'F' alpha helix of around eight residues (Kretsinger and Nockolds, 1973). However, the length of both helices and the binding loop may vary (Gifford et al., 2007). As can be seen in Figure 1.3-1, the helices are arranged in a perpendicular manner, with the loop connecting them forming a circular curve, giving the appearance of a thumb, 3 clenched fingers and pointing forefinger, which inspired its name (Kretsinger and Nockolds, 1973).

Figure 1.3-2, Diagram showing Ca^{2+} binding in an EF hand from calmodulin (PDB code: 1exr, *Wilson and Brunger, 2000)*



TEEQIAEFKEAFALFDKDGDGTITTKELGTVMRSL

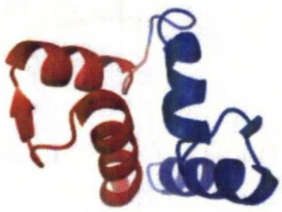
The Ca^{2+} ion (yellow) is bound by four residues (green); three aspartic acids from the loop, and one glutamic acid from the exiting α-helix. The oxygen atoms (light blue) of these residues form a negatively charged sphere surrounding the Ca^{2+} ion. When calcium is not bound, this electrostatically unfavourable configuration is held together by a network of hydrogen bonds.

Three aspartic acids in the loop, together with a fourth, located around three residues from the start of the exiting helix, mediate binding. The oxygen atoms provide the seven coordinating groups required for Ca^{2+} binding (Gifford et al., 2007). The unfavourable proximity of these numerous negatively charged oxygen atoms is stabilised by a network of hydrogen bonds within the structure, allowing a negatively charged sphere to be formed around the positive Ca^{2+} ion (Gifford et al., 2007) (see Figure 1.3-2) .

Figure 1.3-3, Common groupings of EF hands

i) EF hands may occur as single functional units in a motif, such as seen here in the myosin light chain, where the EF hand is highlighted in red

ii) EF hands can also be seen as closely associated pairs, often displaying co-operative binding, as seen here in recoverin, where a closed oval structure is formed between the two EF hands. Here, the first EF hand is red and the second is blue.



iii) Pairs of EF hands are also seen as 'mismatched' pairs, showing differing helical lengths and less close association. Calmodulin is shown here, again, the first EF hand is red the second is blue.

A single EF hand may be seen in a protein domain, for example in the light myosin chain (Herzberg and James, 1988); however, more commonly, multiple EF hands occur in a single domain, typically in pairs (Ikura, 1996). The members of these pairs may be closely associated, forming an oval structure, as seen in recoverin (Flaherty et al., 1993), or further separated, such as in calmodulin (Babu et al., 1985) (see Figure 1.3-3). Often, where a close association between two EF hands is seen, they may bind cooperatively, increasing binding affinity at both sites (Ikura, 1996).

The binding of calcium to the EF hand can result in dramatic conformational changes within the protein, often leading to its activation. For example, in calmodulin, the helices of the EF hand go from an almost anti-parallel, to a much more perpendicular, arrangement upon binding with calcium (Finn et al., 1995).

Other proteins, such as calbindin, however, show very little conformational changes on calcium-binding (Skelton et al., 1994).

The EF hand structures have a number of man-made applications, such as the basis of metal-binding motifs in rational protein design (Lim and Franklin, 2006), and in clinical diagnostics and therapeutic strategies (Schaub and Heizmann, 2008).

EF hands have previously been used in the design of proteins to bind lanthanides. However, as we have seen, the binding loop is stabilised by the interactions of helices that flank it. This structure may not always be suitable when designing a protein; additionally, the affinity of the Ca^{2+}-binding loop to lanthanides showed room for improvement. The EF hand-binding motif has been used as the basis for the design of a 17-amino-acid lanthanide binding tag (Ca/Ln binding site). This synthetic peptide achieves nM-pM affinity with lanthanide ions, but does not rely on the stabilising forces of the helices (Nitz et al., 2004).

Derivatives of EF hand structures have also been used in the design of metallohomeodomains. The Ca/Ln binding site of an EF hand was combined with DNA-binding scaffold, creating a protein that could bind both DNA and lanthanide ions without affecting the overall structure of the protein (Lim and Franklin, 2006).

These developments will be an important tool in the investigation of metalloenzyme function and provide an important building block for rational protein design (Lim and Franklin, 2006).

The EF hand protein calmodulin, in particular, is thought to be involved in many disease processes, including Parkinson's, Alzheimer's, rheumatoid arthritis and certain heart defects. However, no mutations seen in calmodulin have been associated with these diseases (Schaub and Heizmann, 2008). It is thought that mutations of the targets of calmodulin may cause disease symptoms. For

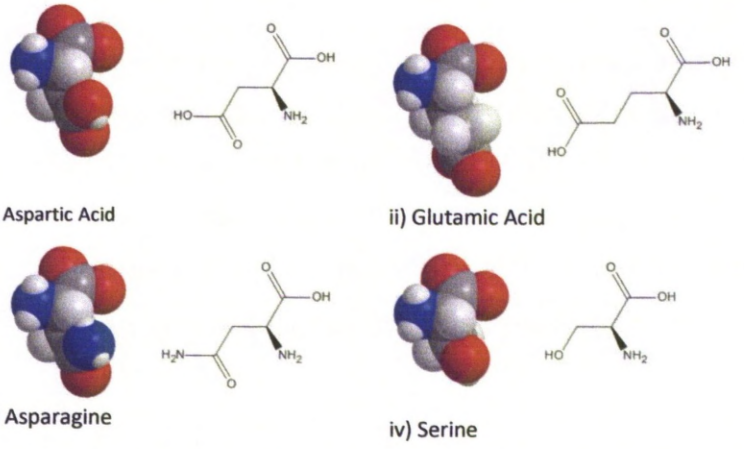
example, calmodulin dependent protein kinase-II may prove a useful target for myocardial dysfunction (Schaub and Heizmann, 2008).

The metal-binding properties so important to all of these applications are a result of three conserved polar residues. These three residues, seen in the binding loop, form a motif often generalised as the DxDxDG motif. This motif, however, has also been shown to be responsible for binding in many other calcium-binding proteins that do not conform to the canonical EF hand blueprint (Rigden and Galperin, 2004).

1.4 The DxDxDG motif and its many guises.

The DxDxDG motif is not solely present in a functional form as part of an EF hand. It seems to have evolved independently a number of times; however, a similar spatial orientation is always seen (Rigden and Galperin, 2004)

Figure 1.4–1, Residues commonly involved in Ca^{2+} binding in the DxDxDG motif



Aspartic acid i) and glutamic acid ii) are very similar in structure; each has a carboxylate group at the end of its side chain. The only difference is that glutamic acid is one carbon longer. Asparagine is also similar to aspartic acid, except it has an amide group in place of the carboxylate. Serine iv) is much smaller, but still has a polarised oxygen at the end of its side chain.

The DxDxDG motif actually shows some variation in its makeup: although the first aspartic acid is always conserved, the second and third may be replaced by the amide derivative asparagine, or the alcohol serine. It is easy to understand why asparagine and serine are suitable replacements for aspartic acid in the second and third positions: they are of similar size, hydrophilic and have regions of high electron density in their side chains (see Figure 1.4-1). In some cases, the glycine is also replaced; Figure 1.4-2 shows the full range of variation as a PROSITE pattern. In addition to the motif, an extra residue, downstream of the motif, completes the negative sphere formed around the Ca^{2+} ion; this is usually either aspartic acid or its amide derivative, asparagine.

Figure 1.4-2, PROSITE pattern for the EF hand binding motif

D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x2-[DE]-[LIVMFYW]

"[ ]" means OR, so "[DS]" could be a single aspartic acid or a single serine.	"{ }" means NOT, so "{IL}" could be any single amino acid except leucine and isoleucine.	"x" means ANY, so "x2" is two amino acids of any type.

PROSITE is a database where protein families and domains are represented using amino acid patterns and profiles; for more details, see section 2.2.2. The "D"s of the motif are highlighted in red, the extra downstream D or E is blue. The above sequence shows pro the PROSITE pattern for the EF hand (PROSITE id: PDOC00018)

The sequence logo view (see Figure 1.4-3) clearly shows the variation in the motif and the conserved glycine. Also apparent are two hydrophobic conserved residues (eight and thirteen on figure 1.4-3, forming a staple) (Dragani and Aceto, 1999; Rigden and Galperin, 2004). This staple helps to stabilise the helix structure following the motif at the N-terminus of the helix.

Figure 1.4–3, Sequence Logo of EF hand Binding motif

Glycine	G	Aqua
Non-polar	AVLIPWFMY	Red
Polar	NCQST	Green
Positive	RHK	Purple
Negative	DE	Blue

This is a sequence logo, generated from the EF hand PROSITE pattern. Each letter represents a conserved amino acid at that position in the sequence; the taller the letter, the more frequently that amino acid occurs (Crooks et al., 2004; Schneider and Stephens, 1990) . Here, the DxDxDG motif can be seen, starting from residue 1. Residue 12 is the extra binding residue from the exiting strand. It can clearly be seen that aspartic acid is conserved at the first D position. However, there is some variation at the second and third, notably for asparagine and serine. The fourth ligand, in the exiting strand, shows a preference for the longer glutamic acid, but aspartic acid is also sometimes present.

+(Crooks et al., 2004; Schneider and Stephens, 1990)

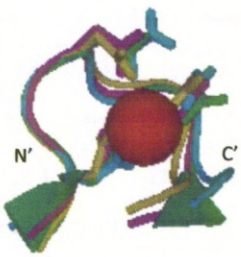
The staple is made up of two hydrophobic residues, one just before and one around four residues into the helix. These pin the helix in place, adding to its

stability (Dragani and Aceto, 1999; Munoz et al., 1995). The first aspartic acid is conserved in all cases. The second and third aspartic acids are known to show some variation, and are most commonly substituted with residues of similar polar character, such as asparagine and serine.

Figure 1.4-3, Variation in the secondary structure of DxDxDG motif-containing proteins.

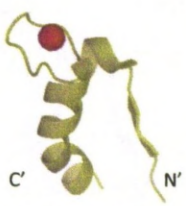
i) 1exr, this is a classic EF hand from calmodulin, the helix-loop-helix structure is clearly seen (Wilson and Brunger, 2000).

ii) 1k1x, shows a section of 4 - alpha – glucanotransferase from *Thermococcus litoralis*, with the binding loop seen as part of a beta turn (Imamura *et al.*,2003).



v) 1EXR, 1K1X, 1GCA and 1LWJ binding loops

iv) 1lwj, 4-alpha-glucanotransferase from *Thermotoga maritima* has a helix-loop-beta-strand structure (Roujeinikova *et al.* 2002).



iii) 1gca, from a periplasmic glucose/galactose receptor shows a beta-strand-loop-helix structure (Zou and Mowbray, 1993).

Here, we have 4 very different versions of calcium-binding structures. All four diagrams are shown here with the binding loop in the same orientation; this serves to highlight the differences in the types of structure and in the directions of the incoming and exiting strands. v) is a close up of the binding loops of these 4, with the three ligating residues shown. It can be seen very clearly that although the surrounding secondary structures i)-iv) are very different, the actual binding loop is in an almost identical conformation.

It has been shown that there is great variation in the different folds that may contain a DxDxDG motif; some examples of these structures can be seen in figure 1.4-4. This amount of variation is interesting, as it is unseen in other metal-binding proteins: for example, the copper-binding proteins, where a great many

types of binding site are known, but no one motif occurs in as many different folds (Herzberg and James, 1988); and the zinc-finger binding site that is seen in a few different folds but not with as much variation as seen with the DxDxDG motif (Krishna et al., 2003).

There is also a huge amount of variation in the secondary structures that may surround a DxDxDG binding loop, with no particular element being essential to its binding properties. Virtually any arrangement of alpha helix, 3_{10} helix and beta strand can be seen surrounding the motif, and, in the case of dockerin, there is no secondary structure in the seven residues preceding the binding loop (Rigden and Galperin, 2004). Despite this marked difference in secondary structures, the spatial orientations of the three binding residues of the DxDxDG motif are strikingly similar (see Figure 1.4-4 v). In fact, the root mean squared deviation (RMSD) between known DxDxDG loops, compared to the loop from calmodulin, is between 0.18 Å and 1.08 Å (mean 0.45 Å), whereas non-calcium-binding DxDxDG proteins range from 1.87 Å to 3.11 Å (Rigden and Galperin, 2004).

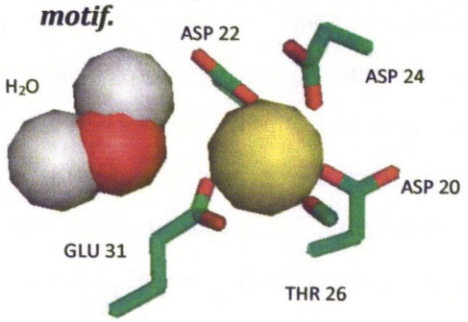
Finally, there is also variation in the extra ligating residue that occurs after the motif. This residue has been seen anywhere from two residues away from the motif, as in *Pseudomonas aeruginosa* alkaline protease 1kap (Baumann et al., 1993), to sixty-five residues away from the motif, as in *Salmonella typhimurium* periplasmic galactose-binding protein 1gcg (Flocco and Mowbray, 1994; Rigden and Galperin, 2004).

The variety of folds in which the DxDxDG motif occurs, along with its short length, the variation in extra ligating residue, and promiscuity across various species and functions, presents a strong case for the convergent evolution of the motif (Rigden and Galperin, 2004).

1.5 Metal binding and additional ligands.

As previously noted, the aspartic acid residues of the DxDxDG motif do not function in isolation. In fact, a calcium ion is typically bound by six or seven groups, consisting of oxygen atoms from side chains, carbonyl groups from the backbone, or from water molecules (Torrance et al., 2007). The typical EF-hand configuration adopts pentagonal bipyramidal geometry around the calcium ion. This consists of three or four oxygens from the three aspartic acid residues of the DxDxDG motif, along with a backbone carbonyl group, a water molecule and an additional one or two oxygens from the side chain of aspartic or glutamic acid downstream of the motif (see Figure 2.2.13-1) (Torrance et al., 2007). This extra residue involved in binding appears to be at a variable distance from the motif itself: although it is most commonly between three to around ten residues downstream, it could potentially be many more than even the sixty-five residue distances observed so far, if the protein is suitably folded (Rigden and Galperin, 2004).

Figure 1.5-1, Pentagonal bipyramidal geometry of the calcium-binding motif.



Pentagonal bipyramidal geometry can clearly be seen in the DxDxDG motif of 1exr. Each of the three aspartic acids provides one oxygen; the downstream glutamic acid provides two; one is provided by the backbone; and one by a water molecule.

1.6 The possible convergent evolution of the DxDxDG motif.

The presumed convergent evolution of the DxDxDG motif needs some context and qualification. Convergent evolution of protein sequences is often used too generally. Convergent evolution implies evolution in which non-homologous entities adapt in a selective manner to become more similar. In the context of proteins, this term could apply to convergence of the protein's gross function, structure, specific functional mechanics or sequence (Doolittle, 1994). In order to decide if convergent evolution occurs at any of these levels, a definition of what we would expect to see from convergent evolution is needed.

Convergent evolution with regard to the function of proteins is thought to have occurred many times. Here, a certain function, such as an enzymic action or specific inhibitor, can be shown to have evolved independently on a number of occasions in proteins with no common ancestor.

An example of this type of convergent evolution is seen in kinases that catalyse the phosphorylation of sugars. It has been shown that proteins with an analogous kinase function have evolved independently on at least three occasions. The hexokinase family, ribokinase family and inosine-guanosine kinase family all catalyse equivalent chemical reactions and display similar specificities. As well as convergence on the catalyzed reaction , convergence of the specificity of members of these protein families is also apparent (Bork et al., 1993) .

Another example is in the binding of bacterial surface proteins to immunoglobulin G antibodies (IgG). Bacterial surface proteins A (from *Staphyloccus aureus*), protein G (from group C and G streptococci) and H (from group A streptococci) appear to interact with the same part of the constant region of IgG. Despite their similar function, no sequence similarity between these proteins could be detected. Additionally, the structure of these three

proteins show marked differences. This evidence strongly suggests that convergent evolution on this binding function has taken place (Frick et al., 1992).

Convergence of protein structure may be seen in the shapes of folds that recur in unrelated proteins. In a clear-cut example of structural convergence, we would expect to see a similar shape to the fold and be unable to detect any signs of sequence similarity.

A plausible example is seen in the formation of alternating α-helices and β-strands, forming a barrel shape. These barrels are seen in many proteins, and often show striking similarity in their tertiary structure and, in some cases, display no detectable sequence similarity. It is possible that many of these barrels have some distant common ancestor; however the diversity of the function of the proteins containing an α-β barrel may also indicate a case for convergence on this structure (Branden, 1991; Farber, 1993; Wilmanns et al., 1991).

Has the sequence of these barrels evolved divergently, leaving the shape conserved, or have unrelated sequences convergently evolved the same advantageous, stable shape? As is clear from the case of the α-β barrel, it is often impossible to distinguish the effects of divergent and convergent evolution.

Convergent evolution concerning the specific mechanics of a proteins function may involve the arrangement of residues that are involved in an interaction. The convergent evolution of a particular arrangement of residues is evident in proteins where the residue arrangement is similar but the sequence and shape of the fold are not homologous; additionally the order that the residues appear may differ.

Guanine binding proteins have a large anionic pocket that facilitates binding of phosphate groups. This pocket consists of a number of glycines and a positive

lysine, which interact with the negative phosphate. Some protein kinases are also known to bind nucleotides in the same way. However, the interacting lycine present in the kinases comes from an entirely different part of the chain; additionally the overall fold and the sequence show no similarity (Knighton et al., 1991; Schulz, 1992).

Another example of this type of convergent evolution is seen in serine proteases. Chymotrypsin and subtilisin both catalyse peptide cleavage reactions using a 'catalytic triad'. Chymotrypsin's triad consists of His57, Asp102 and Ser195; subtilisin's triad is Asp32, His64 and Ser221. The order and spacing of these two sets of catalytic residues is very different. However their 3D arrangements are the same. Furthermore, the general shape, fold and sequence of these two proteins show no similarity (Kraut, 1977). It appears as though, due to certain requirements in the chemistry of the interactions in these two examples, the 3D arrangement of residues has indeed converged on similar patterns. These examples provide strong evidence for convergent evolution based on the chemical properties and arrangements of key residues in an interaction.

Convergent evolution at the sequence level is a subject of great contention. Many possible examples have been sited, however, in many of these cases, it has been argued that processes other than convergent evolution have taken place. In a true case of sequence level convergent evolution, substitutions in two unrelated protein sequences, which lead to increased similarity between the sequences, would be seen. Additionally, these substitutions may lead to beneficial adaptations.

It has been shown that, in certain cases, mutations take place in parallel protein sequences, which lead to substitutions for the same residue. For example, in HIV envelope proteins, the sequence GPGRAF changed to GPGAV in two lineages via different intermediates (Holmes et al., 1992holmes). Looking at the protein sequence as a whole, this type of 'local' adaptive substitution is far

outweighed by the natural tendency towards divergence in amino acid substitutions; therefore the sequence actually may become less similar overall. The similarities between these lineages might be attributed to a local convergent evolution across a few residues, but may also be because of constraints on amino acid mutations (Doolittle, 1994).

Although early examples have been refuted, more recently, the idea of sequence level convergent evolution is gaining greater acceptance. The extendin proteins are components of the toxins seen in poisonous lizards. These proteins appear to have convergently evolved sequence and functional similarities to the vasoactive intestinal peptides, a hormone that regulates vasodilation and smooth muscle contraction (Irwin, 2012).

Convergent evolution of sequence has also been indicated in the DR locus of *Mycobacterium tuberculosis*. Groupings generated using polymorphisims, seen in the DR gene, were compared to the lineage based on the DR gene sequence and showed poor agreement. Additionally, a number of cases showed identical polymorphism variants but were part of different sublineages, suggesting convergent evolution is taking place (Rindi et al., 2012).

There is less controversy in the convergent evolution of short linear motifs; as a result of their short length, they can easily arise convergently (Nevuda et al., 2005). There is evidence of a great number of cases where short linear motifs seem to have evolved independently. Examples of such convergently evolved motifs can be seen in the ELM database; in some cases, many independent evolutions have occurred (Gould et al., 2009).

It is clear that the term convergent evolution is used to describe many differing situations. How do these definitions of convergent evolution relate to the DxDxDG motif and how it may have evolved?

The extent of functional convergent evolution within examples of DxDxDG containing proteins is difficult to determine. There is clearly a great deal of

variation in the functions of calcium-binding proteins that bind using the DxDxDG motif. By definition, however, all of these proteins do share a common function in that they all bind calcium. Looking at the diversity of these calcium-binding proteins, it seems highly likely that examples of non-homologous proteins that share a calcium binding function do occur. However, can evolution of this general trait across many protein families really be considered true convergence on function?

At the structural level, it would appear there is no case for convergent evolution in the folds that contain the DxDxDG motif; in fact the diversity of structural situations seen in section 1.4 would strongly argue against this type of convergence. Additionally, where structural similarities do occur, such as in the EF hand family, there is also strong evidence for homology, based on both structure and sequence.

The type of convergent evolution most strongly evident in the DxDxDG motif is the mechanistic type. There is evidence of both convergence on a particular 3D arrangement of these residues, and the presence of an additional ligating residue at a varying distance from the motif.

This is comparable to the serine protease example; however, some differences exist. The various examples of DxDxDG motif superimpose to a much greater degree than the serine proteases. The DxDxDG example shows some evidence of variation in the distance between residues with respect to the extra ligating residue. This does not rule out the possibility of some kind of unknown loop transfer mechanism acting on solely DxDxDG element. A loop transfer mechanism is, however, not likely in the serine proteases as there is such variation in the spacing of the residues. In contrast the residues of the DxDxDG motif show regular spacing.

However, there are examples of the DxDxDG motif that indicate that a loop transfer is unlikely; Rabbit phosphoglucomutase (PDB code 3pmg) shows structural similarities to other DxDxDG motifs such as calmodulin. However, the

chain of the mutase displays a differing conformation at the G position, has a preference for magnesium and maintains contacts with its flanking domains. It therefore seems likely that the structural similarity, in this example, is more likely to have occurred from convergent evolution than by loop transfer mechanisms.

When specifically considering the convergent evolution of short linear motifs, the literature shows good evidence that this type of evolution occurs. Methods have been developed for identifying over-represented, convergently evolved, short linear motifs within proteins (Edwards et al., 2007).

This lends support to the conclusions that convergent evolution is the mechanism that leads to the DxDxDG motif appearing in many different types of loop (Rigden et al., 2011).

Therefore, the DxDxDG motif shows evidence of convergent evolution at both a mechanistic and local sequence level. It is likely that the striking similarity in the 3D arrangement, of the residues in various examples of the DxDxDG motif, has evolved as a result of a shift towards a mechanistic ideal. The sequence of the DxDxDG motif also shows evidence of convergent evolution.

1.7 Hypotheses

I hypothesise that:

The sequence surrounding a DxDxDG motif contributes to the potential binding properties of the motif, and will dictate if a protein will bind calcium or not.

The characteristics of the sequence surrounding a DxDxDG motif can be used to predict if a protein is likely to bind calcium or not.

In any given genome, there is likely to be a number of proteins that bind calcium through a DxDxDG motif that have not yet been identified.

1.8 Bibliography

Babu, Y., Sack, J.S., Greenhough, T.J., Bugg, C.E., Means, A.R., and Cook, W.J. (1985). Three-dimensional structure of calmodulin. Nature *315,* 37.

Baumann, U., Wu, S., Flaherty, K.M., and McKay, D.B. (1993). Three-dimensional structure of the alkaline protease of Pseudomonas aeruginosa: a two-domain protein with acalcium binding parallel beta roll motif. The EMBO Journal *12,* 3357.

Boehning, D., Patterson, R.L., Sedaghat, L., Glebova, N.O., Kurosaki, T., and Snyder, S.H. (2003). Cytochrome c binds to inositol (1,4,5) trisphosphate receptors, amplifying calcium-dependent apoptosis. Nature Cell Biology *5,* 1051.

Bork, P., Sander, C., and Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, and galactokinase families of sugar kinases. Protein Science 31.

Branden, C. (1991). The TIM barrel the most frequently occurring folding motif in proteins. Current Opinions in Structural Biology *1,* 878.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: A sequence logo generator. Genome Research 1188.

Doolittle, R.F. (1994). Convergent evolution: the need to be explicit. TIBS *19,* 15.

Dragani, B., and Aceto, A. (1999). About the Role of Conserved Amino Acid Residues in the Calcium-Binding Site of Proteins. Archives of Biochemistry and Biophysics *368,* 211.

Edwards, R.J., Davey, N.E., and Shields, D.C. (2007). SLiMFinder: A probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PLoS One

Engman, D.M., Krause, K., Blumin, J.H., Kim, K.S., and Kirchhoff, L.V. (1989). A Novel Flagellar Ca2+-binding Protein Trypanosomes. The Journal of Biological Chemistry *264,* 18627.

Farber, G.K. (1993). An alpha/beta-barrel full of evoultionary trouble. Current Opinions in Structural Biology *3,* 409.

Finn, B.E., Evenas, J., Drakenberg, T., Waltho, J.P., Thulin, E., and Forsen, S. (1995). Calcium-induced structural changes and domain autonomy in calmodulin. Nature Structural Biology *2,* 777.

Flaherty, K.M., Zozulya, S., Stryer, L., and McKay, D.B. (1993). Three-Dimensional Structure of Recoverin, a Calcium Sensor in Vision. Cell *75,* 709.

Flocco, M.M., and Mowbray, S.L. (1994). The 1.9 A X-ray Structure of a Closed Unliganded Form of the Periplasmic Glucose/Galactose Receptor fromSalmonella typhimurium. The Journal of Biological Chemistry *269,* 8931.

Forsburg, S.L., and Nurse, P. (1991). Cell Cycle Regulation in the yeasts Scccharomyces Cerevisiae and Schizosaccharomyces Pombe. Annual Review of Cell Biology *7,* 227.

Frick, I.M., Wikstrom, M., Forsden, S., Drakeneberg, T., Gomi, H., Sjobring, U., and Bjorck, L. (1992). *Convergent evolution among immunoglobulin G-binding bacterial proteins.* Proc. Natl Acad. Sci *89,* 8532.

Gifford, J.L., Walsh, M.P., and Vogel, H.J. (2007). Structures and metal-ion-binding properties of the Ca2+-binding helix-loop-helix EF-hand motifs. Biochem. J. *405,* 199.

Gould, C.M., Diella, F., Via, A., Puntervoll, P., Gemünd, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J.C., Chica, C., *et al.* (2009). ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Research Advance Access 1.

Herzberg, O., and James, N.G. (1988). Refined Crystal Structure of Troponin C from Turkey Skeletal Muscle. J. Mol. Biol. 761.

Holmes, E.C., Zhang, L.Q., Simmonds, P., Ludlam, C.A., and Brown, A.J. (1992). Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proceedings of the National Academy of Sciences *89,* 4835.

Ikura, M. (1996). Calcium binding and conformational response in EF-hand proteins. Trends Biochem. Sci *21,* 14.

Irwin, M. (2012). Origin and convergent evolution of extendin genes. General and Comparative Endocrinology *175,* 27.

Kahl, C.R., and Means, A.R. (2003). Regulation of Cell Cycle Progression by Calcium/Calmodulin-Dependent Pathways. Endocrine Reviews *24,* 719.

Knighton, D.R., Zheng, J., Ten Eyck, L.F., Ashford, V.A., Xuong, N., Taylor, S.S., and Sowadski, J.M. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science *253,* 407.

Kraut, J. (1977). Serine Proteases: Structure and Mechanism of Catalysis. Annual Review of Biochemistry *46,*

Kretsinger, R.H., and Nockolds, C.E. (1973). Carp Muscle Calcium-binding Protein: II. Structure Determination and General Description. Biol. Chem. 3313.

Krishna, S.S., Majumdar, I., and Grishin, N.V. (2003). Structural classification of zinc fingers. Nucleic Acids Research *31,* 532.

Lagasse, E., and Clerc, R.G. (1998). Cloning and Expression of Two Human Genes Encoding Calcium-Binding Proteins That Are Regulated during Myeloid Differentiation. Molecular and Cellular Biology *6,* 240.

Lim, S., and Franklin, S.J. (2006). Engineered lanthanide-binding metallohomeodomains: Designing folded chimeras by modular turn substitution. Protein Science, Vol. 15 *15*, 2159.

Matthews, G.G. (2001). Neural Control of Muscle Contraction. In NEUROBIOLOGY Molecules, Cells and Systems, Blackwell Science Inc.)

Mattson, M.P., and Chan, S.L. (2003). Calcium orchestrates apoptosis. Nature Cell Biology *5*, 1041.

Munoz, V., Blanco, F.J., and Serrano, L. (1995). The hydrophobic-staple motif and a role for loop-residues in alpa-helix stability and protein folding. Nature Structural Biology *2*, 380.

Nevuda, V., Linding, R., Su-Angrand, I., Stark, A., de Masi, F., Gibson, T.J., Lewis, J., Serrano, L., and Russell, R.B. (2005). Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks. PLOS Biology *3*, 2090.

Nitz, M., Sherawat, M., Franz, K.J., Peisach, E., Allen, K.N., and Imperiali, B. (2004). Structural Origin of the High Affinity of a Chemically Evolved Lanthanide-Binding Peptide. Angewandte Chemie *116*, 3768.

Olofsson, M.H., Havelka, A.M., Brnjic, S., Shoshan, M.C., and Linder, S. (2008). Charting calcium-regulated apoptosis pathways using chemical biology: role of calmodulin kinase II. BMC Chemical Biology *8*,

Ordal, G.W. (1977). Calcium ion regulates chemotactic behaviour in bacteria. Nature 66.

Pitta, T.P., Sherwood, E.E., Kobel, A.M., and Berg, H.C. (1997). Calcium is required for swimming by the non-flagellated cyanobacterium Synechococcus strain WH8113. J. Bacteriol. *179*, 2524.

Rigden, D.J., Woodhead, D.D., Wong, P.W.H., and Galperin, M.Y. (2011). New Structural and Functional Contexts of the Dx[DN]xDG Linear Motif: Insights into Evolution of Calcium-Binding Proteins. PLoS One

Rigden, D.J., and Galperin, M.Y. (2004). The DXDXDG Motif for Calcium Binding: Multiple Structural Contexts and Implications for Evolution. J. Mol. Biol. *343*, 971-984.

Rindi, L., Lari, N., and Garzelli, C. (2012). Large Sequence Polymorphisms of the Euro-American lineage of Mycobacterium tuberculosis: A phylogenetic reconstruction and evidence for convergent evolution in the DR locus. Infection, Genetics and Evolution *12*, 1551.

Schafer, B., and Heizmann, C.W. (1996). The S100 family of EF-hand calcium-binding proteins: functions and pathology. TIBS *21*, 134.

Schaub, M.C., and Heizmann, C.W. (2008). Calcium, troponin, calmodulin, S100 proteins: From myocardial basics to new therapeutic strategies. Biochemical and Biophysical Research Communications *369,* 247.

Schneider, T.D., and Stephens, R.M. (1990). Sequence Logos: A New Way to Display Consensus Sequences. Nucleic Acids Res. *18,* 6097.

Schulz, G.E. (1992). Binding of nucleotides by proteins. Current Opinion in Structural Biology *2,* 61.

Skelton, N., Kordel, J., Akke, M., Forsen, S., and Chazin, W. (1994). Signal transduction versus buffering activity in Ca2+-binding proteins. Nature Structural Biology *1,* 239.

Stevens, F.C. (1983). Calmodulin: an introduction. Can. J. Biochem. Cell Biol. *61,* 906.

Tassinary, L.G., and Cacioppo, J.T. (2000). "The Skeletomotor system: surface electromyography". In Handbook of psychophysiology, Tassinary, L. G., Cacioppo, J. T. and Berntson, G. G. eds., (Cambridge: Cambridge University Press)

Tisa, L.S., and Adler, J. (1992). Calcium ions are involved in Escherchia coli chemotaxis. Biochemistry *89,* 11804.

Torrance, J.W., MacArthur, M.W., and Thornton, J.M. (2007). Evolution of binding sites for zinc and calcium ions playing structural roles. Proteins Structure Function Bioinformatics 813.

Wilmanns, M., Hyde, C.C., Davis, D.R., Kirshner, K., and Jasonius, J.N. (1991). Structural Conservation in Parallel Plar-Barrel Enzymes That Catalyze Three Sequential Reactions in the Pathway of Tryptophan Biosynthesis. Biochemistry *30,* 9161.

Wu, Y., Deford, J., Benjamin, R., Lee, M.G., and Ruben, L. (1994). The gene family of EF-hand calcium-binding proteins from the flagellum of Trypanosoma brucei. Biochemical Journal 833.

Section 2:- Searching for the DxDxDG motif in the PDB

2.1 Section Introduction

2.1.1 Section Overview

As has been demonstrated in the previous chapters, the DxDxDG motif is unique, it appears in many non-homologous proteins. Owing to its simple short nature, it is commonly found by chance in both functional and non-functional forms. Therefore, searching for calcium-binding proteins using the motif alone presents a problem.

The topics explored in this section include: how motifs are defined, different ways motifs and domains are used to classify proteins, and how this can help predict protein function. Also covered is how different groups have attempted to overcome the difficulties associated with identification of short linear motifs and metal binding sites. The SPASM software package (Kleywegt 1998) is also described. SPASM takes into account the three-dimensional configuration as well as the identity of the residues contained within the motif, when performing a similarity search.

The method section covers how this software was used to search for occurrences of the DxDxDG motif in known protein structures, and how verification of those that showed binding to a metal group was used to reduce the number of candidates for further analysis. The candidate occurrences of the DxDxDG motif were sorted by their structural characteristics according to SCOP (Murzin et al., 1995) classifications. These were then verified as DxDxDG-type metal-binding proteins by hand. Those proteins identified were used for further analysis in later sections. The similarity of both sequence and spatial arrangement between these proteins was also looked at using clustering software, CLANS (Frickey and Lupas, 2004).

2.1.2 Glossary and abbreviations

Array – a data-structure used to store simple lists.

BLOCKS – a database of motifs based on automated alignments and motif identification.

BLOSUM matrix – substitution matrix used for showing sequence alignment of proteins.

CLANS – software that builds cluster diagrams of sequence similarities.

DILIMOT – linear motif identification server.

ELM – database to search a sequence against a database of linear motifs.

eXtensible Markup Language (XML) – structured data format.

Expect value (E-value) - this refers to the number of matches to a query sequence, you could expect to see in a particular sized database. Low E-values show greater significance.

Hash – a data-structure where each data element is associated with a key.

LSQMAN – software that calculates RMSD between two sets of residues.

Perl – high level interpreted programming language.

Pfam – database of protein families represented as hidden Markov models.

PRINTS – database of protein family fingerprints.

PROSITE – database of protein motifs represented as regex and profiles.

Pseudo atom – a theoretical atom that is used to represent the amino acid side chain, in SPASM.

Regular expression (regex) – a consensus representation of a sequence where some positions show ambiguous identity.

Reverse PSI BLAST (RPS BLAST) – a method of searching for motifs in a given sequence.

SCOP – database that classifies protein structures hierarchically.

SLimDisc – linear motif identification server.

SPASM – software used to search for motifs in a similar 3D arrangement to a given set of residues.

Subroutine – portion of code within a larger program that performs a specific task.

2.1.3 List of Figures

2.2 Preliminaries

2.2.1 Motifs and Domains.

The term motif can have a number of definitions depending on the context. However, it most commonly refers to an element that is present as a functional or structural unit common to a number of related or unrelated proteins. A domain is a similar concept; however, it usually refers to a larger structural or functional unit (Rossmann and Argos, 1981). Domains can be thought of as protein building blocks; the shapes, sizes and functions may vary, but these blocks fit together, along with low-complexity and disordered regions, to form a complete protein (Rossmann and Argos, 1981). A domain may contain one, many or no motifs within it. Collections of motifs and domains often recur in the same pattern, similar to a recurring theme in a piece of music. Motifs and domains, and how they fit together, are therefore a useful tool in the prediction of protein function and classification. These properties have been utilised in a number of databases and search methodologies. Each of these databases has a common aim of helping to assign a function to a novel protein sequence through sequence similarities, but each was constructed and can be searched using a variety of methodologies (Attwood, 2002; Henikoff and Henikoff, 1992; Hulo et al., 2007).

2.2.2 Regular expressions; a way of representing a sequence motif.

Short motifs are often represented as regular expressions of their conserved residues; for example, the EF-hand-binding motif is an example of a motif that contains DxDxDG. This can be represented as **D - {W} - [DNS] - {ILVFYW} - [DENSTG] - [DNQGHRK] - {GP} - [LIVMC] - [DENQSTAGC] - x(2) - [DE] - [LIVMFYW]** (Kawasaki and Kretsinger, 1995).

The PROSITE pattern syntax is particularly suitable for motifs that are short and well conserved. Typically, these include enzyme catalytic sites, prosthetic group-attachment sites, metal-ion-binding sites, disulphide bonds, or other binding sites (Hulo et al., 2007). The basic regular expression syntax used consists of single-letter amino acid abbreviations, including x to represent any amino acid, together with a few other characters. Square brackets '[ ]' are used when a number of possible residues are seen at a given position, each possibility being listed within the brackets; for example, **[DN]** would mean an aspartic acid or an asparagine. Curly brackets '{ }' represent the opposite of square brackets, meaning that a residue may be any but those shown between them; for example, **{GA}** would mean any amino acid but glycine or alanine. Parentheses '( )' contain numbers that allow for repeated terms; for example, **x(3)** is 3 of any amino acid, and **x(2,4)** would mean 2 to 4 of any amino acid . Dashes '-' are used to separate terms to avoid ambiguity.

PROSITE patterns have certain limitations. When they are short in length, there is a high likelihood of a pattern occurring in an unrelated sequence; this means that, for any given search, many false-positive results may be identified.

The regular expression (regex) structure of PROSITE patterns is also quite rigid, meaning that if a novel or uncommon variant of a pattern is present in the target database, it may not be recognised, producing false-negatives. To be sure that as many new examples as possible will be identified, a good range of true-

positive examples must be used to construct the regex; this will ensure that all possible alternatives at a particular residue position are represented.

Where false-positive or false-negative results to a pattern are known to occur, they can be annotated in the database. This type of annotation may not be of use when using the regex to search for novel sequences, unless the regex is revised to include these examples.

Another limitation is the lack of provision for weighting of the variation at a particular position. For example, the last residue in the pattern in the DxDxDG motif is **[LIVMFYW].** However, the sequence logo shows that tryptophan (W) and leucine (L) are far more common than the other amino acids shown (see Figure 1.4-3). This lack of scoring also means that in short patterns the importance of each residue in the sequence is increased.

These factors limit the usefulness of regex as a predictive tool. However, they can still be useful as a quick check for the presence of a well-defined pattern.

2.2.3 PROSITE; a sequence motif database.

The PROSITE database was first distributed in 1988 and contained just 58 patterns (Hulo et al., 2007); since then, it has grown slowly in comparison to other databases, and now contains just over 1500 documentation entries and 1300 patterns (PROSITE, 2009). This slow growth has been mainly owing to the painstaking, by-hand annotation that the database is built upon. Each entry is checked and any predictions or functional annotations are verified. This greatly improves the quality and reliability of the database. The database itself is an annotated protein resource, based around both regular expressions and weight matrices.

PROSITE regex patterns are sensitive to variations that may occur as a result of sequencing errors or sequence divergence. This means that a sequence may be misclassified if just one residue shows a sequencing error. Therefore, PROSITE patterns are generally considered unsuitable for detection of unrelated proteins (Hulo et al., 2007). PROSITE patterns are also an all-or-nothing classification; there is no way to score how well a sequence matches - it either matches or it doesn't.

Owing, in part, to these limitations, many PROSITE entries also now include a sequence profile, and nearly all new PROSITE entries are profiles (Westhead et al., 2002). These profiles are more suited to longer sequences and may characterise sequences over their entire length, not just the most conserved residues. This is done by position-specific scores for each amino acid, along with position-specific penalties for opening or extending an insertion or deletion (Hulo et al., 2007). The format takes the form of a matrix, with each of these scores recorded against their alignment position. These scores take into account not only variation in residues at a position across the alignment, but also allow for likely substitutions that may not be present in the alignment, by weighting these residues accordingly in the matrix. This type of scoring also means that even if a residue that is highly conserved in the alignment shows an error in sequencing or

a mutation, as long as the rest of the sequence is of a high enough similarity, a sequence will still match.

Despite their advantages, the use of PROSITE sequence profiles is not always preferable to using a PROSITE pattern. For example, in cases where a function results from a few highly conserved residues, such as in catalytic sites, patterns are still better discriminators (Sigrist et al., 2002). Also, because of their intelligibility to users and their low computational complexity, they are popular as a first line of investigation (Hulo et al., 2007). Profiles are often more suitable for general structural properties; for example, the relationship of the non-enzymatic haptoglobulin to the trypsin family is detected by the PS50240 profile, even though substitutions are present at the proteolytic active site residues (Kurosky *et al.;* Sigrist *et al.* 2002). This relationship is not detected by the corresponding patterns PS00134 and PS00135, due to their reliance on the conservation of the catalytic residues.

SCANPROSITE is an example of a tool that searches a given sequence, or set of sequences, for a given motif, or against a database of motifs (de Castro et al., 2006). SCANPROSITE, as its name suggests, is specific to the PROSITE database. SCANPROSITE is able to identify which patterns and profiles from the PROSITE database are present within the sequence, or sequences, provided (de Castro et al., 2006). This can assist in the identification of the function of uncharacterised proteins.

The process of designing the patterns that represent a particular protein function or family is manually undertaken. Each entry is also manually annotated and verified before being published, greatly extending the time it takes for new profiles to be added. The small size of this database results from this, time-consuming, manual processes (Westhead et al., 2002).

Even though the size of the database is small and there are limitations to how the patterns are expressed, the quality of this annotation means it is still relevant when looking into the function of a novel protein.

2.2.4 BLAST, PHI-BLAST and PSI-BLAST; methods for searching for protein sequence similarities and sequence motifs.

Many different tools are commonly used to search for similarities within proteins and for occurrences of motifs. A number of these tools are based around the BLAST algorithm. BLAST stands for Basic Local Alignment Search Tool (Altschul et al., 1990). It is a tool used to search a database for sequences that show similarity to a given query sequence, and is often a first point of call when looking to create a multiple alignment for a set of related proteins.

The principle behind BLAST is the assumption that high-scoring alignments are likely to have short stretches of near-identical sequence, referred to as words. After removal of low-complexity regions, the sequence is split up into words (Altschul et al., 1990). Matches to these words are then searched for in the database. A scoring matrix is used to filter the matches that do not meet a certain threshold ("T") (Altschul et al., 1990). An attempt is then made to extend these word matches along the sequence. Each extra letter will change the score, but only those that continue to meet the threshold T will be kept. These alignments are called high-scoring segment pairs (Altschul et al., 1990). Finally, BLAST produces a full dynamic programming alignment of the high-scoring sequences identified. Each sequence is assigned an Expect value (E). This is the likelihood of the sequence similarity occurring at random in a database of the size used. The sequences are then ranked by E value and the best matches are reported (Altschul et al., 1990).

BLAST is a simple, quick search method that is good at identifying similarity between a query protein and sequences in a database. It is often used as an initial method for an outline identification of the likely type and function of a newly sequenced protein. Although a good starting point, a low E value in BLAST should not be taken to be a proof of function or homology. Further investigation is needed to prove the relationship and function of a protein. The BLAST algorithm is good at identifying closely similar and potentially homologous proteins. However, it will often miss more distantly related proteins.

PSI-BLAST, or Position Specific Iterated BLAST, takes an alignment of the top-scoring BLAST results and creates a position specific scoring matrix. This creates a profile of the conserved residues within the alignment (Altschul et al., 1997). This profile is then used to find further hits that may be more distant homologues, and may not show up in a regular BLAST search. This process can be repeated to find more and more distantly related proteins (Altschul et al., 1997).

Again, it should be noted that, although the intention of PSI-BLAST is to highlight potentially distantly related proteins, it actually searches for proteins with low level similarity and, therefore, matches are not necessarily proof of homology. Also, there is no ability to restrict the search to proteins of a particular function. Therefore, it would not be of use if one wanted to search, for example, for all calcium-binding proteins.

Pattern Hit Initiated BLAST, PHI-BLAST, is a variation on PSI-BLAST that combines the matching of regular expressions with alignments surrounding the match (Zhang et al., 1998). Therefore, a BLAST search can be performed where the results are restricted to those with a particular motif. The motif is given to the software in PROSITE format, and this regular expression is used to search the database; any matches found are then aligned using the BLAST algorithm (Zhang et al., 1998). Optionally, as with PSI-BLAST, a profile can then be constructed and further rounds of searches can be completed.

PHI-BLAST is effective at identifying similar proteins that also contain a common motif. The same predictive limitations still apply, however. The presence of a motif does not guarantee function, and similarity does not mean proteins are necessarily homologous.

RPS-BLAST, or Reverse PSI-BLAST, is a tool that uses a given sequence to search for matches within a database of motifs. This is essentially the opposite of PSI-BLAST, which searches a database of sequences with a given profile (Marchler-Bauer et al., 2002). There is an option to give the position in which the

motif occurs in the query sequence. This allows for discrimination if there are several occurrences of a motif in the sequence (Marchler-Bauer et al., 2002).

RPS-BLAST is a different type of tool from the other BLAST searches. Rather than searching for similarity, it searches a sequence for the presence of a number of motifs. This type of search therefore inherits the advantages and disadvantages of the type of motifs that are being matched, be they regex or profile based.

2.2.5 Linear motifs; a special type of motif.

Linear motifs are stretches of sequence within a protein that often facilitate protein-protein interactions. They are linear in that each part of the motif is found in sequence, and a particular structural configuration is not essential for the motif to be evident. Linear motifs are not as easy to identify as domains; domains tend to be large, display structural organisation and may be conserved even in less closely related proteins. Linear motifs tend to be small and seem to be far more commonly found in disordered parts of a protein. As a result, they tend only to be conserved in closely related species (Nevuda et al., 2005).

DILIMOT is a server for finding linear motifs in a set of proteins. The theory behind a DILIMOT search is that proteins that share a binding or interaction partner will share a feature that mediates this binding. When all the shared domains between two proteins that show similar interactions are discarded, what is left is likely to contain the motif that facilitates this interaction (Nevuda et al., 2005). First, a set of proteins that share an interaction partner is obtained. Then, all the regions less likely to contain linear motifs, such as globular domains, coiled coils and any other regions unavailable for interactions, including trans-membrane segments and buried regions, are removed. There are likely to be linear motifs discarded in this step, although the vast majority will not occur in these regions. Next, the sequences are compared, and any homologous regions within the set are discovered and culled, so only one representative remains. This is so that only motifs arising from convergence, and not those resulting from homology, are discovered. Finally, all the motifs of three to eight residues in length are discovered and scored according to their over-representation (Nevuda et al., 2005).

SLimDisc is another method for linear motif discovery and works on similar principles to DILIMOT. The major difference is that SLimDisc does not discard the homologous regions across the sequences prior to motif discovery. Instead, the resulting motifs are weighted according to the relationships of the proteins containing the motif (Davey et al., 2007).

2.2.6 ELM; a linear motif database.

The Eukaryotic Linear Motif resource for Functional Sites in Proteins, or ELM, aims to build on the predictive power of linear motifs represented as regexes (Puntervoll et al., 2003). As discussed previously, simple motifs may occur repeatedly at random, and searching for a functional example using a simple regex produces a great number of false-positives. This over-prediction is compounded when the motif of interest occurs independently of defined structural elements, such as between domains, and additional information cannot be gained by a domain search (Puntervoll et al., 2003).

When a protein sequence of interest is submitted for a motif search, many of the matches may be statistically significant but biologically irrelevant; for example, if a motif is not functional in a particular cellular compartment or species, if a motif occurs buried within the core of a globular domain, or if it appears as part of a secondary structural element, which may disrupt the normal structure of the motif (Puntevoll *et al.*, 2003; Gould *et al.*, 2009).

ELM attempts to overcome these problems by filtering by cell compartment, phylogeny, globular domain clash and structure. This allows many of the false-positives, and motifs that are not of interest, to be discarded. This may reduce the number of matches by an order of magnitude or more, significantly reducing the motifs that need to be considered when trying to identify the function of a protein (Gould et al., 2009).

The ELM database is made up of tables of motifs, along with their associated data, each annotated by hand through use of BLAST searches, multiple alignments, database searches and exchanges with experimentalists (Puntervoll et al., 2003). Additionally, known instances of the motif are linked to the motif's entry and can be browsed, allowing further comparisons to be made (Gould et al., 2009).

Despite the refinement of the linear motif searches provided by ELM, a match cannot be taken as an indicator of a true functional site. Matches must be further verified, and candidates must be tested experimentally (Gould et al., 2009).

2.2.7 Pfam; a protein sequence family database

Pfam is a protein domain and family database; families are represented by hidden Markov models (HMM), and by both seed and full multiple sequence alignments (Bateman et al., 2002). The seed alignment is a manually created alignment of representative members of a protein family. This alignment is used to create a profile HMM to represent the family. The full alignment is then created by searching using the HMM that represents the family and creating an alignment with the results (Finn et al., 2008). The Pfam families are further refined using domain boundary data from the SCOP database (Bateman et al., 2002). There are two versions of the database available, one manually curated and one with automated annotations. This allows for both an accurately verified version and one that has more comprehensive coverage. The format of the database itself is presented in two forms: a flat-file database and a relational database (Finn et al., 2008).

2.2.8 PRINTS and BLOCKS; using multiple sequence motifs to predict protein families.

At a larger scale, collections of motifs found together in a protein can be used to group proteins into families. The parts of a sequence that are common across a protein family are usually the parts of the protein important in the structure or function of the protein; these are the motifs. It makes sense, therefore, just to concentrate on these motifs and how they are arranged within the protein to infer relationships and predict protein function (Attwood, 2002).

PRINTS is a database of protein motif fingerprints. It was established in 1991 with a collection of 29 entries and 116 motifs; it now has over 1900 entries and 11000 motifs represented (Attwood, 2009). A PRINTS protein fingerprint starts off with a manually created multiple alignment. The conserved regions in the alignment are turned into frequency (identity) matrices that represent the motifs. These motifs are then used to search a database of sequences, and any new sequences that match motifs in the correct order are used to improve the fingerprint. Once the process is complete, the fingerprint is manually annotated (Attwood, 2002).

The BLOCKS database is built on similar principles. However, it is fully automated in its construction and its terminology differs slightly. The BLOCKS are constructed using the PROSITE database, keyed against the SWISS-PROT database. The most highly conserved regions in a protein group are identified, and then these initial alignments are used to generate a block. Blocks are created automatically using a substitution matrix. Individual sequences are aligned against candidate motifs by software called MOTIF. Next, the blocks are extended again using a substitution matrix by MOTOMAT. An initial database of blocks is created and then undergoes two rounds of refinement using BLOSUM and PROTOMAT (Henikoff and Henikoff, 1992).

Therefore, the major differences between PRINTS and BLOCKS are: the name of an entry; a fingerprint/print or a block; and the method of construction of the

initial alignment. PRINTS uses manual alignments, whereas BLOCKS uses alignments automatically generated from PROSITE. The method of refinement also differs in that PRINTS aims to expand each entry by adding new sequences to improve each alignment, while BLOCKS aims to reduce redundancy in the database through clustering using BLOSUM.

2.2.9 The Protein Data Bank; a protein structure database.

The Protein Data Bank, or PDB, is a central repository for structural models of proteins, and is provided by the RCSB (Berman et al., 2000). The database contains structural models from NMR spectroscopy and X-ray crystallography, presented in flat-files of a standardised format (Berman et al., 2008; Rose et al., 2012). Each structure is given a four-digit alphanumeric code; for example, as we have seen, calmodulin from *Paramecium tetraurelia* has the PDB accession code 1exr (Wilson and Brunger, 2000).

2.2.10 SCOP; a protein structure family database.

SCOP is another type of protein family database; however, classifications are based on structure instead of sequence, and are arranged in a hierarchical manner.

Figure 2.2.10-1, SCOP Lineage of Calmodulin from *Paramecium tetraurelia*

1. Root: SCOP
2. Class: All alpha proteins [46456]
3. Fold: EF Hand-like [47472]
   core: 4 helices; array of 2 hairpins, opened
4. Superfamily: EF-hand [47473]
   Duplication: consists of two EF-hand units: each is made of two helices connected with calcium-binding loop
5. Family: Calmodulin-like [47502]
   Duplication: made with two pairs of EF-hands
6. Protein: Calmodulin [47516]
7. Species: Ciliate (Paramecium tetraurelia) [TaxId: 5888] [47523]
   SQ Q42478

The SCOP hierarchy consists of 4 broad levels of classification: Class, Fold, Super-family and Family. Additionally seen here are the root, protein and species levels.

The database is divided up based on four layers of hierarchy: the smallest is the family, where each family contains proteins that either have at least 30% sequence identity or show a very similar structure and function (Murzin et al., 1995). Next is the superfamily; these consist of families that have lower sequence identities, but a broadly similar structure and maybe function (Murzin et al., 1995). Then, the common fold groups families and super-families together if they contain secondary structural features in the same arrangement (Murzin et al., 1995). Finally, at the top level, there is the class. This is based on the type of secondary structural elements that make up the gross structure of a fold. The five classes are: all-alpha (a), all-beta (b), alpha and beta (a/b), alpha plus beta (a+b), and multi-domain. The all-alpha and all-beta contain folds consisting principally of α helix or β strand respectively; the a/b contains a mix of α helix and β strand; in a+b, both are present, but regions of α helix are segregated from regions of β strand. Multi domain contains examples that don't conform to these groupings. This hierarchy can be represented in a short form, in the format:

a.39.1.3, where 'a' represents the class (in this case, all-alpha), '39' represents the fold (in this case, EF-hand like), '1' the superfamily (in this case, EF-hand) and '3' the family (in this case, Osteonectin). Additionally, below the family level, structures are grouped by protein, then species (Murzin et al., 1995). The classification of our example of 1exr can be seen in Figure 2.2.10-1.

The database itself is built as a web-based resource; however, a set of parseable files is also available, making automated analysis and protein classifications easier offline (Lo Conte et al., 2002).

2.2.11 SPASM; a method for searching for local structural similarity.

SPASM is software that can be used to perform searches based on spatial arrangement of residues. SPASM uses a query sequence and its associated three-dimensional (3D) positional data to search a database derived from the protein data bank (PDB (Berman et al., 2000)) for target sequences that show a similar 3D arrangement to the query sequence. A SPASM (Kleywegt, 1998) search is made up of three parts: first, a sequence search is performed; then, the positions of the alpha carbons of the protein back bone are matched; finally, the backbone and the side chains are aligned.

The input file for SPASM is based on the PDB file format (Berman et al., 2008). Each of the residues that are to be searched for is listed, with a line for each atom within the residue and its 3D position. If the identity of the residue at any of these positions is expected to be variable, the residue can be listed as "XXX". When the software runs, the user can further specify if there is any flexibility with regard to the identity of a residue, either by specifying substitutions allowed for a particular amino acid, or by the application of a BLOSUM matrix (Henikoff and Henikoff, 1992) set to a user-defined cut-off value.

The software uses its own database, made up of sequences and related structures. The structures stored are reduced to sets of coordinates, both for the alpha carbons of the main chain, and for pseudo-atoms that represent the centre of gravity of the side chains (Kleywegt, 1998).

First, a simple sequence motif search takes place. This reduces the number of potential matches and the time needed for the structural part of the search. The motif pattern used is created from the input file along with any allowed substitutions. Next, an initial structural match is made using the main-chain alpha carbons, the side-chain pseudo-atoms, or both (Kleywegt, 1998). This is quicker than performing a full RMS calculation for each candidate. Finally, the structures of the resulting hits are aligned to the query structure, and a root mean squared value is calculated. The results that fall below a specified RMS cut off are reported (Kleywegt, 1998).

2.2.12 CLANS; software for cluster analysis of sequence similarity and structural arrangements.

One useful way of looking into the relationships within a large set of related sequences is to perform a cluster analysis. Essentially, this technique produces a networked visualisation of the relationships between sequences. Those sequences that show the greatest degree of similarity cluster closely together; those that are more divergent are further apart on the diagram. CLANS (Frickey and Lupas, 2004) can be used to perform this type of analysis.

One advantage of this software is that any dataset that can be translated into an all vs all relation can be used. Therefore, in addition to standard sequence comparisons, it is possible to analyse microarrays, display clusters of the standard amino acids according to a BLOSUM matrix, or to compare the RMSDs of a set of molecular models (Frickey, 2007).

2.2.13 Analysis of motif search methods and how they relate to the DxDxDG motif, metal-binding sites and linear motifs.

The DxDxDG motif is an example of a linear motif. These differ from domains in that they are of much shorter length and are often defined by just a few residues (Neduva and Russell, 2005). The likelihood of a linear motif sequence occurring in a genome, by chance, is high. This is because of their relatively short length. This means that they may be present in many proteins, both in functional and non-functional forms (Neduva and Russell, 2005).

A BLAST, PSI-BLAST or PHI-BLAST search, as noted previously, is effective at finding sequences similar to a query sequence from within a database. This would be useful in obtaining more examples which are similar to proteins we already know to have a functional DxDxDG motif, but may not have been the subject of structural studies. However, as there is known to be no overall sequence similarity between the different families of DxDxDG containing protein, this type of search is unlikely to give examples of new structural environments for the DxDxDG motif.

A search of the PDB for the regex "[DNS]-x-[DNS]-x-[DNS]-G" gives over 10,000 matching sequences. This is over 12% of the database that contains 80,041 sequences. It is likely that a great number of these are non-functional, and many will be unrelated to metal-binding. The large number of candidate sequences, only a small number of which are likely to be functional, and the expectation that new functional examples would probably be related to those already seen, meant that a comprehensive search would be time consuming and unlikely to provide any significant results (Rigden and Galperin, 2004). Recently, the independent evolution of the DxDxDG motif has gained greater acceptance. It is possible, therefore, that functional motifs have not been recognized as such because they are dissimilar from those already known.

The PROSITE profiles represent sequence segments that are usually larger and more explicit than the simple DxDxDG motif. They are too specific for the

problem of searching for new contexts for this motif, as they will only identify a particular, defined motif, domain or family. Restricting the search to one of these profiles will not detect other motifs, domains and families that contain the DxDxDG motif, but which are different in their overall sequence. Even if a search was performed for all of the motifs, domains and families known to contain the DxDxDG motif, in a functional form, any novel examples in the database may still remain undetected.

BLAST methods are effective for identifying proteins that show a similar sequence over the length of the protein. The different classes of DxDxDG containing proteins show no significant similarity in their overall sequence. BLAST searches may be useful in obtaining further examples of DxDxDG proteins that bind calcium, and show similarity to already identified proteins. However, they will be unable to identify new types of DxDxDG proteins as their sequence will likely be dissimilar from known examples.

Pfam and its HMM methods are intended to allow the identification of members of a protein family. As the DxDxDG proteins that are known to bind calcium are members of a number of disparate protein families, the use of Pfam methods may have limited use in the identification of new examples of DxDxDG motifs. It may be possible, however, to utilise HMM methods to produce an effective search for DxDxDG motif proteins. This could be an avenue for further study.

BLOCKS and PRINTS motifs provide a different way of searching for related proteins. Rather than looking for a particular motif or domain, families and super-families are defined by specific elements, including motifs, domains and any other conserved region. As this approach allows some variability in the elements that make up an individual member of a family, it is possible that new combinations of elements containing the motif could be found. It is possible, therefore, that a fingerprint of the DxDxDG containing proteins could be constructed. However, other than the motif itself, there is little sequence

conservation between DxDxDG containing proteins. This is compounded by the lack of multiple examples of many of the families, meaning even conservation within a family is difficult to measure. Additionally, it is not certain that every motif and domain that could potentially make up a functional DxDxDG protein is represented in the dataset. It is unlikely, therefore, that the use of the PRINTS or BLOCKS methodologies would be able to produce a comprehensive search technique.

It is clear that sequence based searches are not suitable in the identification of calcium-binding DxDxDG proteins. A different way of searching for potentially functional occurrences of this motif is needed. Looking at the structure of the motif and possible DxDxDG containing proteins may provide better searches.

Research into the geometry of the residues that bind calcium and zinc has resulted in the identification of four archetypal calcium-binding sites and two archetypal zinc-binding sites (Torrance et al., 2007). The archetypal calcium-binding sites are: trypsin-like sites, EF-hand-like sites, iota-carrageenase-like sites, and collagenase-like sites. The zinc sites are Cys-Cys-Cys-Cys sites and His-Cys-Cys-Cys sites (Torrance et al., 2007).

In an attempt to investigate the evolution of these metal-binding configurations, structural template search methods have been used. In one study, two types of structural template were used: one constructed from the alpha and beta carbons of the residues at the binding site, and one constructed from the atoms directly involved in calcium binding (Torrance et al., 2007). These investigations aimed to discover all the occurrences of the metal-binding sites, but also discovered many relatives of metal-binding proteins that lack the ability to bind calcium. This loss of function seems to result from point mutations of the residues directly involved in calcium binding (Torrance et al., 2007). It has been suggested that the chemistry of metal binding strongly confines the geometry at metal-binding sites, which could explain the convergence evident in the evolution of calcium-binding proteins (Torrance et al., 2007). This research did

not identify many of the novel DxDxDG structures seen in the work by Rigden and Galperin. This was probably because the structural templates were constructed only using models from the PDB with calcium or zinc bound. However, a number of the new examples discovered by Rigden and Galperin were not bound to calcium (Rigden and Galperin, 2004).

Structural-alphabet-based motif-discovery methods have also been used to identify motifs involved in magnesium binding. Magnesium ions are similar to calcium ions, as they both have a double positive charge and, although smaller, are sometimes found interacting with calcium-binding sites (Dudev and Lim, 2007). The structural-alphabet motif-method works by encoding the 3D structure of a five-residue segment. Each segment is represented by one of sixteen single-letter codes. The magnesium-binding motifs can then be identified as recurring structural sequences. This method has revealed that there is significant preference in the secondary structure local to these magnesium-binding motifs, even when there is little sequence similarity (Dudev and Lim, 2007).

As previously noted, the DxDxDG motif shows a well conserved spatial coordination. This feature has been used to conduct a more sensitive search for functional examples of the motif using SPASM (Kleywegt, 1998; Rigden and Galperin, 2004).

A sequence based search for the DxDxDG motif would have a large database to search and provide extensive coverage. However, all the sequence based searches available would either be too selective, or too sensitive. As the DxDxDG motif in its functional form has a highly conserved spatial arrangement, a better way of identifying new examples of this motif is to search using this well-defined structure. However, the molecular structure is not available for many proteins. Therefore, this type of search may not be as comprehensive. A new type of search is needed that can identify new examples of the motif and limit the number of non-functional examples found.

2.3 Methods

2.3.1 Identification of DxDxDG motifs from the PDB using a script that runs SPASM iteratively.

SPASM was used to search for DxDxDG-motif containing proteins within the PDB. A Perl script automatically runs SPASM and collects the output data, then starts another round of SPASM. This iterative searching allows the identification of motifs that are less and less similar to the original structure, but still show a connection through a series of intermediaries.

The starting point for this analysis was the DxDxDG motif found in calmodulin from *Paramecium tetraurelia*, PDB code 1exr. The motif is found between residues 20-29 of the A chain. This is a typical EF hand protein and a good example of the DxDxDG motif.

The seed file used for SPASM is a shortened PDB file containing only the coordinates of the motif residues; in this case, just the coordinates of the three aspartic acids in the 1exr motif. The fourth binding residue has been omitted, owing to its variable relative position.

A RMS cut-off of 1Å was used by SPASM for each round of searching. SPASM was set to allow substitutions of Asn and Ser for the Asp residues at the three coordinating positions. This allows examples of the major variants that have been seen in the motif to be identified, and the appearance of novel forms of the binding domain to be found. The regex for this motif would be [DNS]-x-[DNS]-x-[DNS]. This simplified version of the motif was used to ensure that all possible candidates, and any interesting near-miss-sites, were identified. Near-miss-sites may be those that have lost metal-binding function, or could potentially gain function with minor modifications.

After the initial round of the SPASM search, the output was parsed and the structures of the hits were downloaded from the PDB. New seed files containing

the motifs from these hits were produced and individually fed back into SPASM to search against the database again. The results from these searches were then fed back into another set of searches, and the process continued until all the motifs found by SPASM had been searched once, and no new examples had been found.

At this stage, the results were checked and hits that did not contain an Asp in the first position were filtered out. SPASM allows substitutions for a particular amino acid to be specified, however, it does not allow a set of substitutions to be specified for each position. This means it was necessary to allow substitutions for Ser or Asn to be present at the first Asp position in the search, even though this has been shown to result in a non-functional motif. The PDB files of the hits were also searched for heteroatom entries. Hits that did not show anything within 2Å of at least two of the three residues in the motif were discarded.

The SCOP database (Murzin et al., 1995) was then searched, and each motif assigned a structural group. The PDB code, chain and residue numbers from the hit were used to find the SCOP classification of the part of the protein in which the motif occurs. These SCOP families each represent a different structural context for the motif and, therefore, were the basis for the following phases of the project.

All the hits were then sorted by their SCOP super-families and by the heteroatom bound. For ease of reference and processing, these results were saved into an XML format.

All stages of the iterative searching process, unless specifically stated, were automated using Perl. This automation has allowed a more thorough, reliable and easily repeatable process.

2.3.2 CLANS

As an aside from the main project, data were analysed using a program called CLANS (Frickey and Lupas, 2004). This stands for CLuster ANalysis of Sequences. In its most common usage, this takes a set of sequences and works out an E-value for their similarity using a pair-wise BLAST search, where each sequence is compared to every other in the set. These E-values are then used to construct a two- or three-dimensional network. This enables closely related groups to be picked out. CLANS stores these values in an intermediate matrix file. This file can be filled with other forms of data.

For the CLANS analysis, two sequence based data-sets were used in the regular pair-wise fashion: the entire sequence of the chain where the motif occurs, and the sequence of the motif with the ten residues either side of the motif. Additionally, the network-generation portion of the software was used with the RMSDs between each motif.

The sequence data-sets were fed into the CLANS software in FASTA format. The RMSD values were calculated using LSQMAN (Kleywegt, 1994), and organised into the intermediate matrix file used by CLANS.

2.3.3 Perl

Perl is short for "Practical Extraction and Report Language", and was created by Larry Wall. It was originally designed to create reports from a hierarchy of files for a bug-reporting system, and grew into a general purpose programming language (Schwartz and Christiansen, 1997).

Perl, like any programming language, has certain advantages and disadvantages. It is especially useful for rapid prototyping, and many problems can be solved in far fewer lines than using C or Java. It is portable and can be run on virtually any computer. It is easy to maintain. It is very good at dealing with text strings and flat-files. This makes it particularly suited to bioinformatics, where important biological data, such as structures in the PDB, are stored as flat-files (Tisdall, 2001).

The biggest disadvantage of Perl is its run speed. The code is compiled at run time; for a large program, this would take far too long. The best language for sheer speed is C and can be up to twice as fast a Perl (Tisdall, 2001).

The PDB files, SPASM output, sequences and the SCOP database are all flat text files. This makes Perl a good choice for this project.

2.3.4 XML

XML stands for eXtensible Markup Language, and is a set of rules for encoding data in a way that is machine readable. The principle behind XML is that data should be stored in a structured, easily retrievable, standardised and unambiguous form.

The contents of an XML file are made up of markup and content: markup is anything contained within < >; content is anything else. Markup is used to give the content structure and provide a description of what each piece of data actually represents.

An XML style format is used for all the data files produced throughout this section. This paradigm of self-describing data was chosen to ensure that the output files were easily readable by both humans and computers, and the data was as flexible as possible. The data at this stage will not be accessed in a random order; most operations carried out will involve the use of Perl scripts that will perform operations on all pieces of data in the set.

The format used splits the data into three data files. The first is concerned with each individual hit (see Figure 2.3.4-1 i). Each hit has a unique ID, made up of the PDB ID, the chain ID and the range of residues the motif is found in. Information about the actual residues that are present, the SCOP ID, the binding properties and other characteristics of each hit are stored together.

Figure 2.3.4-1 Example XLM style files.

i) 'hit' file

```
<hit id=1a2x:A:139-143>
<chain>A<\chain>
<binds>CA<\binds>
<>a.39.1.5<\scop>
<res>139<\res>
<res>141<\res>
<res>143<\res>
<\hit>
<hit id=1a2x:A:139-143>
...
...
<\hit>
...
```

ii) 'families' file

```
<scopclass
id=a.39.1.5>
<id>1exr:A:20-24<\id>
<id>1a2x:A:63-67<\id>
<id>1a2x:A:103-
107<\id>
..... .....
<id>5pal:-:51-55<\id>
<id>5pal:-:90-94<\id>
<\scopclass>
.....
```

iii) 'RMS' file

```
<rms id1=1a2x:A:139-143 ID2=1exr:A:20-24>0.34<\rms>
<rms id1=1a2x:A:63-67 ID2=1exr:A:20-24>0.23<\rms>
<rms id1=5pal:-:90-94 ID2=1a2x:A:63-67>0.4<\rms>
```

The second set file (see Figure 2.3.4-1 ii) is concerned with the information associated with the SCOP family groupings. Each family found is listed, and data, such as the structural class and alignments, are stored along with a list of all the hits found that are members of the group. Each hit is represented by its unique ID and is only present in one family.

The third set (see Figure 2.3.4-1 iii) links between hits and has been implemented to reduce the redundancy in the dataset. The RMS distance of each hit was compared with all the others in the data-set and was recorded. Although not as suited to the object like style of the other files, this was necessary because if the list of RMS values for each hit was stored with it in the first file, a great amount of redundancy would occur since each comparison would be recorded twice, once for each hit involved, giving n^2-n (where n is the number of hits found) RMS values.

When it is complete, the data from the files can be parsed and processed easily. The script sortbyclass reads the SCOP classification and sorts the motifs by family. This is then used to create another XML file that lists each family and all the motifs that have been found to be part of it. It is intended that these files can be further processed to include extra data, such as alignments and secondary structural data. These data files were designed to be easily convertible into the various different formats used as input for the AI software packages.

2.3.5 "CaBindSearch" Scripting outline

One main Perl script "CaBindSearch.pl", when launched, automates the process of searching SPASM from the initial searches, to the tracking of what searches have been run, the filtering of results and the creation of the hit-based and familial-based XML files.

The Perl script takes an initial example of the motif in the form of a simplified PDB file, reduced to contain just the atoms that make up the first three residues of the motif.

The main internal data-structure of the script is an array containing one hit per array element. Each array element is made up of a hash-table containing all the data associated with the hit (see Figure 2.3.2-1). Most of the major subroutines in the script act on this data-structure. The initial step, therefore, is to read the input file and represent it in this data-structure. The **initializearray** subroutine creates a new array, places the information found in the input file into a hash, and assigns this as the first element in the **@mastertrylist**. Initially, this array only contains one motif; however, as more hits are found, the list is added to.

The **@mastertrylist** is passed onto the subroutine **integratediterativerun**; this has overall control over how SPASM is run and the results processed.

integratediterativerun extracts the initial motif from the array and passes it to the **runspasm** subroutine. **runspasm** calls **findsubs;** this sets up the appropriate substitutes used by SPASM (in the first case Asp can be substituted with Ser or Asn). These substitutions are passed, along with some user-defined arguments such as the RMS cutoff, to a shell-script that actually executes SPASM. The SPASM result is saved as a file and placed in an array for parsing.

Figure 2.3.5-1 Diagram of the main data structure in the CaBindSearch script

0	motifid	pdbid	iteration	run	chain	binds	resnum	restype
	1exr:A: 20-24	1exr	0	null	A	Ca	20 22 24	ASP ASP ASP
1	motifid	pdbid	iteration	run	chain	binds	resnum	restype
	1alv:A: 180-184	1alv	0	1exr	A	Ca	180 182 184	ASP ASP ASP
n	....							

The main data-structure in the script consists of an array, each element of which contains a hash-table that contains the data associated with each hit. As can be seen, each hash contains a number of elements: pdbid is the PDB code; iteration is the first iteration of the search the hit appeared in; run is the seed that first resulted in this hit; chain is the chain the motif is present in; binds is used to store the atom that the motif binds to; resnum is an array of the residue numbers for the hit; restype is an array of the residue type for the hit.

The results array is parsed by **parsespasm** and the results from SPASM are added to the **@currenthitarray**. At the same time, the PDB files associated with each hit are downloaded, and a seed file for each new motif is created.

The **addresults** subroutine then takes the **@newresultsarray** and filters out any hits that do not bind anything, that don't have Asp in the first residue position, or that don't have a 4th residue close enough to be involved in metal-binding. These results are then saved to file and added to the **@mastertrylist**.

This **@mastertrylist** is then fed back into the **intergrateditterativerun** subroutine and the process starts again. Each time a search is completed, a record is made of the motif ID in an array. This array gets checked by **intergrateditterativerun** each time it is called, to ensure that the same search is not repeated.

2.3.6 CLANS Scripting outline

The rms.pl script is used to create the files containing the RMSD calculations to be used with CLANS. First, the XML files containing the lists of hits from SPASM are read by the **parsexml** subroutine, and the RMSD values for each hit against each other hit are calculated by **makermsfile** using LSQMAN (Kleywegt, 1994) . These data are then saved in another XML file by **saverms.** The family data are obtained from SCOP parsable files by **addscopclass,** which first searches for the SCOP unique identifiers (sun ID) and then the corresponding SCOP identification. **sortbyclass** then generates a list of families and identifies hits that belong to each family; these are then saved in a third XML file. The appropriate input file for CLANS is then generated using the three XML files, one containing hit data, one containing RMSD calculations, and the other relating the hits to their SCOP families.

2.4 Results and Discussions

2.4.1 Motif search

Table 2.4.1-1 – Results from SPASM and filters

SCOP super-family	Bound Metal Ca	Mg	Zn	H_2O	Other	Total	Example PDB code:Chain :Residues	Additional Ligating Residue	Used in Training
True Positives									
a.39.1	83			5	3	**91**	1exr:A:20-24	7	Yes
a.139.1	1					**1**	Calmodulin 1ohz:B:2-6	7	Yes
b.30.5	1					**1**	Dockerin 1k1x:A:392-396	4	Yes
b.69.8	6					**6**	glucanotransferase 1txv:A:297-301	2	Yes
b.80.7	1					**1**	Integrin 1kap:P:446-450	4	Yes
c.1.8	2			3		**5**	Alkaline Protease 1lwj:A:13-17	4	Yes
c.62.1		2		9		**11**	glucanotransferase 2b2x:A:154-158	-	Yes
c.94.1	2		1		1	**4**	RdeltaH I-domain 1j1n:A:171-175	4	Yes
d.3.1	1	1				**2**	Alginate Binding 1vjj:A:301-305	4	Yes
d.92.1	2					**2**	Transglutaminase 3 1g9k:A:49-53	4	Yes
g.75.1	5					**5**	Alkaline Protease 1ux6:A:828-832	7	Yes
k.21.1	4				4	**8**	Thrombospondin 1vrk:A:129-133	7	No
not found	20	2	2	16	3	**43**	Calmodulin		No
False Negatives									
c.84.1		5	1			**6**	2fkm:X:242-246	-	Yes
c.93.1	2			2		**4**	α-d-glucose 1,6-bisphosphate 1gca:-:134-138	-	Yes
f.11.1	2					**2**	Glucose/Galactose Receptor 1acc:-:177-181	4	Yes
							Anthrax Protective Antigen		
False Positives									
b.29.1				1		**1**	1mvq:A:149-153	-	No
b.68.6				1		**1**	lectin 1pjx:A:26-30	-	No
b.152.1				1		**1**	DFPase 1wlg:A:102-106	-	No
c.1.11				1		**1**	FlgE31 2gl5:A:175-179	-	No
							Dehydratase		

As seen previously, the secondary structural elements that surround the motif vary markedly from the typical helix-loop-helix structure of the EF hand,

seen in calmodulin (1exr) and other all-alpha proteins (see Figure 2.4.1-1).

Figure 2.4.1-1, All-alpha proteins.

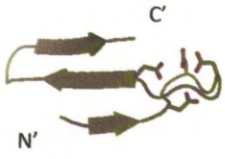
i) a.39.1 (1exr)

ii) a.139.1 (1ohz)

This includes the classic EF hand protein, as can be seen in i), along with variations of different helix and loop length, and even the absence of the E loop as seen in ii).

The motif may be within antiparallel strands of beta sheet, as seen in 4-alpha-glucanotransferase (1k1x) and the other all-beta proteins (see Figure 2.4.1-2.)

Figure 2.4.1-2, All- beta proteins.



i) b.30.5 (1k1x)

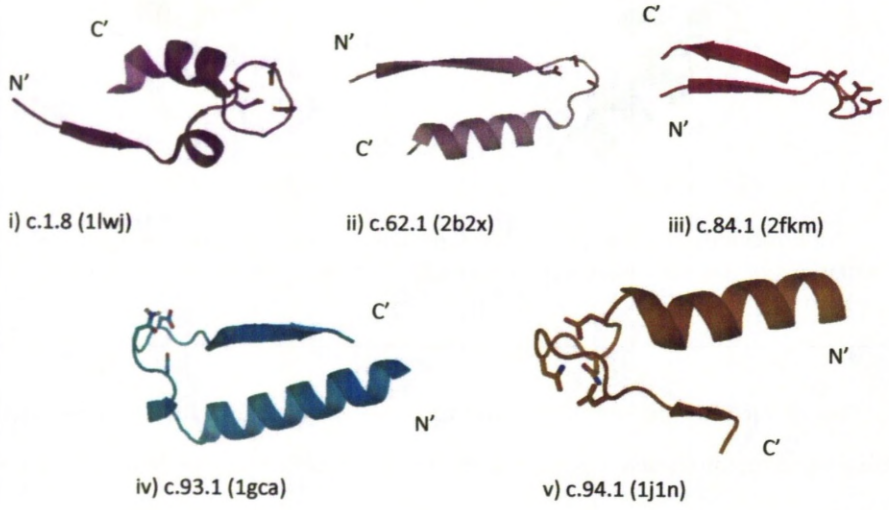
ii) b.69.8 (1txv)

iii) b.80.7 (1kap)

Motifs found in the all-beta classification predominantly seem to be found between two strands of an anti-parallel beta sheet. The lengths of these beta strands, and the number of strands involved in the sheet, vary with each case shown here. The position of the fourth residue involved in bonding varies between the 4th, 7th and 8th residue following the motif.

The motifs can also be seen between an element of alpha helix and beta sheet, as seen in the alpha and beta class of proteins (see Figure 2.4.1-3 and Figure 2.4.1-4).

Figure 2.4.1-3, Alpha and beta proteins (a/b).



i) c.1.8 (1lwj) ii) c.62.1 (2b2x) iii) c.84.1 (2fkm)

iv) c.93.1 (1gca) v) c.94.1 (1j1n)

The alpha and beta proteins (a/b) show a great deal of variation in the secondary structure surrounding the motif. Most examples seem to show a helix and a beta strand flanking the motif; however, the order varies. For example, in ii) 2b2x and iv) 1gca. Also, there is at least one example, iii) 2fkm, that looks more similar to an all-beta case, with the motif loop between two anti-parallel beta strands.

When the motif is seen between alpha helix and beta sheet, the order of the secondary structural elements may vary: the periplasmic glucose/galactose receptor from *Salmonella typhimurium* (1gca) has the alpha helix followed by the binding loop, and then a beta strand, whereas 4-alpha-glucanotransferase from *T. maritima(1lwj)* has a beta strand followed by the binding loop, and then an alpha helix. There are also examples of the rare, and tighter, 3_{10} helix (1kwh 161-184).

Figure 2.4.1-4, Alpha plus beta proteins (a+b).

i) d.3.1 (1vjj)

ii) d.92.1 (1g9k)

The alpha plus beta proteins (a+b) show very little conformity in the secondary structure surrounding the motif. i) shows beta strand either side of the motif; however, they do not seem to associate to form a sheet, as seen with the all-beta examples. ii) shows a beta strand exiting from the motif; however, there is no secondary structure seen nearby before the motif.

Figure 2.4.1-5, Other SCOP Super-families.

i) f.11.1 (1acc)

ii) g.75.1 (1ux6)

The final two examples have been grouped together as they are both from classes with only one superfamily represented in these results. i) 1acc, from the membrane and cell surface proteins class, is an example of a motif that is so close to the start of the chain that only the exiting helix is present. ii) 1ux6, from the small proteins class, shows a number of instances of the motif each surrounded only by loop and no real associated secondary structure. Although members of the same class, the overall structure of these examples varies markedly.

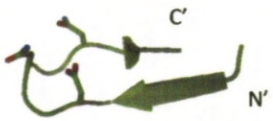
There are also some cases where only one structural element is present, or where there is more than one structural element in close proximity on the same side of the binding loop, as seen in anthrax protective antigen (1acc) (see Figure 2.4.1-4 i). Even in a protein with multiple DxDxDG motifs, there can be variation in the secondary structural elements associated with each motif. In the case of thrombospondin (1ux6), a string of DxDxDG motifs can be seen with no discernible uniformity in their secondary structural elements (see Figure 2.4.1-4 ii).

Some false-negative results were seen, owing to the extra filtering that takes place after the structural motif search. Some results have previously been identified as likely to be true DxDxDG motifs, but were discarded either because they did not have a downstream motif close enough to the ligand found at the binding site, or because there was no ligand found in the PDB file. These were checked using molecular modelling software to ensure their validity, and have been used in further analysis.

Figure 2.4.1-6, Potential false-positive results.

i) b.29.1 (1mvq)

ii) b.30.5 (1pjx)



iii) b.152.1 (1wlg)

iv) c.1.11 (2gl5)

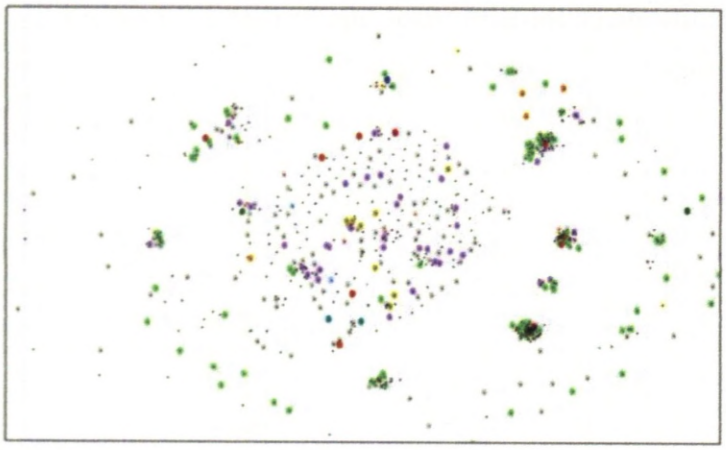
Looking at the various false-positives, it is easy to see why they were picked out by the workflow: each shows the appropriate residues in the correct spatial position; however, each also has the residues in a poor orientation for binding, mostly facing out of the loop. Each of these also does not have any conserved downstream residues to complete the binding sphere.

Similarity, a number of possible false-positive results can be seen to not fit the correct configuration of a calcium-binding motif. Although the residues form the correct spatial pattern, their orientation would not allow for proper binding (see Figure 2.4.1-6). However, it is possible that these proteins may undergo conformational changes that allow them to reorient their binding residues and take part in binding. Additionally, it is known that, in some cases, potential binding proteins have been discarded owing to the lack of additional

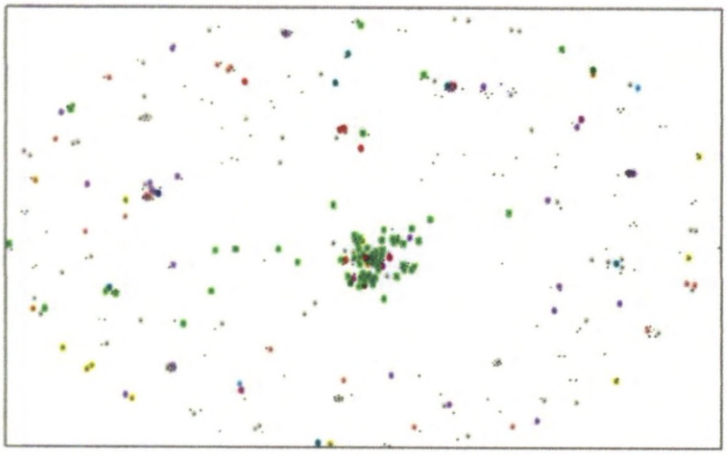
downstream ligating amino acid. However, it has been proposed that, in certain cases, this extra ligating D or E might not be present and the site still remain functional (Rigden and Galperin, 2004). However, because of these ambiguities, unless previously identified, both these sets of proteins were left out of the subsequent steps.

2.4.2 CLANS Analysis

Figure 2.4.2-1, Diagrams produced using CLANS software showing how the hits were obtained using SPASM



i) Results clustered by local sequence similarity. There is a very large, spread out cluster in the centre; this is mainly made up of smaller clusters of different families. The EF hand hits form a number of smaller, tightly packed clusters. Other super-families appear in clusters or are spread across a ring surrounding the rest of the results.



ii) Results clustered by local structural similarity. Most of the hits form a large cluster in the centre; this is mainly made up of the most common hit, the EF hand domain. Other super-families are more evenly spread, with some tightly packed clusters of mixed families appearing across the diagram.

The charts produced with CLANS are difficult to interpret (see Figure 2.4.2-1). Clustering in the sequence-derived diagram shows that the different protein super-families cluster together. This suggests that, despite a lack of homology,

local sequence similarities occur. Additionally, the grouping of proteins from the same superfamily shows that there are variations in their sequence, but they are still detectably related to other members of that superfamily. The diagram that is derived from the RMSD calculations using the position of the residues of the motif, essentially gives an indication of similarity in structure at the motif. There is clustering shown here, but it does not seem to follow division by superfamily as closely as in the sequence diagram. This suggests that, at a structural level, there is a greater convergent pressure than for the sequence. There appears to be some degree of flexibility in the general configuration at the motif, shown by the large central cluster, along with a number of additional much more tightly restricted configurations seen around the central mass. This would hold with the theory that the motif has evolved independently a number of times (Rigden and Galperin, 2004).

2.5 Section Conclusions

Extensive research has been done into ways of identifying and recording functional units in proteins, both as domains and as linear motifs. The linear motif is a more difficult prospect for identification because of its short sequence length.

The SPASM searches and subsequent filtering steps successfully found around 200 individual occurrences of a DxDxDG motif in the correct spatial orientation for the binding of calcium and other metals; 91 of these are, in fact, in the EF hand configuration. SCOP has been used to classify these 91 hits into 14 superfamily groups that show the type of binding we are looking for, as shown in table 2.4.1-1. This is in line with the research already carried out (Rigden et al., 2003)(Rigden et al., 2003). The false-positives and those with no superfamily classification were excluded, as were the members of k.21.1. This was because this is a group of designed proteins. Although these non-naturally occurring proteins may not have had any negative effect on this study, it seemed safer to exclude them. Additionally, the example found was also a calmodulin, which is a group already well represented in the training data.

It was a disappointment to not find any new families displaying the DxDxDG motif using the iterative search methods described here or the more up-to-date June 2010 version of the PDB and significantly larger database of protein structures. This might be because of the highly conserved nature of the binding motif. The search was designed to seek out a set of less closely related structures with greater variation in residue position, in a similar manner to an iterative BLAST search seeking out distant homologues. The highly conserved nature of this motif's structure alludes to this particular spatial arrangement being necessary for binding, and therefore any distortion may lead to a loss of function and loss of conservational pressure. This could be why a search for more distant relatives proved fruitless, as they simply are not viable calcium-binding motifs and lose the structure entirely, and the previous work had already identified all

the most likely variations of the DxDxDG motif currently contained within the PDB. The PDB, however, is a relatively small database, with a bias towards certain types of protein structure. The work and expense involved in structural studies means that proteins of some previously identified function, that also happen to be easily over expressed and easily solubilised or crystallised, are more likely to been fully defined. It is difficult to surmise why the expansion of the database, since the initial investigation, has not provided any extra DxDxDG families. It may be that all the DxDxDG-containing families have already been identified. Potential new examples may also have no well-defined function and, as such, have not been prioritised for structural study. This means that many potential new calcium-binding families may have been overlooked. This project aims to allow new potential calcium-binding proteins to be highlighted without the need for costly, time-consuming lab-based structural studies.

2.6 Section Bibliography

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

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. *215,* 403.

Attwood, T.K. (2009). The PRINTS user guide.

Attwood, T.K. (2002). The PRINTS database: A resource for identification of protein families. Briefings in Bioinformatics *3,* 252.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L.L. (2002). The Pfam protein families database. Nucleic Acids Research *30,* 276.

Berman, H.M., Henrick, K., and Nakamura, H. (2008). Atomic Coordinate Entry Format.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Research 235.

Davey, N.E., Edwards, R.J., and Shields, D.C. (2007). The SLiMDisc server: short, linear motif discovery in proteins. Nucleic Acids Research Web Server Issue *35,* 455.

de Castro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Research Web Server Issue *34,* 362.

Dudev, M., and Lim, C. (2007). Discovering structural motifs using a structural alphabet:Application to magnesium-binding sites. BMC Bioinformatics *8,*

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., *et al.* (2008). The Pfam protein families database. Nucleic Acids Research Database Issue *36,* 281.

Frickey, T. (2007). CLANS.*2009,*

Frickey, T., and Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics *20,* 3702-3704.

Gould, C.M., Diella, F., Via, A., Puntervoll, P., Gemünd, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J.C., Chica, C., *et al.* (2009). ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Research Advance Access 1.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Nat. Acad. Sci. USA 10915.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S., and Sigrist, C.J.A. (2007). The 20 years of PROSITE. Nucleic Acids Research Advance Access

Kawasaki, H., and Kretsinger, R.H. (1995). *Calcium-binding proteins 1: EF-hands.*. Protein Prof. 305.

Kleywegt, G.J. (1998). SPASM.*060804,*

Kleywegt, G.J. (1994). LSQMAN.*071128,*

Lo Conte, L., Brenner, S.E., Hubbard, T.J.P., Chothia, C., and Murzin, A.G. (2002). SCOP database in 2002: Refinements accommadate structural genomics. Nucleic Acids Research *30,* 264.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Research *30,* 281.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. Journal of Molecular Biology 356.

Neduva, V., and Russell, R.B. (2005). Linear motifs: Evolutionary interaction switches. FEBS Letters 3342.

Nevuda, V., Linding, R., Su-Angrand, I., Stark, A., de Masi, F., Gibson, T.J., Lewis, J., Serrano, L., and Russell, R.B. (2005). Systematic Discovery of New Recognition Peptides Mediating Protein Interaction Networks. PLOS Biology *3,* 2090.

PROSITE. (2009). Prosite, Database of protein domains, families and functional sites.*2009,*

Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M.A., Ausiello, G., Brannetti, B., Costantini, A., *et al.* (2003). ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Research *31,* 3625.

Rigden, D.J., Jedrzejas, M.J., Moroz, O.V., and Galperin, M.Y. (2003). Structural diversity of calcium-binding proteins in bacteria: single-handed EF-hands? Trends in Microbiology *11,* 295.

Rigden, D.J., and Galperin, M.Y. (2004). The DXDXDG Motif for Calcium Binding: Multiple Structural Contexts and Implications for Evolution. J. Mol. Biol. *343,* 971-984.

Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., Green, R.K., Goodsell, D.S., Prlić, A., Quesada, M., *et al.* (2012). The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Research *41,* D475.

Rossmann, M.G., and Argos, P. (1981). Protein Folding. Annual Review *50,* 497.

Schwartz, R.L., and Christiansen, T. (1997). Learning Perl (Sebastopol, CA: O'Reilly).

Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002). *PROSITE: a documented database using patterns and profiles as motif descriptors.* Briefings in Bioinformatics 265.

Tisdall, J. (2001). Beginning Perl for Bioinformatics (Sebastopol, CA: O'Reilly).

Torrance, J.W., MacArthur, M.W., and Thornton, J.M. (2007). Evolution of binding sites for zinc and calcium ions playing structural roles. Proteins Structure Function Bioinformatics 813.

Westhead, D.R., Parish, J.H., and Twyman, R.M. (2002). Bioinformatics (Oxford: BIOS Scientific Publishers Limited).

Wilson, M.A., and Brunger, A.T. (2000). The 1.0 A crystal structure of Ca(2+)-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. J. Mol. Biol. *301*, 1237.

Zhang, Z., Schäffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V., and Altschul, S.F. (1998). Protein sequence similarity searches using patterns as seeds. Nucleic Acids Research *26*, 3986.

Section 3:- Data Collection

3.1 Section Introduction

3.1.1 Section Overview

The previous section was concerned with the search for, and analysis of, proteins that bind to metals through the DxDxDG motif. This section concentrates on these selected proteins and the characteristics that confer their binding properties.

Before artificial intelligence algorithms can be used to categorise potential binding motifs, a model of the data needs to be identified that is able to describe characteristics that may influence metal binding through the motif. A number of physical and chemical properties of the amino acids surrounding the protein, which may influence binding efficacy, were identified. These characteristics include the size, hydrophobicity and chemical nature of the amino acids within and surrounding the motif, along with the closest predicted secondary structural elements, presence of a conserved aspartic acid or glutamic acid residue downstream of the motif, and the solvent accessibility of residues local to the motif.

3.1.2 Glossary and abbreviations

Φ - dihedral angle between the nitrogen and α carbon atoms of the main chain of an amino acid.

Ψ - dihedral angle between the α carbon of the main chain and the carbonyl carbon atoms of an amino acid.

Artificial intelligence (AI) – machines and algorithms that have the ability to display intelligent behaviour.

Indel – insertion or deletion of a single or multiple amino acids, this will change the length of the peptide chain and may affect protein folding.

Mview - simple software that converts BLAST results to basic alignments.

Q-Score – a measure of the similarity of two protein structures, often expressed as a %. A Q-score of 100% (or 1) means that they are exactly the same; a Q-score of 50% (or 0.5) would mean that they were less similar; 0% would indicate that they were entirely dissimilar.

SABLE – structural and solvent accessibility prediction software.

PISCES – server used to cull redundancy from a PDB dataset.

Point mutation – a substitution of a single nucleotide in a sequence, which may result in a change in amino acid coded.

PSIPRED – secondary structural prediction software.

PSSCAN – simple software used to search for a motif presented as a regex.

Ramachandran plot – plot showing the backbone conformation of amino acids in a peptide chain.

R Value – R-factor is a measure of agreement between the crystallographic model and the original X-ray diffraction data.

3.1.3 List of Diagrams

3.2 Preliminaries

3.2.1 Amino acid properties

The properties of its constituent amino acids can affect a protein in a number of ways. The major properties of an amino acid are its size, its hydrophilic or hydrophobic character, and its charge. The importance of the physical properties of the side chains comes from the effect these properties have on the folding of a protein and the interactions they facilitate with other proteins and molecules (Stryer, 1995).

Table 3.2.1-1, Amino acid properties

Amino Acid	3-letter code	1-letter code	Hydrophobicity*	Molecular Weight**
Alanine	Ala	A	1.8	89
Arginine	Arg	R	-4.5	174
Asparagine	Asn	N	-3.5	132
Aspartic acid	Asp	D	-3.5	133
Cysteine	Cys	C	2.5	121
Glutamine	Gln	Q	-3.5	147
Glutamic acid	Glu	E	-3.5	146
Glycine	Gly	G	-0.4	75
Histidine	His	H	-3.2	155
Isoleucine	Ile	I	4.5	131
Leucine	Leu	L	3.8	131
Lysine	Lys	K	-3.9	146
Methionine	Met	M	1.9	149
Phenylalanine	Phe	F	2.8	165
Proline	Pro	P	-1.6	115
Serine	Ser	S	-0.8	105
Threonine	Thr	T	-0.7	119
Tryptophan	Trp	W	-0.9	204
Tyrosine	Tyr	Y	-1.3	181
Valine	Val	V	4.2	117

In the group column here, amino acids are grouped according to their charge and polar or non-polar character, with the smallest, glycine, in a group by itself. * (Kyte and Doolittle, 1982), **(Stryer, 1995)

Amino acids vary greatly in size, from glycine that has no side chain and a molecular weight of 75, all the way up to tryptophan, with a molecular weight of 204 (see Table 3.2.1-1). Substitutions of small with large amino acids may be

detrimental to protein folding; a large amino acid at certain positions may interfere with its packing and secondary structure. The size of an amino acid may also be important when considering the properties of a binding site. If the amino acid adjacent to one involved in binding is too large, it may restrict the access of the metal ion to the binding site (Stryer, 1995).

Generally, the distribution of hydrophilic and hydrophobic amino acids dictates the tertiary structure of proteins, with hydrophobic residues tending to group within the core of a globular protein, at transmembrane regions or patches involved in quaternary interactions between chains. If a hydrophobic residue is substituted with a hydrophilic one, or *vice versa*, this may affect the local folding of the protein (Stryer, 1995).

There are a number of different scales for the hydropathy of amino acids, each giving different orders and being constructed in a different manner. The major ways of measuring hydrophobicity are either through the assessment of the physiochemical properties of the amino acids, or by looking at their distribution in proteins of known 3D structure. Both of these methods have their limitations however; the physiochemical properties of amino acids do not always match to their actual distribution in a protein, and the range of 3D structures available is limited, in part, by the solubility of the protein (Kyte and Doolittle, 1982).

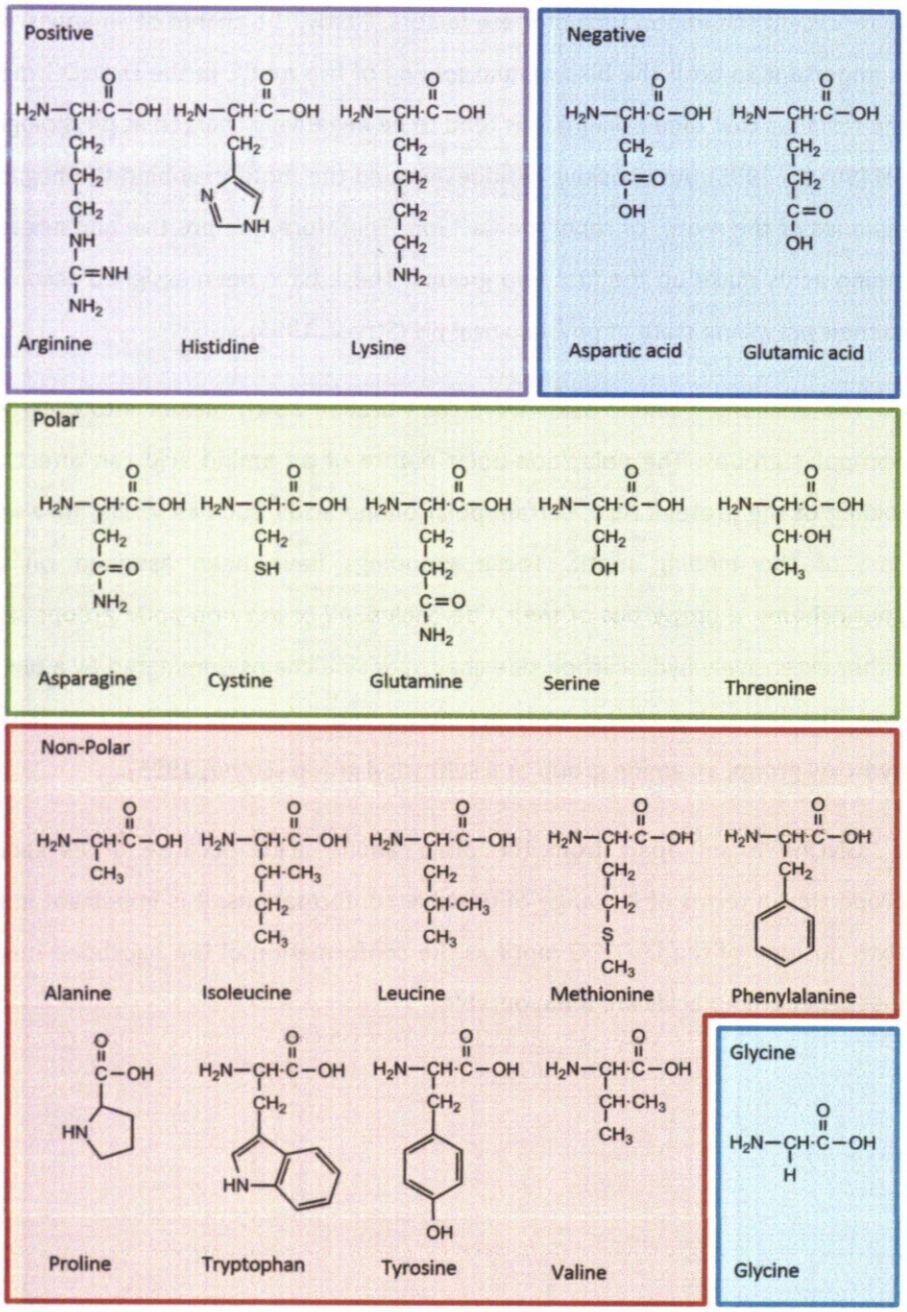
The Kyte and Doolittle scale refines a simple physiochemical property scale by introducing extra information from actual protein sequence. The hydrophobicity of amino acids is first assigned to a sequence. The average hydrophobicity of progressive segments along the sequence is then calculated. The resulting profiles can then be used to adjust the hydrophobicity for each amino acid (Kyte and Doolittle, 1982). The Kyte and Doolittle scale of hydropathy has been widely adopted and can assist in the prediction of protein structure. For example, it is useful in prediction of transmembrane and surface regions in proteins (Kyte and Doolittle, 1982).

The groupings of the amino acids, as seen in Table 3.2.1-1, have been chosen to represent the importance of these factors. Firstly, the charge of an amino acid is important to both the binding and folding of the motif. In the DxDxDG motif, the residues that bind calcium ions tend to be negatively charged at physiological pH (Stryer, 1995). Any positive residues around the motif may bind the negative residues of the motif, or repel the Ca^{2+} ion. Therefore, the positive and negative amino acids make up the first two groups. These have been assigned according to their prevailing state at physiological pH (Stryer, 1995).

The uncharged amino acids were then broken down further into polar and non-polar groups. The polar/non-polar nature of an amino acid can affect the folding of the protein. Also, certain polar amino acids, such as serine, may form part of the binding motif. These groupings have been assigned on the physiochemical properties of their side chains. All of the non-polar groups show either extensively hydrocarbon side chains, or side chains dominated by a phenyl ring. The remaining amino acids are polar, and have side chains that have a hydroxyl group, an amide group or a sulfhydryl group (Stryer, 1995).

Glycine is set apart from the other amino acids because of its special properties in terms of its range of possible conformations. It is important in the sixth position of the DxDxDG motif as the conformation of the backbone means that other amino acids are unfavourable.

Figure 3.2.1-1, Amino acid groupings.



Amino acids were grouped by their physiochemical properties: positive charge, negative charge, polar, non-polar and, glycine, by itself.

(Adapted from Biochemistry (Stryer, 1995))

3.2.2 Downstream conserved residue

As we have seen, Ca^{2+} binding in a typical EF hand involves a fourth glutamic acid or aspartic acid residue, in addition to the three that contribute to the DxDxDG motif (Wilson and Brunger, 2000). This extra binding component is about three residues from the start of the exiting α helix, and completes the negatively charged shell that surrounds and binds the Ca^{2+} ion (Wilson and Brunger, 2000). Either a conserved aspartic acid or glutamic acid is also seen in all the examples of functional DxDxDG-type binding motifs currently found, at various positions downstream of the motif (Rigden and Galperin, 2004). The presence of this conserved residue may be a useful indicator of the likelihood that a particular DxDxDG motif is functional (Rigden and Galperin, 2004).

3.2.3 Secondary Structure prediction

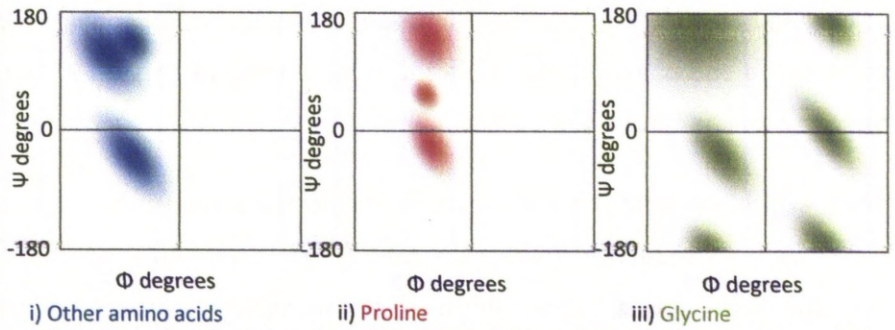
The secondary structure of a protein is how its amino acid backbone is arranged at a local level. The common secondary structural arrangements are alpha helix and beta sheet; the chain may also form less regular loop structures (Stryer, 1995).

The secondary structure is influenced, in part, by the particular residues present in the sequence. Glycine and proline are unique in their structural influence. The conformations that an amino acid can take are influenced by the rotation between the nitrogen and α carbon atoms of the main chain, Φ, and the rotation between the α carbon and the carbonyl carbon atoms, Ψ (Stryer, 1995). Glycine is a particularly flexible amino acid, owing to its lack of side chain. This lack of a side chain gives it a larger range of energetically favourable dihedral angles, Φ and Ψ, which can be seen in a Ramachandran plot as a larger area for glycine (Stryer, 1995) (see Figure 3.2.4-1).

Proline's side chain is unlike all the other amino acids, as it forms a loop with the amino group present in the backbone. This essentially acts to lock the Φ dihedral angle at around -65°. In globular proteins, proline acts as a disruptor to regular secondary structure. Its limited movement is not compatible with an α-helix or β-sheet arrangement (Stryer, 1995). The smaller area in the Ramachandran plot reflects this restricted movement (see Figure 3.2.4-1). Proline is, however, common at the fringes of secondary structural features, often seen pinning the start of an α-helix in place, in the edge strands of β-sheets, or as part of a turn in the peptide chain. Proline is also often seen in transmembrane regions, increasing structural stability (Perálvarez-Marín et al., 2008) or introducing a kink into the helices of channel proteins (Cordes et al., 2002).

The other amino acids all give approximately the same pattern on a typical Ramachandran plot, showing very similar variances in their dihedral bond angles (see Figure 3.2.4-1).

Figure 3.2.4-1, Schematic Ramachandran plots for globular proteins



i) Other amino acids ii) Proline iii) Glycine

These Ramachandran plots show the possible conformations of glycine, proline and the other amino acids. As can be seen here by the larger shaded areas, glycine is much more flexible than any other amino acid. Conversely, proline displays fewer possible combinations of dihedral bond angles (Adapted from Biochemistry (Stryer 1995)).

Despite not showing a great deal of difference in their flexibility, amino acids do, however, seem to show a certain propensity to particular structural states. However, none of these biases are particularly strong, and each amino acid is commonly seen in each type of secondary structural feature.

It would seem that the amino acid present at a particular position therefore only provides limited scope for secondary structural prediction. However, a string of amino acids with a particular structural bias may indicate a likely structure. In isolation, however, this type of prediction has been shown to be of limited predictive power (Jones, 1999).

State-of-the-art prediction methods utilise multiple alignments to add evolutionary information, and significantly improve predictive accuracy. If, in an alignment of similar sequences, the propensity for a particular secondary structural state is conserved, extra evidence is provided for that state being correct (Jones, 1999). Other patterns may also be seen in alignments: α-helices in globular proteins often have a hydrophobic and hydrophilic face; therefore, a

periodic pattern of hydrophilic and hydrophobic amino acids is often indicative of an α-helix.

Insertions, deletions and point mutations can also be seen in alignments. These assist in prediction, by contrasting well conserved regions. These indels and point mutations can be disruptive to secondary structure, leading to misfolding and possible loss of protein function. Owing to this potential loss of function, indels and point mutations are much more likely to be seen in loops rather than other features (Jones, 1999).

Many algorithms aim to try and accurately predict the secondary structure of a protein using its sequence. There have been several methodologies used to achieve this with varying degrees of success. Probabilistic methods used in algorithms, such as the Chou–Fasman method (Chou and Fasman, 1987), show a relatively low accuracy of around 50-60%. Bayesian inference is used in the GOR method (Garnier et al., 1996) and is around 65% accurate. More modern methods achieve accuracies in the 70-80% range and utilise machine learning methods, such as support vector machines, used in YASSPP (Karypis, 2006), and neural networks, used in PSI-PRED (Jones, 1999) and Jpred (Cole et al., 2008).

The method used to obtain secondary structural predictions in this project was PSI-PRED. PSI-PRED uses two feed-forward neural networks to provide a prediction of protein secondary structure with an average Q score of 81.2%. This method was chosen because it is fast, gives accurate results and has been used in similar studies previously. Additionally, the software is easily available, flexible and familiar.

The PSI-PRED process can be divided into three stages: the creation of a sequence profile, the actual secondary structural prediction, and the refinement of this prediction.

The initial profile generation stage takes advantage of the PSI-BLAST search process. As discussed previously, PSI-BLAST performs a multiple alignment and

creates profiles to use in an iterative search. PSI-PRED is designed to use these profiles. This allows the time-consuming multiple alignment to be eliminated. For PSI-PRED to be effective, the sensitivity of the profiles produced is especially important. Therefore, a custom database was used with PSI-BLAST. This database had regions with low information content, transmembrane segments, and regions that form coiled coils removed.

The prediction is then carried out using a standard feed-forward, back-propagation network architecture. A window of 15 rows is used for the prediction. The refinement step is then carried out by a second neural network that filters the output from the main network (Buchan et al., 2010).

3.2.4 Relative solvent accessibility prediction

Solvent accessibility is the area of a residue exposed to solvent, and therefore available for binding interactions. The relative solvent accessibility is expressed as a percentage; the solvent-exposed area of a residue is divided by the maximum possible solvent-exposed area of a residue, and multiplied by 100. A residue with a relative solvent accessibility of 100% would be maximally exposed, and one with 0%, fully buried (see Figure 3.2.4-1) (Adamczak et al., 2004).

Figure 3.2.4-1, Equation for relative solvent accessibility

The RSA of an amino acid residue, i, which will be denoted as RSA_i, is defined as the ratio of the solvent-exposed surface area of that residue observed in a given structure, denoted as SA_i, and the maximum obtainable value of the solvent-exposed surface area for this amino acid, denoted as MSA_i

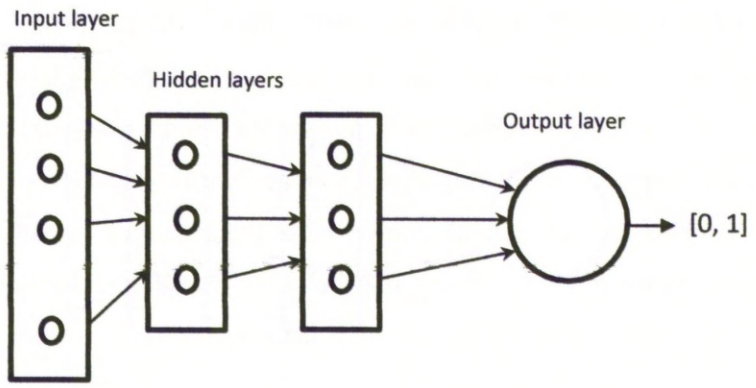
$$RSA_i = 100 \cdot \frac{SA_i}{MSA_i}[\%].$$

Thus, RSA_i adopts values between 0% and 100%, with 0% corresponding to a fully buried and 100% to a fully accessible residue (Adamczak et al., 2004).

Prediction of the relative solvent accessibility can be represented as a classification problem, where a threshold on a variable or set of variables is used to predict if the residue is buried or exposed. Many models have been proposed to attempt to solve this classification problem, including neural networks (Rost and Sander, 1994), Bayesian (Thompson and Goldstein, 1996), substitution-matrix-based and simple baseline approaches (Richardson and Barlow, 1999).

The SABLE software uses a combination of neural network and regression-analysis methods, and shows an increased amount of success, increasing prediction accuracy to 77% (Adamczak et al., 2004). SABLE rates solvent accessibility on a score from one to five, one being the least exposed, five the most (Adamczak et al., 2005).

Figure 3.2.4-2, Schematic diagram of a feed-forward neural network



Schematic representation of a multilayer feed-forward, with a single, real-value (logistic) output node, used here for regression-based prediction of RSA. Subsequent layers are fully interconnected (i.e., each node of a given layer is connected to all nodes in the next layer).

The training data used in the development of the SABLE software was based on the PFAM database; randomly chosen representative structures for each family were obtained. These training data were combined with PSSMs obtained using PSI-BLAST, and used in the training of nine neural network architectures (see Figure 3.2.4-2). These networks were then combined to give a consensus regression-based continuous predictor (Adamczak et al., 2004).

3.3 Methods

3.3.1 Pattern hit initiated BLAST and simple alignments

In the DxDxDG example, we know that there are likely to be a large number of sequences that are not represented in our SPASM results, owing to the limited set of proteins present in the PDB. BLAST searches were used to attempt to pick out these sequences. It is possible that some of the sequences picked out may not be true calcium-binding proteins; the use of PHI-BLAST should help minimise these non-calcium-binding examples in each alignment.

PHI-BLAST (Altschul et al., 1990) was used to obtain related sequences to the representative for each family. This performs a BLAST search using the identified SPASM hit, with the additional restriction that the results must contain the DxDxDG motif. PHIBLAST is given the motif pattern in a file, expressed in PROSITE format as D-X-[DNS]-X-[DNS]-X (Sigrist et al., 2002). Additionally, the exact position in which the motif should be found was also specified.

A small software package called mview (Brown et al., 1998) was used to reformat the BLAST output and produce a simple alignment of the results of this BLAST search.

3.3.2 Amino Acid properties

Amino acids with particular chemical and physical properties may prevent proper folding or binding of the motif. The amino acid composition of each residue within the motif, and four residues either side of the motif, was determined. The amino acids were split up into five groups based on their characteristics: polar, non-polar, positively charged, negatively charged and glycine. At each residue position, the % representation of each group down the alignment was recorded. The amino acids were also grouped by size and hydrophobicity, each possible large/small and hydrophobic/hydrophilic threshold was considered, and the % representation at each position was recorded. The size and hydrophobicity of the amino acid present in the original hit was also recorded.

3.3.3 Conserved downstream D/E

A downstream Asp or Glu almost invariably interacts with the metal ion during DxDxDG-metal binding; as we have seen previously, this completes the negatively charged shell that surrounds the calcium ion. However, it is clear that this interacting residue can be a varying sequential distance from the motif itself, dependent on the local folding of the structure (Rigden and Galperin, 2004). The alignments produced by PHI-BLAST were searched for residues that showed a combined percentage of Asp and Glu greater than 50%. The distance from the motif, and the percentage of Asp and Glu at each of these positions, was recorded.

3.3.4 PSIPRED- Secondary structural prediction

Despite the varying structural contexts, it is possible that the secondary structure plays a role in holding the correct configuration of the DxDxDG motif (Rigden and Galperin, 2004). A PSIPRED secondary structural prediction was performed on the representative sequence for each of our fourteen SCOP super-families (Jones, 1999). PSIPRED first runs a PSI-BLAST search to obtain related sequences, and then uses feed-forward neural networks to predict the secondary structure. It produces a simple output, showing the predicted secondary structure (loop, helix or sheet) at each position in the sequence.

3.3.5 SABLE-solvent accessibility prediction

The accessibility that the solvent and, therefore, the metal ions have to the residues of the motif may determine if a DxDxDG motif is functional. The residues of functional motifs are more likely to be found at the surface of the protein, where they are available for interactions. SABLE (Adamczak et al., 2004; Wagner et al., 2005) was used to predict solvent accessibility. SABLE uses a neural network-based regression model for relative solvent-accessibility prediction. This analysis was performed using the non-redundant protein sequence database with entries from GenPept, Swissprot, PIR, PDF, PDB and NCBI RefSeq (obtained from NCBI June 2007), with the solvent-accessibility prediction only and wApproximator options enabled.

3.3.6 "CaBindData" Scripting outline

The process of collecting the alignments and sequence-derived data from PSI-BLAST, PSIPRED and SABLE was performed by the CaBindingData.pl script. This script calls the **getfamilydata** subroutine part of the families module. The list of positive hits is parsed, and the **phiblast**, **psipred** and **runsable** subroutines are called to run PHI-BLAST, PSIPRED and SABLE respectively. Each of these subroutines used the sequence files associated with representative examples of SCOP families from section 3. These work by parsing the family files and reformatting the data into appropriate input files, then running first, PSIBLAST, then PSIPRED, and finally, SABLE. The output files of these various software packages were then parsed by **processfamilydata**. The attributes needed for the AI analysis were then calculated from these data.

3.3.7 Negative data-set

The AI tools work by identifying patterns that can be used to tell sets of data apart. Therefore, consideration has also been put into providing a data-set of families that show a DxDxDG motif, but are not involved in metal binding. Obtaining a true-negative data-set requires some problems to be overcome. It is important that, first, the data does not show excessive redundancy, but is still representative of the database of protein sequences. Each example used needs to contain a DxDxDG motif but, also, it is important that we can ensure that no examples that could potentially display calcium-binding properties are included.

To obtain a true-negative data-set, first a culled set of sequences from the PDB was obtained using the PISCES server (Wang and Dunbrack, 2003). The PISCES server allows a subset of sequences to be culled from a given list of proteins or the PDB. This culling can be based on the quality of the structural information, or on a maximum permitted sequence identity. The full PDB database was used and culled by entry. A maximum sequence identity of 80% was used, with resolution between 0 and 3Å, and a chain length of 40 to 10,000 residues. Non X-ray and C-alpha-only entries were skipped, and a maximum R value of 0.3 was imposed, meaning those structures where the crystal structure and electron density do not show a good match are culled. This serves to remove a portion of the redundant data in the PDB.

This culled database was then searched using SCANPROSITE (Gattiker et al., 2002) for entries with the DxDxDG motif (PROSITE format D-X-[DNS]-X-[DNS]-X). The resulting list of sequences was then parsed, and put into an XML file of a similar format to those used in the SPASM portion of the project. The sequences were then assigned a SCOP family, and those in a SCOP family that had been already found in the SPASM search (after filtering for metal binding and Asp at the start of the motif) were filtered out, thus ensuring, as much as possible, that no additional false-negatives were included in the true-negative data-set. A random selection of these negative results, comparable in size to the positive set, was then selected to represent the negative data-set.

There was a chance, however, that false-negative results could be included in this set. Therefore, each of the structures selected to represent the negative data-set were checked by hand, to ensure that their motifs did not conform to the expected calcium-binding profile. The associated sequence-derived data for each of these negative examples was then obtained, using similar methods to the positive data-set.

3.3.8 Negative data-set Scripting outline

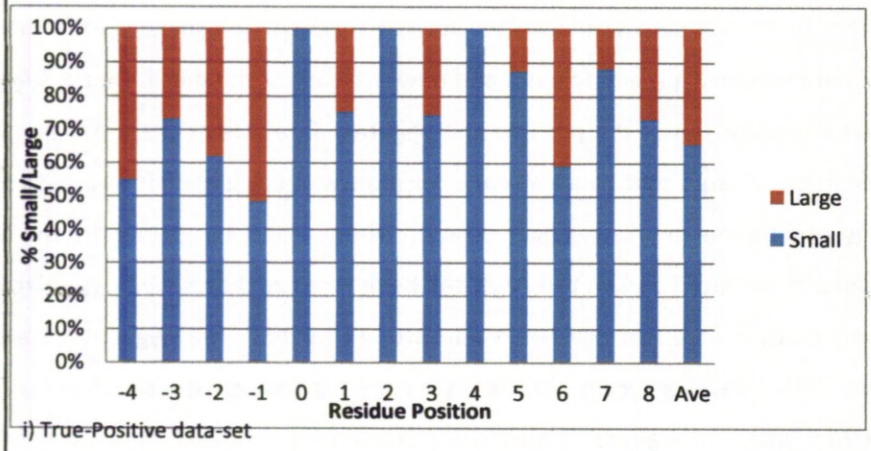
A negative data-set for comparison and training was obtained using the **createnegativeset.pl** script. As detailed above, a representative culled set of proteins was obtained using the PISCES server (Wang and Dunbrack, 2003), and then searched for the DxDxDG motif using PrositeScan (Gattiker et al., 2002). This set then needed to be processed to give a suitably sized negative data-set for the AI tools. First, the file for the culled set, the positive set and its family data were parsed using **parsepsout** and **parsexml** respectively, both called by **createnegativeset.pl**. SCOP family data were then added to the negative data-set using **addscop,** again called from **createnegativeset.pl**. These data-sets were then fed into **sortnegative,** in order to subtract the known positive data from the negative data-set. **filtersuperfamily** from the **clanssubs** module was then used to randomly pick a sample for use with the AI tools from this set.

3.4 Results and Discussions

3.4.1 Amino Acid Size

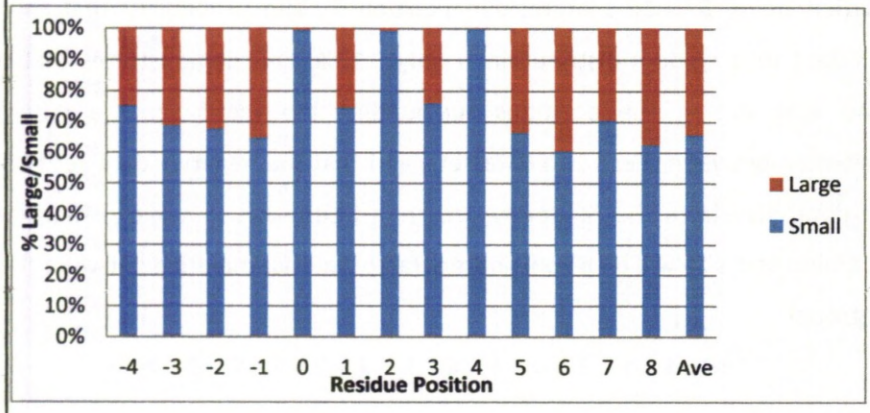
The amino acid size was collected for the residues that make up the motif, and four positions immediately up- and down-stream of the motif. Figure 3.4.1-1 shows a comparison of the positive and negative data-sets at these positions; a percentage of large and small at each position was calculated for each super-family, and the mean taken at each position. The 0, 2 and 4 positions are the Asp residues of the motif and, in this case, can be ignored as they have been selected to only contain Asp, Asn and Ser residues by PHI-BLAST. The negative data-set shows little variation from the average ratio found across all proteins, of 65.4:34.6 small to large, calculated using standard amino acid frequencies, the largest deviation being 75.7:24.3 at residue position 3. The probability that these data are equivalent according to a chi-squared test is 76.7%. The positive data-set, however, shows a greater deal of variation from normal, with the largest deviation being 87.7:12.3 at residue position 7. The probability that this is equivalent to a random distribution is only 0.17%. This demonstrates that the amino acid size at the positions surrounding the motif show a differing distribution between these two data-sets, and that the negative data-set shows little difference from the randomly occurring frequency. It is likely, therefore, that amino acid size will be a contributory factor in discriminating between these two groups.

Figure 3.4.1-1, Percentage of large and small residues at each residue position around the motif



i) True-Positive data-set

The true-positive data-set shows a great deal of variation from the distribution of large and small that would be expected, owing to a random selection of amino acids at their normal frequencies (here, shown in the Ave bar). The 0, 2 and 4 positions should be ignored, in this case, as they have been artificially selected to always show a D, N or S residue.



ii) True-Negative data-set

The true-negative data-set shows far less variation from the distribution of large and small that would be expected, owing to a random selection of amino acids at their normal frequencies (here, shown in the Ave bar). The 0, 2 and 4 positions should be ignored, in this case, as they have been artificially selected to always show a D, N or S residue.

3.4.2 Amino Acid type

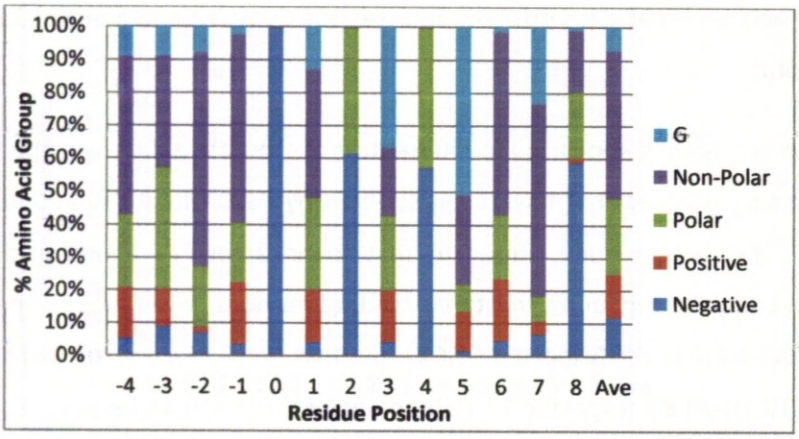
The amino acid type charts (see Figure 3.4.2-1) show the distribution of each amino acid group at each position, four residues upstream and downstream of the motif.

The two graphs appear quite different. In the negative data-set, apart from the binding residues that have been selected by PSSCAN, all residue positions have a distribution quite similar to that found across all amino acids. The negative data-set, ignoring the three binding residues, gives group ranges of: negative, 4.5%-18.3% (expected 11.6%); positive, 5.4%-19.4% (expected 13.3%); polar, 16.3%-24.8% (expected 23.1%); non-polar, 32.9%-59.4% (expected 44.8%); and glycine, 1.8% to 16.1% (expected 7.2%).

The ranges for the positive data-set, however, show a good deal of variation from the average distribution. The positive data-set, ignoring the three binding residue, gives group ranges of: negative, 1.4%-58.6% (expected 11.6%); positive, 1.4%-18.9% (expected 13.3%); polar, 7.7%-37.0% (expected 23.1%); non-polar, 18.5%-64.9% (expected 44.8%); and glycine, 1.1% to 51.0% (expected 7.2%).

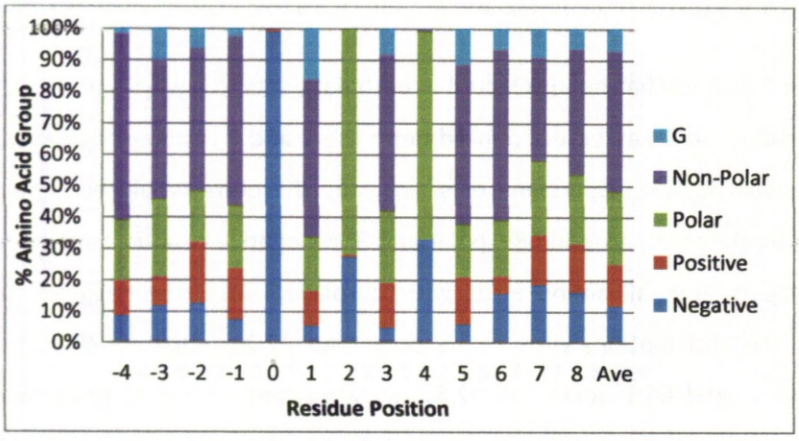
The 0 position has been selected in both data-sets to always show an aspartic acid residue, so it can be discounted here; the 2 and 4 residues, however, show either aspartic acid, asparagine or serine, and therefore would be expected to have a frequency of around 68% polar and 32% negative residues, if their relative frequencies from all amino acids were displayed. The true-negative data-set shows this distribution, with 71.7% polar and 27.2% negative side chains at position 2, and 66% polar and 32.8% negative side chains at position 4. The positive data-set shows 38% polar and 61.4% negative amino acids at position 2, and 42.2% polar and 57.2% negative amino acids at position 4. This means that there is a bias towards the aspartic acid residue at these positions. This has previously been shown in the EF hand motif (see Figure 1.4-3); however, it can now be seen across all known DxDxDG-type proteins.

Figure 3.4.2-1, Percentage of each amino acid group in the residues around the motif



i) True-positive data-set

The true-positive data-set shows variation from the distribution of amino acid groups that would be expected, owing to a random selection of amino acids at their normal frequencies (here, shown in the Ave bar). The 0, 2 and 4 positions make up the motif. The positions surrounding the motif seem to show increased non-polar character. The residues following the D residues of the motif show increased glycine.



ii) True-Negative data-set

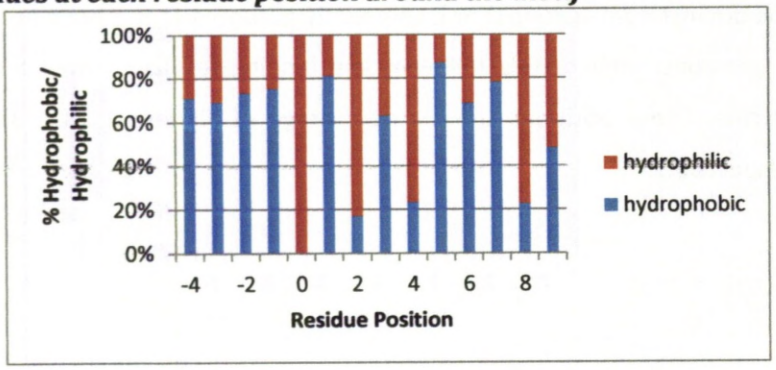
Each of the positions of the negative data-set show similarity to the distribution expected if each amino acid was seen at its normal frequency.

In the positive data-set, the -1, -2, 6 and 7 positions surrounding the motif show a bias towards the non-polar amino acids and away from the negative amino acids. This may be because the arrangement of the negatively charged binding residues so closely together is already energetically unfavourable. The 8th position also shows a bias towards the negative residues. This may be owing to the downstream ligating D or E commonly occurring at this position. The 1, 3 and 5 positions, immediately following the binding residues, show a bias towards glycine. These positions show a percentage of glycine at 13%, 37% and 51%, respectively.

3.4.3 Amino Acid Hydrophobicity

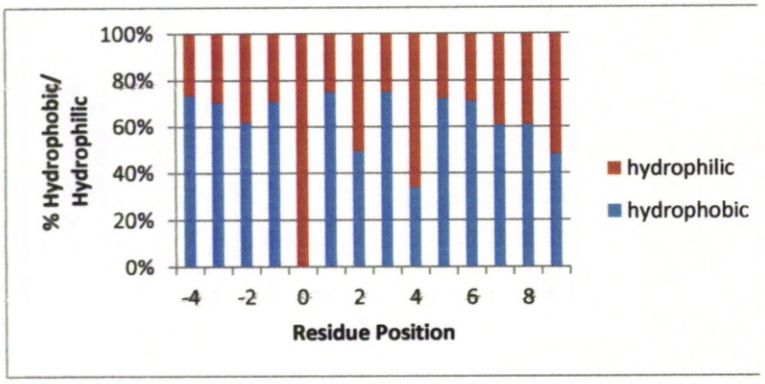
The amino acid hydrophobicity data again show a difference between the positive and negative data-sets.

Figure 3.4.3-1, The percentage of hydrophobic and hydrophilic residues at each residue position around the motif



i) True-Positive Data-set

There is a tendency towards hydrophilic residues at the 8th position. All the other positions are more hydrophobic than normally expected, especially 1, 5 and 7.



ii) True-Negative Data-set

Other than the motif residues, which we would expect to be hydrophilic, all the other positions seem to be around the 60 to 70 percent hydrophobic range.

In the positive data-set, there is a tendency towards hydrophilic residues at the 8th position; this is where the downstream D or E is often found. All the other positions are more hydrophobic than would normally be expected. The range of % hydrophobic residues is 63.1% - 87.1%, excluding the binding motif residues

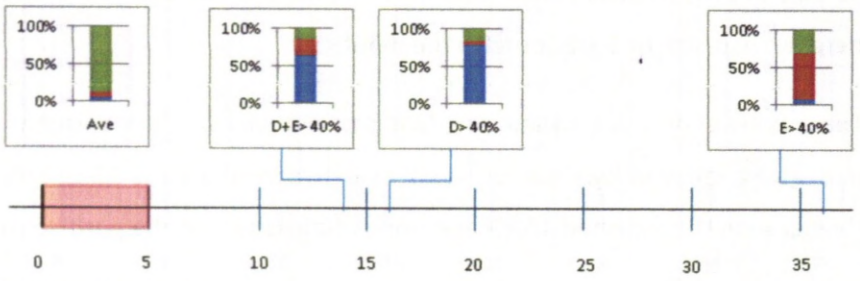
and position 8. This is higher than the % of hydrophobic amino acids seen from their average frequencies of 48.3%.

As expected, hydrophilic residues dominate at the 0, 2 and 4 positions, as this is where the Asp, Asn or Ser is found in the motif.

The negative data-set, when the binding residues and 8^{th} position are ignored, give a range of hydrophobic residues between 60.9% and 75.1%; this is also higher than the expected 48.3%, but not as high as seen in the positive data-set.

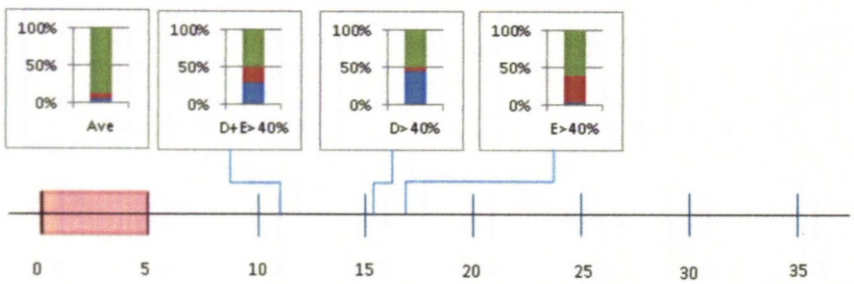
3.4.4 Conserved Downstream D/E

Figure 3.4.4-1, Conservation and distance of first D and E after motif



i) True-positive data-set

Bar charts represent the % of D, E and all other amino acids at a conserved position. The scale shows the position, averaged across the super-families, of the first conserved residue after the motif. Ave is the expected relative amino acid frequencies given no conservation. Where D+E shows >40% conservation, D is favoured. It is interesting that conservation of E seems to be less likely in close proximity to the motif.



ii) True-negative Data-set

At the first conserved residue, the charts in the negative data-set show lower % conservation than in the corresponding positive data. The D/E ratio in the D+E graph is much more evenly distributed. Also, the conserved residues seem to be much more tightly grouped than in the positive data. This may fit with a model of random distribution according to their frequencies.

Figures 3.4.4-1 i) and ii) show that, in both the positive and negative data-sets, there is a greater percentage conservation of aspartic acid than glutamic acid residues. Also, a conserved aspartic acid is more likely to occur closer to the motif than a conserved glutamic acid. The conserved Ds in the positive and negative data-sets tend to be around 10 residues from the motif; however, Es are much further away.

These observations may result from their relative frequencies. However, comparison of the positive and negative sets show a more pronounced bias towards aspartic acid in the positive set and a greater gap between the motif and the first conserved glutamic acid in the negative data-set. This may be an indication of some other pressure acting on the positive set.

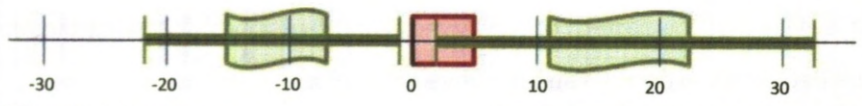
Additionally, there is a greater overall percentage conservation of these residues in the positive data-set, compared with the negative. Also, there seems to be an overall tendency for the first conserved residues to be closer to the motif. This might be explained if, in the true-positive examples, the downstream ligand function is primarily performed by aspartic acid, and any nearby glutamic acid may interfere, providing a negative pressure on its conservation. The increased size of the glutamic acid may also necessitate an increased gap, especially in cases where it is acting as the additional ligand.

It is clear that the pattern of conserved D and E residues, following the motif, differs between the positive and negative data-sets.

3.4.5 Secondary Structure

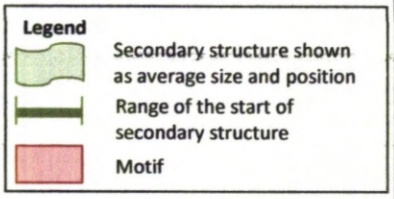
The start and end points of the closest predicted secondary structural element to the motif was recorded from each example, from both the positive and negative data-sets. The outlying values for the start and end points have been used to give a range in which the secondary structure appears. Comparison of the data-sets with regard to secondary structure in any meaningful way is difficult. However, the positive data-set does seem to show some flexibility in the distance between the motif and a secondary structural feature (see Figure 3.4.5-1). The negative data-set, however, appears to have a slightly less confined distance, and, in some cases, the predicted structure encroaches on the motif and across it.

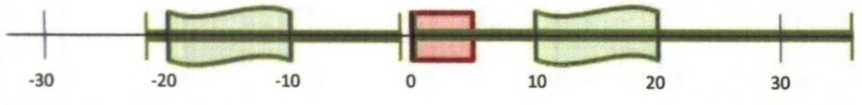
Figure 3.4.5-1, The ranges of secondary structure around the motif



i) True-Positive Data-set

The average size of an upstream structural element in the positive data-set is 8 residues and downstream is 11. As can be seen, the distance of these elements from the motif may vary greatly.





ii) True-Negative Data-set

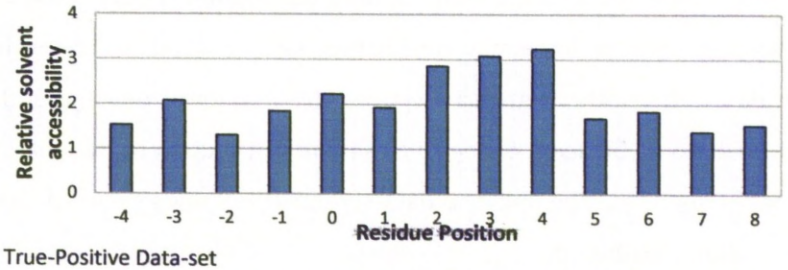
The average size of an upstream structural element in the negative data-set is 11 residues and downstream is 10. The distance from the motif also varies greatly, but downstream elements are more likely to overlap the motif.

3.4.6 Solvent Accessibility

The solvent accessibility predictions show that the area of the motif is likely to display low accessibility to the solvent compared to the protein as a whole. The SABLE software makes predictions using an arbitrary scale between 0 and 9, where 9 is fully exposed and 0 is fully buried. The positive data-set also shows less accessibility than the negative data-set locally to the motif, with around 1 to 3, compared to 3 to 4 in the negative data-set.

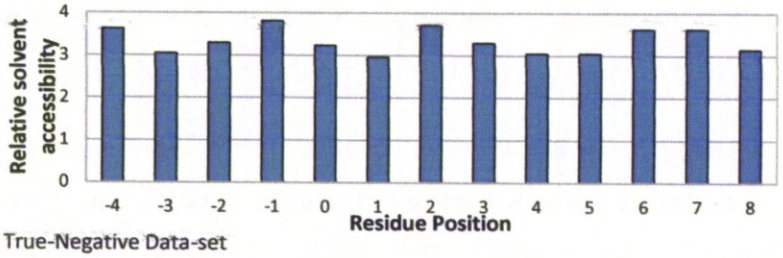
This was not the expected result, as it seems unlikely that residues that are involved in binding an ion in solution would be less accessible to the solvent.

Figure 3.4.6-1, Solvent accessibility of the residues around the motif



i) True-Positive Data-set

Of the residues shown those that make up the motif are the least buried all with a solvent accessibility over 2, the surrounding residues show less solvent accessibility. These data do suggest that all the residues are buried in comparison to their wider surroundings.



ii) True-Negative Data-set

All the residues show around the same solvent accessibility, between 3 and 4 on average, this suggests that all the residues are slightly buried in comparison to their surroundings.

However, the motif itself does show a greater solvent accessibility than its immediate surroundings, indicating that the motif itself may be exposed. The positive data-set shows a low to high to low pattern as we pass the motif (see Figure 3.4.6-1).

3.5 Section Conclusions

Large amounts of ancillary data have been collected about each of the super-families. This includes an alignment and information about the physico-chemistry of the residues surrounding the motif. A number of representatives from each super-family have been selected, and a BLAST search has been completed in order to obtain an alignment that represents the super-family. Various other software packages, such as PSIPRED and SABLE, have been used to obtain information about characteristics that may contribute to the binding at the motif.

The size of the amino acids found around the motif has been looked at, as amino acids that are too large may impede binding. The X^2 test indicates that the positive data-set deviates significantly from the frequency of large and small residues that would be seen, owing to a random distribution of the amino acids at normal frequencies. The negative data-set, however, does seem to show this type of random distribution.

The amino acid type also appears to deviate from the expected distribution in the positive set, but less so in the negative group. There especially seems to be an increase in the non-polar amino acids found immediately surrounding the motif; this could be important in differentiating the two sets of data.

The solvent accessibility is less conclusive, as the motif residues of the positive results show a greater tendency to being buried within the protein than the negative results. However, the relative solvent accessibility to the surrounding of the motif may still be useful in the discrimination of binding and non-binding examples since the positive results show a difference between the accessibility of the motif and its surrounding residues, whereas the negative data-set showed approximately equal accessibility between the motif and its surroundings.

The appearance of downstream conserved Asp and Glu residues is difficult to interpret and present. A downstream Asp or Glu is known to interact with the calcium ion at the site of binding: this can be easily seen in the 3D representations of the proteins found in the PDB. However, identifying these residues using only sequence data has proven difficult. There does seem to be some difference between the positive and negative data-set, and optimum distance ranges can be seen in the charts; however, information on the accessibility of these residues, in combination with their conservation in the alignment, might give more conclusive results.

The work done so far seems to support the supposition that it will be possible to differentiate between sequences that display the DxDxDG motif and bind calcium, and those that display the motif but have no binding function.

The project will now move into a phase of data analysis, where the data obtained will be fed into AI algorithms to try to devise an accurate and efficient way of distinguishing binding and non-binding occurrences of the DxDxDG motif.

3.6 Section Bibliography

Adamczak, R., Porollo, A., and Meller, J. (2005). Combining Prediction of Secondary Structure and Solvent Accessibility in Proteins. Proteins: Structure, Function and Bioinformatics *59*, 467.

Adamczak, R., Porollo, A., and Meller, J. (2004). Accurate Prediction of Solvent Accessibility Using Neural Networks Based Regression, Proteins: Structure, Function and Bioinformatics *56*, 753.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. *215*, 403.

Brown, N.P., Leroy, C., and Sander, C. (1998). MView: A Web compatible database search or multiple alignment viewer. Bioinformatics *14*, 380.

Buchan, D.W., Ward, S.M., Lobley, A.E., Nugent, T.C., Bryson, K., and Jones, D.T. (2010). Protein annotation and modelling servers at University College London. Nucleic Acids Research *38*, W563.

Chou, P.Y., and Fasman, G.D. (1987). Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol *47*, 45.

Cole, C., Barber, J.D., and Barton, G.J. (2008). The Jpred 3 secondary structure prediction server. Nucleic Acids Research *35*, W197.

Cordes, F.S., Bright, J.N., and Sansom, M.S. (2002). Proline-induced distortions of transmembrane helices. Journal of Molecular Biology *323*, 951.

Garnier, J., Gibrat, J.F., and Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. Methods in Enzymology *266*, 540.

Gattiker, A., de Castro, E., and Gasteiger, E. (2002). ProSite Scan.*1.67*,

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 195.

Karypis, G. (2006). YASSPP: Better kernels and coding schemes lead to improvements in protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics *64*, 575.

Kyte, J., and Doolittle, R.F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. Journal of Molecular Biology *157*, 105.

Perálvarez-Marín, A., Lórenz-Fonfría, V.A., Simón-Vázquez, R., Gomariz, M., Meseguer, I., Querol, E., Padrós, E., and Padrós, E. (2008). Influence of Proline on the Thermostability of the Active Site and Membrane Arrangement of Transmembrane Proteins. Biophysical Society Journal *95*, 4384.

Richardson, C.J., and Barlow, D.J. (1999). The bottom line for prediction of residue solvent accessibility. Protein Engineering *12,* 1051.

Rigden, D.J., and Galperin, M.Y. (2004). The DxDxDG Motif for Calcium Binding: Multiple Structural Contexts and Implications for Evolution. J. Mol. Biol. *343,* 971-984.

Rost, B., and Sander, C. (1994). Combining evolutionary information and neuralnetworks to predict protein secondary structure. Proteins *19,* 55.

Sigrist, C.J.A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002). *PROSITE: a documented database using patterns and profiles as motif descriptors.* Briefings in Bioinformatics 265.

Stryer, L. (1995). Biochemistry (New York: W.H. Freeman and Company).

Thompson, M.J., and Goldstein, R.A. (1996). Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins *25,* 38.

Wagner, M., Adamczak, R., Porollo, A., and Meller, J. (2005). Linear Regression Models for Solvent Accessibility Prediction in Protein. Journal of Computational Biology. *12,* 355.

Wang, G., and Dunbrack, R.L. (2003). PISCES: a protein sequence culling server. Bioinformatics *19,* 1589.

Wilson, M.A., and Brunger, A.T. (2000). The 1.0 A crystal structure of Ca(2+)-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. J. Mol. Biol. *301,* 1237.

Section 4:- Artificial intelligence analysis

4.1 Section Introduction

4.1.1 Section Overview

Up to this point, a set of metal-binding and non-metal-binding occurrences of the DxDxDG motif have been found. The metal-binding set was found by searching the PDB using SPASM, and filtering hits to those likely to bind metals. In a similar fashion, the non-binding set was found by searching the PDB using SCANPROSITE to find all occurrences, and filtering out those that do bind metals.

In order to quantify differences between these data-sets, and identify patterns that set them apart, a number of attributes that can be derived from the protein sequence were then identified, and their associated data collected.

The next step is to process these data, searching for patterns that enable us to separate out data from each of these sets. Decision trees and Support Vector Machines were both used for this step, and their respective usefulness and limitations in this problem have been compared.

4.1.2 Glossary and abbreviations

To avoid confusion, a slight change in terminology should be noted during this section: the representative examples from the SCOP super-families from section two, for clarity, will be referred to as cases. Additionally, the amino acid properties or characteristics, such as amino acid size and hydrophobicity, analysed in section three, will be referred to as attributes.

C4.5 – software used to construct decision trees.

Decision Trees (DT) – method of classifying data, using branching questions.

Genetic algorithms – method of solving optimisation problems using evolutionary methods.

False-positive rate – the cases that were incorrectly predicted to be positive (calcium-binding) by the AI tools, expressed as a percentage of the predicted positive data-set.

False-negative rate – the cases that were incorrectly predicted to be negative (non-calcium-binding) by the AI tools, expressed as a percentage of the predicted negative data-set.

Hidden Markov Models (HMM) – simple dynamic Bayesian network, used to predict the next outcome in a sequence.

Known positive misclassified rate – the percentage of the training data that is known to be positive (calcium-binding) and was misclassified.

Known negative misclassified rate – the percentage of the training data that is known to be negative (non-calcium-binding) and was misclassified.

MHC – Major histocompatibility proteins are cell surface molecules that mediate the interactions of leukocytes with other cells.

NP Complete – nondeterministic polynomial time complete, meaning that there is known method to solve the problem to completion in a reasonable time.

Strong AI – what is typically thought of as AI; machines displaying human characteristics.

Support Vector Machines (SVM) – method of classifying data, using a hyper plane to divide points in a multi-dimensional space.

SVMLite – software used to generate SVMs.

Weak AI – a method of problem solving using computer algorithms but do not intend to match human capabilities, sometimes called pattern recognition.

4.1.3 List of Diagrams

4.2 Preliminaries

4.2.1 Artificial intelligence

Artificial intelligence or AI is often divided into two distinct areas. Strong AI is what most associate with the term, and refers to the ability of software to show human-like intelligence and be able to act in an intelligent way, no matter what task it is given. Strong AI, for the moment at least, is confined to science fiction. Weak AI concentrates on more limited problems, and is concerned with the use of software to study or solve a specific defined problem, sometimes aimed at mimicking human behaviour, but often beyond the scope of human cognitive abilities.

What constitutes a strong AI is essentially the question of how intelligence is defined. There is currently no definition that satisfies everyone as to how intelligence is defined and can be tested. However, there are certain general aspects that have been agreed that an AI would have to be able to perform. These include the ability to reason, plan, learn, represent knowledge and communicate in natural language, but also to be able to integrate these skills towards a common goal.

One of the most famous tests proposed is the Turing test. The set-up of this test involves a person (the judge) having a conversation with both a computer and another person (the subject). The judge and the subject are put in separate rooms, and only allowed to communicate through text using a keyboard and screen. The judge then has to decide which they think is the computer, and which the human. If a computer is able to fool a sufficient number of judges into thinking it is human, the computer has passed the test.

Much of the work on strong AI and what constitutes intelligence remains theoretical, as the problem of simulating human-level intelligence has proven a far more difficult undertaking than initially thought.

Weak AI is more concerned with solving specific problems using computer algorithms that are often beyond the cognitive capacity of the human brain. Many different general tools have been used to construct models to solve these problems, including: search and optimisation, probabilistic methods, and statistical learning methods.

Probabilistic methods are used when problems have an uncertain element, and the probability of certain factors needs to be utilised to solve the problem. Bayesian networks and HMM are typical examples of this type of tool.

Search and optimisation tools use automated searching to find the best solution from many possible solutions to a problem. This may also include a process of optimisation to incrementally improve a solution. Examples of this kind of AI include genetic programming. The number of possible solutions often makes it impossible to provide an exhaustive search of all potential solutions; therefore, heuristics must be employed to achieve a best guess.

Classifiers and statistical methods are employed when data can be split up into groups; pattern matching is used to sort out these data. For example, in our case, the data can be sorted into groups of binding and non-binding proteins. The most common tools used to construct these models of classification include neural networks, SVMs and decision trees. However, no particular classifier has been shown to be optimal for all types of problem, and a degree of trial-and-error is necessary in determining how best to solve a given problem.

Elements of each of these concepts are often used together to solve problems; for example, search and optimisation heuristics are used in the construction of decision trees.

4.2.2 Markov and Hidden Markov models – an example of a probabilistic method

A Markov chain model describes a system where the next observation in a sequence is determined by a random variable or state. The state of the system can be inferred from the previous observation in the chain. Hence, the next result can be predicted with knowledge of the previous result in a chain (Rabiner, 1989).

An HMM is a situation where the state cannot be entirely inferred from the previous result; the state is hidden. This can be described as a model where a sequence of observations occurs as a result of the influence of a number of possible hidden states. The probability of a particular result varies, dependent on the state (Rabiner, 1989). For example, if someone visiting a coffee shop had a latte more often when they felt happy and an espresso when they felt sad, the owner might try to guess his or her mood from what he or she were drinking each day, and try to anticipate his or her next order. However, the person's mood is not certain; it is only more likely, given a particular drink chosen. The hidden state is the mood of the person, and the drink chosen is the observation.

Hidden Markov models have a number of advantages:

- The theory is easily understood, leading to easier analysis and development.
- They scale well, allowing for incremental learning.
- They have been proven to be effective in solving a number of different bioinformatics problems (Bateman et al., 2002; Henderson et al., 1997; Krogh et al., 1994; Smith, 2002).
- They are useful in different types of problems: they can be used either to evaluate the probability of an observed sequence, given a trained model, or to find the most likely path through a model, given an observed sequence (Rabiner, 1989).

They also have a number of disadvantages:

- A large amount of training data is necessary to give a reliable model.
- Only positive data are used in training, resulting in high false-positive rates.
- A very large number of parameters need to be set.
- A number of assumptions need to be made about the data. For example, states are assumed to be independent, but this is not always the case.
- They can be slow, as all paths through the model need to be tried.

Figure 4.2.2-1, Mathematical definition of an HMM.

An HMM may be defined as taken from Smith (Smith, 2002) and Rabiner (Rabiner, 1989);

1. Set S of N states, $S = S_1S_2 \ldots S_N$
2. Set V of M observation symbols, the output alphabet. $V = v_1v_2 \ldots v_M$.
3. Set A of state transition probabilities, $A = a_{ij}$, where a_{ij} is the probability of moving from state i to state j.

$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$, $1 \leq i, j \leq N$

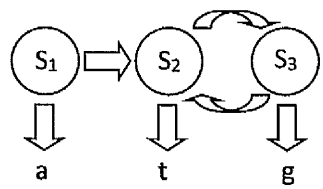
4. Set B of observation symbol probabilities: at state j, $B = b_j(k)$, where $b_j(k)$ is the probability of emitting symbol k at state j.

$b_j(k) = P(v_k \mid q_t = Sj)$; $1 \leq j \leq N$, $1 \leq k \leq M$.

5. Set π, the initial state distribution $\pi = \pi_i$, where π_i is the probability that the start state is i.

$\pi_i = P(q_1 = S_i)$; $1 \leq i \leq N$.

Given the definitions above, the notation of a model is $\lambda = (A,B, \pi)$.



A simple HMM = (A,B, π), where N = 3, M = 3, a_{12}, a_{23}, a_{32} are non-zero,

$b_1(a)$, $b_2(t)$, $b_3(g)$ = 1 and π= 1, 0, 0. Note that states can be 'null' states that do not emit any symbol.

HMMs have been used to tackle a number of biological problems, including the modelling of Pfam protein families (Bateman et al., 2002), protein modelling (Krogh et al., 1994), and finding genes in DNA (Henderson et al., 1997); (Smith, 2002).

4.2.3 Decision Trees

Decision trees provide a simple, efficient way of classifying data by sequentially dividing it into smaller data-sets with questions that have a limited number of answers, eventually leading to a classification. These tree-like schema are trained using large amounts of data, and can then be used to classify an input based on its characteristics (Mitchell, 1997; Winston, 1992).

Conceptually, a decision tree is easy to understand; data are classified using a set of questions. Each question aims to split these data such that one group has the maximum population of one class. Ideally, one question would divide all data into the classes. If this is not possible, further questions are asked until all data are classified. In decision trees, the questions are called nodes; if a question results in another question, that is a branch; if a question results in a classification, it is a leaf (Kingsford and Salzberg, 2008).

There are a number of advantages to the use of decision trees:

- They are easy to understand and interpret. Decision nodes relate directly to the characteristics, and the effect each characteristic has on the result can easily be seen (Kingsford and Salzberg, 2008).
- They are able to use numerical data in addition to categorical data; this is usually achieved using a threshold value. Other types of analysis often specialise in one data type.
- Trees are robust in that they are not unduly affected by outlying data (Mitchell, 1997).
- They perform well with a large amount of data in a short time (Kingsford and Salzberg, 2008).

Decision trees also have a number of limitations:

- There are a large number of possible tree configurations with even a limited number of characteristics. The decision-tree learning problem is said to be NP complete in computational complexity theory. This means

that there are no known algorithms able to complete these problems in a reasonable amount of time. Therefore, to be practical, heuristics need to be used in decision-tree learning. These heuristics tend to return locally optimal decisions for each node; this does not guarantee the optimal decision tree (Mitchell, 1997; Winston, 1992).

- They have a tendency towards over-fitting, leading to overly complex trees that describe the data rather than act as a useful tool for classification (Mitchell, 1997; Winston, 1992).

Figure 4.2.3-1, Mathematical representation of entropy.

Taken from Kingsford (Kingsford and Salzberg, 2008):

Suppose we are trying to classify items into m classes using a set of training items, E.

Let: $p_i\ (i = 1,\ldots,m)$

be the fraction of the items of E that belong to class i.

The entropy of the probability distribution

$$(p_i)_{i=1}^{m}$$

gives a reasonable measure of the impurity of the set, E.

The entropy,

$$-\sum_{i=1}^{m} p_i log p_i$$

is lowest when a single p_i equals 1 and all others are 0,

whereas it is maximised when all p_i are equal.

The C4.5 algorithm uses the concept of information entropy. Essentially, the algorithm works by the data splitting dependent on a characteristic that leaves the resulting groups with a maximally increased proportion of one of the classes. This is then repeated for each of the groups until classification is complete (Quinlan, 1993). The mathematical formalisation of entropy and how it is minimised can be seen in Figure 4.2.3-1.

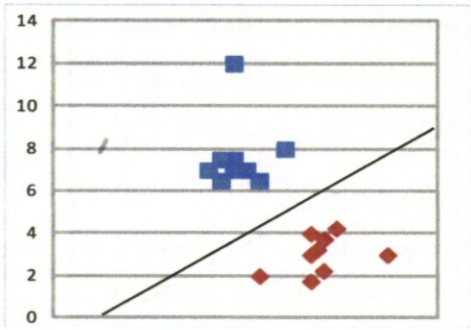
Decision trees have been used to solve various biological problems, including: the identification of potential MHC class I peptide epitope motifs (Savoie et al., 1999), and the discovery of motif-based protein function classifiers (Wang et al., 2003).

4.2.4 Support Vector Machines

Support vector machines (SVMs) plot all data points, with each characteristic as an axis on a multi-dimensional graph. Classification is then attempted by trying to find planes and hyper-planes that separate grouped data-points (Joachims, 2002). The formula that represents this hyper-plane can then be used to sort new examples into these groups.

For the purpose of explanation, this concept can be looked at in two dimensions. This is equivalent to finding the optimal line in a graph that separates two collections of data-points (see Figure 4.2.4-1). However, this has to be imagined not just in three dimensions, but in a multi-dimensional space equal to the number of attributes that describe each data-point. In addition, the line separating the data-points becomes a hyper-plane, and does not have to be linear in any dimension (Joachims, 1998).

Figure 4.2.4-1, Optimal line through 2 collections of data



The line separates the two sets of data optimally. The equation that the line represents could be used to classify these data into the two groups.

There are a number of advantages to the use of SVMs:

- They produce accurate classifiers.
- They are robust to noise. In other words, they are less prone to over-fitting (Joachims, 1998).

There are also a number of disadvantages:

- SVMs are binary classifiers. To do a multi-class classification, pair-wise classifications can be used.
- It is difficult to understand and interpret the output.
- They are computationally expensive and can be slow to run (Joachims, 1998).

Mathematically, the training of an SVM can be expressed as a quadratic optimisation problem (see Figure 4.2.4-2).

Figure 4.2.4-2, Mathematical representation of SVM optimisation.

Taken from Joachims (Joachims, 1998). If the number of examples (cases) used in training is, α is a vector of l variables, and each component of α_i is a training example (x_i, y_i), then the SVM optimisation problem can be represented as:

Minimise: $W(\alpha) = -\sum_{i=1}^{l} \alpha_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$

Subject to: $\sum_{i=1}^{l} y_i \alpha_i = 0$

And: $\forall i: 0 \leq \alpha_i \leq C$

SVMs have been used to solve a number of biological problems, including: multi-class protein-fold recognition (Ding and Dubchak., 2001), and gene selection for microarray data (XinZhou, et al., 2008).

4.2.5 Genetic algorithms - an example of a search and optimisation method

Genetic programming uses an evolutionary approach to try to find the algorithm most fit for the purpose.

This is done by the swapping of virtual chromosomes that contain potential elements of a solution, followed by the selection of the 'fittest' algorithms and further swapping until a suitably accurate solution is found. It is important to make a good model of the goals one is trying to achieve in order to get an accurate result (Mitchell, 1998).

4.3 Methods

4.3.1 Data Normalisation

The data presented in section 2 shows a large bias towards the EF-hand super-families: 91 in the EF-hand super-family SCOP classification, a.39.1, compared to the other super-families that may only contain a few members. It was suggested, therefore, that in order to compensate for this bias, only one representative from each super-family group should be used for the AI analysis. As only fourteen super-families were identified, this would mean only fourteen positive cases for the AI analyses. This small amount might to be too few to reach an accurate classification. It was decided that five calcium-binding proteins from each super-family would be randomly selected to give a greater number of examples for use in AI training. In families where there were not five examples to choose from, all of the proteins were used. This resulted in 38 positive cases being selected for use in the training of the AI tools; these are listed in Table 4.3.1-1.

Table 4.3.1-1, List of positive cases used for AI training

Super-family	Motif	Super-family	Motif
a.139.1	1daq:A:8-12	**c.84.1**	2fkm:X:242-246
a.39.1	1exr:A:20-24	**c.93.1**	1gca:-:134-138
	1ht9:A:54-58	**c.94.1**	1y3n:A:171-175
	1qls:A:66-70		1j1n:A:171-175
	1g8i:A:73-77		1r9l:A:124-128
	1ooj:A:57-61		1xoc:A:346-350
b.30.5	1k1x:A:392-396	**d.3.1**	1vjj:A:301-305
b.69.8	1txv:A:426-430	**d.92.1**	1h71:P:49-53
	1txv:A:297-301		1g9k:A:49-53
	1txv:A:365-369	**f.11.1**	1acc:-:177-181
	1jv2:A:284-288	**g.75.1**	1ux6:A:828-832
	1jv2:A:413-417		1ux6:A:843-847
b.80.7	1kap:P:446-450		1ux6:A:866-870
c.1.8	1lwj:A:13-17		1ux6:A:879-883
	1m53:A:63-67		1ux6:A:915-919
	1wza:A:44-48		
	1gzj:A:202-206		
	1h1n:A:202-206		
c.62.1	2b2x:A:154-158		
	1dzi:A:151-155		
	1mq9:A:137-141		
	1shu:X:50-54		
	2ica:A:137-141		

4.3.2 Decision Trees

Decision-tree software called C4.5 was used for the construction of decision trees from the recovered data. First, the appropriate input files had to be constructed; these consisted of a names file and a data file for each experiment (see Figure 4.3.2-1). The names file consists of each of the attributes used in the experiment listed in order; the data file then lists each case on a separate line, with the value of each attribute in order, separated by a comma. These files were constructed automatically (for further details see section 4.3.4).

Figure 4.3.2-1, Examples of the files used as input for C4.5

List of possible classifications; B+ for binding, B- for non-binding.

The name of the variables, each on a separate line.

```
B+,B-.

ATG-4P: continuous.
ATG-3P: continuous.
ATG-2P: continuous.
ATG-1P: continuous.
ATG0P: continuous.
ATG1P: continuous.
...
ATG8P: continuous.
ATG-4N: continuous.
...
```

A list of possible values for each variable. Continuous is used if the variable is numeric.

i) C4.5 *.names file

```
0,96,0,0,0,0,0,0,1,0,100,0,100,100,...0,0,0,B+
0,24,22,1,0,23,68,26,17,24,24,8,35,... 0,0,0,B+
0,26,30,1,0,28,75,37,9,33,30,8,19,... 44,0,0,0,B+
0,57,33,0,0,51,75,17,26,0,68,0,35,... 97,0,0,0,B+
0,100,0,0,0,4,0,3,98,0,98,0,0...0,0,0,0,B+
...

1,100,0,10,0,1,78,24,55,1,1,90,0...96,0,1,0,B-
4,0,0,0,0,22,3,2,92,15,0,8,0,95...0,0,7,100,B-
```

The actual data: each line is a different case, and the data are listed in the same order specified in the names file.

Each line ends with the classification of the case; here, either B+ or B-.

ii) C4.5 *.data file

The decision-tree software was run with its default settings, using both the "c4.5" command to generate a tree file, and the rules command to generate a set of "c4.5rules" that can be used to classify further examples.

Each experiment was repeated using a miss-one-out strategy to allow verification of the results. For this, one family was missed out in each repeat, meaning, in fact, one to five cases were missed in each repeat. These missed examples can then be used as testing data against the trees produced.

The tree files were parsed, and the error rates of each tree were recorded, along with the general tree structure, and presented in a tabulated format. Analysis of % correct classification for the tree, when using the training data, was used to compare the tree's effectiveness.

Also, each of the cases was checked to see how it was classified by each of the decision trees produced. This allowed the significance of each case to be looked at, and allowed cross-checking when different characteristics were used. To achieve this, the rules files were parsed and used to repeat the classification of each case. Then, each case could be assigned as true-positive, true-negative, false-positive and false-negative, further helping comparison of tree effectiveness and validity.

4.3.3 Support Vector Machines

SVMlite was used to conduct the SVM experiments. The data were converted into the appropriate format for input into the SVM software. These files include the code for each attribute represented, and the value of the attribute, in a strict order. These files were constructed automatically (see Figure 4.3.3-1).

Figure 4.3.3-1, Examples of the files used as input for SVMlite

```
1 000002:0 000003:100 ... 120112:1 120113:0 120114:0#ID:1exr:A:20-24
1 000002:0 000003:6 ... 120112:87 120113:0 120114:76#ID:1ht9:A:54-58
1 000002:0 000003:12 ... 120112:94 120113:0 120114:87#ID:1qls:A:66-70
...
-1 000002:0 000003:1 ... 120112:12 120113:0 120114:65#ID:1g8i:A:73-77
-1 000002:0 000003:3 ... 120112:1 120113:0 120114:96#ID:1ooj:A:57-61
-1 000002:19 000003:12 ... 120112:65 120113:76 120114:96#ID:1k1x:A:392-396
...
```

The classification of the data: 1 for binding, -1 for non-binding. Each case is on a separate line.

The actual data in the format: `label:value`. The label and the data value must both be a number.

The label for the case is at the end of the line in the format: `#ID:label`

4.3.4 "CaBindTrain" Scripting Outline

The automation for the AI algorithms centres around, first, producing the input files for C4.5 and SVMlite, then running the software as a miss-one-out experiment, then collecting the statistics from each run and presenting them in a unified file. MakeSVM.pl and MakeDT.pl are responsible for these processes; each works in a similar manner. The attribute data were collected and parsed by **processfamilydata. saveSVMdata** and **saveDTdata** translated these data into the appropriate format for SVMlite and C4.5. **classifybysvm** and **classifybydt** then ran the actual software. Finally, **getclassresults, getSVMerrors** and **getDTerrors** produced summary files of the results.

4.4 Results and Discussions

Note, in this section, experiments are presented in a "lab book" style. Each experiment shows a list of the attributes used, a summary of its aims, results and a short discussion. After each set of experiments, the results are summarised in a more formal manner. For each result, a range showing the % error across all the miss-one-out runs is given, followed by the average across these runs; finally, "all" is the % error from the tree where no cases were missed. The error rates are calculated using all of the cases.

4.4.1 Decision Trees, All-attributes

All-attributes experiment 1

Attributes used: - amino acid type (ATG), amino acid size (AAS), solvent accessibility (SOL), conserved residues (CON) secondary structure (SS).

Experiment description: - initial experiment to give indication of usefulness of methodologies and ideas for further experiments.

Results: -

All: 1.7%-3.7% error, ave 3.41%, all 3.4%.

Known positive misclassified rate: 5.3%. Known negative misclassified rate: 1.6%.

Discussion: - seems that only the AAT attribute contributes to the decision-making process; this is probably because this is the attribute with the most informative data.

All-attributes experiment 2

Attributes used: - amino acid size (AAS43114443893291), amino acid type (AAT), solvent accessibility (SOL), conserved residues (CON) secondary structure (SS), amino acid hydrophobicity (AAHYD_R4).

Experiment description: - experiment using the optimised attributes.

Results: -

All: 1.7%-3.7% error, ave 3.41%, all 3.4%.

All2: 1.7%-7,4% error, ave 5.6%, all 6.8%.

Known positive misclassified rate: 5.7%. Known negative misclassified rate: 5.7%.

Discussions: - seems that both the AAT and AAS attributes contribute to the decision-making process; these are probably the attributes with the most informative data, with respect to classification.

The initial experiment using all the attributes was to give an indication of the usefulness of decision trees, and to help with assessment of further experiments. Only the amino acid type data, where amino acids have been sorted by group, appear in the resulting decision tree; this is a good indication that this characteristic may be the best candidate for further optimisation. The error rate in the training data from this tree was 3.3%. After the optimisations mentioned below for each attribute, the experiment was repeated with all attributes. However, this time, decision nodes included both the amino acid size and type attributes. This gave a worse error rate of 5%. This highlights the nature of the decision-tree algorithms; the best tree is not necessarily picked from the data if there is a better classifier at a particular node. It is interesting to note that when the error rate is broken down into false-positives and false-negatives, the first experiment shows a bias towards classifying as a positive result. This is shown by the high false-negative rate of 5.3%, compared to the false-positive rate of 1.6%. In the second experiment, the rate of each type of error is exactly the same; this means there is no bias towards classification as either positive or negative.

4.4.2 Decision Trees, Amino acid type optimisation

Amino acid type experiment 1

Attributes used: - amino acid type (ATG).

Experiment description: - amino acid type used by itself; each group of amino acids is represented by a % down the columns of the BLAST alignment.

Results: -

AAT: 1.7%-3.7% error, ave 3.41%, all 3.4%.

Known positive misclassified rate: 5.3%. Known negative misclassified rate: 1.6%.

Discussion: - AAT: gives the best results for the initial set of attributes, of around 3.4% error; positions 3, 5, 4 and 6 are used in the tree.

Amino acid type experiment 2

Attributes used: - amino acid type (ABST).

Experiment description: – the amino acid type, AAT, and amino acid size, AAS, attributes are determined as a percentage down a BLAST search (e.g., an entirely conserved residue might be AAT-100% G and AAS-100% small). The effect of using only the value of the hit initially found with SPASM and, therefore, discrete as opposed to continuous values for the decision trees, was investigated.

Results: -

AAT: 1.7%-3.7% error, ave 3.41%, all 3.4%.

ABST: 3.7%-9.3% error, ave 7.61%, all 8.5%.

Discussions: - This gave a worse result for the ABST vs AAT; therefore, it was decided to continue with the current scheme.

Taken individually, amino acid type gave the same trees as seen in the all-attributes experiment. A number of different ways of representing the data, including using individual amino acids instead of groups, and using just the value of the initial hit as opposed to the score across the BLAST alignment, were used to attempt to improve the error rate, but failed.

4.4.3 Decision trees, Amino acid size optimisation

Amino acid size experiment 1

Attributes used: - amino acid size (AAS).

Experiment description: -

Results: -

AAS: 12.1%-16.7 error, ave 13.99%, all 13.6%.

Discussions: - AAS: unexpectedly poor results, of around 14% error, suggested that there might be better ways of expressing the size, such as simply using the molecular mass of the amino acid, or using a different threshold, as the 133 used may not be optimal.

Amino acid size experiment 2

Attributes used: - amino acid size (ABSS), amino acid size (AAS).

Experiment description : - the amino acid size, AAS, attribute is determined as a percentage down a BLAST search (e.g., an entirely conserved residue might be AAS-100% small). The effect of using only the value of the hit initially found with SPASM and, therefore, discrete as opposed to continuous values for the decision trees, was investigated.

Results: -

AAS(115): L 0%-1.8% error, ave 0.13%, all 0%.

ABSS: not done.

Discussions: - this method gave a far worse result for the ABST vs AAT; therefore, it was decided not to run this experiment .

Amino acid size experiment 3

Attributes used: - amino acid size (AST)

Experiment Description: - for the AAT attribute, it was decided that the threshold chosen, a molecular mass of 133 Da, may not be the most appropriate. Each possible threshold was used to divide the amino acids into large and small, in order to try to improve on the result.

...

Results: -

S <= 75 < L 3.4%-6.9% error, ave 5.36%, all 5.1%

S <= 89 < L 1.9%-3.7% error, ave 3.27%, all 3.4%

S <= 105 < L 0%-1.8% error, ave 0.13%, all 0%

S <= 115 < L 0%-1.8% error, ave 0.13%, all 0%

S <= 117 < L 0%-3.7% error, ave 0.51%, all 0%

S <= 119 < L 0%-3.5% error, ave 1.77%, all 1.7%

S <= 121 < L 0%-3.5% error, ave 1.77%, all 1.7%

S <= 131 < L 0%-3.5% error, ave 1.89%, all 1.7%

S <= 132 < L 3.7%-9.3% error, ave 5.74%, all 8.5%

S <= 133 < L 12.1%-16.7% error, ave 13.99%, all 13.6%

S <= 146 < L 13%-16.7% error, ave 17.49%, all 18.6%

S <= 147 < L 1.9%-25.9% error, ave 13.99%, all 18.6%

S <= 149 < L 3.7%-26.3% error, ave 17.49%, all 9.16%

S <= 155 < L 5.2%-13.3% error, ave 7.5%, all 6.8%

S <= 165 < L 1.7%-6.9% error, ave 5.09%, all 5.1%

S <= 174 < L 10.3%-24.1% error, ave 17.38%, all 11.9%

S <= 181 < L 18.2%-25.9% error, ave 23.34%, all 23.7%

Discussions: - the best threshold seems to be around 105 to 117; however, a dip at the higher thresholds suggests something else may be going on, and that some positions may have a different best threshold. Additionally, there is quite a bit of variation in the structure of the trees produced in each miss-one-out experiment; this suggests that the low error rates may be owing to over-fitting to the data.

Amino acid size experiment 4

Attributes used: - amino acid weight (SIZE), amino acid size (AAS), amino acid (AA).

Experiment Description: - as such improvement was made by varying the threshold for the amino acid size (AAS), it was decided that a greater amount of freedom might show further improvement; the obvious way to do this would be to allow the decision-tree algorithm to decide its own threshold for each residue position. To achieve this, the amino acid weight of the hit (only those selected by SPASM, not by BLAST) was used. This seemed to be similar to simply specifying each amino acid separately (AA). However, this will be treated differently by the decision-tree software, as there is no continuous scale, only discrete values, so this was also attempted for comparison.

...

Results: -

AAS (115): 0%-1.8% error, ave 0.13%, all 0%.

SIZE: 1.7%-7.4% error, ave, all 5.1%.

AA: 6.9%-11.1% error, ave 8.76%, all 8.5%.

Discussion: - the size give worse results when compared with the AAS values. Without the AABS data, it is impossible to tell how much this is affected by the change from an across-BLAST system to a hit-only-one. Averaging the molecular mass across the BLAST search may improve this situation. This may also be a result of giving the decision tree too many variables.

Amino acid size experiment 5

Attributes used: - amino acid size (AAS), amino acid weight (SIZE).

Experiment description: - there is some doubt that a single threshold used for all the amino acids is the best model. There are two weights where error rates improve; around 105 and 165 in the previous experiment. This experiment was devised to assess the optimal threshold for each position individually.

In order to do this, the data were set up so that each threshold could be varied independently, then all the positions except one were set to 115, this being the previously found best value from amino acid size experiment 3. For the remaining position, the amino acid weight was used; therefore, at this position only, the software could essentially pick its own threshold. This was then repeated, setting each position as an amino acid weight in turn. The experiment was then repeated using the poor threshold, 133, as the constant value.

Results: -

Position	115	133	Best	Notes
-4	-	-	115	
-3	105, 89	105, 131	105	89,131 >errors
-2	181	75, 146	146	181, 146 >errors
-1	147, 155	133	115	147,155,133 >errors
0	-	-	115	
1	131	-	115/131	similar results
2	132, 105	105	105	132 >errors
3	131	-	115/131	similar results
4	132	132, 105	132	105 >errors
5	89, 75	105, 75	75/89/105	similar results
6	131	89, 181, 165	89	131,181,165 >errors
7	-	132, 131	132	131 >errors
8	75	75, 131, 133	75/131	similar results

Discussions: - the resulting trees can be analysed to see what positions cause the amino acid weight to be used as a discriminator in the tree, and what threshold is picked. The error rate of that tree then allows us to decide if that threshold is more or less effective than 115 and 133. From the analysis, it was possible to identify the best threshold for most of the residue positions. To ensure the optimum threshold has been found for the remaining positions, further experiments need to be done.

Amino acid size experiment 6

Attributes used: - amino acid size (AAS).

Experiment description: - continuing from the previous experiment, there are still 4 positions that did not have clear optimum thresholds. This experiment will attempt to establish the optimum threshold by using each proposed threshold for each of the four positions. The other positions each use the established optimum.

Results: -

Position 5

AAS43114483891291: 0%-3.7% error, ave 0.9%, all 0%.

AAS43114483892291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114483893291: 0%-1.8% error, ave 0.1%, all 0%.

Position 8

AAS43114483892291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114483892298: 0%-3.5% error, ave 1.7%, all 1.7%.

AAS43114483893291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114483893298: 0%-3.5% error, ave 1.7%, all 1.7%.

Position 3

AAS43114483892291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114483492291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114483893291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114483493291: 0%-1.8% error, ave 0.1%, all 0%.

Position 1

AAS43114483892291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114443892291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114483893291: 0%-1.8% error, ave 0.1%, all 0%.

AAS43114443893291: 0%-1.8% error, ave 0.1%, all 0%.

Known positive misclassified rate: 1%. Known negative misclassified rate: 0.3%.

Discussion: - for the 5th residue, one of the 3 thresholds was ruled out, and experiments were continued with both of the remaining thresholds at this position. The 8th residue's threshold was successfully discovered; for the remaining 2 residues, both thresholds were equally effective.

By itself, amino acid size gave a 13.6% error in the training data; this was unexpectedly poor. The high error may be owing to a poor model being used in assessing size. The threshold chosen for the amino acid size threshold was an arbitrary 133Da, based on the median across the 20 amino acids. Optimisation of this threshold showed that a threshold of between 105 and 117 gave significantly

better results, giving a 0% error across the training data. However, there is also a decreased percentage error at higher masses, suggesting a threshold that varies between residues may optimally represent the data. Variation across the tree structures suggests that the low error rates may also be a result of over-fitting to the data. A series of experiments was carried out to assess the optimal threshold for each residue position, but these experiments did not improve the error rates for classifying the data, as all cases were already correctly classified. However, all the trees produced showed a similar structure, and used the same attributes for classification. This indicates that the tree has described the underlying biology effectively, and was not just fitting the data.

4.4.4 Decision trees, Solvent accessibility optimisation

Solvent accessibility experiment 1

Attributes used: - solvent accessibility (SOL).

Experiment description: - solvent accessibility score from SABLE was used with the decision trees.

Results: -

SOL: 1.9%-7% error, ave 5.6%, all 6.9%.

Discussions: - SOL: a middling result of around 5%; no way could be thought of to improve on this model.

Solvent accessibility gave a mid-range error rate of 5.6%; however, no means of improving this was discovered.

4.4.5 Decision trees, Secondary Structure optimisation

Secondary structure experiment 1

Attributes used: - secondary structure (SS).

Experiment description: - secondary structure was represented by the distance from the motif upstream and downstream, the size of the element upstream and downstream, and the number of residues within the motif that were predicted to have secondary structure.

Results: -

SS: 24.1%-35.2% error, ave 31.75% all 30.5.

Discussions: - very poor results were expected, as there is very little information given for each case. This might be improved on by adding secondary structure type, to go with the distance, length and score. This may still, however, be useful as a broad discriminator, as the presence of secondary structure at the site of the motif binding is unlikely.

Secondary structure experiment 2

Attributes Used: - secondary structure (SS2).

Experiment Description: - the description of the secondary structure around the motif uses a very limited set of values; the distance from the motif and length of the upstream and downstream predicted secondary structure for the protein, along with a score for the amount of secondary structure predicted within the motif - a total of 5 attributes per case. Here, the type of secondary structure, helix or sheet, is added to improve the results.

Results: -

SS: 24.1%-35.2% error, ave 31.75%, all 30.5%.

SS2: 22.4%-33.3% error, ave 30.21%, all 30.5%.

Discussion: - an improvement is seen; however, it is not clear how further improvement could be made to this attribute's results. Not much hope was put on secondary structure being a good discriminator as, although there does seem to be a need for some amount of structure to hold the motif in the correct conformation for binding, there is little to differentiate this from secondary structure that may appear unrelated to the motif.

Secondary structure gave a very high error rate of 31.7%; this was expected, as only the presence of the absence of secondary structure, and distance from the motif, were represented. This was improved marginally, by addition of secondary structure type, to 30.5%.

4.4.6 Decision trees, Conserved Residue

Conserved residue experiment 1

Attributes used: - conserved residues (CON).

Experiment description: - the sequence downstream of the motif was divided into ranges; if an over 50% conserved D or E reside is found within a range, it returns true.

Results: -

CON: 7.7%-9.3%, ave 9%, all 8.5%.

Discussions: - a poor result of around 9% error. After discussion, it was decided that the amino acid boxes (a box returned true if there was a conserved residue in a certain range) used in this experiment were too large, and a set of smaller boxes was decided on.

Conserved residue experiment 2

Attributes used: - conserved residue (CON2).

Experiment description: - the conserved residue attribute did not give very good results; therefore, it was suggested that the scheme be improved. It was decided that the 10-residue window was too large to pick up any differences there may be between the two groups, as the conserved D or E seen downstream of the motif is, in most cases, within a few residues of the motif; however, it can also be further away. A scheme where the first 10 residues were considered individually, then the next 10 as a group, followed by groups of 20, up to 100, then all residues over 100, was used.

Results: -

CON: 7.7%-9.3%, ave 9.01%, all 8.5%.

CON2: 7.4%-9.3%, ave 8.5%, all 8.5%.

Discussions: - changing the model to reflect the variation in the distance of the conserved residue slightly improved the % error rate.

Conserved residue information gave a surprisingly low error rate of 9% considering the limited amount of information represented. This was improved upon by changing how the data were represented. Rather than using boxes of 10 amino acids, a scheme where the first 10 residues were considered individually, then the next 10 as a group, followed by groups of 20, up to 100, then all residues over 100, was used. The error rate improved slightly to 8.5%.

4.4.7 Decision trees, Amino acid hydrophobicity

Amino acid hydrophobicity experiment 1

Attributes used: - hydrophobicity (HYD).

Experiment description: - initially, the hydrophobicity of each amino acid was given to the decision trees.

Results: -

HYD: 1.7%-6.9%, ave 3.8%, 3.4%.

Known positive misclassified rate: 5.3%. False-positive rate: 1.6%.

Discussions: - this gives a low error rate; however, in the amino acid size experiments, a threshold was shown to give better results. This might prove true with hydrophobicity as well.

Amino acid hydrophobicity experiment 2

Attributes used: - hydrophobicity threshold (HYDv).

Experiment description: - similar to the threshold experiments for amino acid size, a series of experiments was done to find the best threshold for hydrophobicity.

Results: -

R0 HPI <= -4.5 < HPO 7.3%-15.5% error, ave 13.3%, all 13.6%

R1 HPI <= -3.9< HPO 1.9%-3.7% error, ave 3.27%, all 3.4%

R2 HPI <= -3.5 < HPO 0%-27.8% error, ave 11.7%, all 25.4%

R3 HPI <= -3.2< HPO 0%-3.7% error, ave 1.9%, all 1.7%

R4 HPI <= -1.6< HPO 0%-1.9% error, ave 0.4%, all 0%

R5 HPI <= -1.3< HPO 0%-1.9% error, ave 0.5%, all 0%

R6 HPI <=-0.9 < HPO 0%-3.5% error, ave 1.9%, all 1.7%

R7 HPI <= -0.8 < HPO 1.7%-3.5% error, ave 2.2%, all 1.7%

R8 HPI <= -0.7 < HPO 3.4%-6.9% error, ave 5.3%, all 5.1%

R9 HPI <= -0.4 < HPO 1.7%-5.6% error, ave 3.4%, all 3.4%

R10 HPI <= 1.8 < HPO 3.%-5.6% error, ave 4.9%, all 5.1%

R11 HPI <= 1.9 < HPO 0%-5.5% error, ave 2.2%, all 1.7%

R12 HPI <= 2.5 < HPO 0%-8.8% error, ave 4.3%, all 3.4%

R13 HPI <= 2.8 < HPO 0%-3.7% error, ave 3.1%, all 3.4%

R14 HPI <= 3.8 < HPO 0%-3.4% error, ave 1.8%, all 1.7%

R15 HPI <= 4.2 < HPO 1.7%-3.7% error, ave 3.2%, all 3.4%

R16 HPI <= 4.5 < HPO 0%-7.4% error, ave 2%, all 1.7%

Discussions: - a threshold of -1.6 gives the lowest error rate from the training data. However, two further dips are seen at higher levels of hydrophobicity, suggesting again that one threshold may not be the best for all amino acid positions.

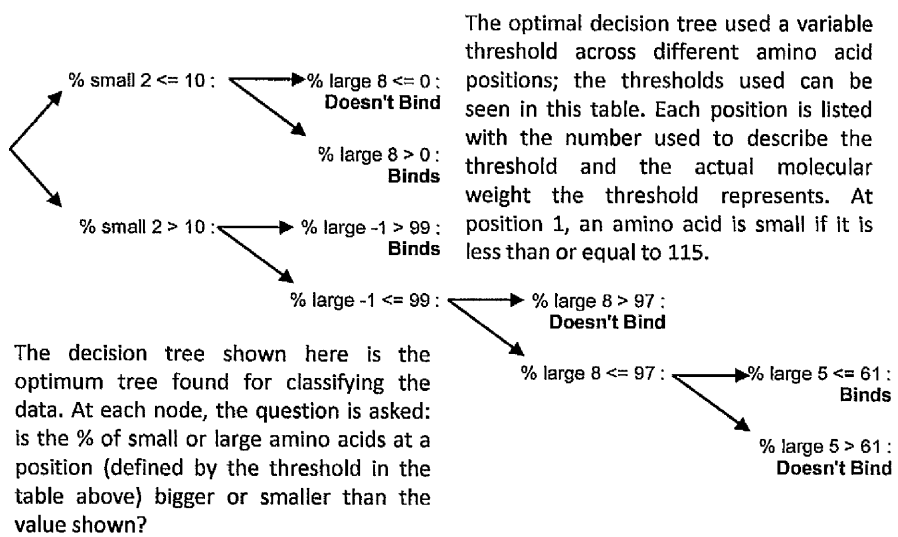
As the weight and type attributes of the amino acids showed low error rates, other chemical properties were considered, such as hydrophobicity. Initially, the simple hydrophobicity score was used; this gave low error rates of 3.4%. The model was then improved, based on the thresholds model used with amino acid size, which greatly improved the % error to 0%. It is possible that the model could have been further improved by using a variable threshold across each amino acid position, as used for amino acid size.

4.4.8 Best Decision Tree found and verification

The best tree was found using a variable threshold across each residue position. This set of trees gave the lowest % error rate, with none of the training data incorrectly classified. Additionally, the miss-one-out verification showed a high, but not complete, agreement in the structure of the trees. In each of the runs, the same nodes were present, but in a different configuration. The tree constructed when all the cases were used will be used for classification in the next section (see Figure 4.4.1-1).

Figure 4.4.8-1, The optimum decision tree found, with table describing the attributes used to create it.

Position	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8
Value	4	3	11	4	4	8	3	8	9	2	2	9	1
Weight	S≤115	S≤105	S≤146	S≤115	S≤115	S≤131	S≤105	S≤131	S≤132	S≤89	S≤89	S≤132	S≤75



The optimal decision tree used a variable threshold across different amino acid positions; the thresholds used can be seen in this table. Each position is listed with the number used to describe the threshold and the actual molecular weight the threshold represents. At position 1, an amino acid is small if it is less than or equal to 115.

The decision tree shown here is the optimum tree found for classifying the data. At each node, the question is asked: is the % of small or large amino acids at a position (defined by the threshold in the table above) bigger or smaller than the value shown?

The thresholds used ranged from 75 to 131. Position 8 appears twice in the decision tree; it has a threshold of 75, which is essentially asking if glycine is present at this position. Position 2 (the second motif position), we know, is D, S or N; the threshold here is 105; therefore, the question here could be rephrased as: is serine present at this position? Positions -1 and 5 both flank the motif; here, the size of the amino acid may lead to steric interference; interestingly, at -1, if an amino acid larger than proline is present, then this leads to binding. At position 5, the question is whether an amino acid smaller than serine is present.

4.4.9 Support Vector Machines, All attributes

All-attributes experiment 1

Attributes used: - amino acid size (AAS43114443893291), amino acid type (AAT), solvent accessibility (SOL), conserved residues (CON) secondary structure (SS) and amino acid hydrophobicity (AAHYD_R4).

Experiment description: - all the attributes were used with the improved versions of the attribute models, including the amino acid size, where the optimum profile of thresholds was used.

Results: -

All: 3.8%-15.7% error, ave 12.6%, all 12.6%.

Known positive misclassified rate: 10.9%. Known negative misclassified rate: 11.7%.

Discussions: - a middling result, far poorer than the corresponding DT experiment.

All-attributes experiment 2

Attributes used: - amino acid size (AAS), amino acid type (AAT), solvent accessibility (SOL), conserved residues (CON) secondary structure (SS) and amino acid hydrophobicity (AAHYD_R4).

Experiment description: - all the attributes were used with the improved versions of the attribute models, except for the amino acid size, which used a uniform threshold of 115 across all the residues.

Results: -

All: 1.9%-12.5% error, ave 9.7%, all 11.3%.

Known positive misclassified rate: 8%. Known negative misclassified rate: 10.1%.

Discussions: - Omitting the optimum thresholds seemed to increase the accuracy of the SVM.

Similar to the decision-tree experiments, an all-attributes experiment was used to give indication of the usefulness of the methodologies and ideas for further experiments. All the attributes were used with the improved versions of the attribute models. This gave a relatively poor error rate of between 3.8% and 15.7%, with an average of 12.6%. In an attempt to improve the error rate, the experiment was revised using a uniform amino acid size threshold of 115 across all the residues. This improved the error rate, giving an average error of only 9.3%. The SVMs seem biased towards classifying the positive results correctly, with 8% misclassified, in comparison with the negative results, where 10.1% were misclassified.

4.4.10 SVMs, Amino acid type

Amino acid type experiment 1

Attributes used: - amino acid type (ATG).

Experiment description: - amino acid type used by itself; each group of amino acids is represented by a % down the columns of the BLAST alignment.

Results: -

AAT: 0%-9.2% error, ave 5.8%, all 7.4%.

Known positive misclassified rate: 4.2%. Known negative misclassified rate: 8.3%.

Discussions: - AAT: gives the best results for classification by SVM, of around 5% errors.

The amino acid type was run in a similar manner to the best of the decision-tree experiments, with a percentage representation of each group at each residue down the BLAST search used. This gave by far the best results for the SVMs, with an average misclassification rate in the training data of just 5.8%.

4.4.11 SVMs, Amino acid size

Amino acid size experiment 1

Attributes used: - amino acid size (AST).

Experiment Description: - for the AAT attribute, it was decided that the threshold chosen, a molecular mass of 133 Da, may not be the most appropriate. Each possible threshold was used to divide the amino acids into large and small, in order to try to improve on the result.

Results: -

S <= 75 < L 10.1%-30.5% error, ave 20.3%, all 30.5%

S <= 89 < L 10.1%-30.5% error, ave 25.6%, all 30.5%

S <= 105 < L 3.4%-23.7% error, ave 0.13%, all 20.9%

S <= 115 < L 3.4%-27.1% error, ave 24%, all 25.4%

S <= 117 < L 5.1%-30.5% error, ave 28.4%, all 30.5%

S <= 119 < L 5.1%-30.5% error, ave 27.6%, all 30.5%

S <= 121 < L 6.8%-30.5% error, ave 28.2%, all 30.5%

S <= 131 < L 6.8%-15.3% error, ave 9.6%, all 10.2%

S <= 132 < L 8.5%-13.6% error, ave 11.3%, all 8.5%

S <= 133 < L 28.8%-35.6% error, ave 34.2%, all 35.6%

S <= 146 < L 33.9%-35.5% error, ave 35.5%, all 35.6%

S <= 147 < L 35.6%-35.6% error, ave 35.6%, all 35.6%

S <= 149 < L 35.6%-35.6% error, ave 35.6%, all 35.6%

S <= 155 < L 35.6%-35.6% error, ave 35.6%, all 35.6%

S <= 165 < L 35.6%-35.6% error, ave 35.6%, all 35.6%

S <= 174 < L 35.6%-35.6% error, ave 35.6%, all 35.6%

S <= 181 < L 35.6%-35.6% error, ave 35.6%, all 35.6%

Discussions: - the best threshold seems to be at 131. This differs substantially from the results given with the decision tree methods.

Amino acid size experiment 2

Attributes used: - amino acid size (AAS).

Experiment description: - this experiment uses the optimum thresholds established using the decision trees, in order to compare how well it performs against the uniform threshold from the previous experiment.

Results: -

S <= 131 < L 6.8%-15.3% error, ave 9.6%, all 10.2%.

False-negative rate: 0%. False-positive rate: 31.2%.

AAS43114443893291: 8%-26.1% error, ave 22.3%, all 25.5%.

Known positive misclassified rate: 14%. Known negative misclassified rate: 25.4%.

Discussion: - The amino acid size with a variable threshold using SVMs shows a far greater error rate than using DT. It is possible that this is not the optimum set of thresholds for use with SVMs. The best result was given using the uniform threshold.

Again, a series of experiments were carried out to determine the best way of representing the amino acid size. The amino acid size thresholds that were established using the DT methods were used with SVMs. The results given using a variable threshold, similar to the one used with the decision trees, gave a 22.3% rate of misclassification in the training data. This is a much higher error rate than given using decision trees. It is possible that either this was not the optimum set of thresholds for use with SVMs, or that this characteristic was not well suited to use with the SVMs.

The best error rate, however, was given by the uniform threshold of a molecular mass of 131, giving an error rate of between 6.8% and 15.3%, with an average of 9.6%.

4.4.12 SVMs, Solvent accessibility

Solvent accessibility experiment 1

Attributes used: - solvent accessibility (SOL).

Experiment description: - solvent accessibility score from SABLE was used with the SVMs.

Results: -

SOL: 18%-31.1% error, ave 26.4%, all 31.1%.

Known positive misclassified rate: 0%. Known negative misclassified rate: 14%.

Discussions: - SOL: a poor result of around 26%; again, no way could be thought of to improve on this model.

There were no improvements from the decision trees for solvent accessibility, so the same scheme was used again. However, it gave a very poor score of 26.4% misclassification error in the training data.

4.4.13 SVMs, Secondary structure

Secondary structure experiment 2

Attributes Used: - secondary structure (SS2).

Experiment Description: - the improved version of the secondary structural model was used, including the distance from the motif, length of the upstream and downstream, predicted secondary structure for the protein, along with a score for the amount of secondary structure predicted within the motif, and the type of secondary structure, helix or sheet.

Results: -

SS2: 48% error, ave 48%, all 48%.

Known positive misclassified rate: 0%. Known negative misclassified rate: 100%.

Discussion: - another poor result for the SVMs; here, the SVMs simply assign all results as positive, hence the 100% known negative misclassified rate.

Again, a very low score was found for the secondary structure data - 55% of the training data was misclassified. This would be approximately the distribution expected if the data was sorted randomly.

4.4.14 SVMs, Conserved residue

Conserved residue experiment 1

Attributes used: - conserved residue (CON2).

Experiment description: - since it gave improved results for the decision trees, the scheme was used where the first 10 residues were considered individually, then the next 10 as a group, followed by groups of 20, up to 100, and then all residues over 100.

Results: -

CON2: 11.3%-20.4%, ave 18%, all 18%.

Known positive misclassified rate: 16.8%. Known negative misclassified rate: 9.8%.

Discussions: - this scheme gave an 18% error rate with the SVMs. Unusually, this method seems to give a bias towards false-negatives, with 16.8% of the known positive cases classified as negative and only 9.8% of the known negative cases classified as positive.

The conserved residue data gave poor results for SVMs, showing an 18% error across the training data. It is interesting to note that this scheme seems to favour the assignment of cases as negative, rather than positive as most of the other methods have. This could be because of the negative selection that was expected as a result of secondary structure deforming a potential calcium-binding motif.

4.4.15 SVMs, Amino Acid Hydrophobicity

Amino acid hydrophobicity experiment 1

Attributes used: - hydrophobicity (HYD).

Experiment description: - the hydrophobicity of each amino acid was used with the SVMs.

Results: -

HYD: 5.2%-21%, ave 26.8%, all 31.1%.

Known positive misclassified rate: 8.6%. Known negative misclassified rate: 42.6%.

Discussions: - this gives another high error rate; again, a bias is seen towards classification as positive.

Hydrophobicity shows a similar trend to the other SVM results, in being poorer than the decision-tree version, and gives an error of 26.1%.

4.4.16 Best support vector machine and verification

The best SVM result was the classification by amino acid type, with an error of just 5%. This was verified by the use of miss-one-out verifications. An excerpt from the actual SVM model can be seen in Figure 4.4.2-1. This is essentially just a string of numbers representing support vectors and is very difficult to interpret.

Figure 4.4.16-1 Excerpt from optimum SVM model file

SVM-light Version V6.02
0 # kernel type
3 # kernel parameter -d
1 # kernel parameter -g
1 # kernel parameter -s
1 # kernel parameter -r
empty# kernel parameter -u
10414 # highest feature index
58 # number of training documents
40 # number of support vectors plus 1
0.1795504 # threshold b, each following line is a SV (starting with alpha*y)

-1.10798953116771808550091241163 39e-05 10002:0 10003:14 10004:0 10005:14 10006:0 10007:14 10008:14 10009:57 10010:0 10011:14 10012:14 10013:57 10014:14 10102:100 10103:71 10104:43 10105:14 10106:0 10107:29 10108:0 10109:29 10110:14 10111:86 10112:0 10113:

-1.10798953116771808550091241163 39e-05 10002:95 10003:1 10004:1 10005:1 10006:0 10007:5 10008:98 10009:2 10010:85 10011:1 10012:2 10013:1 10014:2 10102:1 10103:95 10104:99 10105:99 10106:0 10107:92 10108:0 10109:8 10110:0 10111:9 10112:96 10113:98 10114:0

-1.10798953116771808550091241163 39e-05 10002:3 10003:3 10004:2 10005:7 10006:0 10007:7 10008:64 10009:80 10010:48 10011:12 10012:0 10013:4 10014:4 10102:95 10103:3 10104:2 10105:74 10106:2 10107:46 10108:0 10109:8 10110:0 10111:16 10112:96 10113:16 10114:

...

...

9.12678861533178990665671204141 06e-06 10002:54 10003:99 10004:0 10005:1 10006:0 10007:90 10008:36 10009:32 10010:97 10011:0 10012:59 10013:5 10014:0 10102:29 10103:0 10104:87 10105:2 10106:0 10107:4 10108:0 10109:31 10110:0 10111:97 10112:37 10113:94 10114:1 10202:13 10203:0 10204:1 10205:0 10206:100 10207:2 10208:64 10209:2 10210:3 10211:0 10212:0 10213:1 10214:99 10302:2 10303:0 10304:0 10305:97 10306:0 10307:1 10308:0 10309:10 10310:0 10311:2 10312:4 10313:0 10314:0 10402:1 10403:0 10404:11 10405:0 10406:0 10407:1 10408:0 10409:25 10410:0 10411:0 10412:0 10413:0 10414:0 #ID:1vjj:A:301-305

1.10798953116771808550091241163 39e-05 10002:50 10003:4 10004:20 10005:0 10006:0 10007:8 10008:23 10009:8 10010:8 10011:0 10012:15 10013:20 10014:4 10102:19 10103:31 10104:72 10105:72 10106:4 10107:85 10108:4 10109:85 10110:4 10111:19 10112:23 10113:16 10114:4 10202:8 10203:23 10204:0 10205:4 10206:96 10207:4 10208:73 10209:0 10210:88 10211:0 10212:12 10213:48 10214:92 10302:19 10303:35 10304:4 10305:16 10306:0 10307:4 10308:0 10309:8 10310:0 10311:4 10312:46 10313:16 10314:0 10402:4 10403:8 10404:4 10405:8 10406:0 10407:0 10408:0 10409:0 10410:0 10411:77 10412:4 10413:0 10414:0 #ID:1k1x:A:392-396

1.10798953116771808550091241163 39e-05 10002:1 10003:66 10004:0 10005:7 10006:1 10007:54 10008:15 10009:39 10010:94 10011:10 10012:85 10013:59 10014:19 10102:2 10103:13 10104:96 10105:87 10106:0 10107:30 10108:0 10109:20 10110:1 10111:3 10112:6 10113:7 10114:48 10202:1 10203:7 10204:1 10205:0 10206:99 10207:4 10208:85 10209:1 10210:6 10211:0 10212:1 10213:2 10214:23 10302:96 10303:13 10304:3 10305:4 10306:0 10307:6 10308:0 10309:6 10310:0 10311:1 10312:4 10313:28 10314:2 10402:0 10403:1 10404:0 10405:2 10406:0 10407:6 10408:0 10409:34 10410:0 10411:86 10412:5 10413:4 10414:7 #ID:1h1n:A:202-206

4.5 Section Conclusions

There were a number of different methodologies of AI tools considered, each with its own advantages and disadvantages. As the problem was essentially that of a classification of likely-to-bind against those-not likely-to-bind, classification and machine-learning methods seemed most appropriate - specifically used were SVMs and decision trees.

Decision trees and SVMs have both been successfully used to create rules for the classification of binding and non-binding DxDxDG motifs. The best decision tree used amino acid size attributes for its prediction. The trees used a threshold that varied by amino acid position to determine the % large or small down a column of the BLAST generated alignment.

Table 4.5-1 Table showing the error rates associated with the best decision tree and best SVM

	% of known positive training data misclassified	% of known negative training data misclassified	Error rate	False-positive rate (%)	False-negative rate (%)
Decision Trees					
Amino Acid Size (AASvDT)	1%	0.30%	0.10%	0.20%	1.70%
Amino Acid Type by group (ATGDT)	5.30%	1.60%	3.40%	0.90%	8.60%
Support Vector Machines					
Amino Acid Size (AASvSVM)	14%	25.40%	22.30%	14.30%	24.70%
Amino Acid Type by group (ATGSVM)	4.20%	8.30%	5.80%	4.60%	7.20%

Decision trees showed some bias towards classification of a particular case as negative over positive, whereas SVMs showed bias towards classification of a particular case as positive over negative. This is especially prevalent where the overall error rates were high. In the case of decision trees, this may be due to the

pruning stages of the process, designed to limit over-fitting to the data. The widening of the gap with higher error rates is probably due to the particular characteristics being poor at discriminating between positive and negative results and, therefore, simply classifying a case as the larger of the two data-sets.

The decision trees and SVMs both performed well in the task of classifying the data. The decision trees achieved lower error rates than the SVMs in every equivalent experiment attempted. However, more time was taken to optimise the attribute models for the decision trees than for the SVMs.

It was generally assumed that the model that best fitted the decision trees would also be most effective in the SVMs. This assumption seemed reasonable as the improvements were based on modelling the known biology; however, it may be useful to attempt a broader range of experiments to ensure these assumptions are correct.

As some of the attributes considered were predictions, with attached probabilities, probabilistic models might also be an area for further investigation into this problem.

Genetic algorithms may also prove to be a suitable methodology for analysis of the problem of classifying proteins as calcium-binding and non-calcium-binding. Genetic algorithms tend to show greater success when elements of the best solution have already been discovered. A greater understanding of this problem has been achieved through analysis of sequence characteristics and the application of SVMs and decision trees. This knowledge is a solid basis from which to further refine the solution, potentially by the incorporation of genetic algorithms.

4.6 Section Bibliography

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L.L. (2002). The Pfam protien families database. Nucleic Acids Research *30,* 276.

Ding, C.H.Q., and Dubchak., I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics *17,* 349.

Henderson, J., Salzberg, S., and Fasman, K.H. (1997). Finding Genes in DNA with a Hidden Markov Model. Journal of Computational Biology. *4,* 127.

Joachims, T. (2002). SVM light software package.*4.00,*

Joachims, T. (1998). Making Large-Scale SVM Learning Practical. In Advances in Kernel Methods - Support Vector Learning, Burges, C. J. C., and Smola, A. J. eds., MIT Press)

Kingsford, C., and Salzberg, S.L. (2008). What are decision trees? Nature Biotechnology *26,* 1011.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994). Hidden Markov Models in Computational Biology Applications to Protein Modeling. Journal of Molecular Biology *235,* 1501.

Mitchell, M. (1998). An introduction to genetic algorithms. Cambridge, Mass., London : MIT press.

Mitchell, T. (1997). Decision Tree Learning. In Machine Learning. The McGraw-Hill Companies, Inc. pp. 52-78.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. Morgan Kauffman.

Rabiner, L.R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE *77,* 257.

Savoie, C.J., Kamikawaji, N., and Sasazuki, T. (1999). Use of BONSAI decision trees for the identification of potentialMHC Class I peptide epitope motifs. Pacific Symposium on Biocomputing *4,* 182.

Smith, K. (2002). Hidden Markov Models in Bioinformatics with Application to Gene Finding in Human DNA. In Machine Learning Project,

Wang, X., Schroeder, D., Dobbs, D., and Honavar, V. (2003). Automated data-driven discovery of motif-based protein function classifiers. Information Sciences *155*, 1.

Winston, P. (1992). Learning by Building Identification Trees. In *Artificial Intelligence*, Addison-Wesley Publishing Company) pp. 423-442.

XinZhou, Wu, X.Y., Mao, K.Z., and Tuck, D.P. (2008). Fast Gene Selection for Microarray Data Using SVM-Based Evaluation Criterion. BIBM. '08.

Section 5 – Searching for Calcium-Binding Proteins

5. 5.1 Section Introduction

5.1.1 Section Overview

The focus of the project has been to enable the identification of new cases of calcium-binding proteins that contain the DxDxDG motif. AI algorithms have been used to created classification schemes that are able to distinguish between proteins that contain the DxDxDG motif that are likely or not likely to bind calcium or other metals.

Now an initial scan for these proteins in actual genomic data will take place. Two complete genomes have been selected: *E. coli* as a well-studied model organism, and *Bacillus coahuilensis,* for its interesting calcium metabolism.

The best predictor from the decision trees was use of amino acid size with a varying threshold, and, for SVMs, was use of amino acid groups. As such, these are the methods those that will be concentrated on for this section. As was discussed in the last chapter, a combination of these predictions is likely to produce the best true-positive set. The positive results from all four of the best algorithms have been used to predict occurrences of the DxDxDG motif most likely to bind metal. The predicted calcium-binding and predicted non-calcium-binding data-sets have been examined to see if they the exhibit structural and functional characteristics that are consistent with DxDxDG-type binding proteins.

5.1.2 Glossary and abbreviations

Bacillus coahuilensis - bacterium from a calcium-rich environment.

CLUSTALW – general purpose multiple sequence alignment program.

Escherichia coli – well-studied model organism.

FASTA - simple sequence format.

GDT_TS - Global Distance Test — Total Score describes percentage of well-modelled residues in the model with respect to the target.

LSQMAN - software that calculates RMSD between two sets of residues, using the coordinates from a PDB file.

M4T server - server that produces a structural prediction of a target protein using a template.

MODELLER – program used to produce homology models of protein tertiary structures.

Multiple mapping method (MMM) – algorithm to align a target sequence to a structural template.

MUSCLE - MUltiple Sequence Comparison by Log Expectation, another general purpose multiple sequence alignment program.

PSSCAN – SCANPROSITE, used to search a sequence for PROSITE patterns or profiles.

5.1.3 List of Figures

5.2 Preliminaries

5.2.1 *Escherichia coli* as a model organism

E. coli has long been used as a model organism for the study of bacteria. It is particularly appropriate for this purpose owing to its ease of culture; suitable media is readily available and ideal growth conditions are easily maintained. Also, *E. coli* is harmless if proper protocols are observed.

As a model organism, *E. coli* has been widely studied, and a great deal is known about its metabolism and biochemical pathways. It has also proven to be a useful tool in molecular biology, where it is used extensively as an expression vector. *E. coli's* genome was one of the first to be sequenced because of its importance and relatively short size.

The K-12 strain of *E. coli* is commonly used in microbiology and molecular biology research, both as a tool and as a model organism. This strain has a genome of 4,639,221 base pairs, which contains 4,288 protein-coding genes (Blattner et al., 1997).

5.2.2 Bacillus coahuilensis: ***an organism from a calcium-rich environment***

Bacillus coahuilensis was isolated in 2003 in the Cuatro Cienegas valley in Coahuila, Mexico. Biochemical tests have shown that there are differences in its carbon metabolism and fatty acid make-up, in comparison to related strains in its genus (Cerritos et al., 2008). This bacterium shows an interesting evolutionary history, caused by a changing local environment.

It is believed that the formation of a desiccation lagoon has resulted in the evolution of this species from a marine bacterium, in order to cope with the resultant low NaCl, high sulphate and high Ca^{2+} (Alcaraz et al., 2008). In particular, this bacterium shows evidence of adaptation to the limited levels of phosphorous in its surroundings, thought to have come about through horizontal gene transfer. It is possible that further modifications have occurred, to take advantage of, or to compensate for, the richness of calcium in its environment.

B. coahuilensis has a slightly shorter genome compared to *E. coli*, consisting of around 3,350,000 base pairs, these coding for 3,640 proteins. The conditions of the desiccation lagoon may have resulted in a greater proportion of these coding genes having calcium-binding properties compared to *E. coli*. Also, the isolated situation could mean that unique and interesting calcium-binding proteins have emerged in *Bacillus coahuilensis*.

5.2.3 M4T server

This is a structural prediction server that tries to predict by comparative modelling the complete 3D structure of a given sequence, including secondary structural elements and the overall tertiary structure.

The general process of comparative modelling consists of three major steps. First, a template or templates are identified; second, the template(s) are aligned to the target; finally, a model is produced of the target using the template(s). The M4T approach has been developed to minimise the errors in the first two steps of this process (Fernandez-Fuentes et al., 2007; Fernandez-Fuentes et al., 2007; Rykunov et al., 2009).

The M4T server first performs sequence analysis and searches for known structures with high sequence similarity to a target. These structures then act as templates. The need to determine the optimal set of templates, rather than simply finding as many as possible, increases the complexity of this process. The M4T server starts this process by performing three iterations of a PSI-BLAST search against the PDB, using an E-value cut-off of 0.0001. These results are then filtered, to ensure that there is the maximum sequence overlap between the target domains and the PDB chain. The resulting potential templates are then subjected to an iterative clustering process, by which they are selected to provide the optimal coverage in the fewest templates.

An iterative implementation of the Multiple Mapping Method (MMM) is used to align these template sequences to the target sequence, thus producing an MMM profile (Rai and Fiser, 2006).

The first step of this process involves using PSI-BLAST and BlastProfiler to produce profiles of both the target and templates (Rai et al., 2007).

These profiles are then aligned using CLUSTALW and MUSCLE to produce a clustalw_d_profile, clustalw_m_profile and muscle_profile. The profiles are then combined to produce the final MMM alignment (Edgar, 2004; Thompson, 1994).

The final structural model is then generated using MODELLER in its default settings. The inputs for this software are the templates and alignments from the previous stages (Fiser and Sali, 2003).

The accuracy of this method varies greatly, depending on the target, and has been shown to give a global distance test total score ranging between 40 and 95. This means that between 40% and 95% of the residues are in the correct position with respect to the experimentally determined target structure. It is difficult to be sure how accurate a particular prediction is without knowing the actual structure, although a high % sequence identity would suggest a more accurate prediction and a low sequence identity a poor prediction.

5.3 Methods

5.3.1 The genomes and other data

The complete genomes for *E. coli* (K-12) and *B. coahuilensis* were obtained from the NCBI Microbial genomes database. The genomes were downloaded as FastA files containing entries for each protein encoded in the genome.

These genomes were searched using PSSCAN for the DxDxDG motif and processed in a similar fashion to the negative data-sets used for AI training (see 3.3.7). A number of changes to the process used for the negative data-sets were necessary. It could not be assumed that all of the sequences to be classified would have associated structural data; because of this, the sequences could not be organised by super-family. Also, the sequence used for the data collection was obtained directly from the genome file - again, as it could not be assumed that PDB files would be available for these sequences. Otherwise, the process continued as described in section 3.3. First, the sequences that contained the PROSITE pattern D-x-[DNS]-x-[DNS] were found and put into an XML file. Then, the data associated with these sequences were collected from BLAST, PSIPRED and SABLE using automated scripts. From this data, the input files for the decision tree and SVM software were then created.

The proteins could then be classified as binding or non-binding by a Perl script that classifies using tree data, and by the SVM software; any annotations found in the genomic FastA files were also collected and stored with the results.

5.3.2 "CaBind" Scripting Outline

The **CaBind.pl** script takes the genome sequence as a list of protein sequences in FastA format. First, the **runpscan** subroutine uses PSSCAN to search for instances of the simplified DxDxDG motif, expressed as Dx[DSN]x[DSN]; all examples containing this motif were then compiled into a .smh file containing the name of the coding sequence, the position of the motif residues, and their identity. At the same time, individual FastA files for each sequence were produced. These files were fed into a modified version of **getfamilydata,** which, again, is responsible for collecting the characteristic data to allow the AI algorithms to identify those proteins likely to contain a functional form of the motif. The data-set obtained here was limited to those characteristics proven to have performed best in the previous analyses. This collection of sequence-derived data was then used, in conjunction with the .smh file, to classify the sequences based on the rules obtained by the AI tools. This classification was performed by **classifybydt** and **classifybysvm.** Each occurrence of the DxDxDG motif was sorted as binding or non-binding.

5.3.3 RPS-BLAST

RPS-BLAST was then also used on the genome to search for domains associated with known functional DxDxDG motifs. RPS-BLAST was used on the training data to pick out domains that show an association with functional DxDxDG motifs. All the proteins picked out by PSSCAN as having a DxDxDG motif were then also searched by RPS-BLAST. This allowed us to ensure that all the known DxDxDG proteins predicted by a motif search were classified correctly as calcium-binding, providing a simple method of verification of the results. This search used the CDD database (Marchler-Bauer et al., 2011), and included NCBI-curated domains and data imported from Pfam (Finn et al., 2008), SMART (Schultz et al., 1998), COG (Tatusov et al., 2001), PRK (Klimke et al., 2008), and TIGRFAM (Haft et al., 2003).

5.3.4 Secondary Structural Prediction

A secondary structural prediction was carried out by PSIPRED as part of the classification process of the gene products. This prediction followed the same format as the data collection steps carried out in section 3. The secondary structure predictions of the proteins, identified by the AI tools as having functional DxDxDG motifs, have been retrieved. In this section, these secondary structural predictions have been used in analysis of how well the AI tools have performed. See sections 3.2.3 and 3.3.4 for more details.

5.3.5 M4T Server

The top predictions from the AI tools, those that gave a positive result with all 4 methods, were submitted to the M4T server using the default settings. This is a quick and easy method for obtaining template-based structural predictions. Additionally, a sample of a similar number of the proteins, predicted as non-calcium-binding by the AI tools, were submitted to the M4T. These structural predictions have then been analyzed and compared at both the fold and motif level, in order to assess the reliability of the AI tool's predictions.

5.3.6 Calculation of RMS values

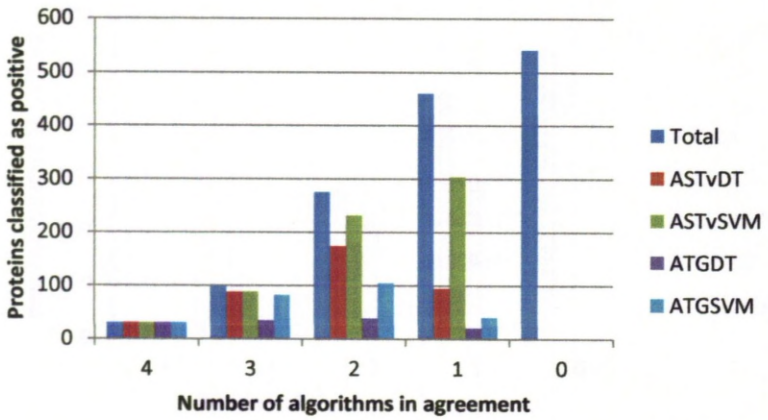
LSQMAN was used to calculate the RMSD of each of the structures predicted by the M4T server against the 1EXR DxDxDG motif. The proteins predicted as calcium-binding by all the AI methods, along with a random sample of similar size from the negative data, as detailed above, was used.

5.4 Results and discussions

5.4.1 *E. coli*

General Findings

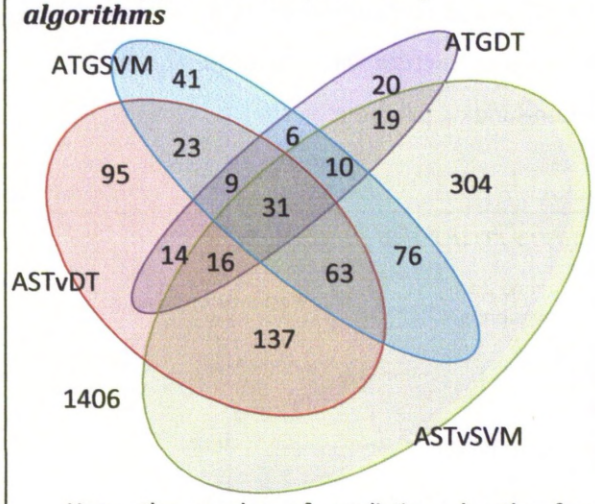
Figure 5.4.1-1, Chart showing the agreement between the four algorithms



There are 31 examples predicted as positive by all 4 algorithms in *E. coli.* Variable threshold amino acid size using decision trees (ASTvDT) and the amino acid group using SVMs (ATGSVM) show the best agreement. Variable threshold amino acid size using SVMs (ASTvSVM) over-estimates the number of binding examples, whereas amino acid group using decision trees under-estimates them (ATGDT).

As can be seen in figure 5.4.1-1, there were 31 proteins classified as binding by all four AI algorithms and 98 predicted by three of the algorithms. The variable threshold amino acid size using decision trees (ASTvDT) and the amino acid group using SVMs (ATGSVM) were seen previously to give the lowest error rates with the training data. Comparatively few were predicted by all four algorithms; it is likely that this figure has been limited by the propensity of the decision trees using amino acid groups to produce false-negatives.

Figure 5.4.1-2, Venn diagram showing overlap of predictions from the four algorithms



Here, the overlap of predictions by the four algorithms is represented schematically. The areas approximately represent the number of predictions made by each algorithm.

The best two algorithms showed a good amount of similarity, agreeing on the classification of 72.4% of the proteins from *E. coli*, although this may be less significant than it appears, as a large number of cases are predicted negative by all the algorithms. The algorithm that seems to over-estimate the most is the SVM using only amino acid size. The least likely to predict a positive result and, therefore, most likely to produce false-negatives, is decision trees using amino acid groups. This is not unexpected, as these two algorithms produce the highest % errors of the four, in the training data in section 4. Figure 5.4.1-2 shows the overlap of predictions between the four algorithms. ASTvSVM and ASTvDT have the largest amount of overlap; this is probably for two reasons: first, they both use the same data; second, ASTvSVM shows the least specificity and it shows a large overlap with all the other algorithms, in proportion to their own specificity.

Secondary Structural Predictions from PSI-PRED

Overall, the secondary structural prediction for the examples seen in *E. coli* is difficult to interpret; both the negative and positive data-sets show some overlapping of the predicted secondary structure with the motif, and certain similarities with the known calcium-binding proteins.

Figure 5.4.1-3, Key to secondary structural diagrams

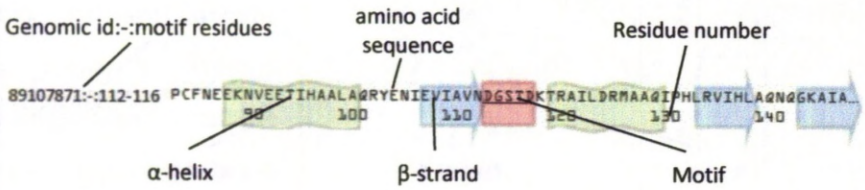
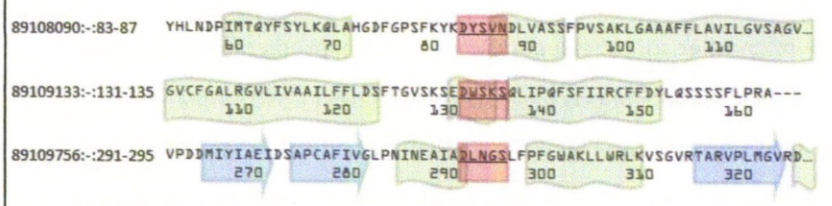


Figure 5.4.1-4, Predicted alpha helix-motif- alpha helix

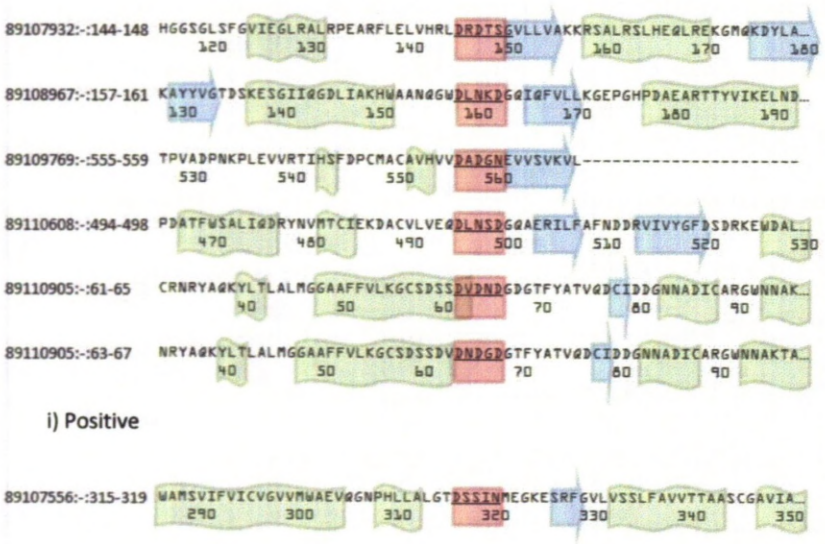
i) Positive



ii) Negative

The examples predicted as positive show a variety of different arrangements of secondary structure around the motif. Two examples show overlap of one or two residues; however, this does not necessarily mean that they can be discounted as true examples. The negative examples similarly show variation; however, one example shows extensive overlap of the secondary structure with the motif.

Figure 5.4.1-5, Predicted alpha helix – motif – beta strand



i) Positive

ii) Negative

In the examples seen here, there is no real difference immediately obvious between the predicted secondary structures of the positive and negative data-sets.

For ease of comparison, the secondary structural predictions have been separated out into classifications based on the nearest upstream and downstream elements to the motif. True SCOP classification, as used previously, would be impossible without more structural details provided by structural prediction or lab-based structural studies.

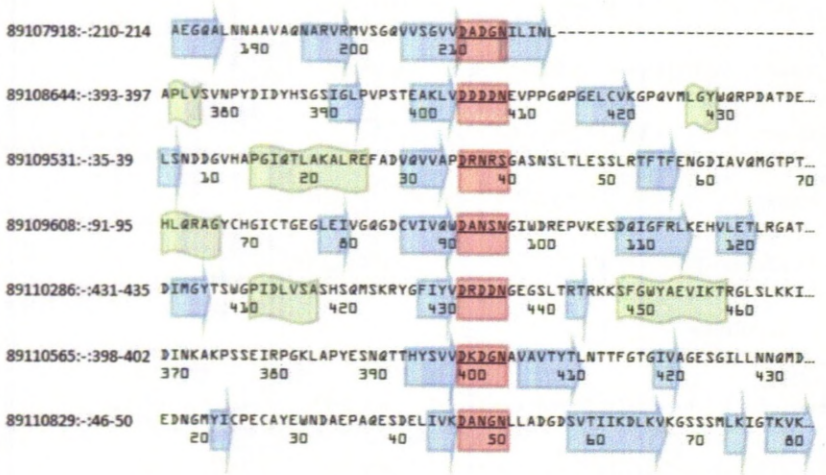
All possible combinations of secondary structure are seen in *E. coli*: alpha helix-motif-alpha helix (see Figure 5.4.1-4); alpha helix-motif-beta strand (see Figure 5.4.1-5); beta strand-motif-alpha helix (Figure 5.4.1-6); beta strand-motif-beta strand (Figure 5.4.1-7); and the motif flanked only on one side by either structural element (see Figure 5.4.1-8).

Figure 5.4.1-6, Predicted beta strand – motif – alpha helix

The predicted secondary structure of the positive data-set shows strong similarities to known DxDxDG proteins, with closely associated secondary structure that does not overlap the motif. The negative set shows less convincing secondary structure, with some overlapping, and one example where there is predicted beta strand in the middle of the motif.

There are 2 cases of secondary structure overlapping two or more residues into the motif region in the positive data-set, and 6 in the negative data-set. This shows a general trend in the negative data-set for secondary structural elements to be more likely to extend into the motif. This is likely to lead to restrictions in the spatial arrangement of the binding residues and may mean that they are incorrectly oriented for binding. Additionally, one negative example has a beta strand predicted within the motif. This also may lead to distortion of the spatial arrangement of the binding residues and reduced binding efficacy.

Figure 5.4.1-7, Predicted beta strand – motif – beta strand

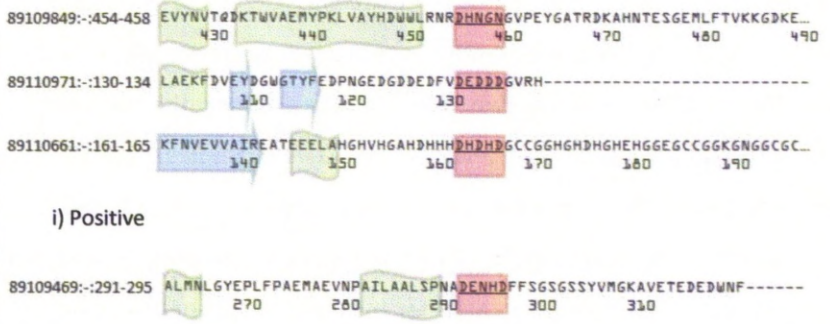


i) Positive

ii) Negative

None of the positive examples here show predicted secondary structure that might prevent binding. Two of the negative examples do show predicted secondary structure that might affect the motif's binding ability.

Figure 5.4.1-8, Predicted other arrangements



i) Positive

ii) Negative

Little difference is seen here between the positive and negative examples. One of the positive examples does not show an additional downstream ligand.

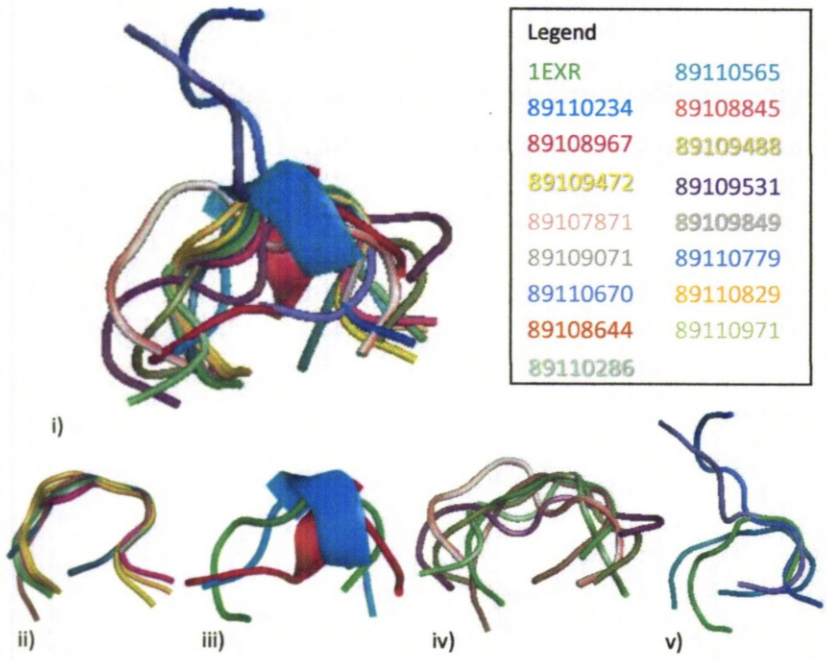
RPS Blast

None of the domains discovered in positive training data by RPS-BLAST were seen in either the positive or negative data-sets from *E. coli*.

Structural predictions from M4T and RMS calculations

There were only a few examples of structures successfully predicted by the M4T server; in other cases, no suitable template was found and the server returned no match. On initial observation, these predictions did not seem to show any dramatic differences between the positive and negative data-sets. However, it should be noted that a few of the positive examples, such as 89109472 and 89110565, showed a very similar structure to a true DxDxDG motif. None of the structures predicted by the negative examples showed this same characteristic structure. The RMSD for the orientations of the 3 binding residues from the predictions, and the 3 binding residues from the 1EXR motifs, are a little more convincing, however.

Figure 5.4.1-9, Comparison of top results to 1EXR



When all the top results are superimposed, as seen in i), the tangle does not appear to show any strong correlation to the typical DxDxDG motif (1EXR). When separated out, a stronger resemblance can be seen. All those in ii) show an identical bend to 1EXR. Those in iii) show some similarities, but a predicted helix caused the bend to deform. iv) shows examples that show a similar overall shape to a typical bend. v) shows a region of similarity, but then a fork that deviates from the typical binding loop.

The positive results were between 0.33Å and 2.14Å, with an average of 0.79Å. The negative results were between 0.71Å and 3.14Å, with an average of 2.24Å. This would seem to suggest that the predicted structures of the motifs from the positive data-set are indeed more similar to a verified DxDxDG motif than those predicted from the negative set.

Figure 5.4.1-10, Comparison of negative results to 1EXR

i) Comparison of negative results to 1EXR

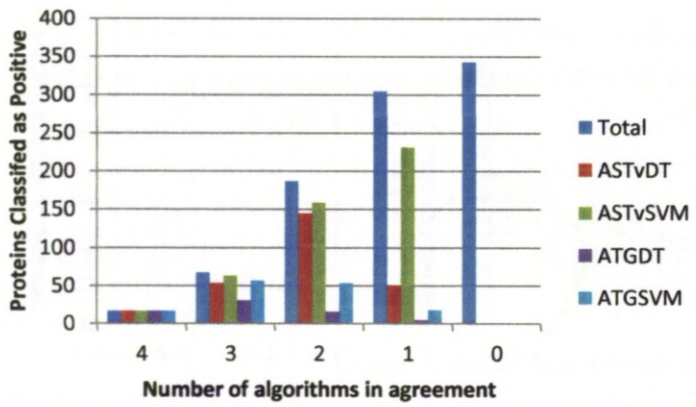
Legend
1EXR
89108108
89109223
89108294
89108862
89109415
89109469

Few, if any, similarities can be seen when the negative set is superimposed on the 1EXR binding motif; many of the motifs are much straighter. None show the identical bend seen in 1EXR.

5.4.2 Bacillus coahuilensis

General findings

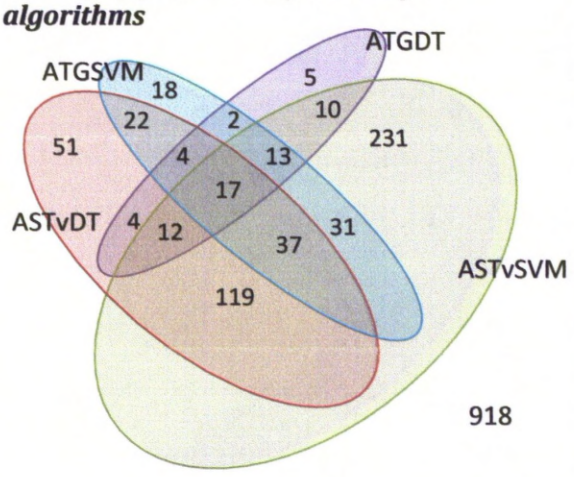
Figure 5.4.2-1, Chart showing the agreement between the four algorithms



There are 17 examples predicted as positive by all 4 algorithms in *B. coahuilensis*. The algorithm that seems to over-estimate by the most is the SVM using only amino acid size. The least likely to predict a positive result and, therefore, most likely to produce false-negatives, is decision trees using amino acid groups.

As can be seen in Figure 5.4.2-1, there were just 17 predicted to be DxDxDG motifs by all four AI algorithms; this is far fewer than found in *E. coli*. There were 68 predicted by three of the algorithms.

Figure 5.4.2-2, Venn diagram showing overlap of predictions from the four algorithms

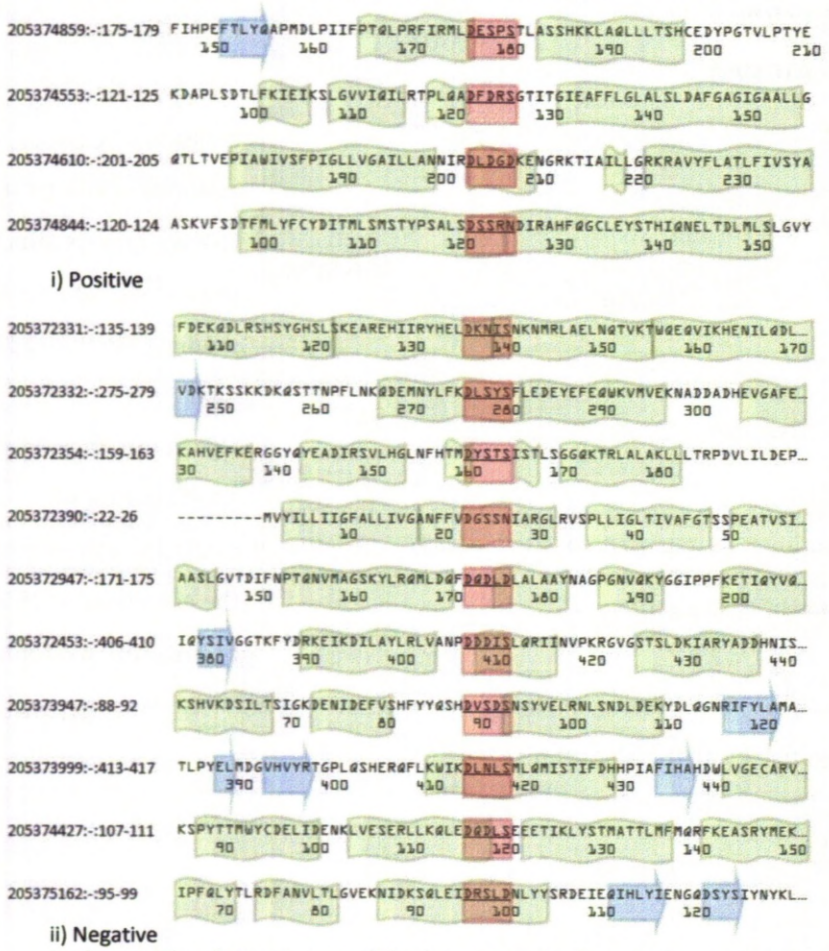


Here, the overlap of predictions by the four algorithms is represented schematically. The areas approximately represent the number of predictions made by each algorithm.

Again, comparatively few were predicted by all four algorithms. Interestingly, the decision trees using amino acid groups seem to have far less effect on the number of examples predicted in this case. Again, the best two algorithms showed a good amount of similarity, agreeing on the classification of 73% of the proteins from *B. coahuilensis*; however, again, it should be noted that many cases were predicted as negative by all the algorithms.

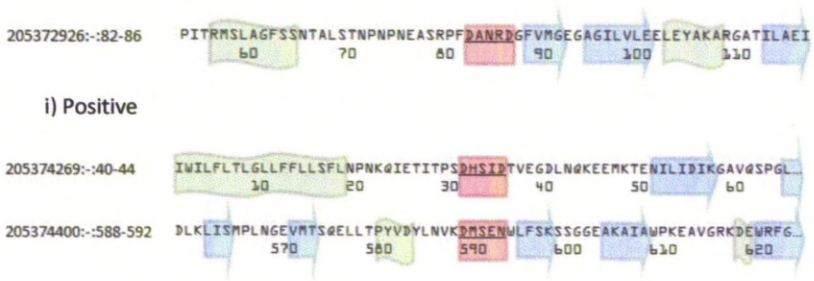
Secondary Structural Prediction from PSI-PRED

Figure 5.4.2-3, Predicted alpha helix-motif- alpha helix



In the positive data-set, two of the four examples have helix crossing the motif predicted by PSI-PRED. If these predictions are correct, it is possible that the AI tools have made false-positive predictions; secondary structure across the motif would likely disrupt binding. These secondary structural predictions, however, are not entirely reliable. PSI-PRED has predicted helix encroaching on the motif in virtually all of the examples from the negative data-set.

Figure 5.4.2-4, Predicted alpha helix – motif – beta strand

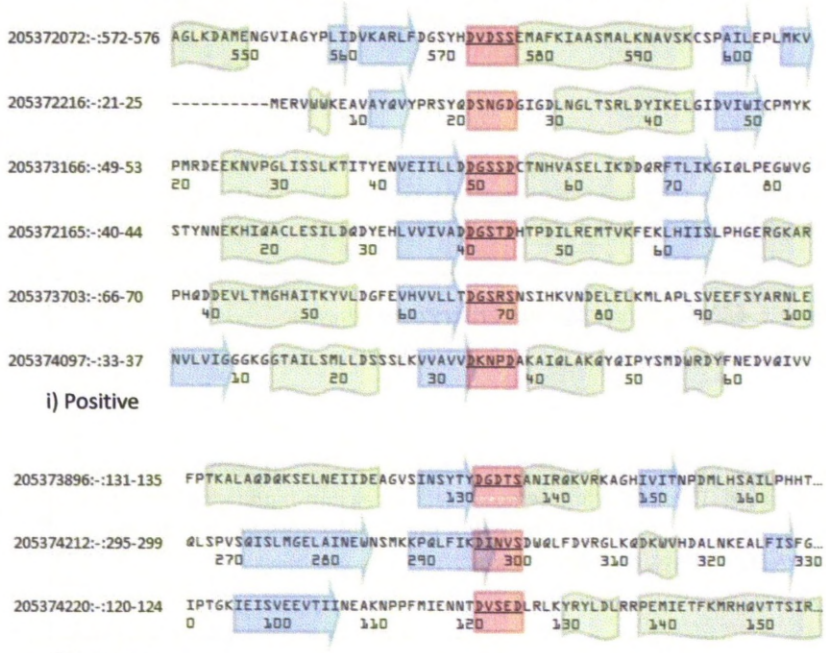


i) Positive

ii) Negative

The secondary structure predicted shows little difference between the positive and negative data-set. None of the examples seen can be discounted as able to bind.

Figure 5.4.2-5, Predicted beta strand – motif – alpha helix

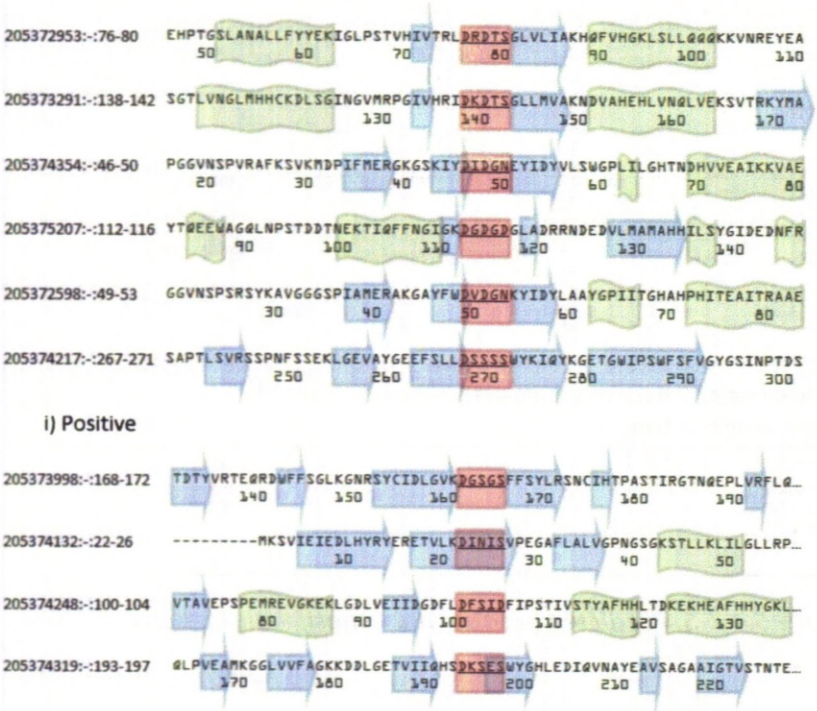


i) Positive

ii) Negative

All of the positive examples show a close association of the downstream alpha helix; this is often seen in the training data. None have predicted secondary structure encroaching on the motif. One of the negative examples shows beta strand overlapping with the motif.

Figure 5.4.2-6 Predicted beta strand – motif – beta strand



i) Positive

ii) Negative

Again, the secondary structural prediction shows little immediately obvious differences between the two sets of data. The negative set, however, has one case where beta strand is predicted to cross the motif.

Figure 5.4.2-7, Predicted other arrangements

i) Positive

No examples of the chain ending close to the motif were seen in the negative data-set. The one example predicted as calcium-binding seems plausible as there is no secondary structure predicted across the motif, and downstream D and E residues are present.

Again, there is a general trend in the negative data-set for secondary structural elements to be more likely to extend into the motif region, and for there to be secondary structural elements predicted within the motif.

However, there is even less differentiation between the negative and positive sets seen in the secondary structural prediction than in *E. coli*.

RPS Blast

Domains found using RPS BLAST from the sequences, predicted positive and negative by the AI tools, were cross-matched against domains found in the training data in an attempt to further verify their validity. The *B. coahuilensis* results, as can be seen in table 5.4.2-1 and 5.4.2-2, showed a number of domains, both in the data-set predicted as calcium-binding from the *B. coahuilensis*, and the known examples of calcium-binding motifs from the training data.

Table 5.4.2-1 Domains found by RPS-BLAST in both the training and positive data-set.

	Number of algorithms predicted +ve	PFAM ID	Domain
205372216:-:21-25	4	pfam00128	Alpha-amylase,
205374347:-:81-85	3	pfam00092	VWA, von Willebrand factor type A domain
205374347:-:982-986	3	pfam00092	VWA, von Willebrand factor type A domain
205372903:-:135-139	2	pfam00128	Alpha-amylase,
205372906:-:53-57	2	pfam00128	Alpha-amylase,
205372303:-:123-127	1	pfam01547	SBP_bac_1, Bacterial extracellular solute-binding
205372405:-:581-585	1	pfam01841	
205372604:-:108-112	1	pfam00496	SBP_bac_5, Bacterial extracellular solute-binding
205374346:-:151-155	1	pfam00092	VWA, von Willebrand factor type A domain
205371983:-:140-144	1	pfam00092	VWA, von Willebrand factor type A domain

This indicates that domains that occur with a known functional calcium-binding site are also found associated with binding sites predicted to bind calcium by the AI tools.

Table 5.4.2-2 Domains found by RPS-BLAST in both the training and negative data-set.

Negative E Coli Set	Number of algorithms predicted +ve	matching PFAM ID	
205371983:-:158-162	0	pfam00092	VWA, von Willebrand factor type A domain
205372906:-:282-286	0	pfam00128	Alpha-amylase,
205373998:-:168-172	0	pfam03065	Glyco_hydro_57, Glycosyl hydrolase family 57
205373999:-:413-417	0	pfam03065	Glyco_hydro_57, Glycosyl hydrolase family 57
205374346:-:293-297	0	pfam00092	VWA, von Willebrand factor type A domain

However, a number of domains predicted by the AI tools not to bind calcium also matched the same domains from the database. This could suggest that these examples have been falsely-predicted negative by the AI tools, or that these matching domains may have no association with the calcium-binding function. In the training data, these domains may simply co-occur with the calcium-binding motif.

Structural predictions from M4T and RMS calculations

Figure 5.4.2-8, Comparison of top results to 1EXR

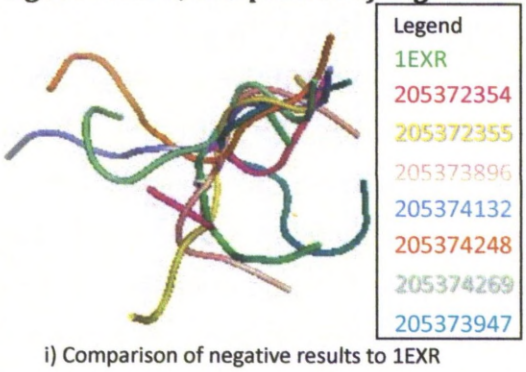
Legend
1EXR
205372072
205372165
205372216
205374354
205372955
205373291
205375207

All the bends are in the same general shape as the bend from 1EXR. Five of the seven examples seen here are almost identical.

i) Comparison of negative results to 1EXR

The *B. coahuiliensis* showed similar trends to the *E. coli.* There were no standout differences in the structures, but the statistical analysis showed that the positive data produced structural predictions more likely to be similar in spatial orientation to the true DxDxDG motif. The RMS range for the positive data was 0.52Å to 2.45Å, with an average of 0.88Å; for the negative data-set, the range was 0.6Å to 3.1Å, and the average was 1.4Å.

Figure 5.4.2-9, Comparison of negative results to 1EXR



Again, little similarity can be seen when the negative set is superimposed on the 1EXR binding motif; many of the motifs are in much straighter sections of the protein. None show a bend identical to the one seen in 1EXR.

i) Comparison of negative results to 1EXR

5.5 Section conclusions

The general findings from the different methods of prediction show that there is good agreement between the best methods from both algorithms. In both organisms, 73% of the proteins are classified the same by both decision trees and SVMs. When all four variations are used for classification, it is likely that a high false-negative rate is seen owing to the low sensitivity of the amino acid type by group decision tree (ATGDT).

The algorithms that produced the least agreement of the four were the SVMs using only amino acid size (ASTvSVM), and the decision trees using amino acid groups (ATGDT). The ASTvSVM seemed to have a tendency to over-estimate, and the DT to under-estimate. In section 4, this ASTvSMV showed an 11.5% false-positive rate, and an 8.5% false-negative rate, indicating that it was also prone to over-estimation in prediction of the training data. The ATGDT showed 0% false-positive rate, and a 3.4% false-negative rate, showing a tendency to under-estimate in the training data. These algorithms seem to show similar relative prediction rates as seen in the training data. This is a good indication that the training data were representative of the range of DxDxDG motifs, both binding and non-binding.

Little extra information could be confirmed by the PSI-PRED secondary structural prediction; both the motifs predicted as calcium-binding by the AI tools, and the randomly selected examples from those predicted as non-calcium-binding, show few significant differences. However, 37.5% of the negative examples from *E. coli* displayed secondary structure overlapping at least two residues into the motif, compared to only 6.5% in the positive set. Similarly, in *B. coahuilensis,* 55% of examples predicted as non-calcium-binding showed overlap, with only 17.6% in the positive set. This shows there is a greater propensity for secondary structure to be predicted by PSI-PRED within, or encroaching on, the motif in the data-set predicted as non-calcium-binding.

There were no cross-matching domains predicted by RPS BLAST in the *E. coli* and in the training data. In *B. coahuiliensis*, those domains that were matched by RPS BLAST in the calcium-binding data also cropped up in the non-calcium-binding data. It seems, therefore, that there is unlikely to be any association between the prediction of a certain domain from RPS BLAST and the occurrence of a true DxDxDG motif. RPS BLAST therefore proved to be of limited use in the verification of the methods used here.

The structural prediction performed by the M4T server on the proteins shows that predicted calcium-binding proteins seem to be far more likely to give a structure where binding is possible. The proteins predicted as non-calcium-binding are more likely to give a linear loop structure around the motif, whereas in the calcium-binding examples, the motif is more likely to appear in a bend. The RMS deviations of the structural predictions from the calcium-binding data-set and 1EXR average at 0.79Å and 0.88Å (for *E. coli* and *B. coahuilensis*, respectively). However, the RMS deviations of the structural predictions from the non-calcium-binding data-set and 1EXR set average at 2.24Å and 2.45Å. This means that the predicted structures of the motifs that the AI tools have predicted as binding calcium are more similar to 1EXR than those the AI tools have predicted will not bind calcium.

Overall, the methodologies used here do seem useful in highlighting proteins that are worthy of further study to verify their binding properties. This is a partial confirmation of the proposed theory that these proteins can be predicted using sequence data alone.

There are examples that seem unlikely to be true calcium-binding proteins, owing to secondary structure being predicted to overlap the motif and unfavourable predicted structures, even in the examples predicted by all four algorithms. In the negative set explored here, however, it does also seem likely that none of the examples seen are true binding proteins either. A more thorough analysis of the examples predicted as non-calcium-binding would allow

any probable false-negatives to be picked out. This could be done using structural prediction, and then matching using SPASM. It is possible that further refinements to these techniques, or a different combination of the characteristics and algorithms, may produce better selectivity and specificity in the classifications. In the next chapter, a more detailed analysis of a number of individual positive examples will take place.

5.6 Section Bibliography

Alcaraz, L.D., Olmedo, G., Bonilla, G., Cerritos, R., Hernández, G., Cruz, A., Ramírez, E., Putonti, C., Jiménez, B., Martínez, E., *et al.* (2008). The genome of Bacillus coahuilensis reveals adaptations essential for survival in the relic of an ancient marine environment. PNAS *105*, 5803.

Blattner, F.R., Plunkett, G.3., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., *et al.* (1997). The complete genome sequence of Escherichia coli K-12. Science. *277*, 1432.

Cerritos, R., Vinuesa, P., Eguiarte, L.E., Herrera-Estrella, L., Alcaraz-Peraza, L.D., Arvizu-Go´ mez, J.L., Olmedo, G., Ramirez, E., Siefert, J.L., and Souza, V. (2008). Bacillus coahuilensis sp. nov., a moderately halophilic species from a desiccation lagoon in the Cuatro Cie´negas Valley in Coahuila, Mexico. International Journal of Systematic and Evolutionary Microbiology *58*, 919.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 113.

Fernandez-Fuentes, N., Madrid-Aliste, C.J., Rai, B.K., Fajardo, J.E., and Fiser, A. (2007). M4T: a comparative protein structure modelling server. Nucleic Acids Research Web Server Issue 363.

Fernandez-Fuentes, N., Ra, B.K., Madrid-Aliste, C.J., Fajardo, J.E., and Fiser, A. (2007). Comparative protein structure modelling by combining multiple templates and optimizing sequence-to-structure alignments. Bioinformatics *23*, 2558.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., *et al.* (2008). The Pfam protein families database. Nucleic Acids Research Database Issue *36*, 281.

Fiser, A., and Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol *374*, 461.

Haft, D.H., Selengut, J.D., and White, O. (2003). The TIGRFAMs database of protein families. Nucleic Acids Research *31*, 371.

Klimke, W., Agarwala, R., Badretdin, A., Chetvernin, S., Ciufo, S., Fedorov, B., Kiryutin, B., O'Neill, K., Resch, W., Resenchuk, S., *et al.* (2008). The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Research *37*, D216.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., Deweese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., *et al.* (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research *39*, D225.

Rai, B.K., Madrid-Aliste, C.J., Fajardo, J.E., and Fiser, A. (2007). MMM: a sequence-to-structure alignment protocol. Bioinformatics *22*, 2691.

Rai, B.K., and Fiser, A. (2006). Multiple mapping method: a novel approach to the sequence-to-structure alignment problem in comparative protein structure modelling. Proteins *63,* 644.

Rykunov, D., Steinberger, E., Madrid-Aliste, C.J., and Fiser, A. (2009). Improved scoring function for comparative modelling using the M4T method. Journal of Structural and Functional Genomics *10,* 95.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998). SMART, a simple modular architecture research tool: Identification of signalling domains. PNAS *95,* 5857.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonina, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Research *29,* 22.

Thompson, J.D. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position- specific gap penalties and weight matrix choice. Nucleic Acids Research *22,* 4673.

Section 6 – Examples of Predicted Calcium Binding Proteins

6.1 Section Introduction

6.1.1 Section Overview

The process of classification using AI algorithms has been used successfully in the identification of potential calcium-binding proteins in two genomes of interest. These searches have identified a number of potentially novel and interesting new examples of calcium binding through the DxDxDG motif.

In this section, a closer look will be taken at some of the predicted calcium-binding proteins. We will look individually at the proteins that have been predicted by all four of the top AI algorithms to display calcium-binding characteristics. Our coverage here is also shortened to only those that produced a structural prediction from the M4T server; the extra information provided allows for comparison of the arrangements of the predicted binding motifs with true DxDxDG-type proteins.

Each of these proteins has been analysed with respect to its predicted function, predicted secondary structure and predicted overall structure.

6.1.2 List of Figures

6.2 Preliminaries

6.2.1 The relative reliability of structural predictions

The structural prediction methods used are based around differing methodologies, and aim to make predictions at differing levels of detail. PSIPRED uses a custom database to produce a model of the secondary structural elements associated with particular amino acids and peptide sequences. A profile of the target sequence is generated using PSI-BLAST, and a prediction is made using a feed-forward neural network. PSIPRED focuses on prediction of just the secondary structure. The M4T server differs in its approach, as each target is matched to sequences from the PDB. Matching sequences are used as templates, and used to generate alignments with the target. These alignments are then used to generate the structural model. M4T sever produces a model that encompasses both the secondary structure and the tertiary structure of a protein.

Both of these methodologies have their advantages and limitations.

PSIPRED uses a generalisation of sequence properties; therefore, it is able to produce a prediction, even if no structures of a high sequence similarity to the target are found.

PSIPRED can only produce a model of the secondary structure of a protein. This gives no indication of how closely the DxDxDG motifs, predicted as binding calcium by the AI tools, match the spatial configuration of a true calcium-binding motif.

The prediction accuracy is generally high and has few results of particularly poor accuracy, with a range between 60% to 95% of correct predictions and an average of around 82%.

M4T, on the other hand, is unable to successfully produce a model if no structure of high sequence similarity is found. This limitation results from the need for a reliable template for prediction. However, when M4T is given a suitable template, is able to produce a full structure that can be compared by means of RMSD calculations to a true calcium-binding motif.

The accuracy of the M4T server is dependent on the template used. If a template sequence shows 100% identity to the target, the structure may already have been solved experimentally, and the correct structure will be given. However, if there are no significantly similar sequences, then no structural model will be produced. The accuracy of the prediction can therefore vary markedly. This has been shown to be between 40% and 95% of the residues in the correct position, when compared to experimentally determined structures. This gives an average of around 78% of residues correctly positioned.

Even when the target and template display strong similarity, some regions of the protein may give a poor alignment. The model may be less accurate in these regions than in other parts of the protein.

The M4T and PSIPRED predictions may not always agree when it comes to the prediction of secondary structural features. It is difficult to assess from the results given which prediction method is likely to give the best secondary structural prediction; this is partially due to the methods using different measures to assess accuracy. The prediction that is more likely to be accurate varies with the particular sequence in question. Generally, PSIPRED is likely to give a more reliable prediction; however, if a good match is found to the target sequence, with a high level of identity surrounding the motif, the M4T server may produce the better model.

6.3 Methods

6.3.1 Identification of new calcium-binding proteins

As has been shown, the most reliable method in terms of sensitivity and selectivity for identifying calcium-binding proteins is to combine the results of the two best methods: the size-based method from the decision trees, and the type-based method from SVMs. These methods give 204 predictions in total. Tables 6.3.1-1 and 6.3.1-2 list the proteins predicted as binding calcium from *E. coli;* Tables 6.3.1-3 and 6.3.1-4 list the proteins predicted as binding calcium from *B. coahuilensis*.

Table 6.3.1-1, List of proteins from E. coli predicted as calcium binding by all of the four best AI algorithms.

89110608:-:494-498	hypothetical protein	89107873:-:615-619	predicted outer membrane protein
89110670:-:585-589	protein chain elongation factor EF-G	89107872:-:358-362	predicted enzyme associated with biofilm formation
89109873:-:73-77	predicted inner membrane protein	89108644:-:393-397	acyl-CoA synthetase
89109037:-:92-96	hybrid sensory kinase in two-component regulatory system with RcsB and YojN	89108967:-:157-161	methyl-galactoside transporter subunit
89110905:-:63-67	hypothetical protein	89109531:-:35-39	broad specificity 5'(3')-nucleotidase and polyphosphatase
89110829:-:46-50	predicted phosphonate	89109872:-:77-81	predicted inner membrane protein
89110905:-:61-65	hypothetical protein	89108845:-:172-176	N-(5'-phospho-L-ribosyl-formimino)-5-amino-1-(5'-phosphoribosyl)-4-imidazolecarboxamide isomerase
89109608:-:91-95	predicted inner membrane protein	89107062:-:175-179	tetraacyldisaccharide-1-P synthase
89110779:-:687-691	ATPase and DNA damage recognition protein of nucleotide excision repair excinuclease UvrABC	89109800:-:490-494	DNA topoisomerase IV, subunit B
89109472:-:145-149	glycine betaine transporter subunit	89109849:-:454-458	predicted glycosyl hydrolase
89109488:-:237-241	membrane-bound lytic murein transglycosylase B	89109769:-:555-559	hydrogenase 2, large subunit
89110971:-:130-134	hypothetical protein	89110661:-:161-165	FKBP-type peptidyl prolyl cis-trans isomerase
89107932:-:144-148	23S rRNA pseudouridylate synthase	89110565:-:398-402	gamma-glutamyltranspeptidase
89109071:-:48-52	undecaprenyl phosphate-L-Ara4FN transferase	89110234:-:359-363	transcription termination factor
89107918:-:210-214	assembly protein for flagellar basal-body periplasmic P ring	89110286:-:431-435	cryptic phospho-beta-glucosidase B
89107871:-:112-116	predicted glycosyl transferase		

Table showing the proteins predicted by all four methods as binding calcium, with associated genomic coordinates. The examples that produced a structural prediction using M4T are highlighted in blue.

Table 6.3.1-2, List of proteins from E. coli predicted as calcium binding by both of the two best AI algorithms.

89110213:-:25-29	frataxin, iron-binding and oxidizing protein	89106938:-:706-710	exported protein required for envelope biosynthesis and integrity
89109939:-:241-245	phosphoglucosamine mutase	89106946:-:223-227	L-arabinose isomerase
89108846:-:222-226	imidazole glycerol phosphate synthase, catalytic subunit with	89108110:-:341-345	component I of anthranilate synthase
89107867:-:192-196	hypothetical protein	89106947:-:96-100	L-ribulokinase
89108847:-:80-84	fused phosphoribosyl-AMP cyclohydrolase and phosphoribosyl-ATP pyrophosphatase	89107605:-:132-136	3-deoxy-D-arabino-heptulosonate-7-phosphate synthase, phenylalanine repressible
89110396:-:43-47	predicted glycosyl transferase	89108172:-:176-180	L-Ala-D/L-Glu epimerase
89108717:-:26-30	hypothetical protein	89108804:-:264-268	adhesin
89107006:-:85-89	hypoxanthine phosphoribosyltransferase	89107172:-:187-191	attaching and effacing protein, pathogenesis factor
89110206:-:248-252	DNA-dependent ATPase I and	89109848:-:101-105	hypothetical protein
89110314:-:498-502	DNA gyrase, subunit B	89109925:-:230-234	hypothetical protein
89107030:-:31-35	fused glycosyl transferase and transpeptidase	89107979:-:146-150	tRNA (5-methylaminomethyl-2-thiouridylate)- methyltransferase
89110160:-:10-14	fused DNA-binding response regulator in two-component regulatory system with GlnL,	89110493:-:53-57	predicted DNA-binding response regulator in two-component regulatory system
89107124:-:161-165	predicted DNA-binding transcriptional regulator	89107555:-:550-554	potassium translocating ATPase, subunit B
89110540:-:22-26	periplasmic protein	89107977:-:270-274	adenylosuccinate lyase
89110099:-:204-208	1,4-dihydroxy-2-naphthoate	89110885:-:34-38	oligoribonuclease
89110213:-:23-27	frataxin, iron-binding and oxidizing protein	89107185:-:264-268	betaine aldehyde dehydrogenase, NAD-dependent
89109089:-:132-136	predicted peptidase	89109670:-:350-354	peptide chain release factor RF-2
89108362:-:174-178	altronate oxidoreductase, NAD-dependent	89106898:-:477-481	chaperone Hsp70, co-chaperone with DnaJ
89108559:-:105-109	threonyl-tRNA synthetase	89107578:-:294-298	citrate synthase
89110155:-:268-272	predicted sugar phosphate	89107298:-:61-65	protoheme IX farnesyltransferase
89108623:-:256-260	hypothetical protein	89106888:-:183-187	threonine synthase
89108004:-:35-39	5-methylcytosine-specific restriction endonuclease B	89109240:-:103-107	predicted oxidoredutase, sulfate metabolism protein
89109454:-:267-271	succinate-semialdehyde	89108711:-:230-234	predicted methyltransferase
89110951:-:113-117	fructose-1,6-bisphosphatase I	89107852:-:161-165	hypothetical protein
89111089:-:22-26	2-deoxyribose-5-phosphate aldolase, NAD(P)-linked	89109355:-:425-429	hypothetical protein
89109800:-:471-475	DNA topoisomerase IV, subunit B	89108026:-:37-41	hypothetical protein
89110989:-:59-63	hypothetical protein	89110360:-:442-446	ATP-dependent DNA helicase
89109070:-:107-111	uridine 5'-(beta-1-threo-pentapyranosyl-4-ulose diphosphate) aminotransferase,	89110034:-:12-16	fused DNA-binding response regulator in two-component regulatory system with ZraS
89109576:-:126-130	predicted flavoprotein	89110524:-:203-207	hypothetical protein
89109671:-:79-83	ssDNA exonuclease, 5'-_ 3'-specific	89110314:-:500-504	DNA gyrase, subunit B
89109761:-:202-206	hypothetical protein with nucleoside triphosphate hydrolase	89110829:-:88-92	predicted phosphonate metabolizing protein
89110074:-:254-258	phosphoenolpyruvate carboxylase	89109441:-:81-85	predicted antirestriction protein
89110516:-:197-201	predicted SAM-dependent	89108456:-:100-104	predicted outer membrane porin

89109088:-:222-226	hypothetical protein	89109330:-:41-45	hypothetical protein
89109074:-:520-524	4-amino-4-deoxy-L-arabinose transferase	89108910:-:44-48	galactitol-specific enzyme IIA component of PTS
89108868:-:245-249	phosphomannomutase	89107265:-:89-93	manno(fructo)kinase
89110996:-:128-132	hypothetical protein	89107071:-:16-20	hypothetical protein
89108879:-:42-46	predicted glycosyl transferase	89107064:-:401-405	DNA polymerase III alpha subunit
89108159:-:53-57	predicted oxidoreductase, Zn-dependent and NAD(P)-binding	89108850:-:243-247	gluconate-6-phosphate dehydrogenase, decarboxylating
89109178:-:180-184	hypothetical protein	89107514:-:40-44	hypothetical protein
89109837:-:566-570	RNA polymerase, sigma 70 (sigma D) factor	89107772:-:290-294	Involved in chromosome partioning, Ca2+ binding protein
89111099:-:113-117	fused predicted transporter subunits and ATP-binding components of ABC superfamily	89109682:-:625-629	glycine decarboxylase, PLP-dependent, subunit (protein P) of glycine cleavage complex
89107105:-:89-93	predicted antitoxin of the YafO-YafN toxin-antitoxin system	89109946:-:50-54	GTPase involved in cell partioning and DNA repair
89110074:-:252-256	phosphoenolpyruvate carboxylase	89110051:-:460-464	RNA polymerase, beta prime
89109170:-:41-45	bactoprenol glucosyl transferase	89107033:-:239-243	iron-hydroxamate transporter
89109160:-:139-143	phosphohistidine phosphatase	89109610:-:85-89	thymidylate synthetase
89107758:-:305-309	5-enolpyruvylshikimate-3-phosphate synthetase	89109881:-:77-81	pyruvate formate-lyase 4/2-ketobutyrate formate-lyase
89107610:-:98-102	UDP-galactose-4-epimerase		

Table showing all the calcium-binding proteins predicted by the two best algorithms for classification. Each is labelled by genomic identifier and residue position.

Table 6.3.1-3, List of proteins from B. coahuilensis predicted as calcium binding by all of the four best AI algorithms.

205372953:-:76-80	pseudouridine synthetase	**205374859:-:175-179**	hypothetical protein Bcoam_17645
205373703:-:66-70	hypothetical protein Bcoam_10760	205372216:-:21-25	alpha-glucosidase
205372165:-:40-44	glycosyltransferase	**205372598:-:49-53**	glutamate-1-semialdehyde aminotransferase
205374217:-:267-271	cell-wall amidase lytH precursor	205372072:-:572-576	elongation factor G
205375207:-:112-116	cell wall endopeptidase	205373291:-:138-142	YlyB
205374553:-:121-125	hypothetical protein Bcoam_15851	**205373166:-:49-53**	putative glycosyl transferase
205374844:-:120-124	hypothetical protein Bcoam_17570	**205372926:-:82-86**	3-oxoacyl-(acyl carrier protein) synthase II
205374354:-:46-50	glutamate-1-semialdehyde aminotransferase	**205374097:-:33-37**	PAS modulated sigma54 specific transcriptional regulator
205374610:-:201-205	1,4-dihydroxy-2-naphthoate octaprenyltransferase		

Table showing all the calcium-binding proteins predicted by all of the four best algorithms for classification. Each is labelled by genomic identifier and residue position. The examples that produced a structural prediction using M4T are shown in blue.

Table 6.3.1-4, List of proteins from B. coahuilensis predicted as calcium binding by both of the two best AI algorithms.

205373142:-:258-262	aldehyde dehydrogenase	205372620:-:103-107	hypothetical protein Bcoam_04005
205373489:-:156-160	outer spore coat protein	205375560:-:223-227	serine protease Do
205374758:-:241-245	glucosylceramidase	205374419:-:68-72	hypothetical protein Bcoam_15026
205374683:-:329-333	hypothetical protein Bcoam_16630	205372779:-:37-41	YcgL
205372564:-:51-55	DNA recombinase, putative	205372115:-:163-167	Tn7-like transposition protein D
205374609:-:410-414	MenF	205372109:-:240-244	phosphoglucosamine mutase
205373044:-:85-89	protoheme IX farnesyltransferase	205372092:-:71-75	2-C-methyl-D-erythritol 2,4-
205374679:-:218-222	1-pyrroline-5-carboxylate	205372912:-:95-99	hydrolase
205374026:-:62-66	required for exogenous DNA-binding	205372557:-:65-69	hypothetical protein Bcoam_03590
205374211:-:88-92	single-stranded-DNA-specific	205375597:-:189-193	YceG
205373609:-:77-81	hypothetical protein Bcoam_10165	205375051:-:135-139	hypothetical protein Bcoam_18687
205373638:-:503-507	DNA topoisomerase IV subunit B	205372277:-:199-203	FAD-dependent pyridine nucleotide-disulphide oxidoreductase
205374227:-:147-151	tRNA	205375160:-:230-234	binding-protein-dependent transport systems inner membrane
205372775:-:276-280	aldehyde dehydrogenase	205372019:-:10-14	hypothetical protein Bcoam_00636
205374343:-:112-116	hypothetical protein Bcoam_14554	205372076:-:449-453	DNA-directed RNA polymerase subunit beta'
205375016:-:28-32	hypothetical protein Bcoam_14554	205373874:-:32-36	hypothetical protein Bcoam_11760
205373339:-:104-108	3-oxoacyl-[acyl-carrier-protein	205372937:-:264-268	hypothetical protein Bcoam_05990
205374242:-:134-138	uridine kinase	205372937:-:266-270	hypothetical protein Bcoam_05990
205374470:-:173-177	nicotinate	205374014:-:116-120	ROK family glucokinase
205375066:-:311-315	hypothetical protein Bcoam_18782	205372749:-:85-89	BNR repeat-containing protein
205375115:-:51-55	imidazole glycerol phosphate synthase subunit HisF	205373782:-:274-278	resB protein
205375116:-:74-78	bifunctional phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase protein	205374187:-:99-103	hypothetical protein Bcoam_13694
205374934:-:14-18	fosfomycin resistance protein FosB	205372091:-:68-72	glutamyl-tRNA synthetase
205373799:-:87-91	hypothetical protein Bcoam_11325	205374047:-:731-735	ribonucleotide-diphosphate reductase subunit alpha
205374972:-:42-46	hypothetical protein Bcoam_18252	205373423:-:965-969	DNA polymerase III PolC
205375366:-:26-30	hypothetical protein Bcoam_18252	205372788:-:212-216	pyridine nucleotide-disulphide oxidoreductase dimerisation region
205373489:-:154-158	outer spore coat protein	205375253:-:79-83	hypothetical protein Bcoam_19814
205373489:-:161-165	outer spore coat protein	205374558:-:85-89	S-adenosylmethionine
205375127:-:168-172	ATP-dependent Clp protease proteolytic subunit	205374532:-:380-384	DNA polymerase III DnaE
205375260:-:51-55	hypothetical protein Bcoam_19849 [Bacillus coahuilensis m4-4]	205374809:-:303-307	methyl-accepting chemotaxis protein
205374014:-:39-43	ROK family glucokinase		

Table showing all the calcium-binding proteins predicted by the two best algorithms for classification. Each is labelled by genomic identifier and residue position.

However, 204 motifs are far too many results to cover in a detailed analysis. Therefore, only the results picked by all four of the methods analysed in the previous section have been used here. Of these 48 examples, only those 23 that successfully produced a structural prediction using the M4T server were used to undertake a more detailed analysis, including a look at the predicted structure around the motif, a literature search and RMSD calculations.

6.3.2 Analysis methods

RPS-BLAST

The RPS-BLAST search used in section 5 has been further analysed, relating the identified motifs and domains to individual protein chains. Please see section 5 for more details.

PSIPRED secondary structural predictions

The PSIPRED predictions used in section 5 have been further analysed, relating predictions to individual protein chains. Please see sections 3.3.4, 5.3.1, 5.4.1 and 5.4.2 for more details.

M4T Server

The M4T server predictions used in section 5 have been further analysed, relating the predictions to the individual protein chains. Additionally, these predictions have been visualised and compared to the example DxDxDG motifs shown in section 2.4; the most similar example was then chosen for the Figures presented. Please see sections 5.2.3, 5.3.5, 5.4.1 and 5.4.2 for more details.

RMSD

The RMSD calculations used in section 5 have been used to compare the similarity of the binding motif of 1EXR to individual protein chains. The alpha carbons of the two the target and 1EXR motif are compared. Please see sections 5.3.6, 5.4.1 and 5.4.2 for more details.

Protein Data Bank

A keyword and sequence search of the PDB was used to ascertain if any of the proteins under investigation, or any related proteins, were present in the PDB.

BLAST

A simple BLAST search was performed using each sequence to check for related proteins with known calcium-binding function or any other connections with calcium binding.

Additional Literature Search

The annotation in the genome file was used with a general literature search using PubMed, and a web search using Bing, to identify any other publications that may associate each class of protein with a calcium-binding function.

Presentation of results

For each protein investigated, a diagram has been constructed, showing: i) the secondary structure predicted by PSIPRED, ii) the 3D structure of the surrounding chain as predicted by the M4T server, and iii) a superimposition of the predicted bend of the motif by the M4T server, with the bend of a known example of a DxDxDG-type binding-protein. The example of a known DxDxDG-type binding-protein was selected based on its having a similar secondary structure to the predicted calcium-binding protein. RPS-BLAST, RMSD, PDB, BLAST and literature searches are referred to in the text where appropriate.

6.4 Results and discussions

6.4.1 E. coli gene 89107871, predicted glycosyl transferase

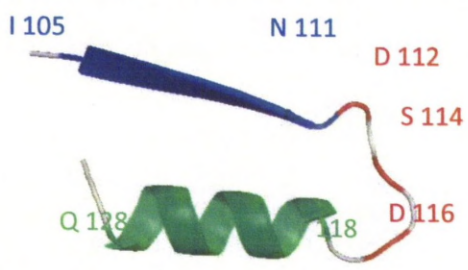
89107871 is listed in the genome as a predicted glycosyl transferase. This is confirmed by RPS-BLAST, as this predicts a glycosyl transferase domain (Pfam 00535) with a high degree of confidence. This particular protein was not present in the PDB; additionally, there were no other glycosyl transferases, from *E. coli,* in the PDB. However, glycosyl transferases from other organisms were identified both in the PDB and by BLAST. The annotations of these results showed no indication of known calcium-binding function. Additionally, no literature was found relating glycosyl transferases with calcium binding. As a result of these searches, we can be confident that this class of proteins has not been previously identified as calcium binding.

Looking at the predicted secondary structure and predicted folding (see Figure 6.4.1-1 i), both seem to show some similarities to the known DxDxDG protein 2b2x, as seen in Figure 2.4.1-3 ii. However, the predicted arrangement of the binding residues (see Figure 6.4.1-1 iii) displays a structure that deviates from the typical arrangement. The M4T server selected the 3bcv protein, a putative glycosyl transferase from *Bacteroides fragilis*, as a basis for this prediction.

The general shape of this prediction is similar to 2b2x. Also, there may be some plasticity in the orientation of the residues, meaning a closer agreement could be formed. It is possible that, in the presence of calcium, this protein may undergo conformational change, and the arrangement may then more closely resemble a typical DxDxDG motif. There is also a downstream D, at position 123, that could act as an additional ligating amino acid.

Figure 6.4.1-1, 89107871

i) The predicted secondary structure from PSIPRED shows an upstream beta strand (blue) and a downstream alpha helix (green), both closely associated with the motif (red), typical of an a/b protein, such as 2b2x.



ii) The spacing and layout of the predicted secondary structure around the site of the motif displays similarities to known a/b proteins, as seen in Figure 2.4.1-3 ii.

iii) Binding resides of 89107871 compared to 2b2x. It is possible that this could be an unbound conformation and the residues fold more tightly in the presence of Ca^{2+}.

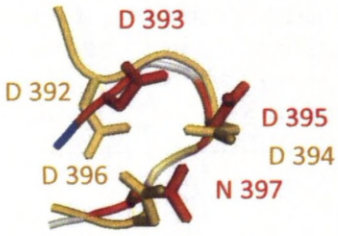
6.4.2 E. coli gene 89108644, acyl-CoA synthetase

RPS-BLAST finds an AMP-binding enzyme domain (Pfam 00501) with an expect score of e-106. The genome lists 89106844 as acyl-CoA synthetase; the literature search did not find any association between this and any known calcium-binding proteins or previously identified DxDxDG proteins. When used to search the PDB, this sequence could not be found; however, another predicted acyl-CoA synthetase (3DMY) from *E. coli* was found. The BLAST search also found many examples of acyl-CoA synthetases. The annotations gave no indication of a known calcium-binding function.

Figure 6.4.2-1, 89108644

i) The binding loop seems to be more extended than in other strand-loop-strand proteins, with extra residues downstream of the motif.

ii) The extra residues in the loop are evident in the perpendicular orientation of the motif to the sheet.



iii) Binding residues of 89108644 compared to 1k1x.

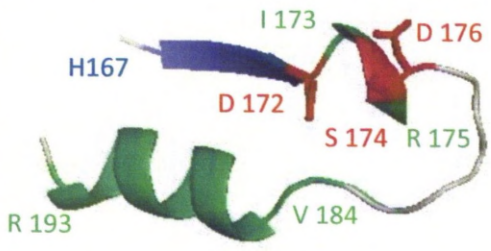
The folding of the secondary structure is very similar to the all-beta protein examples, seen in Figure 2.4.2-1; however, the loop is extended, leading to a variation in the folding and orientation in relation to the beta strands (see Figure 6.4.2-1 ii). The M4T structural prediction uses 1pg4, 2p2m and 3c5e, which are all acetyl coA synthetases, and 2d1s, which is a luciferase. The layout of the binding residues is very similar to the typical DxDxDG proteins, and they do seem to be grouped around a point; however, the orientation of the residues is a little different (see Figure 6.4.2-1 iii). Looking at the sequence, there are 3 potential downstream ligating amino acids at positions 398, 423 and 426. This looks like a likely candidate for binding, and certainly could be investigated further.

6.4.3 E. coli gene 89108845, N-(5'-phospho-L-ribosyl-formimino)-5-amino-1-(5'-phosphoribosyl)-4-imidazolecarboxamide isomerase

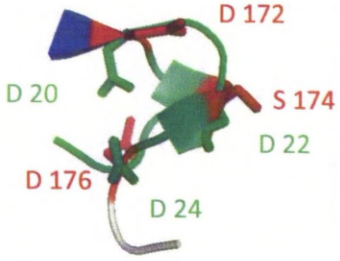
89108845 is recorded in the genome as an N-(5'-phospho-L-ribosyl-formimino)-5-amino-1- (5'-phosphoribosyl)-4-imidazolecarboxamide isomerase. A search of the PDB did find imidazolecarboxamide isomerases from *E. coli,* but not this specific sequence.

Figure 6.4.3-1, 89108845

i) Again, the loop seems to be extended in comparison to typical DxDxDG motifs. However, the layout is still plausible, and there are a number of possible additional downstream ligands to assist in binding.



ii) The structural prediction server, M4T, disagrees with the secondary structural prediction by PSI-PRED, both in the length of the beta strand and the prediction of an additional helix within the motif.



iii) Binding resides of 89108644 compared to 1EXR. Notice that the predicted helix in the motif distorts the binding arrangement.

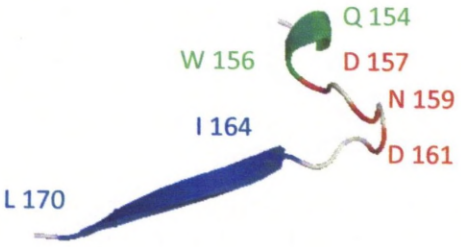
The secondary structural elements up- and down-stream of the motif are consistent with the examples seen previously, especially relative to the 2b2x protein. Also, there are a couple of Es in the downstream sequence that could act as additional ligating amino acids.

In the M4T structural prediction, a secondary structural element has been predicted within the motif; this helix was not present in the secondary structural prediction by PSI-PRED. Despite this additional helix, the layout of the binding residues is still very similar to the canonical EF-hand arrangement, if perhaps a little distorted in the orientation, owing to the restriction of the helix. The disagreement between these two predictions is interesting, as it is likely that, if the helix was found not to be present, an even better match would be seen.

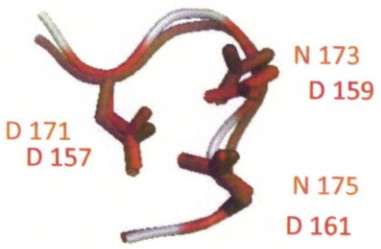
6.4.4 E. coli gene 89108967, methyl-galactoside transporter subunit

Figure 6.4.4-1, 89108967

i) A long helix followed by an extended loop and a short beta strand is seen here; this is similar to the secondary structure of 1j1n (see Figure 2.4.1-3 v).



ii) Disagreement is seen on the secondary structural prediction; however, both are plausible arrangements for calcium binding. This structure shows similarity with the motif seen in 1gca (see Figure 2.4.1-3 iv).



iii) An almost perfect match is seen between the binding residues of 89108967 and 1j1n.

RPS-BLAST identifies a periplasmic binding protein domain; this is often associated with sugar transport, and is consistent with the genomic listing as a methyl-galactoside transporter. Methyl-galactoside transporter is a known calcium-binding protein, with the calcium acting to stabilise the folding of the protein (Herman et al., 2005), and is present in the PDB listed as 2gpf (Vyas et al., 1988). This example was identified by SPASM; however, no SCOP classification was found. Therefore, this protein was not included in the training data.

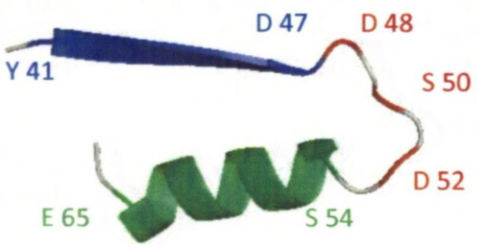
As would be expected from a known calcium-binding protein, the predicted secondary structure and arrangement of the motif is very similar to examples we have already seen, such as 1gca and 1j1n (see Figure 6.4.4-1).

6.4.5 E. coli gene 89109071 undecaprenyl phosphate-L-Ara4FN transferase

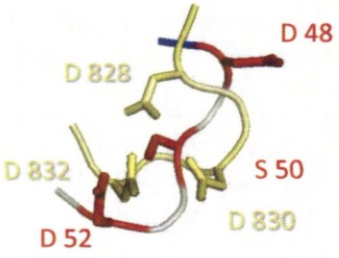
RPS-BLAST identifies a glycosyl transferase domain in this protein with an expect score of 3e-25. This protein is listed in the genome as undecaprenyl phosphate-L-Ara4FN transferase. As mentioned above, there were no examples of glycosyl transferase found in the PDB from *E. coli,* but a number from other species.

Figure 6.4.5-1, 89109071

i) The predicted secondary structure shows an upstream beta strand (blue) and a downstream alpha helix (green), both closely associated with the motif (red), typical of an a/b protein such as 2b2x.



ii) This shows a similar layout to the examples of a/b proteins seen in Figure 2.4.1-3, especially 2b2x.



iii) Binding resides of 89107871 compared to 1ux6. This seems to be a more open version of the motif seen in 89107871 (Figure 6.4.1-1).

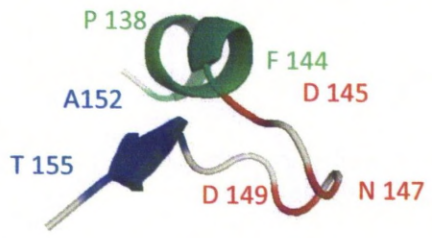
Looking at the predicted secondary structure and predicted folding (see Figure 6.4.5-1 i & ii), both seem to show similarities to the known DxDxDG protein, 2b2x, as seen in Figure 2.4.1-3 ii. The structural prediction was again based on 3bcv, and therefore does not show much similarity to 2b2x or, as shown here, 1ux6, in its arrangement. Also, there are two glutamic acids downstream of the motif, one of which may assist in binding. Again, it is possible that the arrangement seen is that of an unbound conformation and may more closely resemble the typical DxDxDG motif when bound to calcium.

6.4.6 *E. coli gene 89109472, glycine betaine transporter subunit*

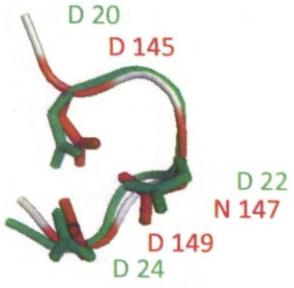
89109472 is listed in the genome as a glycine betaine transport subunit. RPS-BLAST identifies a Pfam 04069 domain; this is an ABC-type glycine substrate binding domain. This protein is present in the PDB as 1r9l, and does seem to contain a functional motif. However, the PDB file lists the ligand as unknown.

Figure 6.4.6-1, 89109472

i) PSIPRED predicts a series of short alpha helices, then the motif, then an extended loop before another alpha helix. The predicted downstream helix may not interact with the motif as it is too far downstream.



ii) The structural prediction server, however, predicts a beta strand downstream of the motif and also much closer than the helix predicted by PSIPRED.



iii) The motif of 89109472 shows a very strong similarity to the motif found in 1EXR.

This was identified by SPASM as a potential binding motif, but may have been overlooked due to the unknown ligand.

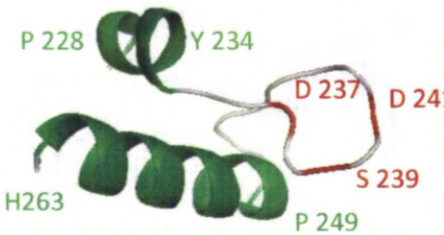
It is of note that there is a good amount of disagreement between the secondary structural prediction PSIPRED and the structural prediction from the M4T server around the region of the motif (see Figure 6.1.6-1 i & ii). The structural "prediction" is based on 1r9l, and therefore is the actual structure. PSIPRED predicts a series of helices with a distal downstream helix unlikely to interact with the motif. The structure shows an upstream helix and a much closer downstream beta strand. As would be expected, the binding motif shows a very similar arrangement to 1EXR.

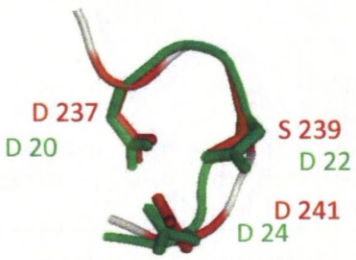
6.4.7 E. coli gene 89109488, membrane-bound lytic murein transglycosylase B

The genome file lists 89109488 as membrane bound lytic murein transglycosylase B. However, RPS-BLAST did not identify any motifs with a high enough score. This is a known calcium-binding protein, with an identified DxDxDG motif present. This protein is present in the PDB; however, an alternate PDB file, 1qus, was identified by SPASM. 1qus was identified as having nothing bound and was discarded. The bound form is also present in the PDB, but was missed because a non-redundant set was used.

Figure 6.4.7-1, 89109488

i) The predicted secondary structure is a similar arrangement to a classic EF hand, with a short upstream helix and a longer downstream helix.





ii) The methods of structural prediction agree on the positioning of both secondary structural elements, but do not predict an EF-hand shape. However, the fold shows similarities to the 1lwj motif.

iii) The arrangement of the binding residues of 89109488 is almost identical to 1EXR, with just a slight deviation between S239 and D241.

The secondary structure predicted here shows a slightly extended binding loop. However, other than a slight deviation between the second S239 and third D241 binding residues, the arrangement is very similar to the typical 1EXR motif.

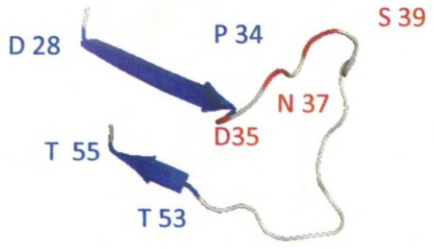
6.4.8 E. coli gene 89109531, broad specificity 5'(3')-nucleotidase and polyphosphatase

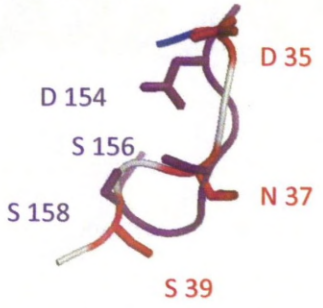
This is listed in the genome file as broad specificity 5'(3')-nucleotidase and polyphosphatase. An RPS-BLAST search reveals that a Pfam 01975 motif, SurE, Survival protein SurE, is found within this protein, with an expect score of 2e-79.

From the literature, and a search of the PDB, there does not appear to be any connection between these protein classes and calcium-binding function or any structures of this protein.

Figure 6.4.8-1, 89109531

i) The downstream secondary structure is far away from the motif, which may mean that the motif may not be held in the correct conformation.





ii) The long loop between the two beta strands does not seem to be beneficial for proper folding of the motif. However, some flexibility may mean a conformational change takes place in the presence of a Ca2+ ion.

iii) The binding residues of 89109531 are in similar positions to 2b2x; however, the rest of the backbone deviates significantly.

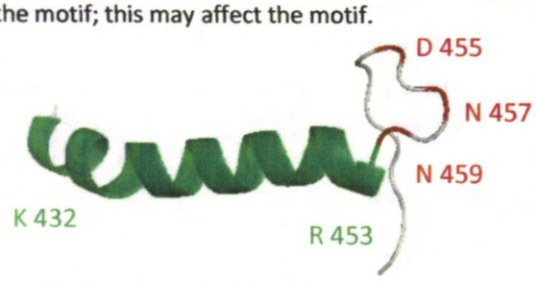
The long binding loop does not conform to the typical layout and results in a flattened grouping of the binding residues. However, increased flexibility may mean that the loop is able to bend into a more suitable arrangement. It may also allow for the downstream E to come in and ligate the metal.

6.4.9 E. coli gene 89109849, predicted glycosyl hydrolase

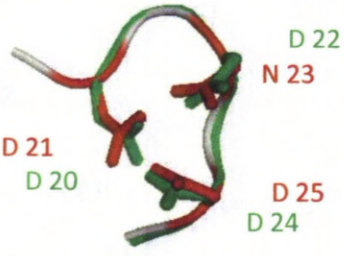
The genome file lists this as a predicted glycosyl hydrolase. This class of protein is known to have an association with calcium binding. This protein is seen in the PDB as 3c67, and has identified as calcium binding through a DxDxDG motif (Kurakata et al., 2008). RPS-BLAST finds a trehalase motif, a mannose oligosaccharide glycosidase motif and a bacterial alpha-L-rhamnosidase; all these predictions show low certainty.

Figure 6.4.9-1, 89109849

i) There does not seem to be any secondary structural feature downstream of the motif; this may affect the motif.



ii) The lack of downstream secondary structure does not seem to affect the arrangement of the residues in the motif.



iii) The predicted binding site of 89109849 is almost identical to that of 1EXR.

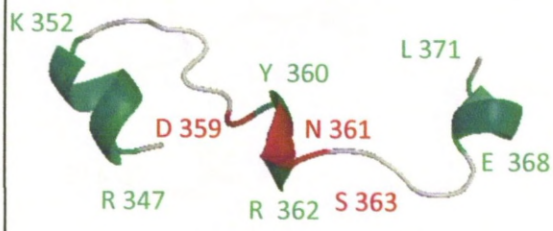
Despite the termination of the chain soon after the motif, the overall structure and arrangement of the motif is very similar to that of 1EXR. There is even a potential downstream ligating amino acid seen a few places away from the motif.

6.4.10 E. coli gene 89110234, transcription termination factor

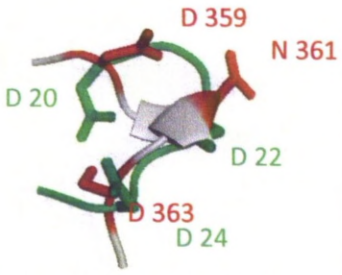
89110234 is listed in the genome file as a transcription termination factor. The motifs found by RPS-BLAST confirm this: an ATP synthase, Rho termination factor RNA binding, Rho termination factor N-terminal domain, and a ribonuclease B OB domain are all identified with low expect scores. This transcription termination factor from *E. coli* is present in the PDB as 1pvo. However, this example was not identified by SPASM, most probably because the alpha helix distorts the motif significantly.

Figure 6.4.10- 1, 89110234

i) Two short helices are seen flanking the motif at a distance of around 6 residues. This gives a long binding loop and may mean a poor arrangement of the residues.



ii) The structural prediction and secondary structural prediction disagree on the presence of a helix within the motif.



iii) Binding residues of 89100234 compared to 1EXR. Notice that the predicted helix in the motif distorts the binding arrangement.

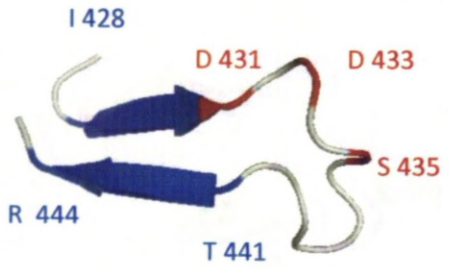
The M4T server has predicted a secondary structural element within the motif; this, however, was not present in the secondary structural prediction. Despite this additional helix, the layout of the binding residues is still similar to a typical DxDxDG motif, such as 1EXR, if perhaps a little distorted, owing to the restriction caused by the helix. It is likely, therefore, that if the helix was found not to be present, an even better match would be seen. However, as the structural prediction is based on the actual structure, the helix is present. Despite the distortion, the motif still gives a low RMSD of 0.33Å when compared with the alpha carbons of motif of 1EXR.

6.4.11 E. coli gene 89110286, cryptic phospho-beta-glucosidase-B

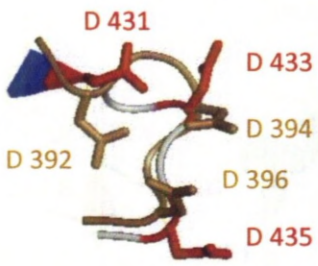
89110286 is listed as a cryptic phospho-beta-glucosidase B in the genome file. RPS-BLAST identifies a glycosyl hydrolase. This protein is not seen in the PDB; however, there are phospho-beta-glucosidases present from other species. However, the annotations and literature search did not show any associations with calcium binding.

Figure 6.4.11-1, 89110286

i) The downstream element seen here is a little further from the motif than typical; a downstream E is seen that could act as an extra ligand.



ii) The structure predicted here does not seem to promote potential binding. The downstream loop may interfere with the binding residues.



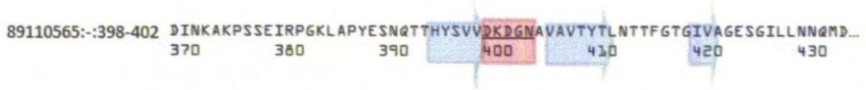
iii) The arrangement of the binding residues of 89110286 is similar to that of 1k1x.

The structural predictions seen in Figure 6.4.11-1 show there is a similar general arrangement of the binding residues to a typical calcium-binding motif. Furthermore, their side chains could orient to allow binding. The surrounding structure may not favour calcium binding, as the downstream loop may block access to the binding residues. The RMSD against 1EXR is 0.8Å; this is in the middle of the range of those predicted to be calcium binding. 1pbg, 1wcg, 2pbg and 1ug6 were used by the M4T server to produce the predicted structure.

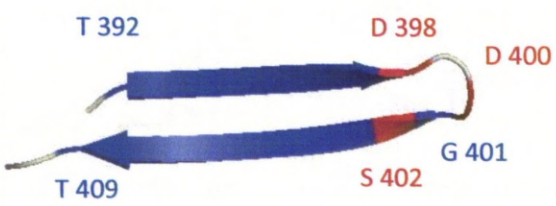
6.4.12 E. coli gene 89110565, gamma-glutamyltranspeptidase

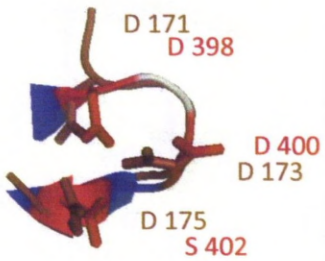
89110565 is listed as gamma-glutamyltranspeptidase in the genome file. This has many structures from diverse species in the PDB; however, no calcium is seen to be bound in the version from *E. coli* (Wada et al., 2008). This protein was not picked out by SPASM as a potential calcium-binding protein. However, it has been shown that gamma-glutamyltranspeptidase is regulated by calcium in some cases (Raulf et al., 1985) .

Figure 6.4.12- 1, 89110565



i) The closely associated, but not overlapping, structure seen here is typical of true DxDxDG binding motifs.





ii) This tight bend is very similar to a number of false-positive examples from SPASM. However, it still might be a true example in this case, especially considering the looser arrangement predicted by PSIPRED.

iii) Although the fold is tight, the binding residues of 890110565 are still in a very similar orientation to 1j1n.

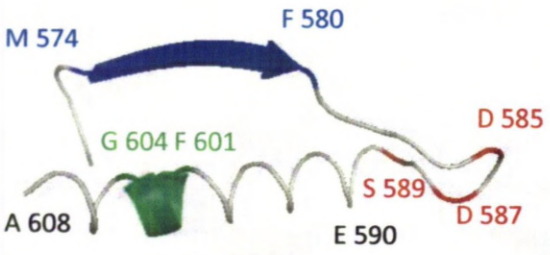
There is some disagreement between the PSIPRED and M4T server predictions here, regarding the start positions of the beta strands. Despite this, the M4T server predicts a binding motif of very similar shape to that of 1j1n, which shows a similar arrangement of secondary structure and is known to bind calcium (see Figure 6.4.12-1 iii). When compared to 1EXR, an RMSD of 0.5Å is calculated. It is likely that, if the PSIPRED prediction is correct, an even better match would be seen.

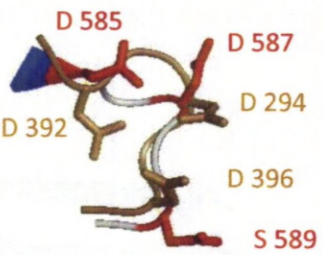
6.4.13 E. coli gene 89110670, protein chain elongation factor EF-G

89110670 is listed as protein chain elongation factor EF-G in the genome file. This protein is seen in the PDB as 2rdo; the annotation does not suggest any associated calcium-binding function.

Figure 6.4.13-1, 89110670

i) The secondary structure prediction shows a good match to typical calcium-binding proteins; the upstream elements, however, may be slightly distal.





ii) It seems likely that errors have been made in the marking of secondary structure here, as an obvious helix can be seen starting immediately downstream of the motif.

iii) Some similarity in general shape is seen between the binding residues of 8911067 and 1k1x.

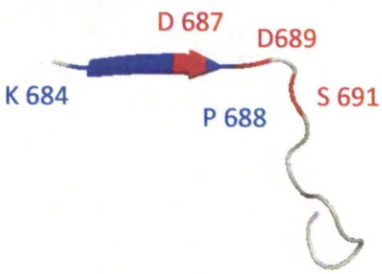
The secondary structural prediction from PSIPRED shows a good agreement with the typical DxDxDG motif; the M4T server disagrees with this prediction. However, an error seems to have occurred, as a helix is seen, but has not been labelled as such. It is possible that the arrangement of the motif would allow for a binding function, but is by no means convincing evidence, especially considering the RMSD of 1.4 Å when compared to 1EXR. The additional E just after the motif may also assist in binding.

6.4.14 E. coli gene 89110779 ATPase and DNA damage recognition protein of nucleotide excision repair excinuclease UvrABC

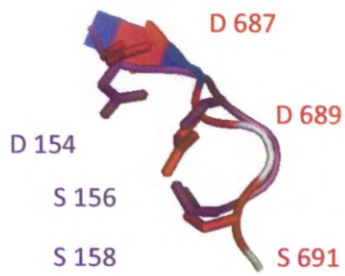
89110779 is listed in the genome file as ATPase and DNA-damage recognition protein of nucleotide excision repair excinuclease UvrABC. A number of proteins from the UvrABC system can be identified in the PDB. A number of these, for example, 2r6f and 2ygr, display Zn^{2+} binding. This may be an indication of a potential DxDxDG, as Zn^{2+} may be an alternate ligand.

Figure 6.3.14- 1, 89110779

i) The secondary structure prediction seen here and the structural prediction disagree greatly; it is, therefore, difficult to draw any solid conclusions from these data.



ii) This motif is seen near the end of the peptide chain. The arrangement seen here looks unlikely to correspond to a true DxDxDG motif; however, the binding residues show some correlation to known calcium-binding proteins.



iii) Binding residues of 89110779 compared to 2b2x. It is possible that this could be an unbound conformation, and that the residues fold more tightly in the presence of Ca^{2+}.

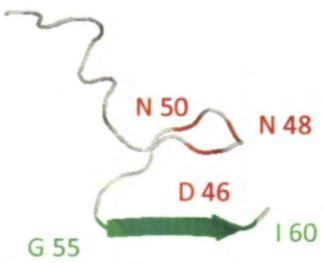
The predicted arrangement of the binding residues is in close agreement with the known DxDxDG motif, 2b2x, and the RMSD, when compared to 1EXR, is 2.1Å. Not much can be concluded from the PSIPRED secondary structural prediction (see Figure 6.3.1-14 i) compared with the M4T structural server prediction (see Figure 6.3.1-14 ii) as they disagree greatly. The P in the position usually occupied by a G is a possible reason for this distortion. 2r6f was used as a template for the structural prediction.

6.4.15 E. coli gene 89110829, predicted phosphonate metabolising protein

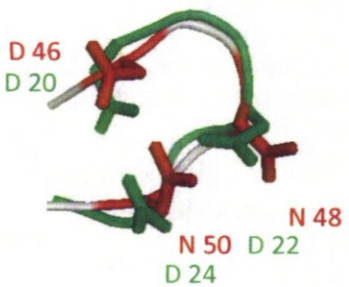
The genome file identifies this protein as a predicted phosphonate metabolising protein. No structure for this was found in the PDB; however, 2akl shows significant similarity, and is known to bind Zn^{2+}.

Figure 6.4.15-1, 89110829

i) The secondary structural prediction is in line with the prediction of a calcium-binding protein. Potential secondary ligands, downstream of the motif, can also be seen.



ii) This structural prediction shows little correspondence with the secondary structure predicted.



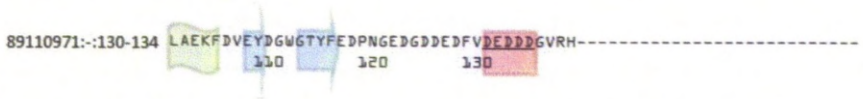
iii) The predicted arrangement of the motif from 89110829 very closely resembles the typical EF hand, 1EXR.

Again, there are differences between the PSIPRED and M4T predictions. However, the motif does seem to show good similarity between the typical arrangement of 1EXR and a low RMSD of 0.8Å, and there are a number of aspartic acids downstream that could assist in metal binding.

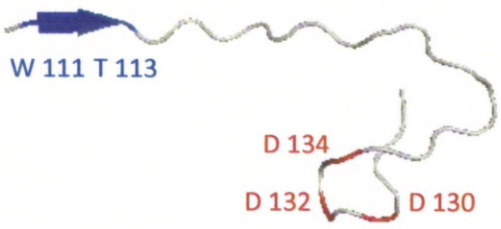
6.4.16 E. coli gene 89110971, hypothetical protein

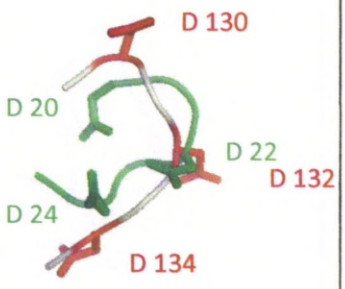
This is a hypothetical protein, indicating it may not actually be expressed in *E. coli*. The BLAST hit with the greatest significance is a provisional RNase inhibitor protein, with an e-value of 5.55e-42.

Figure 6.4.16-1, 89110971



i) The motif appears to be far away from any secondary structures, and resides at the end of the chain.





ii) This motif is seen right at the end of the chain with little secondary structure close by. It seems unlikely this protein is a true example of a DxDxDG from this prediction.

iii) The binding residues of 89110971 show almost no correlation to the 1EXR motif.

The termination of the protein chain soon after the motif may affect its ability to keep its shape, although this might be stabilised at binding. There is, additionally, no secondary structure predicted close to the motif upstream, which may add to the instability of the chain. The predicted arrangement of the motif appears very open, and does not show similarity to the motif of 1EXR; the RMSD is 1.4Å. Additionally, there is no potential downstream extra ligating amino acid. However, the presence of multiple Ds and Es upstream of the residue may provide the potential for an upstream ligand.

6.4.17 E. coli results summary

Table 6.4.17-1, E. coli results summary

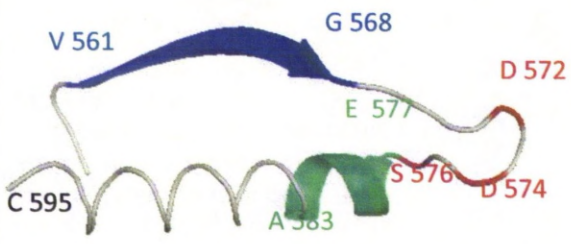
Genome LD ID and residue position	Identified as in genome	PSIPRED Secondary structure overlap	RMSD using M4T prediction and 1exr	M4T template	Potential ligand	Comments	Confidence in calcium-binding prediction
89110670:-:585-589	protein chain elongation factor EF-G	None	1.4	2rdo	614	Present in PDB as 2rdo	
89110829:-:46-50	predicted phosphonate	None	0.8	2akl	64, 66,72		High
89110779:-:687-691	ATPase and DNA damage recognition protein of nucleotide excision repair excinuclease UvrABC	None	2.1	2r6f	713	Secondary structure predicted overlapping motif by M4T	Intermediate
89109472:-:145-149	glycine betaine transporter subunit	None	0.1	1r9l	153, 164, 175	Present in PDB as 1r9l with unknown ligand	Very High
89109488:-:237-241	membrane-bound lytic murein transglycosylase B	None	0.2	1qus	248, 251, 269		Very High
89110971:-:130-134	hypothetical protein	None	1.4	1nxi	None	potential upstream ligand?	Good
89109071:-:48-52	undecaprenyl phosphate-L-Ara4FN transferase	None	2.4	3bcv	60, 65		Intermediate
89107871:-:112-116	glycosyl transferase	None	2	3bcv	123	Not previously identified as calcium-binding	Intermediate
89108644:-:393-397	acyl-CoA synthetase	None	0.7	1pg4, 2p2m, 3c5e, 2d1s	398, 423, 426	Not previously identified as calcium-binding	High
89108967:-:157-161	methyl-galactoside transporter subunit	None	0.2	2fvy	172, 177, 179	Known calcium-binding protien	Confirmed
89109531:-:35-39	broad specificity 5'(3')-nucleotidase and polyphosphatase	None	2.3	1l5x, 1j9l	48, 57, 60	flexiblity of binding loop may lead to tighter fit on binding	Intermediate
89108845:-:172-176	imidazolecarboxamide isomeraise	None	0.1	1qo2, 2vep	188, 189		Very High
89109849:-:454-458	predicted glycosyl hydrolase	None	0.1	3d3i	462, 468, 475		Very High
89110565:-:398-402	gamma-glutamyltranspeptida	None	0.5	2e0w	422, 432		High
89110234:-:359-363	transcription termination factor	None	0.4	1pv4	368, 369, 375, 376		High
89110286:-:431-435	cryptic phospho-beta-glucosidase B	None	0.8	1pbg, 1wcg, 1ug6, 2pbg	437, 453		High

6.4.18 B. coahuilensis gene 205372072, elongation factor G

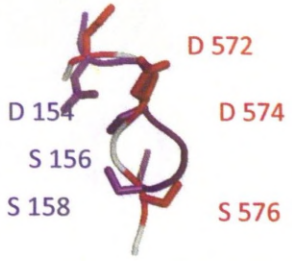
205372072 is listed as a elongation factor G in the genome file, and elongation factor in *E. coli* (89110670) has also been predicted to contain a DxDxDG motif. RPS-BLAST identifies four motifs in this protein, all associated with elongation factors. These include: elongation factor Tu GTP binding domain (Pfam 00009); elongation factor Tu domain 2 (Pfam 03144); elongation factor G, domain IV (Pfam 03764); and elongation factor G C-terminus (Pfam 00679). This protein is not seen in the PDB from *B. coahuilensis*, although it is present from other species. There has been evidence that elongation factors in other species may be regulated by calcium, but not through direct binding (Liu and Gelli, 2008).

Figure 6.4.18-1, 205372072

i) The secondary structural prediction has a short beta strand, a gap, then the motif and a helix.



ii) The model seems to contain an error in that the helix seems to have been predicted but has not been marked as such. The position of the motif seems very late in the bend, making the second half of the motif appear "straightened-out".



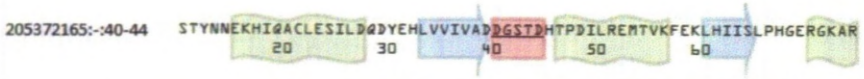
iii) The predicted binding site of 205372072 appears to deviate from that of 2b2x markedly.

The predicted binding motif shows quite a bit of deviation from the 2b2x motif used for comparison, and the RMSD with 1EXR is 1.2Å. The motif predicted appears to be more linear than would be expected of a calcium-binding protein. The secondary structure predictions are in general agreement; however, there seem to be errors in the labelling of the downstream helix.

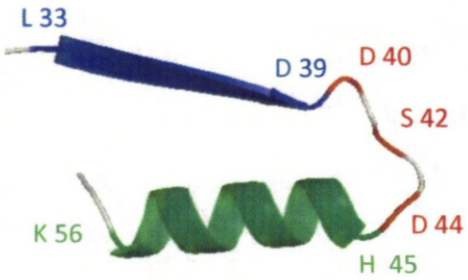
6.4.19 B. coahuilensis gene 205372165, glycosyltransferase

205372165 is listed in the genome as a glycosyltransferase; this class of protein was also predicted to contain a calcium-binding motif in *E. coli* (89110670). There was no glycosyltransferase from *B. coahuilensis* identified in the PDB. However, as mentioned above, they are present from other species.

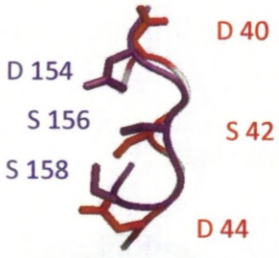
Figure 6.4.19-1, 205372165



i) The predicted secondary structure is closely associated with the motif, slightly cramping the binding loop. There is, however, a potential 4th ligand downstream of the motif.



ii) The overall shape is similar to a true DxDxDG example; however, the binding loop is much shorter, leading to a slight deviation.



iii) The binding motif of 205372165 is less curved than that of 2b2x; this may be owing to the close association of the secondary structural elements.

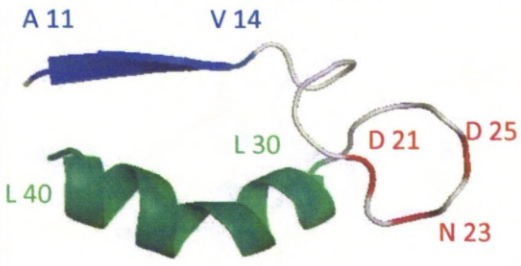
The structural predictions show similarities to known DxDxDG motif examples; however, the binding loop is slightly shortened, leading to a slightly flattened out binding motif, and a high RMSD with 1EXR of 2.5Å. The motif is, however, still quite similar to the 2b2x motif used for comparison.

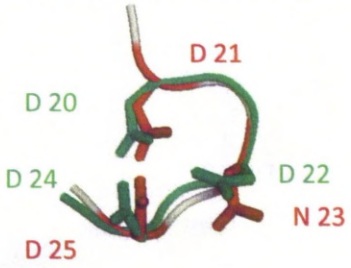
6.4.20 B. coahuilensis gene 205372216

205372216 is listed in the genome as an alpha-glucosidase. There was no alpha-glucosidase from *B. coahuilensis* identified in the PDB. Examples from other species are present, however, and, in some cases, display Ca^{2+}-binding properties (for example, 2ze0 and 1lwj).

Figure 6.4.20-1, 205372216

i) The upstream strand is further away from the motif than in most cases of the motif. The position of the motif in relation to the start of the chain is just one residue different from that of 1EXR.





ii) The predicted motif loop is extended, but the arrangement and presence of a downstream aspartic acid make this a strong candidate for calcium binding.

iii) The predicted binding site of 205372216 is very similar to 1EXR, showing only slight deviations.

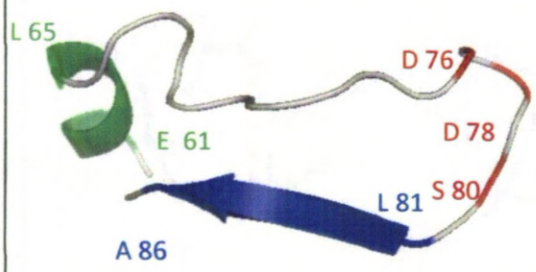
Although the binding loop is greatly extended compared to a typical EF hand, the predicted binding motif shows an almost perfect match to the 1EXR motif used for comparison, giving an RMSD of just 0.15Å. The structure is based on the models of 1uok and 2ze0.

6.4.21 B. coahuilensis gene 205372953, pseudouridine synthetase

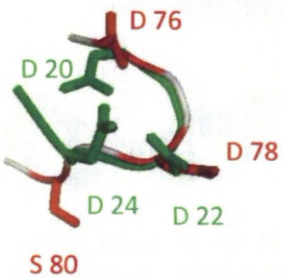
205372953 is listed in the *B. coahuilensis* genome file as pseudouridine synthetase. This particular protein does not appear in the PDB, although examples from *E. coli* are present. Their annotation, and the literature search, gave no indication of known calcium-binding functions.

Figure 6.4.21-1, 205372953

i) The spacing of the predicted secondary structural elements from the motif, and the length of the binding loop, are similar to known DxDxDG motifs.



ii) The predicted structure seen here shows a very loose fold, especially upstream of the motif. Downstream, there is a better correlation to the typical motif; however, there does not seem to be any viable additional ligand.



iii) The predicted motif of 205372953 shows some deviation in the third residue from the 1EXR motif. Additionally, the binding residues point in different directions.

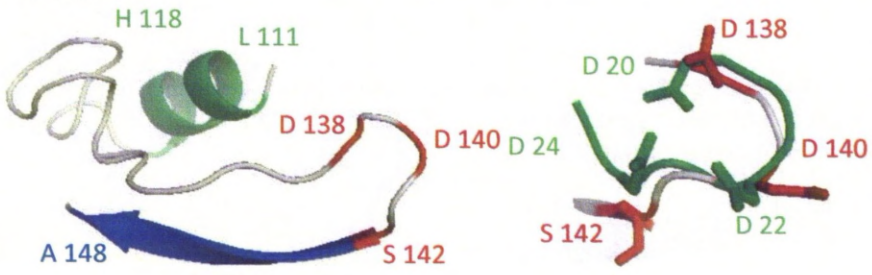
PSIPRED predicts an additional upstream secondary structural element close to the motif; this prediction is closer to the typical EF hand spacing. However, the M4T server still predicts a good match between the binding motif of 205372953 and the binding motif of 1EXR, with an RMSD of 0.5Å. The secondary structure model was based on 2ist, RluD from *E. coli*; this, however, also seems not to have a defined calcium-binding function. There are also potential downstream extra ligating amino acids. This is interesting, as a convincing structure has been predicted from a protein that is not known to bind calcium.

6.4.22 B. coahuilensis gene 205373291, YlyB

This protein is listed in the genome file as YlyB; this is a gene of unknown function. RPS-BLAST predicts a glycosyl transferase motif (Pfam 00535), with an expect score of 1e-20. A protein with this motif was also identified as calcium binding in *E. coli* (see 89109071).

Figure 6.4.22-1, 205373291

i) The secondary structure predicted here shows a good separation from the motif, close downstream and slightly father away upstream, as seen in many known DxDxDG motifs.



ii) The secondary structure prediction disagrees here, both in the upstream element and in the beginning residue of the downstream strand. It is possible that the shape of the motif would be better if these differences were taken into account.

iii) The shape of the motif in 20537291 is generally similar to that of 1EXR. However, the early starting beta strand leads to a deviation in the position of the third residue.

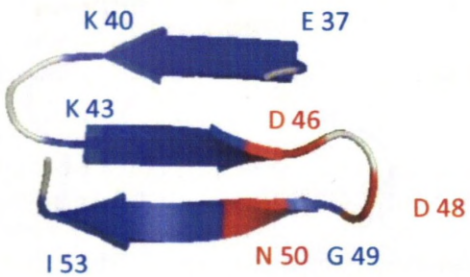
The M4T server prediction of an early start to the downstream beta strand seems to lead to a slight deviation of the binding motif from the motif of 1EXR used for comparison. The secondary structural prediction, however, shows a later start to the strand; this may result in a better fit to the typical binding motif, if this prediction is correct. Despite this, the RMSD between 205373291 and 1EXR is just 0.5Å.

6.4.23 B. coahuilensis gene 205374354, glutamate-1-semialdehyde aminotransferase

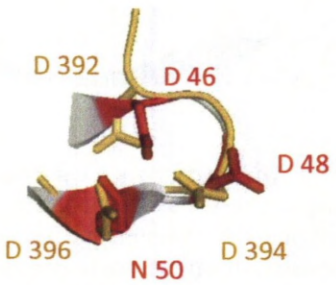
205374354 is identified in the genome file as glutamate-1-semialdehyde aminotransferase; this has been shown to be regulated by calcium and calmodulin (Im et al., 1996). RPS-BLAST identifies an amino transferase class III motif with an expect score of 1e-71, and amino transferase class I and II motifs with an expect score of 6e-05, in this protein. Neither of these types of protein from *B. coahuilensis* are represented in the PDB.

Figure 6.4.23-1, 205374354

i) The upstream beta strand encroaches slightly on the motif; this, however, will probably not affect the motif too much, as the rest of the residues are not restricted by secondary structure.



ii) Again the predictions disagree on the starting residue of the downstream element. This gives a very tight turn. However, the distinctive stack of three strands in a sheet is also seen in known DxDxDG protein, 1k1x.



iii) Even with the tight turn and the encroaching beta strand, the motif of 205374354 is very similar to that of 1k1x.

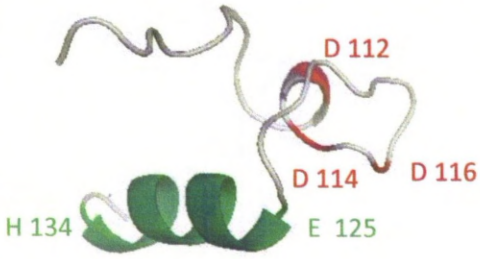
This is another example where there is disagreement between the prediction of the M4T server and PSIPRED on the length of a secondary structural element. The binding loop is much shortened in this case; however, looking at the comparison with the motif of 1k1x, the arrangement of the binding residues is not adversely affected, and the RMSD with 1exr is 0.6Å.

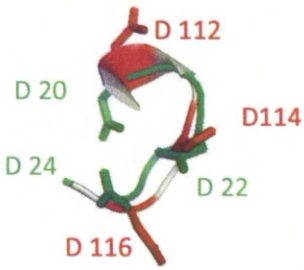
6.4.24 B. coahuilensis gene 205375207, cell wall endopeptidase

The genome file lists this protein as cell wall endopeptidase; this protein, in other organisms, is known to be modulated by calcium (Oliveira et al., 2005). This particular protein, however, was not present in the PDB. RPS-BLAST identified a peptidase family domain, Pfam 01551, with an expect score of 7e-22.

Figure 6.4.24-1,205375207

i) PSIPRED predicts an unusual helix-strand, motif, strand arrangement, with very short beta strands and a shortened binding loop.





ii) The secondary structural predictions here disagree greatly. The helix predicted in the motif would be likely to greatly affect the shape of the motif.

iii) The binding motif of 205375207 shows some similarity to 1EXR. The predicted helix causes the first residue to point in the wrong direction.

Here, the two secondary structural predictions differ greatly, both in the proximity of the structural elements to the motif and in the element type. It is difficult to tell what effects the structure predicted by PSIPRED might have if applied to the M4T server prediction. It is likely that, if the short helix predicted within the motif is false, the match with 1EXR would be much better. The RMSD of the prediction shown here is 0.7Å. The structural prediction was based on 2gu1 and 2b0p, both of which bind Zn^{2+}.

6.4.25 B. coahuilensis results summary

Genome ID and residue position	Identified as in genome	PSIPRED Secondary structure overlap	RMSD using M4T prediction and 1EXR	M4T template	Potential ligand	comments	Confidence in calcium-binding prediction
205372953 :-:76-80	pseudouridine synthetase	None	0.5	2ist	105, 107, 110	No previous evidence of calcium binding	High
205372165 :-:40-44	glycosyltransferase	None	2.5	3bcv, 1qgq	48, 52, 69		Intermediate
205375207 :-:112-116	cell wall endopeptidase	None	0.7	2b0p, 2gu1	120, 124	evidence of calcium regulation	High
205374354 :-:46-50	glutamate-1-semialdehyde aminotransferase	1 Residue	0.6	2e7u, 2epj	54	evidence of calcium regulation	High
205372216 :-:21-25	alpha-glucosidase	None	0.15	1uok, 2ze0	29	Some examples in other species bind calcium	High
205372072 :-:572-576	elongation factor G	None	1.2	2bm0	577		Good
205373291 :-:138-142	YlyB	None	0.5	1v9f, 2ist	151	Known to bind calcium in E. coli	Very High

6.5 Section conclusions

As a result of their short length and the flexibility in their residue composition, it was expected that many proteins in each of the genomes would contain a DxDxDG-like motif. However, not all of these motifs would display functional calcium-binding properties. The two best AI algorithms used in chapter 5 were applied, in order to screen out poor matches and reduce the number of potential candidates down to 204. An individual analysis of each of these predictions would be too extensive to provide here; however, it is both practical and prudent to take a sample of the positive predictions for further analysis. Twenty-three proteins, predicted positive by all four of the top AI algorithms that also produced a structure using the M4T server, were analysed in greater detail. Sixteen of these were from *E. coli* and seven from *B. coahuilensis.*

The literature searches have shown that there appears to be three proteins that are either known to bind calcium themselves, or have analogues that bind calcium. These are: methyl-galactoside transporter subunit (89108967:-:157-161); alpha-galucosidase (205372216:-:21-25); and YlyB (205373291:-:138-142).

In addition to these known calcium-binding proteins, the literature searches also highlighted a number of proteins that are believed to be regulated by calcium. Regulation by calcium does not necessarily mean that these proteins bind directly to calcium ions. Regulation may be mediated by interaction with another protein, or some other process.

There were two examples from *E. coli* where the completed structure was present in the PDB: glycine betaine transporter subunit (89109472:-:145-149) and protein chain elongation factor EF-G (89110670:-:585-589). Interestingly, these structures were omitted from AI training, as there was either no ligand present at the binding site, or a ligand was present but its identity was unknown.

The RPS-BLAST searches generally show agreement in the inferred protein function found in the genomic file; this is probably because of the use of BLAST

searches in the assignment of protein functions to the genome. It is also of note that glycosyltransferase and elongation factors seem to appear multiple times across the two species, suggesting possible calcium regulation of these proteins.

Table 6.5-1, List of proteins with motifs that give an RMSD of less than one when compared to the 1EXR motif

Genomic coordinates and residue position of motif	Identity of protien as listed in genome	RMS
89109472:-:145-149	glycine betaine transporter subunit	0.1
89108845:-:172-176	imidazolecarboxamide isomeraise	0.1
89109849:-:454-458	predicted glycosyl hydrolase	0.1
205372216:-:21-25	alpha-glucosidase	0.15
89109488:-:237-241	membrane-bound lytic murein	0.2
89108967:-:157-161	methyl-galactoside transporter subunit	0.2
89110234:-:359-363	transcription termination factor	0.4
89110565:-:398-402	gamma-glutamyltranspeptidase	0.5
205372953:-:76-80	pseudouridine synthetase	0.5
205373291:-:138-142	YlyB	0.5
205374354:-:46-50	glutamate-1-semialdehyde	0.6
89108644:-:393-397	acyl-CoA synthetase	0.7
205375207:-:112-116	cell wall endopeptidase	0.7
89110829:-:46-50	predicted phosphonate metabolizing	0.8
89110286:-:431-435	cryptic phospho-beta-glucosidase B	0.8

Of those twenty-three that display a predicted binding motif, fifteen show a very close match to a known DxDxDG-type motif (when the alpha carbons of the three predicted binding residues are compared to the three known binding residues of 1EXR using LSQMAN, the RMSD is less than 1Å) (see Table 6.5-1). The remaining examples, although they do not resemble the 1EXR configuration as closely, cannot be ruled out as calcium-binding proteins; the structural predictions are not entirely accurate, and all seem to show the other features required of a true calcium-binding protein. Five examples in particular show either a possible error in secondary structural prediction or an open enough binding loop to allow for conformational changes in the presence of a metal to bind.

It is interesting to note the differences in the predictions between the PSI-PRED secondary structural prediction and the M4T server. This highlights the

difficulties in structural prediction, and that it is not a good idea to rely on one model when looking at predicted structure and how this may affect function.

There are a number of examples seen here where a secondary structural element that is predicted in a different place, or is of a different length, could greatly affect the folding of the motif, and, hence, the predicted function of the protein. As seen at the start of this chapter, each case of a discrepancy between the two predictions may be as a result of inaccuracies from either, or both, methods. Neither method shows a great degree of accuracy in every case; this mostly results from the variation of the accuracy of the M4T servers predictions, which are greatly dependant on the quality of the templates used.

It is of note that so few of the proteins had been previously identified as calcium-binding, although some scepticism about the accuracy of these predictions is prudent as many lack the G at the end of the motif. According to the literature searches, only one protein (methyl-galactoside transporter subunit 89108967:-:157-161) has been previously shown to bind calcium.

All of the proteins identified by the AI tools, and analysed further here, show some potential for calcium-binding properties. Three can almost be confirmed as calcium-binding through their known structures or homologues in other species. Twelve of them seem very likely indeed to be calcium-binding proteins, and show a strong resemblance in features and predicted structure. The remaining eight all show so many features of calcium-binding proteins that it seems unwise to discard them based primarily on structural predictions that may be inaccurate.

6.6 Section Bibliography

Herman, P., Vecer, J., Barvik, J., I., Scognamiglio, V., Staiano, M., de Champdore', M., Varriale, A., Rossi, M., and D'Auria, S. (2005). The Role of Calcium in the Conformational Dynamics andThermal Stability of the D-Galactose/D-Glucose-Binding Protein From Escherichia coli. PROTEINS: Structure, Function, and Bioinformatics *61*, 184.

Im, C., Matters, G.L., and Beale, S. (1996). Calcium and Calmodulin Are Involved in Blue Light Induction of the gsa Gene for an Early Chlorophyll Biosynthetic Step in Chlamydomonas. The Plant Cell *8*, 2245.

Kurakata, Y., Uechi, A., Yoshida, H., Kamitori, S., Sakano, Y., Nishikawa, A., and Tonozuka, T. (2008). Structural insights into the substrate specificity and function of Escherichia coli K12 YgjK, a glucosidase belonging to the glycoside hydrolase family 63. Journal of Molecular Biology *381*, 116.

Liu, M., and Gelli, A. (2008). Elongation Factor 3, EF3, Associates with the Calcium Channel Cch1 and Targets Cch1 to the Plasma Membrane in Cryptococcus neoformans. Eukaryotic Cell *7*, 1118.

Oliveira, V., Garrido, P.A.G., Rodrigues, C.C., Colquhoun, A., Castro, L.M., Almeida, P.C., Shida, C.S., Juliano, M.A., Juliano, L., Camargo, A.C.M., *et al.* (2005). Calcium modulates endopeptidase 24.15 (EC 3.4.24.15) membrane association, secondary structure and substrate specificity. FEBS Journal *272*, 2978.

Raulf, M., Stoning, M., and Konig, W. (1985). Metabolism of leukotrienes by L-y-glutamyl-transpeptidase and dipeptidasefrom human polymorphonuclear granulocytes. Immunology *55*, 135.

Vyas, N.K., Vyas, M.N., and Quiocho, F.A. (1988). Sugar and signal-transducer binding sites of the Escherichia coli galactose chemoreceptor protein.*242*, 1290.

Wada, K., Hiratake, J., Irie, M., Okada, T., Yamada, C., Kumagai, H., Suzuki, H., and Fukuyama, K. (2008). Crystal Structures of Escherichia coli γ-Glutamy-trans-peptidase in Complex with Azaserineand Acivicin: Novel Mechanistic Implication for Inhibition by Glutamine Antagonists. Journal of Molecular Biology *380*, 361.

Section 7 - Project Conclusions

7.1 Section Overview

This closing chapter aims to bring together all the results and conclusions from the preceding sections.

The relation of these findings to the work that has gone before, and to the current state of research in this area, is considered. Furthermore, the limitations of the project are discussed, and how further work may be done to overcome these shortcomings.

Finally, the thesis is drawn to a close with a general summary of the project and its results. The aims of the project and hypothesis are reflected upon, and it is discussed how successfully these aims were met and whether the hypothesis was validated.

7.2 The work completed and its relevance

This project covers three main areas, for each of which a new process for solving a specific problem has been developed. The first was to refine a search using the 3D structure of a motif, to identify all examples in the PDB. The second dealt with identifying characteristics that could be used to differentiate between true occurrences of the motif and those that are inert. Third was the problem of how to use these characteristics as an effective predictive tool, able to identify putative unidentified examples of the motif. The finally all these steps were drawn together, to use the predictive tools developed on two bacterial species, to both validate the process and to discover new information about the organisms concerned.

An iterative search was implemented using the SPASM software package; each result was parsed and fed back into the software as a new search query, until no further new examples were found. This search built on the work done by Rigden and Galperin. In this paper, the PDB was searched using the spatial orientation of the residues in the DxDxDG motif (Rigden and Galperin, 2004). These were fed into the SPASM software and those below a RMSD threshold of 1Å were retained. Using this scheme, an increase in this threshold gives an increase in the number of results. However, there is no guarantee that the extra results would be of a similar arrangement to any of the other results. In other words, there may be little similarity between an outlying result and its nearest neighbour giving a large RMSD. This would lead to results where the conformations are not incremental variations on a theme, but an entirely different arrangement. The work completed by Torrance *et al.* is similar in many ways to this process; however, it centred more on identifying different arrangements of binding motifs (Torrance et al., 2007). It was acknowledged in the paper that motifs were missed, owing to reliance on searches for specific bound metals rather than the motifs themselves.

The iterative search used here was shown to be effective in filtering for a true-positive set of proteins with similar arrangements to the initial seed motif, as opposed to a less restrictive search that may have given a larger set, but of proteins with less favourable arrangements. In section 2.4.1, a number of new examples from the 13 classes of DxDxDG proteins were successfully identified.

It has been shown that there are indeed differences between the sequence-derived characteristics of the true examples of the motif and the characteristics displayed in the negative examples. This can be seen from the data in section 3.4, where the amino acid size, amino acid type, and hydrophobicity all showed statistically significant differences, and the secondary structure and solvent accessibility showed clear observable differences.

This builds on the postulate that the occurrence of active motifs could be predicted by sequence alone, as proposed in Rigden and Galperin (Rigden and Galperin, 2004). The most effective differential characteristics were shown to be amino acid size and amino acid type.

Two AI algorithms were used to analyse the differences seen in these two data-sets, to create a set of rules that would allow further potential examples of a functional DxDxDG motif to be identified automatically from a large data-set. Decision trees and SVMs were used on the sequence-derived data, and a number of sets of characteristics were examined to identify the optimal set of rules for classification of DxDxDG proteins. As was expected, given the results of section 3, amino acid size and amino acid type were found to be the best characteristics for discriminating binding from non-binding proteins. The two best methods (variable threshold amino acid size for decision trees, and amino acid-type SVMs), gave error rates for the training data, of 0% and 5%, respectively.

This work has built on motif-based search methods by allowing for structural and chemical characteristics, inferred from protein sequence, to enhance classification.

Twenty-three strongly predicted functional DxDxDG candidates have been identified in *E. coil* and seventeen in *B. coahuilensis.* A more detailed look at twenty-three of these proteins (see section 6) has shown that 15 of those identified are highly likely to be functional. The other 8 are less likely to be calcium-binding, but do show many of the properties expected in true calcium-binding proteins. Of particular interest is the protein 89109472, a glycine betaine transporter subunit, that is likely to be a true calcium-binding protein, but was not used in the training of the AI algorithms. This protein can be seen in the PDB (Berman et al., 2008; Rose et al., 2012) as 1r9l, and is bound to an unknown ligand, which could be calcium or simply another ligand occupying a calcium-binding site.

This final search aims to improve on current motif search methods, such as SCANPROSITE (de Castro et al., 1996; Gattiker et al., 2002), and those that search Pfam (Bateman et al., 2002) and PRINTS (Attwood, 2002). These methods use sequence similarity and regexs to identify possible domains and motifs.

As seen in sections 1 and 2, the current similarity search methods are of limited value in the detection of these types of motif. This is partly because their aim is to detect similarity between sequences undergoing divergent evolution, rather than the presence of convergently evolved motifs. This focus on homologs arising from divergent evolution places certain limitations on these search methods. The whole premise of these searches is based on what substitutions are likely, in order to conserve essential functions. However, when looking for convergently evolved examples, the model needs to be adjusted to expect

substitutions that are directed towards the emergence of a specific form or function.

Also established in section 2, was that the particular characteristics of this type of motif contribute to limiting the success of these conventional tools, when looking at short linear motifs in general and the DxDxDG motif in particular.

In the methodologies used here, rather than just using simple sequence data, further information that can be derived from this data is also considered, creating a more informed search model. The extra information allows the effects of a substitution to be compared to a model of the function potentially being converged upon, therefore more closely representing the process of convergent evolution.

It is possible that the overall procedure could be generalised to allow its use with other examples of small motifs, or highly conserved binding sites, that may display convergent evolution. The methodologies used here could be extended to other similar motifs. More work would have to be done, however, to establish if the same characteristics used here could be used in other motif examples. The amino acid size and type characteristics were shown to be most effective in the DxDxDG motif; these characteristics are general and may be applicable to all other types of motif. However, the pattern of thresholds for the amino acid size surrounding the motif is likely to be particular to the DxDxDG motif, although an automated method of optimization of these thresholds may be possible using training data for each motif investigated.

It is difficult to compare the relative frequencies of the DxDxDG motif, compared to other linear motifs, in any meaningful way. By definition, these motifs are short and of low complexity. It is likely, therefore, that their occurrence by random chance would far outweigh any conservative pressure and occlude the effects of convergent evolution. Although the simple frequency of

such short linear motifs could be compared, without knowing the proportion of these that were functional, little could be determined about the evolutionary pressures, or if a particular motif has been selected more often than another. It is the additional structural features and the highly conserved spatial arrangement that allow the DxDxDG motif to function. This project has shown it is these elements that are the target of the convergent evolution displayed by the DxDxDG motif. Therefore, for a true comparison of frequency, a similar study to this one, but on another motif, would need to be undertaken.

Further evidence has been provided to support the assumption of the convergent evolution of this motif. Although no new structural contexts were identified, the CLANS diagrams, based on the RMSDs of the proteins (see section 2.4.2), seem to indicate closely related clusters of very similar spatial configurations that would be most easily explained by the action of convergent evolution.

7.3 Project Limitations and Potential Improvements

Although much has been learned from this work, there are several places where the methods could be further refined to produce improved sensitivity and selectivity.

The first phase of the project used SPASM to search for the DxDxDG motif in an iterative way. Part of the basis of the identification relied on an assumption that all proteins were properly classified in the SCOP database. Sometimes the classification was incorrect or simply not present. Given more time, a method for automated assignment and validation of SCOP classification could have been developed.

A number of examples may also have been missed, owing to the filtering steps used. Rather than searching for a bound metal or molecule, which may not have been present in some cases, a pseudo-ligand could have been used to assess the likelihood that the residues in a motif were in a suitable configuration for binding. The requirement for an additional ligating residue may have also caused some examples to be missed; this step could be modified or omitted entirely, however, all verified examples have been shown to display this ligand.

A random subset of each superfamily was used to represent the entire family. This was primarily to prevent bias towards those groups where a vastly greater number of examples, compared to other groups, were seen in the PDB, such as in the case of the EF-hand superfamily. This bias could be the result of an increased interest in these proteins and their preferential analysis, rather than being a true representation of how often they occur; this could be checked against an automatically generated family database, such as Pfam. Such a database gives much greater coverage than the PDB, and should also show less bias towards the investigation of particular "interesting" proteins and those proteins that are good candidates for structural study.

As a result of the normalisation process, important variation within these groups may have been lost. This could further be refined by searching through the sequence databases for similar examples, and scaling up the representation of the smaller families accordingly. However, the sequence-derived characteristics data would have required a significantly greater amount of time to obtain.

The second part of the project aimed to identify characteristics that could be derived from sequence alone, and could distinguish a functional example from the many randomly occurring instances of this simple motif. BLAST searches were used to attempt to represent more members of a family than are found in the PDB. This, however, may have resulted in weakening of the true-positive signal, where related but non-functional proteins were used to construct a model of residue conservation and properties. A more detailed analysis of the BLAST results, and the selection of likely positive sequences, might have prevented this, but may also introduce investigator bias to the alignments. Rather than relying upon AI tools to optimise their own suitability, each characteristic's model could have been investigated at this stage, in order to compare those most suited to this analysis.

The AI analysis used decision trees and SVMs to determine a set of rules to classify proteins with a DxDxDG motif as binding or non-binding. The most obvious improvement to this section would be to undertake more comprehensive trials of different AI algorithms; the use of different techniques may have provided even better results. Some of the results from the genomic searches were clearly inaccurate; classification of these examples might have been improved with filtering of the proteins before their presentation to the AI algorithms. For example, there are a few cases where binding is unlikely, owing to secondary structural elements being predicted across the motif. These could have been filtered out after the secondary structural predictions had taken place.

7.4 Summary of Conclusions

The approach used for this project, supplementing motif searches with amino acid properties and structural data inferred from the sequence, has been successful in identifying examples of DxDxDG-type binding proteins from a genome-wide search.

Although the structural motif searches described in section 2 did not produce vastly improved results from previous non-iterative searches, this may be more to do with the database than the methodology. It is hoped that the effectiveness of this method of searching will become more evident as the PDB grows and becomes more representative of all proteins.

The sequence-derived data, collected from the amino acid sequences surrounding known binding motifs and non-binding motifs, have been shown to be a useful indicator of motifs that do and don't bind, without resorting to full structural studies. Although further analysis is needed, it appears that increased specificity can be gained by enhancing a motif search with amino acid properties.

The AI algorithms proved successful in making use of the extra data inferred from the sequence, allowing the training data to be classified as binding or non-binding with few errors and a good amount of agreement across miss-one-out verifications.

Finally, two genomes of interest, *E.coli* and *B.coahuiliesis*, were successfully searched, and proteins highly likely to be involved in calcium binding identified. Further analysis of individual proteins from these classifications have confirmed that known calcium-binding proteins have been identified, and also highlighted interesting variants of the known families.

Reflecting back to the original three hypotheses:

The sequence surrounding a DxDxDG motif contributes to the potential binding properties of the motif, and will dictate if a protein will bind calcium or not.

Section 3 provided evidence that particular features of the sequence surrounding the DxDxDG motif do seem to affect the binding properties.

The characteristics of the sequence surrounding a DxDxDG motif can be used to predict if a protein is likely to bind calcium or not.

Section 4 showed that, through the use of AI tools, these features can be used to make predictions as to whether a particular protein containing a DxDxDG motif binds calcium or not.

In any given genome, there is likely to be a number of proteins that bind calcium through a DxDxDG motif that has not yet been identified.

In sections 5 and 6, a number of previously unidentified potential calcium-binding proteins were identified.

Overall, the project has been successful in its aim of using sequence data from known calcium-binding motifs to devise a method for identification of further motifs using only sequence-derived data.

7.4 Section Bibliography

Attwood, T.K. (2002). The PRINTS database: A resource for identification of protein families. Briefings in Bioinformatics *3,* 252.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L.L. (2002). The Pfam protein families database. Nucleic Acids Research *30,* 276.

Berman, H.M., Henrick, K., and Nakamura, H. (2008). Atomic Coordinate Entry Format.

de Castro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., and Hulo, N. (1996). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Research Web Server Issue *34,* 362.

Gattiker, A., de Castro, E., and Gasteiger, E. (2002). PS Scan.*1.67,*

Rigden, D.J., and Galperin, M.Y. (2004). The DXDXDG Motif for Calcium Binding: Multiple Structural Contexts and Implications for Evolution. J. Mol. Biol. *343,* 971-984.

Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., Green, R.K., Goodsell, D.S., Prlić, A., Quesada, M., *et al.* (2012). The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Research *41,* D475.

Torrance, J.W., MacArthur, M.W., and Thornton, J.M. (2007). Evolution of binding sites for zinc and calcium ions playing structural roles. Proteins Structure Function Bioinformatics 813.

Appendix – paper in support of thesis

Paper in support of thesis,

New Structural and Functional Contexts of the Dx[DN]xDG Linear Motif: Insights into Evolution of Calcium-Binding Proteins: The data supporting this paper were in part contributed from the experiments carried out in section 2 of this thesis. Additionally the analysis presented in this thesis, was used in the writing of the following paper.

OPEN ACCESS Freely available online

PLoS one

New Structural and Functional Contexts of the Dx[DN]xDG Linear Motif: Insights into Evolution of Calcium-Binding Proteins

Daniel J. Rigden[1]*, Duncan D. Woodhead[1], Prudence W. H. Wong[2], Michael Y. Galperin[3]

1 Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom, **2** Department of Computer Science, University of Liverpool, Liverpool, United Kingdom, **3** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

Abstract

Binding of calcium ions (Ca^{2+}) to proteins can have profound effects on their structure and function. Common roles of calcium binding include structure stabilization and regulation of activity. It is known that diverse families – EF-hands being one of at least twelve – use a Dx[DN]xDG linear motif to bind calcium in near-identical fashion. Here, four novel structural contexts for the motif are described. Existing experimental data for one of them, a thermophilic archaeal subtilisin, demonstrate for the first time a role for Dx[DN]xDG-bound calcium in protein folding. An integrin-like embedding of the motif in the blade of a β-propeller fold – here named the calcium blade – is discovered in structures of bacterial and fungal proteins. Furthermore, sensitive database searches suggest a common origin for the calcium blade in β-propeller structures of different sizes and a pan-kingdom distribution of these proteins. Factors favouring the multiple convergent evolution of the motif appear to include its general Asp-richness, the regular spacing of the Asp residues and the fact that change of Asp into Gly and vice versa can occur though a single nucleotide change. Among the known structural contexts for the Dx[DN]xDG motif, only the calcium blade and the EF-hand are currently found intracellularly in large numbers, perhaps because the higher extracellular concentration of Ca^{2+} allows for easier fixing of newly evolved motifs that have acquired useful functions. The analysis presented here will inform ongoing efforts toward prediction of similar calcium-binding motifs from sequence information alone.

Citation: Rigden DJ, Woodhead DD, Wong PWH, Galperin MY (2011) New Structural and Functional Contexts of the Dx[DN]xDG Linear Motif: Insights into Evolution of Calcium-Binding Proteins. PLoS ONE 6(6): e21507. doi:10.1371/journal.pone.0021507

Editor: Vladimir B. Bajic, King Abdullah University of Science and Technology, Saudi Arabia

Received November 25, 2010; **Accepted** June 2, 2011; **Published** June 24, 2011

Funding: Biotechnology and Biological Sciences Research Council studentship (DDW); National Institutes of Health Intramural Research Program at the National Library of Medicine (MYG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: drigden@liv.ac.uk

Introduction

Calcium-binding proteins (CaBPs) regulate a variety of cellular processes, including cell division, differentiation, motility and apoptosis [1–3]. In addition, Ca^{2+} ions serve as cofactors in a number of (mostly hydrolytic) enzymes [4]. Sequence and structural comparisons identified a number of different Ca^{2+}-binding sites [5–8] that coordinate Ca^{2+} ions with 6 or 7 coordination bonds [6]. The best known Ca^{2+}-binding motif is a helix-loop-helix structure, referred to as the EF-hand [9–12]. In the canonical EF-hands, Ca^{2+} ions are coordinated by oxygen atoms from the side chains of the first, third, and fifth residues from the loop (which are usually Asp residues - the third and, less frequently, the fifth residue can alternatively be Asn). Additional coordination bonds are provided by the backbone oxygen atom of the seventh loop residue (which can be any residue), a water molecule coordinated by the side chain of the ninth loop residue (which is usually D, E, S, T or N), and the side chain of an acidic (usually Glu) residue in the 12th position from the beginning, which is typically located at the start of the second helix [9,10,12,13]. Additional conserved residues include Gly in the sixth position and a hydrophobic residue (Ile, Leu or Val) in the eighth position of the loop [14]. As a result, the first 10 residues of the Ca^{2+}-binding loop of the EF-hands structure typically form a Dx[DN]xDGx[ILV][DSTN]x sequence pattern, see [15].

We have previously studied the distribution of the DxDxDG-containing loop among proteins of known structure and found this loop in an impressive variety of non-EF-hand structural contexts [15–17]. In contrast to the helix-loop-helix EF-hand structure, these included helix-loop-strand, helix-loop-turn, strand-loop-helix, strand-loop-strand, and several structural contexts without a regular secondary structure element either before or after the DxDxDG-containing loop [15]. In each of these cases the loops demonstrably bound Ca^{2+} ions and the calcium-binding ligands superimposed extremely well. Furthermore, insertion of such a DxDxDG-containing, Ca^{2+}-binding loop between two β-strands of rat CD2 protein proved sufficient to create a new Ca^{2+}-binding site [18,19].

These data clearly demonstrated that the DxDxDG-containing Ca^{2+}-binding loop was a separate well-defined structural element and raised the question as to how it arose in such similar forms in so many unrelated protein folds. Two hypotheses were put forward to explain the diversity of the DxDxDG-containing calcium-binding loops: 1) a putative novel mechanism involving transplant of 10–12 residue Ca^{2+}-binding loops between different protein contexts or 2) local convergent evolution within an existing loop structure leading to the emergence of the DxDxDG motif [15].

Here we report and analyse further instances of the Ca^{2+}-binding DxDxDG loop revealed by rapidly expanding knowledge of the protein structure universe. Given sequence trends at the third position, not only in EF-hands but also in the novel examples, we introduce here the Dx[DN]xDG name, although it must be noted that as a strict regular expression, Dx[DN]xDG covers most but not all of the calcium-binding motifs characterized here. We further consider the evolutionary mechanisms that are responsible for the origin and maintenance of the Ca^{2+}-binding sites. The results have important implications for the prediction and interpretation of similar motifs in protein sequence databases.

Results and Discussion

General description

The new data presented here show four entirely new folds to harbour Dx[DN]xDG calcium-binding loops that superimpose very closely on the archetypal EF-hand motifs (Table 1, Figs. 1 and 2). These new folds are all-α (the α/α toroid of *E. coli* glycoside hydrolase YgjK [20]), all-β (the supersandwich of a glycoside hydrolase from *Bifidobacterium longum* [21]; and the galactose-binding domain-like fold of a *Porphyromonas* adhesin [22]) or mixed α+β (*Thermococcus* subtilisin [23]). The similarities between these calcium-binding loops and those of EF-hands or other instances of the Dx[DN]xDG motifs have not been reported previously. These new examples significantly expand the range of the Dx[DN]xDG motifs, currently visible in 16 different structural contexts. Yet more examples may await discovery.

We previously noted the Dx[DN]xDG motif in the extracellular β-propeller of integrin. Here we report similar motifs in differently sized propeller domains of two bacterial proteins, *Bacillus subtilis* rhamnogalacturonan lyase [24] and *Pseudomonas aeruginosa* PilY1 [25] and a fungal lectin [26]. The resemblance of the motif of the last to the EF-hand has not previously been noted. The relationships between and distributions of the propeller-borne motifs, here named calcium blades, are considered later. Asn is present at the third position of the motif with a frequency approaching that of Asp, hence the change in nomenclature from the DxDxDG to the Dx[DN]xDG motif.

For the newly described structures, calcium binding is crystallographically observed in all cases except for the *Bifidobacterium* endo-β-N-acetylgalactosaminidase. In that crystal structure manganese is bound to the Dx[DN]xDG motif but calcium may be considered as a stronger candidate for *in vivo* binding due to its much higher concentration in the environment. Calcium is bound at this position in the homologous (48% sequence identity) enzyme from *Streptococcus pneumoniae* [27]. Confirmed calcium-binding proteins such as EF-hands have been crystallized in complex with a variety of metals including manganese.

The newly discovered motif examples recapitulate the remarkable local structural homogeneity in the vicinity of the motif (Fig. 1; Table 1). This was assessed quantitatively through measuring root mean square deviations (RMSD) of corresponding atoms following superposition of the new six amino-acid motifs on the first EF-hand of *Paramecium tetraurelia* calmodulin (PDB code 1exr [28]), this latter employed as a reference. Since the amino acids varied, detailed side chain comparisons were not possible and the measurements were based on 'extended main chain atoms' (i.e. main chain N+Cα+C+O plus Cβ - virtual Cβ in the case of Gly). The resulting RMSD values were no more than 0.55 Å indicating that the new motifs superimposed extremely well on this reference EF-hand structure. For comparison, the other calcium-binding motifs in *Paramecium tetraurelia* calmodulin yield RMSD values of up to 0.42 Å.

In each of the new motif examples, the backbone carbonyl of the residue immediately following the motif contributes to metal binding (Fig. 1). As before, Asp residues, with occasional substitution by Asn, predominate at the D positions of the motif, justifying the continued use of the name. However, an interesting novelty is present in the *Psathyrella velutina* lectin structure [26]) where the second D position is occupied by Thr. This residue was not previously observed in one of the key positions of the motif, although Ser was twice seen at the second D position in our earlier examples [15]. Inspection of the crystal structures shows that both the Ser and Thr residues ligate the metal through lone pairs on their side chain oxygen atoms. For example, the separation of the Oγ1 atom Thr345 and bound calcium in lectin structure is 2.4 Å, a figure that may be compared to a typical calcium-H_2O interaction distance of 2.39 Å [29].

As previously, the side chain interactions from the D positions and the main chain interaction with the bound Ca^{2+} ions are supplemented by the interaction of side chains from at least one further acidic residue (or, occasionally an amide residue). Remarkably, all the new examples follow precedent in positioning the additional residue(s) later in the protein sequence: in not a single example from 16 different folds positions does the additional residue occur before the motif. We previously observed striking variation in the separation of the Dx[DN]xDG motif and the additional residue, from a minimum of two intervening residues to a maximum of 65. With the exception of the *Bifidobacterium* glycoside hydrolase, which has a separation of 5 residues, the new examples presented here have hitherto unseen separations of 4, 7, 36 and 112 residues (Table 1, Fig. 3). Curiously, naturally observed binding geometries do not, so far, include that of the artificially engineered EF-hand variant which was designed to include direct side chain interactions by residues separated by 2 or 5 residues, respectively, from the Dx[DN]xDG motif [30].

Most of the new examples conform to the previously common pattern in which the Dx[DN]xDG motif is positioned in a loop flanked by elements of regular secondary structure (Fig. 3). As before, the upstream and downstream secondary structures may equally well be β-strands or α-helices. The exception to this trend is the subtilisin structure in which the Dx[DN]xDG motif is part of a 25-residue, irregular insertion into the subtilisin fold that is stabilised by binding of four Ca^{2+} ions.

We previously discovered homologous binuclear calcium-binding motifs involving Dx[DN]xDG sequences in anthrax protective antigen (PDB code 1acc [31]) and human thrombospondin (PDB code 1ux6 [32]). One of the new structures, that of *Thermococcus* subtilisin shows a different kind of binuclear centre in which the second and third D positions of the Dx[DN]xDG motif, and one of the two additional residues contribute to the binding of a second Ca^{2+} ion. A further Asp residue, exclusive to the second site, completes the binding. When the Dx[DN]xDG motifs of subtilisin and thrombospondin are superimposed, the second calcium ions also superimpose perfectly, yet the differences elsewhere, including the fact that two more calcium ions are bound nearby in subtilisin, show that the subtilisin binuclear site is not homologous to the others.

The sequence conservation of the motifs was assessed in two ways. Motif conservation was first measured in the set of proteins retrieved in a simple database search with phmmer [33,34] (see Methods and Table 1). This shows the motifs in calcium blades (see below) to be well conserved but, in contrast, the motif to be present in only a tiny fraction of subtilisin-like sequences. Other motif instances exhibit intermediate conservation. Motif frequency was also assessed with respect to Pfam families or, where unavailable, the results of iterative database searches (Table 1). The frequency of predicted functional motifs tends to be lower in these sets of broader homologues, as expected. For example, the motif in *Escherichia coli* YgjK is conserved

Table 1. Novel families containing Dx[DN]xDG calcium-binding loops.

Representative[a,b]	SCOP class; fold of domain containing Dx[DN]xDG loop	PDB code, reference; position of the first D of Dx[DN]xDG	Distribution of proteins containing Dx[DN]xDG loop	Frequency of the Dx[DN]xDG loop in		R.m.s. fit of Dx[DN]xDG to the first calmodulin motif	Distance between Dx[DN]xDG and later Ca^{2+} ligands (aa)	Function of bound calcium	Broader molecular function, shared by Ca^{2+}-binding proteins and non-binding homologues
				homologous proteins[c]	phhmer neighbors[d]				
Novel structural contexts									
Thermococcus kodakaraensis subtilisin	α+β; Subtilisin-like[d]	2z2x [92]; 212	Some thermophilic archaea	<1% (Pfam PF00082)	<1%	0.12	4/7	Folding, in the context of a 25-residue insertion [23]	Proteolysis
Bifidobacterium longum endo-α-N-acetylgalactosaminidase	All β; Supersandwich[e]	2zxq [21];601	Bifidobacteria	Approx 7%	34%	0.15	5	Possibly structural [21]	Carbohydrate digestion
Escherichia coli YgjK, glycoside hydrolase family GH65	All α; α/α-toroid[e]	3c68 [20]; 431	Some gamma-proteobacteria	Approx 7% (Pfam PF01204)	37%	0.12	2/112	Not known	Carbohydrate digestion
Porphyromonas gingivalis gingipain adhesin domain	All β; Galactose-binding domain-like[e]	3km5 [22]; 1179	*Porphyromonas*, *Flavobacterium*	Approx 90% (Pfam PF07675)	83%	0.35	36	Possibly structural [22]	Carbohydrate binding
Calcium blades									
Human integrin αVβ3	All β; 7-bladed β-propeller	1jv2 [93]; 284; 349; 413	Eukaryotes	100% of 3 sites (Pfam PF01839)	90%, 100%, 94%	0.61–1.07	2	Potentially regulatory [90]	No shared broad molecular function
Bacillus subtilis rhamnogalacturonan lyase	All β; 8-bladed β-propeller[e]	2z8r [24]; 158; 222; 369	[g]	[g]	100%, 100%, 96%	0.22–0.36	2	Not known (further calcium-binding site required for activity) [24]	
Pseudomonas aeruginosa pilus biogenesis factor PilY1	All β; 7-bladed β-propeller[e]	3hx6 [25]; 851	Bacteria	Approx 75% (Pfam PF05567)	100%	0.37	2	Regulation of pilus biogenesis and motility [25]	
Psathyrella velutina lectin	All β; 7-bladed β-propeller[e]	2bwr [26]; 177; 233; 343[f]	[g]	[g]	100%, 81%, 95%	0.48–0.55	2	Possibly structural [26]	

[a]A version that includes previously reported families is provided as Table S1.
[b]All these proteins have been experimentally demonstrated to bind calcium ions.
[c]As defined by Pfam, SMART or by full-length matches in PSI-BLAST (E-value of 0.0001) run until convergence.
[d]Proteins from UniRef90 with e-value<0.001. See Methods for details.
[e]Based on the entry for a homologous protein or the authors' description.
[f]The motif commencing residue 233 is not bound to calcium in the deposited structure but crystal soak data show that it is capable of doing so [26].
[g]A distinct group could not be defined with PSI-BLAST.
doi:10.1371/journal.pone.0021507.t001

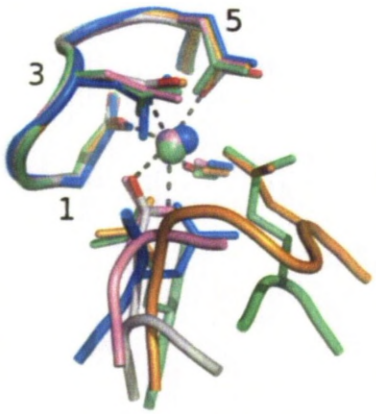


Figure 1. Comparison of Dx[DN]xDG calcium-binding motifs in calmodulin and the new structural contexts presented here. The metal (sphere) is bound by the side chains of the Dx[DN]xDG motif (labelled 1, 3, 5) and the carbonyl group of the residue immediately following the motif. These, and the entire motif backbone, superimpose very well, while additional contributions to binding from later residues vary hugely in spacing and number (see text, Table 1 and Fig. 2). The representative calmodulin (PDB code 1exr) is coloured by atom type, with carbon white, oxygen red and bound calcium in purple. Other structures and their bound calcium ions are coloured uniformly with *T. kodakaraensis* subtilisin (PDB code 2z2x) in orange, endo-α-N-acetylgalactosaminidase (PDB code 2zxq) in pink, *E. coli* YgjK (PDB code 3c68) in green and the *Porphyromonas* adhesion domain (PDB code 3km5) in blue. Interactions of calmodulin with bound metal are shown as dotted lines.
doi:10.1371/journal.pone.0021507.g001

in functional form in 37% of phmmer homologues but in only 7% of the large trehalase Pfam family (PF01204).

When compared with the previous set of Dx[DN]xDG structural contexts, the new examples are generally of narrower phyletic distribution. The most extreme example is that of the gingipain adhesion domain where, in the current sequence databases, the Dx[DN]xDG motif is confined to *Porphyromonas gingivalis*. This may reflect the increasingly complete coverage of large pan-phyla families in the PDB, at least among soluble proteins. Among our previous set of motifs, instances in archaea were rather rare, being confined to a few EF-hands and dockerin domains plausibly originating from lateral gene transfers. It is interesting, therefore, to see in the new results an archaea-specific Dx[DN]xDG motif found in a few thermophiles. This suggests that there may not be an intrinsic bias against evolution of the motif in archaea, rather a simple under-representation of their sequences in the current databases.

Interestingly, it has become increasingly clear that known examples of the Dx[DN]xDG motif have a strong bias towards periplasmic or cell surface localization or secretion. The only proven exceptions so far appear to be the EF-hands, an isolated member of the transglutaminase family [35] and some calcium blades (see below). This may reflect the fact the extracellular concentrations of calcium are much higher than generally found inside cells [36,37] so that newly generated motifs are 'fixed' more often in the extracellular milieu through acquisition of useful functions.

Propeller-borne Dx[DN]xDG motifs: the calcium blades

Remarkably, as Table 1 shows, there are now four distinct examples in the PDB of calcium-binding Dx[DN]xDG motifs found at the tips of the blades of β-propeller folds. First seen in integrin [15], they are now also visible in two bacterial proteins and in a fungal lectin. This immediately raises the question of whether the four instances share a common evolutionary origin. As Fig. 4 shows, metal binding geometries in the four proteins are very similar and in each case the separation of motif and additional side chain interaction is two residues (Fig. 3). The orientation of the motif with respect to its flanking β-strands is similar for all cases except PilY1 but the difference in the latter still appears compatible with a shared common origin of them all. Equally, the fact that the propellers differ in the number of blades – seven except for the eight in rhamnogalacturonan lyase – is not strong evidence against homology since it is known that propellers can readily evolve through duplication of an entire blade [38].

Using the modern, sensitive database searches of the HMMER3 package [33,34], connections between the four calcium blades are readily demonstrated. We took the region comprising the motif and downstream additional residue – Dx[DN]xDG–[D/E] – along with six flanking residues both before and after. Database searches with the JackHMMER program [39] in the nr protein database [40] of up to 30 iterations were carried out using e-values of either 0.01 or 0.001. As Fig. 5 shows, even at the more stringent e-value the Dx[DN]xDG motifs of the four different propellers could be connected by statistically significant relationships. Importantly, at $e = 0.001$, the search results were uncontaminated by non-propeller instances of the Dx[DN]xDG motifs. At the more permissive $e = 0.01$, EF-hands were occasionally picked up by the searches, but were inevitably discarded in later iterations and therefore absent from the final results.

Importantly, the likely homology of calcium blades is not evident from browsing current domain databases. Integrin is represented by the FG-GAP (PF01839) or Int_alpha (SM00191) domains in Pfam and SMART, respectively, both of which entries inform that some members contain a calcium-binding site. The fungal lectin and rhamnogalacturonan lyase match no domains with default search parameters, although raising the e-value cut-off produces weak matches to the FG-GAP domain. The PilY1 protein matches the Neisseria_PilC entry in Pfam (PF05567) with no indication of a propeller fold.

Since the FG-GAP and Int_alpha domain entries span whole propeller blades and contain many blades that lack Dx[DN]xDG motifs it appears that the calcium blades map awkwardly onto present domain databases, only being present in a subset of FG-GAP matching regions, but simultaneously existing in proteins not matching the FG-GAP domain. This prompted us to search for further instances of this type of Dx[DN]xDG motif in the human genome. Using the results of the iterative database searches described above various integrins and integrin-like proteins were retrieved, as expected, along with the related domains known to be present in phosphatidylinositol -specific phospholipase D [41] and the motifs recently described in cartilage acidic protein [42]. Three novel proteins containing two motifs each (Table 2) were also recovered with significant e-values; proteins that were independently confirmed to be β-propellers by profile-profile matching. These examples are poorly visible in databases – UniProt entries reveal just a single FG-GAP domain in T cell immunomodulatory protein (TIP), while sequence searches at Pfam produce FG-GAP hits (three) for only kaptin. The secreted or cell-surface TIP has been characterised as an immunomodulatory protein that stimulates T-cells to secrete several cytokines [43]. The *Caenorhabditis elegans* orthologue of TIP is implicated, by RNAi experiments catalogued in WormBase [44], in reproduction, embryonic and larval development. Interestingly, a related protein in *Cryptococcus neoformans* that shares about 26% sequence identity with TIP, is a known virulence factor of that fungal pathogen [45]. Most

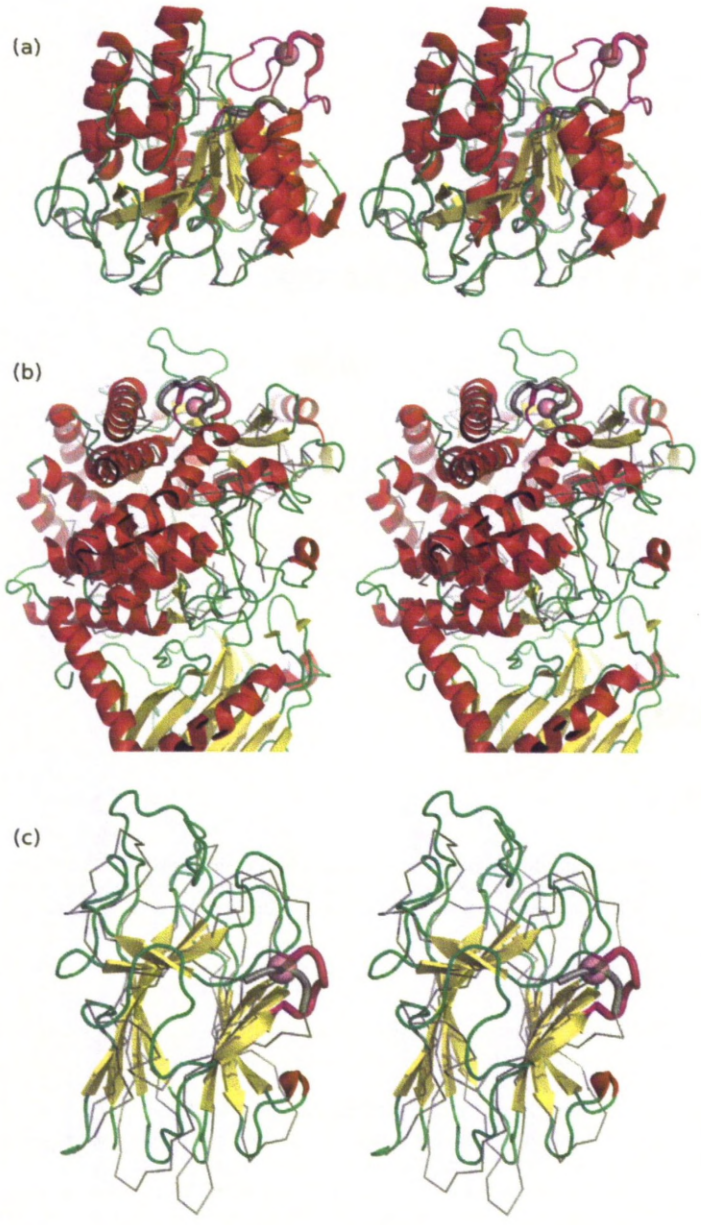


Figure 2. Stereo structure superpositions of novel Dx[DN]xDG calcium-binding motifs with nearest non-calcium binding structural neighbours. Panel a) shows *T. kodakaraensis* subtilisin (PDB code 2z2x), b) *E. coli* YgjK (PDB code 3c68) and c) the *Porphyromonas* adhesion domain (PDB code 3km5). In each case the Dx[DN]xDG motif is shown as a thick magenta cartoon with bound calcium in pink and the remainder of the calcium binding protein coloured by secondary structure. In a) the Dx[DN]xDG motif is positioned in a larger insertion binding four calcium ions which is also shown in magenta. Structural neighbours (*Bacillus lentus* subtilisin (PDB code 1c9m) in a), a predicted hydrolase from *Thermus thermophilus* (PDB code 2z07) in b), and an adhesion domain from human Tyr phosphatase mu (PDB code 2v5y) in c) are in grey with the portion aligning to the calcium binding region shown as thick cartoon. Note that the fourth novel context (2zxq) has no non-calcium binding structural neighbour in the present PDB.
doi:10.1371/journal.pone.0021507.g002

intriguingly, kaptin and Bardet-Biedl syndrome 2 protein (BBS2) are both intracellular proteins in contrast to the exclusively extracellular calcium blades previously characterised. Kaptin is an actin-binding protein [46,47] localized at the tips of stereocilia in cochlea [48], bodies related to the mechanotransduction of sound. This, and the location of its gene near a known deafness locus, strongly implicate the protein in audition [48]. BBS2 and another protein, BBS4, localise to cellular structures associated with motile cilia and which are required, not for their synthesis, but for the structural integrity and function of the mature cilia [49]. Interestingly, more recent data also implicate BBS proteins in signalling via the leptin receptor [50]. While the role of calcium binding to these proteins remains to be confirmed, it is probably significant that the D174E mutation in BBS2, which is in general a

Figure 3. Secondary structure context of the Dx[DN]xDG motifs, highlighting additional metal-binding residues (Table 1). Residues binding to metal using side chains are in red (direct interaction with calcium) or purple (through-water interaction). Secondary structure as defined by STRIDE [78] is indicated as follows: α-helices, blue shading; β-strands, yellow shading; turns, brackets. A version including previously reported families is included as Figure S1.
doi:10.1371/journal.pone.0021507.g003

well-accepted substitution [51] but in this case predicted to abolish calcium binding to one of its motifs, is associated with the disease [52].

It is interesting to note the functional parallels between stereocilia and cilia with which kaptin and BBS2, respectively, are associated, particularly since the relationship between the two proteins, in statistical terms, is at best borderline significant. For example, bending of both stereocilia and cilia results in entry of calcium into the cell through ion channels [53,54]. More broadly, it is perhaps more than coincidence that bacterial PilY also contains a calcium blade: historically, the homology between BBS8 and the bacterial PilF protein, involved in pilus assembly and twitching, provided an initial clue that Bardet-Biedl syndrome could be related to defects in cilia function [55].

Elsewhere, the distribution and abundance of calcium blades seems to vary widely. Model organisms *Escherichia coli* and *Saccharomyces cerevisiae* lack the motif entirely, but it is present in some archaea, in two proteins from *Methanosarcina acetivorans* and one from *Archaeoglobus fulgidus*, but not in *Sulfolobus solfataricus*. The ease with which propeller blades duplicate [38] and structural plasticity of the results [56] are probably responsible for some spectacular tandem duplications of the motif evident in sequence databases. Currently, the most extensive is a protein coded by locus Npun_R4253 in the cyanobacterium *Nostoc punctiforme* in which there appear to be three tandem, seven-bladed propellers formed largely of calcium blades.

Function of the new Dx[DN]xDG motifs

Broadly speaking, functions of our previously reported set of Dx[DN]xDG motif proteins could be divided into structural or regulatory roles. In the former, an essentially permanent metal interaction with protein was considered to stabilise the protein fold. In contrast, regulatory roles involve variation in the calcium binding status of the protein according to prevailing local calcium concentration with functional implications. Among the new structural contexts (Table 1) the literature shows that structural functions of bound calcium have been tentatively proposed in two cases. More interestingly, experimental data indicate a novel function for bound calcium in the case of *Thermococcus kodakaraensis* subtilisin (Tk-subtilisin): an essential role in the folding of the protein. Subtilisins are of interest as model systems for studying the thermodynamics and kinetics of protein folding since the final structure of the mature protein strongly depends on the propeptide portion ([56]). Unusually, and in contrast to bacterial subtilisins, Tk-subtilisin requires calcium for proper folding, even in the presence of its propeptide sequence [57] which, atypically, is not required for folding [58]. This calcium requirement has been assigned to the four-calcium insertion containing the Dx[DN]xDG

Figure 4. Comparison of calcium blades and their flanking β-strands. Backbone is shown as ribbon, side chains that interact with metal as sticks and the metal ions as small spheres. The structures are coloured as follows: integrin (PDB code 1jv2; three examples) in shades of pink, lectin (2bwr; three examples) in shades of green, rhamnogalacturonan lyase (2z8r; three examples) in shades of blue and PilY1 (3hx6) in orange.
doi:10.1371/journal.pone.0021507.g004

motif [58]: an insertion-less mutant failed to fold. An attempt was made to specifically eliminate the Dx[DN]xDG calcium site: the mutant could fold, but interpretation of the role of the bound calcium was complicated by compensatory structural changes [59]. While folding requires the whole insertion, with its four calcium sites, this is still the first clear example of the involvement of Dx[DN]xDG-bound calcium in the protein folding process. Earlier data on mutants of glycosylphosphatidylinositol-specific phospholipase D with reduced metal binding to its propeller-borne Dx[DN]xDG sites showed dramatically reduced expression. An effect on protein folding would be one explanation, but the reduction could equally well result from impaired intracellular transport or secretion [60].

As mentioned above, a single substitution in one of the propeller-type motifs in BBS2 is enough to lead to disease suggesting that calcium plays an important role in its function. Experimental data also clearly show the importance of calcium

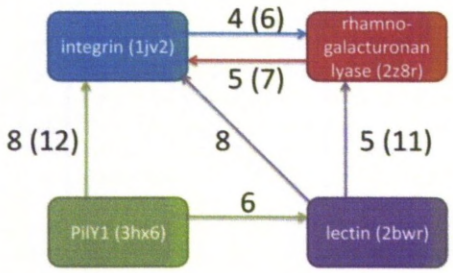


Figure 5. Schematic representation of statistically significant relationships between calcium blades revealed by JackHMMER [39] iterative database searches. Arrows indicate retrieval of a given motif by a query, with numbers indicating the number of iterations required at e-values of 0.01 or, bracketed, 0.001.
doi:10.1371/journal.pone.0021507.g005

binding to the related motifs in PilY1 protein [25]. Chelation of calcium or mutation of the Dx[DN]xDG motif each leads to loss of *Pseudomonas* twitching motility through elimination of surface pili. Surprisingly, the functions of the propeller-bound calcium ions in integrin remain mysterious [61]. Nevertheless, although not all Dx[DN]xDG motifs have been experimentally probed, it is already apparent that at least a large proportion of these motifs have structural and/or functional importance to their respective proteins.

Very recently, structural and dynamic analysis of metal-binding proteins has demonstrated their particular suitability for signal propagation, a property possibly related to the relative rigidity of the sites themselves [62]. This finding may go some way to explain the frequency with which signalling and regulatory functions are associated with Dx[DN]xDG motif calcium-binding proteins (Table 1 and Table S1).

Evolution of Dx[DN]xDG motifs

We previously argued that the unrelated structural contexts in which superimposable motifs were found implied their arising by either an as-yet uncharacterised splicing of loops from one protein to another, or multiple convergent evolution. Since then the awareness of the scientific community of the power of convergent evolution has increased significantly. Not only do enzymes exhibit convergently evolved mechanism but, more relevant to the present work, large numbers of convergently evolved linear motifs have been characterised, methods for their prediction produced [63,64] and a database set up [65]. In the light of this literature, it appears that convergent evolution is the more likely explanation for the Dx[DN]xDG motifs, but the question still arises as to why it has evolved so frequently. In order to assess this frequency in comparison to other linear motifs, we examined the number of unrelated proteins known to contain examples of other motifs in a benchmarking subset of the ELM database [65] (see Table 1 of [64]). The mean number of motif instances in unrelated proteins for this set of 17 motifs was 9.2, but this value falls to 7.8 for motifs with four defined positions. Summing the present data with previously characterised Dx[DN]xDG motifs (see Table S1) produces at least 16 instances in unrelated proteins. Clearly, the Dx[DN]xDG motif has evolved more often than most well-characterised linear motifs.

As we have previously shown, there are many examples where homologous proteins differ in possession of the Dx[DN]xDG motif: one protein has a short motif-less loop between secondary structure elements while in a related protein a longer loop harbours a functional motif. Such differences in length can arise from various sources including slipping during replication resulting in single or double amino-acid repeats [66] or meiotic recombination events that can produce larger repeats [67].

Two characteristics of the Dx[DN]xDG sequence may facilitate its formation: its sequence bias, being Asp-rich, and its regularity. The possible contributions of each are now explained. The Dx[DN]xDG motif typically contains two or three Asp residues and, furthermore, the additional interactions required for metal binding may be provided by another Asp separated from the motif by as few as two residues. Clearly, generally acidic regions will be predisposed to form the motif, particularly as Glu may provide the later interaction. Thus, slippage mechanisms generating tandem single amino acid repeats [68], in this case of Asp residues, could be part of the explanation of the frequency of Dx[DN]xDG motif appearance. An interesting parallel can be drawn with the DxxDxxxD motif, convergently evolved multiple times for binding in partners of yeast protein phosphatase 1 [63]. As examination of Fig. 4 of Neduva et al. [63] illustrates, in that case as well many of

Table 2. Novel putative calcium blades in human proteins.

UniProt ID	Protein name	Length (residues)	Subcellular localization	Number of predicted binding motifs[a]	Motif sequences
BBS2_HUMAN	Bardet-Biedl syndrome 2 protein (BBS2)	721	Cytoplasm/cilium membrane [94]	2	170 DFDGDGKKE 178[b] 251 DLNSDGVNE 259
KPTN_HUMAN	Kaptin	436	Cytoplasm/actin filaments [46,47]	2	317 DVDLDGRPE 325 373 DLTGDGLQE 381
TIP_HUMAN	T cell immunomodulatory protein (TIP)	612	Extracellular and/or transmembrane [43]	2	266 DFDGDGHMD 274 338 DYNMDGYPD 346

[a]Estimated conservatively: substitutions at the key positions of the DxDxDG motif are only allowed if precedents exist in Figure 3.
[b]An Asp174→Glu mutation has been identified in a patient with Bardet-Biedl syndrome [52].
doi:10.1371/journal.pone.0021507.t002

the functional motifs evolved in generally acidic regions. It is also worth noting that seven out of the nine residues forming a different recently-described mode of calcium binding, the calcium bowl [69], are Asp residues although only two of their side chains interact with the metal.

A second notable characteristic is the regular nature of the motif: $(Dx)_3$. In many instances of the motif one or other of the x positions, particularly the second, is occupied by Gly (Fig. 3). For example, in the *Porphyromonas* lectin, the motif sequence is DADGDG while in *Thermotoga maritima* 4-α-glucanotransferase it is DGDLDG. Thus, the slipping mechanism for repeat expansion, operating on a hexanucleotide sequence, could easily generate a nascent motif from a single instance of DG. Again, other comparable examples exist: methylated $(RG)_n$ repeats bind to the Tudor domain [70] while $(RS)_n$ motifs are common in the RS domains of SR (serine/arginine-rich) proteins and function in protein-protein interactions [71].

Finally, we note that only single nucleotide changes, of the more common transition type, separate Gly (coded by GGN in the genetic code) and Asp (GAT or GAC). This could ease the introduction of Gly into Asp-rich tracts or vice versa. Curiously, a single mutation, albeit a less common transversion, also separates Arg (AGA or AGG) and Ser (AGC or AGT), the components of the RS domain repeat mentioned above. Taken together, it seems likely that the biased composition – Asp richness – and regularity of the motif, along with the coding proximity of Asp and Gly, are at least significantly responsible for the anomalous frequency of the Dx[DN]xDG motif. Naturally, not every evolved Dx[DN]xDG motif will be structurally capable of adopting the characteristic metal-binding conformation. However, two factors may increase the proportion of Dx[DN]xDG motifs that are. First, the motif is indifferent to varied or absent flanking secondary structure, appearing simply to require a suitable structural separation of its beginning and end. Secondly, the additional residues required for metal interaction – acidic or amide group (Fig. 3) – are naturally abundant at the protein surface.

If the modes of evolution proposed above indeed played a role in producing the present day set of convergently evolved Dx[DN]xDG motifs then sequences resembling ancestral evolutionary intermediates might be present in current sequence databases. We therefore looked at motif presence or absence in the context of sequence clustering trees. Unfortunately, several factors conspired to limit the usefulness of the analysis including the fact that motifs in families of sequences tend to be either rare eg subtilisin or near universal eg the gingipain adhesion domain (Table 1). Furthermore, it is difficult to root trees composed of bacterial sequences, for example, given the lack of an external clock. Finally, the diversity of sequences in families led to a relative lack of well-supported nodes after bootstrapping analysis. Nevertheless, some features in well-supported structures of the tree derived from PilY1 (represented by PDB code 3hx6; see Table 1) and related proteins in Pfam family PF05567 (Fig S2) may shed light on modes of motif evolution. A group of four sequences from *Xanthomonas campestris* or *Stenotrophomonas* sp. SKA14 (marked with A in Fig S2) groups reliably with a set of *Xylella fastidiosa* sequences but lack the presumed functional motif DtDgDGlvD of the latter. Instead the four proteins have a longer Asp- and Gly-rich sequence such as DrwGGasqtDGvrDGyaD (in the protein with UniProt code Q4UW82). This may represent an Asp-rich, Gly-rich, ancestral-like protein or, alternatively, could be the relic of a motif inactivated by insertion. Another, acidic-rich, Gly-rich sequence positioned correspondingly to functional motifs elsewhere is found in a *Desulfuromonas acetoxidans* protein (Q1JW99; B in Fig S2) – DDGaGEk. Again, unfortunately, it is not possible to determine whether this is ancestral-like or simply the degraded result of a mutated, previously functional motif. Finally, examples of proteins containing single DG units are found in distinct parts of the tree in proteins from *Herminiimonas arsenicoxydans* (A4G7L9; C in Fig S2) and *Legionella pneumophila* (Q5X7C3; D in Fig S2): it is possible these resemble an ancestral-like sequence from which the motif evolved by duplication as outlined above although, of course, other scenarios can be imagined. It may be that this kind of analysis will be more productive in future, larger sequence databases which would lead to more confidently structured trees.

Conclusions

The new instances highlighted here reinforce how exceptional the Dx[DN]xDG calcium-binding motif is. We are aware of no other comparable motif that has apparently convergently evolved so many times: shared general themes of 3D interactions with metals and small molecules are common (e.g. [72,73]), but not the near structural uniformity observed for this linear motif (Fig. 1). Furthermore, the Dx[DN]xDG motif, unlike so many functional linear motifs [63], does not appear in regions of intrinsic protein disorder: indeed, our approach depends on the determination of motif structure by crystallography. We have highlighted, for the first time, specific features that are likely to have facilitated the appearance of the Dx[DN]xDG motif in so many structural contexts: consideration of these features may be relevant to future motif prediction efforts. Efforts are underway to exploit sequence trends – both in specific amino-acids and in broader physico-chemical characteristics - and other information, such as appearance and spacing of predicted secondary structure elements, for the prediction of functional Dx[DN]xDG motifs from sequence alone. Given the widening and deepening understanding

of the roles of calcium-binding Dx[DN]xDG motifs, such a method could contribute significantly to genome annotation.

Methods

In order to search for new structural contexts for calcium-binding DxDxDG loops, searches were done, as before [15], using SPASM 3.7.3 [74]. A minimal query using only the D positions of the first such motif of *Paramecium tetraurelia* calmodulin (PDB code 1exr, sequence DKDGD [28]) was employed. Position-specific allowed residues were used based on the typical composition of such motifs: Asp was required at the first D position, at the second any of Asp, Asn, Ser or Thr was allowed while only Asp or Asn could be present at the third position. SPASM matches motifs based on two pseudoatom positions per residue, one each representing main chain and side chain, respectively. A SPASM library file containing PDB structures available as at June 2010 was generated locally using the MKSPAZ utility (http://xray.bmc.uu.se/usf/) and searched. The results were visually screened for bound metal. All the metal-binding motif hits contained Gly at the G position of the motif and shared the typical main chain loop conformation (Figs. 1,3). LSQMAN [75] was used for local structural superpositions including quantitative comparison of newly discovered motifs with a reference structure, first EF-hand of *Paramecium tetraurelia* calmodulin (PDB code 1exr [28]). Since sequences varied RMSD measurements were based on 'extended main chain atoms' (i.e. main chain N+Cα+C+O plus Cβ - virtual Cβ in the case of Gly). SSM [76] and DALI [77] were employed for fold comparisons e. g. to compare Dx[DN]xDG loop-containing structures with their nearest non-calcium-binding structural neighbours. These latter searches were done on the respective servers (http://www.ebi.ac.uk/msd-srv/ssm/; http://ekhidna.biocenter.helsinki.fi/dali_server/) using default parameters. Structures were visualised and manipulated in PyMOL (http://www.pymol.org). STRIDE [78] was used for secondary structure assignment in order to examine the position of the Dx[DN]xDG loop with respect to nearby secondary structure elements. Structural classifications were browsed in the SCOP [79] database and sequence domains in Pfam [80] and SMART [81].

Programs of the HMMER3 suite (http://hmmer.org; [33,34]) were used for iterative database searching (JackHMMER [39] in order to discover distant sequence homologues in the nr sequence databases [40]; up to 30 iterations with e-value 0.01 or 0.001 were allowed. Genome mining was done using the resulting Hidden Markov Models (hmmsearch; e-value 0.001). Genome databases were obtained from UniProt (human; [82]) or the NCBI [40]. Motif occurrence in near sequence neighbours was evaluated as follows. Homologous sequences in the UniRef90 database [83] were obtained with phmmer [33,34] using an e-value cut-off of 0.001. The queries in these cases were the structural domains containing the motifs or, in the case of calcium blades, the strand-turn-strand sequence in which the motif was embedded. The results were aligned with MUSCLE [84] and the occurrence of functional motifs assessed by search for a motif of the form Dx[DNST]x[DN][GADN]xx[DE]using the ps_scan software [85]. In this motif definition, the separation of Dx[DN]xDG motif and later calcium-binding residue(s) was required to match that seen in the crystal structures (Table 1) with the exception of large separations (>30 residues) where the later acidic reside was omitted from the motif definition. Profile-profile matching was done with HHPRED [86] employing default parameters and searching PDB [87] and/or Pfam databases [80]. This was done to sensitively annotate the Pfam domain structure of predicted calcium blade-containing sequences and to provide independent support for their containing β-propeller folds. Sequence alignments were visualised and manipulated with Jalview 2 [88]. A bootstrapped, neighbour-joining tree for the members of Pfam family PF05567 (Figure S2) was produced with MEGA4 [89–91] in order to assess their evolutionary relationship. Presumably due to the internal symmetry of the propeller structure the Pfam entry contains a large number of partial alignments. The sequences in the family were realigned with MUSCLE [84] and truncated down to the portion common to most members. This corresponded to residues 724–875 of the *Pseudomonas aeruginosa* protein of known structure (Table 1) – approximately the last three blades of the propeller.

Supporting Information

Figure S1 Secondary structure context of the Dx[DN]xDG motifs, highlighting additional metal-binding residues (Table 1). The figure includes those motifs described in [15], Rigden & Galperin (2004) The DxDxDG motif for calcium binding: Multiple structural contexts and implications for evolution. J Mol Biol 343(4): 971–984. Residues binding to metal using side chains are in red (direct interaction with calcium) or purple (through-water interaction). Secondary structure as defined by STRIDE [74] is indicated as follows: α-helices, blue shading; β-strands, yellow shading; 3_{10} helices, green shading; turns, brackets.
(PDF)

Figure S2 Bootstrapped, neighbour-joining tree made with MEGA4 [90] using sequences edited and realigned from Pfam entry PF05567. Nodes with less than 50% bootstrap support have been collapsed. Individual sequences and groups mentioned in the text are labelled as follows: A, PilY1 sequences from *Xanthomonas campestris* and *Stenotrophomonas* sp.; B, *Desulfuromonas acetoxidans* PilY1-like protein Dace_0383 (UniProt: Q1JW99); C, *Herminiimonas arsenicoxydans* protein HEAR2375 (UniProt: A4G7L9); D, *Legionella pneumophila* protein Lpp0682 (UniProt: Q5X7C3).
(PDF)

Table S1 Families containing Dx[DN]xDG calcium-binding loops, including those in [15].
(PDF)

Author Contributions

Conceived and designed the experiments: DJR PWHW MYG. Performed the experiments: DJR DDW. Analyzed the data: DJR DDW PWHW MYG. Contributed reagents/materials/analysis tools: PWHW. Wrote the paper: DJR MYG.

References

1. Smith RJ (1995) Calcium and bacteria. Adv Microb Physiol 37: 83–133.
2. Carafoli E (2002) Calcium signaling: A tale for all seasons. Proc Natl Acad Sci U S A 99: 1115–1122.
3. Carafoli E, Klee CB, eds. Calcium as a cellular regulator. New York: Oxford University Press.
4. Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM (2008) Metal ions in biological catalysis: From enzyme databases to general principles. J Biol Inorg Chem 13: 1205–1218.
5. McPhalen CA, Strynadka NC, James MN (1991) Calcium-binding sites in proteins: A structural perspective. Adv Protein Chem 42: 77–144.

6. Pidcock E, Moore GR (2001) Structural characteristics of protein binding sites for calcium and lanthanide ions. J Biol Inorg Chem 6: 479–489.
7. Torrance JW, Macarthur MW, Thornton JM (2008) Evolution of binding sites for zinc and calcium ions playing structural roles. Proteins 71: 813–830.
8. Baumann U, Wu S, Flaherty KM, McKay DB (1993) Three-dimensional structure of the alkaline protease of pseudomonas aeruginosa: A two-domain protein with a calcium binding parallel beta roll motif. EMBO J 12: 3357–3364.
9. Gifford JL, Walsh MP, Vogel HJ (2007) Structures and metal-ion-binding properties of the Ca^{2+}-binding helix-loop-helix EF-hand motifs. Biochem J 405: 199–221.
10. Grabarek Z (2006) Structural basis for diversity of the EF-hand calcium-binding proteins. J Mol Biol 359: 509–525.
11. Kretsinger RH (1976) Calcium-binding proteins. Annu Rev Biochem 45: 239–266.
12. Strynadka NC, James MN (1989) Crystal structures of the helix-loop-helix calcium-binding proteins. Annu Rev Biochem 58: 951–998.
13. Kawasaki H, Nakayama S, Kretsinger RH (1998) Classification and evolution of EF-hand proteins. Biometals 11: 277–295.
14. Dragani B, Aceto A (1999) About the role of conserved amino acid residues in the calcium-binding site of proteins. Arch Biochem Biophys 368: 211–213.
15. Rigden DJ, Galperin MY (2004) The DxDxDG motif for calcium binding: Multiple structural contexts and implications for evolution. J Mol Biol 343: 971–984.
16. Rigden DJ, Jedrzejas MJ, Moroz OV, Galperin MY (2003) Structural diversity of calcium-binding proteins in bacteria: Single-handed EF-hands? Trends Microbiol 11: 295–297.
17. Rigden DJ, Jedrzejas MJ, Galperin MY (2003) An extracellular calcium-binding domain in bacteria with a distant relationship to EF-hands. FEMS Microbiol Lett 221: 103–110.
18. Ye Y, Shealy S, Lee HW, Torshin I, Harrison R, et al. (2003) A grafting approach to obtain site-specific metal-binding properties of EF-hand proteins. Protein Eng 16: 429–434.
19. Ye Y, Lee HW, Yang W, Shealy SJ, Wilkins AL, et al. (2001) Metal binding affinity and structural properties of an isolated EF-loop in a scaffold protein. Protein Eng 14: 1001–1013.
20. Kurakata Y, Uechi A, Yoshida H, Kamitori S, Sakano Y, et al. (2008) Structural insights into the substrate specificity and function of *Escherichia coli* K12 YgjK, a glucosidase belonging to the glycoside hydrolase family 63. J Mol Biol 381: 116–128.
21. Suzuki R, Katayama T, Kitaoka M, Kumagai H, Wakagi T, et al. (2009) Crystallographic and mutational analyses of substrate recognition of endo-alpha-N-acetylgalactosaminidase from *Bifidobacterium longum*. J Biochem 146: 389–398.
22. Li N, Yun P, Nadkarni MA, Ghadikolaee NB, Nguyen KA, et al. (2010) Structure determination and analysis of a haemolytic gingipain adhesin domain from *Porphyromonas gingivalis*. Mol Microbiol 76: 861–873.
23. Takeuchi Y, Tanaka S, Matsumura H, Koga Y, Takano K, et al. (2009) Requirement of a unique Ca^{2+}-binding loop for folding of Tk-subtilisin from a hyperthermophilic archaeon. Biochemistry 48: 10637–10643.
24. Ochiai A, Itoh T, Maruyama Y, Kawamata A, Mikami B, et al. (2007) A novel structural fold in polysaccharide lyases: *Bacillus subtilis* family 11 rhamnogalacturonan lyase YesW with an eight-bladed beta-propeller. J Biol Chem 282: 37134–37145.
25. Orans J, Johnson MD, Coggan KA, Sperlazza JR, Heiniger RW, et al. (2010) Crystal structure analysis reveals *Pseudomonas* PilY1 as an essential calcium-dependent regulator of bacterial surface motility. Proc Natl Acad Sci USA 107: 1065–1070.
26. Cioci G, Mitchell EP, Chazalet V, Debray H, Oscarson S, et al. (2006) Beta-propeller crystal structure of *Psathyrella velutina* lectin: An integrin-like fungal protein interacting with monosaccharides and calcium. J Mol Biol 357: 1575–1591.
27. Caines ME, Zhu H, Vuckovic M, Willis LM, Withers SG, et al. (2008) The structural basis for T-antigen hydrolysis by *Streptococcus pneumoniae*: A target for structure-based vaccine design. J Biol Chem 283: 31279–31283.
28. Wilson MA, Brunger AT (2000) The 1.0 A crystal structure of Ca^{2+}-bound calmodulin: An analysis of disorder and implications for functionally relevant plasticity. J Mol Biol 301: 1237–1256.
29. Harding MM (2006) Small revisions to predicted distances around metal sites in proteins. Acta Crystallogr D Biol Crystallogr 62: 678–682.
30. Nitz M, Sherawat M, Franz KJ, Peisach E, Allen KN, et al. (2004) Structural origin of the high affinity of a chemically evolved lanthanide-binding peptide. Angew Chem Int Ed Engl 43: 3682–3685.
31. Petosa C, Collier RJ, Klimpel KR, Leppla SH, Liddington RC (1997) Crystal structure of the anthrax toxin protective antigen. Nature 385: 833–838.
32. Kvansakul M, Adams JC, Hohenester E (2004) Structure of a thrombospondin C-terminal fragment reveals a novel calcium core in the type 3 repeats. EMBO J 23: 1223–1233.
33. Eddy SR (1998) Profile hidden markov models. Bioinformatics 14: 755–763.
34. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23: 205–211.
35. Adany R, Bardos H (2003) Factor XIII subunit A as an intracellular transglutaminase. Cell Mol Life Sci 60: 1049–1060.
36. Gangola P, Rosen BP (1987) Maintenance of intracellular calcium in *Escherichia coli*. J Biol Chem 262: 12570–12574.
37. Bronner F (2001) Extracellular and intracellular regulation of calcium homeostasis. ScientificWorldJournal 1: 919–925.
38. Chaudhuri I, Soding J, Lupas AN (2008) Evolution of the beta-propeller fold. Proteins 71: 795–803.
39. Johnson LS, Eddy SR, Portugaly E (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 11: 431.
40. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35: D5–D12.
41. Li JY, Hollfelder K, Huang KS, Low MG (1994) Structural features of GPI-specific phospholipase D revealed by proteolytic fragmentation and Ca^{2+} binding studies. J Biol Chem 269: 28963–28971.
42. Redruello B, Louro B, Anjos L, Silva N, Greenwell RS, et al. (2010) CRTAC1 homolog proteins are conserved from cyanobacteria to man and secreted by the teleost fish pituitary gland. Gene 456: 1–14.
43. Fiscella M, Perry JW, Teng B, Bloom M, Zhang C, et al. (2003) TIP, a T-cell factor identified using high-throughput screening increases survival in a graft-versus-host disease model. Nat Biotechnol 21: 302–307.
44. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, et al. (2010) WormBase: A comprehensive resource for nematode research. Nucleic Acids Res 38: D463–D467.
45. Liu OW, Chun CD, Chow ED, Chen C, Madhani HD, et al. (2008) Systematic genetic analysis of virulence in the human fungal pathogen *Cryptococcus neoformans*. Cell 135: 174–188.
46. Bearer EL (1992) An actin-associated protein present in the microtubule organizing center and the growth cones of PC-12 cells. J Neurosci 12: 750–761.
47. Bearer EL, Abraham MT (1999) 2E4 (kaptin): A novel actin-associated protein from human blood platelets found in lamellipodia and the tips of the stereocilia of the inner ear. Eur J Cell Biol 78: 117–126.
48. Bearer EL, Chen AF, Chen AH, Li Z, Mark HF, et al. (2000) 2E4/Kaptin (KPTN)–a candidate gene for the hearing loss locus, DFNA4. Ann Hum Genet 64: 189–196.
49. Shah AS, Farmen SL, Moninger TO, Businga TR, Andrews MP, et al. (2008) Loss of Bardet-Biedl syndrome proteins alters the morphology and function of motile cilia in airway epithelia. Proc Natl Acad Sci USA 105: 3380–3385.
50. Seo S, Guo DF, Bugge K, Morgan DA, Rahmouni K, et al. (2009) Requirement of Bardet-Biedl syndrome proteins for leptin receptor signaling. Hum Mol Genet 18: 1323–1331.
51. Bordo D, Argos P (1991) Suggestions for "safe" residue substitutions in site-directed mutagenesis. J Mol Biol 217: 721–729.
52. Hoskins BE, Thorn A, Scambler PJ, Beales PL (2003) Evaluation of multiplex capillary heteroduplex analysis: A rapid and sensitive mutation screening technique. Hum Mutat 22: 151–157.
53. Beurg M, Fettiplace R, Nam JH, Ricci AJ (2009) Localization of inner hair cell mechanotransducer channels using high-speed calcium imaging. Nat Neurosci 12: 553–558.
54. Praetorius HA, Spring KR (2001) Bending the MDCK cell primary cilium increases intracellular calcium. J Membr Biol 184: 71–79.
55. Mykytyn K, Sheffield VC (2004) Establishing a connection between cilia and Bardet-Biedl syndrome. Trends Mol Med 10: 106–109.
56. Yadid I, Kirshenbaum N, Sharon M, Dym O, Tawfik DS (2010) Metamorphic proteins mediate evolutionary transitions of structure. Proc Natl Acad Sci USA 107: 7287–7292.
57. Shinde U, Inouye M (1996) Propeptide-mediated folding in subtilisin: The intramolecular chaperone concept. Adv Exp Med Biol 379: 147–154.
58. Tanaka S, Saito K, Chon H, Matsumura H, Koga Y, et al. (2007) Crystal structure of unautoprocessed precursor of subtilisin from a hyperthermophilic archaeon: Evidence for Ca^{2+}-induced folding. J Biol Chem 282: 8246–8255.
59. Takeuchi Y, Tanaka S, Matsumura H, Koga Y, Takano K, et al. (2009) Requirement of a unique Ca^{2+}-binding loop for folding of Tk-subtilisin from a hyperthermophilic archaeon. Biochemistry 48: 10637–10643.
60. Li JY, Low MG (1999) Studies of the role of the integrin EF-hand, Ca^{2+}-binding sites in glycosylphosphatidylinositol-specific phospholipase D: Reduced expression following mutagenesis of residues predicted to bind Ca^{2+}. Arch Biochem Biophys 361: 142–148.
61. Gahmberg CG, Fagerholm SC, Nurmi SM, Chavakis T, Marchesan S, et al. (2009) Regulation of integrin activity and signalling. Biochim Biophys Acta 1790: 431–444.
62. Dutta A, Bahar I (2010) Metal-binding sites are designed to achieve optimal mechanical and signaling properties. Structure 18: 1140–1148.
63. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3: e405.
64. Edwards RJ, Davey NE, Shields DC (2007) SLiMFinder: A probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PLoS One 2: e967.
65. Gould CM, Diella F, Via A, Puntervoll P, Gemund C, et al. (2010) ELM: The status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res 38: D167–D180.
66. Wells RD (1996) Molecular basis of genetic instability of triplet repeats. J Biol Chem 271: 2875–2878.
67. Richard GF, Paques F (2000) Mini- and microsatellite expansions: The recombination connection. EMBO Rep 1: 122–126.
68. Mar Alba M, Santibanez-Koref MF, Hancock JM (1999) Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. J Mol Evol 49: 789–797.

69. Yuan P, Leonetti MD, Pico AR, Hsiung Y, MacKinnon R (2010) Structure of the human BK channel Ca^{2+}-activation apparatus at 3.0 Å resolution. Science 329: 182–186.
70. Sprangers R, Groves MR, Sinning I, Sattler M (2003) High-resolution X-ray and NMR structures of the SMN Tudor domain: Conformational variation in the binding site for symmetrically dimethylated arginine residues. J Mol Biol 327: 507–520.
71. Graveley BR (2000) Sorting out the complexity of SR protein functions. RNA 6: 1197–1211.
72. Torrance JW, Macarthur MW, Thornton JM (2008) Evolution of binding sites for zinc and calcium ions playing structural roles. Proteins 71: 813–830.
73. Brakoulias A, Jackson RM (2004) Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: An automated all-against-all structural comparison using geometric matching. Proteins 56: 250–260.
74. Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285: 1887–1897.
75. Kleywegt GJ (1996) Use of non-crystallographic symmetry in protein structure refinement. Acta Crystallogr D Biol Crystallogr 52: 842–857.
76. Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60: 2256–2268.
77. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233: 123–138.
78. Heinig M, Frishman D (2004) STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32: W500–W502.
79. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) Data growth and its impact on the SCOP database: New developments. Nucleic Acids Res 36: D419–D425.
80. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–D222.
81. Letunic I, Doerks T, Bork P (2009) SMART 6: Recent updates and new developments. Nucleic Acids Res 37: D229–D232.
82. UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38: D142–D148.
83. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: Comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23: 1282–1288.
84. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
85. Gattiker A, Gasteiger E, Bairoch A (2002) ScanProsite: A reference implementation of a PROSITE scanning tool. Appl Bioinformatics 1: 107–108.
86. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951–960.
87. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, et al. (2011) The RCSB protein data bank: Redesigned web site and web services. Nucleic Acids Res 39: D392–D401.
88. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191.
89. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783–791.
90. Kumar S, Nei M, Dudley J, Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform 9: 299–306.
91. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
92. Tanaka S, Matsumura H, Koga Y, Takano K, Kanaya S (2007) Four new crystal structures of Tk-subtilisin in unautoprocessed, autoprocessed and mature forms: Insight into structural changes during maturation. J Mol Biol 372: 1055–1069.
93. Xiong JP, Stehle T, Diefenbach B, Zhang R, Dunker R, et al. (2001) Crystal structure of the extracellular segment of integrin αVβ3. Science 294: 339–345.
94. Nachury MV, Loktev AV, Zhang Q, Westlake CJ, Peranen J, et al. (2007) A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. Cell 129: 1201–1213.