Large-scale modelling as a route to multiple surface comparisons of the CCP module family

Dinesh C.Soares1,2, Dietlind L.Gerloff3, Neil R.Syme1, Andrew F.W.Coulson1, John Parkinson3 and Paul N.Barlow2,4

1Biocomputing Research Unit, Michael Swann Building and 2Biomolecular NMR Unit, Joseph Black Chemistry Building, University of Edinburgh, The King’s Buildings, Edinburgh EH9 3JH, UK and 3Programs in Genetics and Genomic Biology/Structural Biology and Biochemistry and Departments of Biochemistry/Medical Genetics and Microbiology, University of Toronto, Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

4To whom correspondence should be addressed
E-mail: paul.barlow@ed.ac.uk

Numerous mammalian proteins are constructed from a limited repertoire of module-types. Proteins belonging to the regulators of complement activation family—crucial for ensuring a complement-mediated immune response is targeted against infectious agents—are composed solely of complement control protein (CCP) modules. In the current study, CCP module sequences were grouped to allow selection of the most appropriate experimentally determined structures to serve as templates in an automated large-scale structure modelling procedure. The resulting 135 individual CCP module models, valuable in their own right, are available at the online database http://www.bru.ed.ac.uk/~dinesh/ccp-db.html. Comparisons of surface properties within a particular family of modules should be more informative than sequence alignments alone. A comparison of surface electrostatic features was undertaken for the first 28 CCP modules of complement receptor type 1 (CR1). Assignments to clusters based on surface properties differ from assignments to clusters based on sequences. This observation might reflect adaptive evolution of surface-exposed residues involved in protein–protein interactions. This illustrative example of a multiple surface-comparison was indeed able to pinpoint functional sites in CR1.

Keywords: CCP modules/comparative modelling/complement system/electrostatic surface analysis/protein function prediction

Introduction

A sizable fraction of the vertebrate proteome consists of proteins that are composed from a limited repertoire of domain- or module-types (Bork et al., 1996). In the cases of the most commonly occurring types of modules there are now numerous experimentally-derived 3D structures (Copley et al., 2002). These provide the opportunity to predict structures for the remaining examples of a particular module-type through comparative modelling. The resolution attained with this widely used technique (also known as homology modelling) is high enough to advocate the use of such models as approximations of the true 3D-structures (Baker and Sali, 2001; Peitsch, 2002). In turn, a well-populated database of 3D structures allows, within a class of modules, comparisons based upon structural features rather than sequences. Such an approach should provide a richer source of information for the design and rationalisation of mutagenesis experiments; it might also shed light on phenotypes associated with specific polymorphisms. Ultimately structure-based comparisons could be a useful tool in the prediction of function and identification of functional sites (Nicholls et al., 1991; Blomberg et al., 1999; Pawlowski and Godzik, 2001). The most obvious features to use in a comparison of module structures of the same type are their surfaces. In this report we describe large-scale comparative modelling of CCP modules and by way of example, surface-comparisons amongst modules within the complement receptor-type 1 (CR1, CD35) (Klickstein et al., 1987).

Complement receptor-type 1 and other members of the regulators of complement activation (RCA) family consist almost entirely of multiple examples of the CCP module (Figure 1). This module-type has also been referred to as the ‘sushi domain’, ‘short consensus repeat’ or SCR (Reid and Day, 1989). A chain of between four and 30 CCP modules, joined by linking sequences of three-to-eight residues, are found within each member of the RCA family. These proteins, which include membrane co-factor protein (MCP, CD46), the factor H (fH) family, C4b-binding protein (C4BP), decay accelerating factor (DAF, CD55), CR1 and complement receptor-type 2 (CR2, CD21), are expressed by a cluster of genes located on the long arm of chromosome 1 (1q32) in humans. With the apparent exception of CR2, their role is to ensure a complement-mediated immune response is both directed against the infectious agent and is proportionate. Complement receptor-type 2 is generally regarded as a member of the RCA family even though it is not involved in complement regulation. The other RCA proteins interact, via binding sites that involve between two and four CCP modules, with the C3b and C4b components of the C3 and C5 convertase complexes (Medof et al., 1982; Krych et al., 1991; Brodbeck et al., 1996; Hardig et al., 1997; Jokiranta et al., 2000; Liszewski et al., 2000; Blom et al., 2001; Kuttner-Kondo et al., 2001). By preventing assembly of new convertase complexes, accelerating the dissociation of already-formed convertases and acting as cofactors for proteolytic degradation of the dissociated convertase components, the RCA proteins negatively regulate the complement cascade. All surfaces of self-cells exposed to serum have RCA proteins embedded in, attached to or associated with them, as vital protection against attack by the complement system (review by Walport, 2001a,b).

In addition to their preponderance amongst the RCA family, CCP modules are present in several other proteins within the complement system that interact with C3b and/or C4b. These include C1r, C1s, C2, factor B, mannan-binding lectin-associated serine protease (MASP) 1, MASP 2, C6 and C7. All of these contain two or three CCP modules. This set of
non-RCA complement proteins is more typical of extracellular multiple-module proteins in that each contains a mixture of several module-types and is said to be ‘mosaic’ (Bork et al., 1996). There are, in addition, a very diverse range of non-complement mammalian proteins that contain putative CCP modules (see Figure 1 and http://smart.embl-heidelberg.de/).

The 3D structures of a wide range of CCP modules—25 in total, including 14 CCP modules from human RCA proteins—have been determined experimentally over the last decade (http://www.rscb.org/pdb/). Each approximately 60-residue CCP module is characterised by a compact hydrophobic core wrapped in a β-sheet framework, held together by two strictly conserved disulfide bridges (Kirkitadze and Barlow, 2001). Superposition of solved CCP module structures on the structure of fH/C2416 using the program CE (Combinatorial Extension) (Shindyalov and Bourne, 1998) resulted in values for root mean square deviation (r.m.s.d.) for equivalent Cα atoms ranging between 1.9 and 3.1 A.

Because it is time consuming to solve 3D structures by NMR or X-ray crystallography, many attempts to model CCP module structures have been reported (Kuttner-Kondo et al., 1996; Villoutreix et al., 1998; Liszewska et al., 2000; Ranganathan et al., 2000; Aslam and Perkins, 2001; Perkins and Goodship, 2002). In most cases, however, the CCP module structures within these models were based on a smaller set of structural templates than are now available. Consequently, there is currently scope for improving the quality of modelled CCP structures.

The extensive use and structural diversity of CCP modules presumably reflects the versatility of a structural scaffold that has been adapted by evolution to suit many purposes: both, ‘architectural’, i.e. bestowing on specific proteins an appropriate reach and level of flexibility or rigidity; and ‘functional’, i.e. providing specific surfaces for molecular recognition and binding (Kirkitadze and Barlow, 2001). Module deletion and site-directed mutagenesis approaches have revealed the identity of numerous functional CCP modules (Medof et al., 1982; Krych et al., 1991; Brodbeck et al., 1996; Hardig et al., 1997; Jokiranta et al., 2000; Liszewska et al., 2000; Blom et al., 2001; Kuttner-Kondo et al., 2001). This list, however, is restricted mainly to the RCAs and is not exhaustive even for this set of proteins. Moreover, at the level of individual side chains, functional site mapping through experiment is still incomplete in most cases.

While computational, sequence-based methods exist that screen for residues important for function on the basis of sub-family specific conservation (Lichtarge et al., 1996), such analyses are hampered if the location of functional patches on different modules is not equivalent (with reference to the common structural scaffold). For CCP modules, such
non-equivalence is strongly suggested by the observation that in most cases several neighbouring modules are implicated in interactions with a single asymmetric partner protein.

One approach to suggesting the location of functional sites in this case is to examine the properties of CCP module surfaces. In this paper, we present a large-scale modelling strategy targeted at individual modules, with the aim of providing surface information for a large number of modules with the highest possible precision. The strategy makes optimal use of the current set of available experimentally derived CCP module structures. The database of models was subsequently used in a multiple surface comparison of 28 CCP modules from CR1. Functionally critical surface residues are likely to be under different evolutionary pressures compared with non-critical residues. Despite highly conserved sequences, the two functional sites in CR1 display radically different surface features. This observation is entirely consistent with previous studies of CR1 that show distinct functional profiles for the two sites (Krych et al., 1994, 1998). Further analysis demonstrated that in this example, surface comparison readily highlighted functionally important surfaces.

**Materials and methods**

**Sequence-based clustering of CCP modules**

Previous work (Kirkitaizde and Barlow, 2001) revealed that the residue before the first cysteine and two or three residues after the fourth and last cysteine of the consensus sequence commonly contribute to the 3D structure of a CCP module. Consequently, in the current study, one residue prior to the first consensus cysteine and three residues following the fourth consensus cysteine were included in the sequence selected to represent each CCP module.

Using this approach, sequences representing 243 CCP modules from 48 proteins [47 human proteins plus one from vaccinia virus (VCP)] were extracted from the SMART database (http://smart.embl-heidelberg.de/; Schultz et al., 1998; Letunic et al., 2004). The vaccinia virus protein, VCP, with four CCP modules, was included along with the human sequences in this process because its structure has been experimentally determined (Wiles et al., 1997; Henderson et al., 2001; Murthy et al., 2001; Ganesh et al., 2004). It therefore provides valuable additional templates for modelling purposes. The sequences were then classified by means of a clustering procedure. Initial cluster assignments were produced according to the ‘unweighted paired-group method with arithmetic mean’ (UPGMA) using a program that implements Corpet’s MULTALIN algorithm (Corpet, 1988), concomitantly with rounds of multiple sequence alignment and subsequent manual removal of individual sequences that impeded convergence. Nine stable groups or clusters, (labelled \( D \)-I) were identified in this manner. A hidden Markov model (HMM) was then built for each of these nine clusters using the software package HMMer version 2.0 (http://hmmer.wustl.edu/; Durbin et al., 1998). All sequences (including the initial set) were then scanned with the HMMs and module sequences were assigned to clusters using cut-off expectancy values (E-values) of \( 10^{-10} \) for strong and \( 10^{-5} \) for weak assignments. Sequences that failed to be matched to an HMM were labelled as unassigned. Weak HMM assignments were further investigated by pairwise similarity comparisons of the respective module sequences with the stably assigned set using MPSrch (Collins and Coulson, 1990). They were accepted only if corroborated by predominance of the HMM assignment in the MPSrch list of the five most similar module sequences. If sequences were matched to more than one HMM with E-value \(< 10^{-10} \), they were either: labelled as ambiguous if the E-value of the second match was within one order of magnitude, or assigned to the cluster with the higher score. Of 243 module sequences, 169 were strongly assigned by HMM, 30 were weakly assigned and six were ambiguously assigned, while 38 could not be assigned to a cluster at all (Figure 1).

Each cluster was subsequently aligned separately using ClustalX version 1.81 (Thompson et al., 1997). The automated alignments were edited manually placing strong emphasis on conserving the ‘Cys–Cys–Cys–Trp–Cys’ signature pattern (Kirkitaizde and Barlow, 2001) and positioning alignment gaps plausibly considering the experimentally solved 3D structures within the cluster.

**Automated comparative modelling of each cluster**

A sequence of programs to carry out the step-wise modelling task as described below were called from a PERL script as summarised in Figure 2. Eighty-three sequences out of the original set of 243 were not modelled because either they were not assigned to a cluster or they belong to clusters \( D \), \( E \) and \( I \) for which no templates are yet available.

The script first employs an option within ClustalW version 1.83 (Thompson et al., 1994) to convert a multiple sequence alignment from FASTA to PIR format, as required for Modeller version 6v2 (Sali and Blundell, 1993). Each sequence is then considered individually and the relevant cluster is screened for sequences corresponding to modules of known structures, to serve as templates for modelling. If no template sequence is found, no model is created. On the other hand, if one or more template sequences do exist, the script extracts the target sequence along with the template sequence(s) directly from the alignment and prepares a standard input file for Modeller. Modeller 6v2 is then called upon to calculate ten 3D models of the target. The model with the lowest value of the Modeller ‘objective function’ is selected and subjected to stereochemical evaluation using the program Procheck version 3.5.4 (Laskowski et al., 1993). For the RCA CCP modules, out of the 63 modelled structures, 61 had >90% of residues in combined ‘core’ and ‘allowed’ regions of the Ramachandran plot and the other two had >85% of residues in these regions.

**Surface electrostatic analysis**

The CCP modules of CR1 were selected to illustrate the utility of a surface-comparison approach because they have been the subject of extensive functional and mutagenesis studies in the past (Krych et al., 1994, 1998). The set of experimentally determined and modelled 3D structures represent the N-terminal 28 CR1 CCP modules (out of 30 in total) was pulled from the database and subjected to a comparative analysis of electrostatic properties using a combination of the programs; PIPSA version 1.0 (Blomberg et al., 1999), NMRClust version 1.2 (Kelley et al., 1996) and GRASP (Nicholls et al., 1991).

A structural alignment was obtained using Multitprot (Shatsky et al., 2004) to ensure that surfaces at equivalent locations were compared. The aligned structures were then analysed using PIPSA, to find similarities within this set of CCP modules based on their surface electrostatic properties.
PIPSA computes the molecular potentials of the model surfaces analytically as a multipole expansion that permits comparison of large datasets (Blomberg et al., 1999). The modules were then clustered by submitting the dissimilarity matrix generated by PIPSA to NMRClust. Surface electrostatic property-based cluster diagrams were derived manually by successive joining of closest neighbours based on similarity threshold values provided to NMRClust. To help assess the validity of the clusters derived in this way, electrostatic surface images were also generated by GRASP and inspected and grouped manually.

Results and discussion

Clustering helps to optimize the choice of templates for modelling

A reliable procedure for cluster assignment enhances the value of the large-scale modelling procedure used in this study because it ensures that the most appropriate set of templates will be employed in each case. Using an implementation of the hierarchical cluster assignment method of Corpet (1988), which was extended through subsequent sequence comparisons using hidden Markov models and exhaustive similarity comparisons, a total of 205 out of the original set of 243 CCP module sequences were each assigned to one of nine clusters (labelled A–I). Standard phylogenetic methods have drawbacks when applied to this family as alternative approaches to clustering, as is often observed with shorter protein sequences (Rokas et al., 2003). The main problem is that estimates of evolutionary distances are at lower levels of precision than with longer sequences. This ‘low signal-to-noise ratio problem’ is addressed more satisfactorily by a protocol such as the one we applied here to produce the initial set of clusters, since only stable assignments are reported. The additional HMM-based assignments (shown in Figure 1) are more tentative, but all of them are corroborated independently by sequence comparisons.

The clusters to which the modules of the RCA proteins are assigned are shown in Figure 1. In such a representation, the four heptad or ‘long homologous repeats’ (LHRs)—HCAFACA—that comprise the N-terminal 28 modules of CR1 were recognised previously (Klickstein et al., 1987, 1988; Hourcade et al., 1988) and are thought to have an evolutionary origin involving exon
duplication and shuffling (Hourcade et al., 1990). The triad CAF additionally appears twice in CR2 and once each in C4BP and MCP for a total of eight occurrences. Also apparent in Figure 1 are imperfect tetrad repeats in CR2: from the N-terminus—AFAC, AFAX, AF'C and AFAF [X = unassigned, ↑ is the site of an inserted module in the 16-module splice variants (Moore et al., 1987; Barel et al., 1998), which in independent sequence comparisons using MPSrch, displays strong similarity with other A-cluster members]. With the exception of factor H and factor H-related proteins, CCP modules from just a few clusters—A, F, C and H—account for all but one of the assigned modules of the RCA family (MCP has a G-cluster member at its N-terminus).

Interestingly, the majority of N-terminal CCP modules from amongst the RCA proteins either remained unassigned (fH, VCP and C4BPβ), or were only weakly assigned (MCP and C4BPα) to a cluster. This might imply varying rates of evolution and, intriguingly, coincides with a requirement of these modules for functional viability of the respective proteins. Moreover, in the first, second and third LHRs of CR1 and in DAF, C4BPα and MCP, a triad of modules comprise a binding site for components within the convertase enzymes of complement and facilitates the dissociation or destruction of these proteolytic complexes (Medof et al., 1982; Krych et al., 1991; Brodbeck et al., 1996; Hardig et al., 1997; Liszewski et al., 2000; Blom et al., 2001; Kuttner-Kondo et al., 2001). The cluster assignments for these functional units all seem to follow a motif XCA (where X = A, G, H or unassigned).

The unit XCA is not, however, seen in fH even though this protein regulates the convertases in a similar way to the other RCAs and its N-terminal three CCP modules are critical for this function (Jokiranta et al., 2000). Factor H contains six modules not assigned to clusters, consecutive runs of six B-cluster members and three modules of cluster I. Members of these clusters are unique to fH and the fH-related proteins amongst the RCA proteins, but blood-clotting factor FXIIIb also contains B- and I-cluster members. These proteins are closely linked and located within 650 kb–2.2 Mb in chromosome 1 (Rey-Campos et al., 1990; Skerka et al., 1995). This observation of a clear distinction between fH and the other RCAs is consistent with one made by Krushkal et al., who constructed a phylogenetic tree for 132 individual module sequences of the RCA gene cluster. Factor H is thought to have diverged from CR1, CR2, MCP and DAF at an early point in evolutionary history (Krushkal et al., 2000).

**Modelling**

Three-dimensional (3D) structure modelling was accomplished for 135 individual human CCP module sequences, in each case based on the most similar homologues or set of homologues for which experimentally determined structures were available. At the start of this procedure (outlined schematically in Figure 2), each of the nine sequence clusters was aligned separately.

**Fig. 3.** DAF−1,2,3,4 test comparison. (a) Ribbon trace of the modelled module structures of DAF−1−4 (blue) created by the automated modelling procedure, compared with corresponding modules from the highest resolved crystal structures (red) (PDB ID, 1OK3 for modules 1 and 2; PDB ID, 1H03 for modules 3 and 4). R.m.s.d. values (Cα atoms) are indicated in each case. (b) Comparison of the modelled (on the right) versus the experimental structure of DAF−3 showing (i) electrostatic surfaces (negative charge coloured red, positive charge coloured blue, ranging from −5 to +5 kT) generated with GRASP (Nicholls et al., 1991) and (ii) lipophilic surfaces (ranging from hydrophilic coloured blue, to lipophilic or hydrophobic coloured brown) generated using MOLCAD (Heiden et al., 1993).
The multiple alignments were subsequently used to guide the automated modelling of individual CCP modules of unknown structure using the program Modeller (Sali and Blundell, 1993). When the large-scale modelling process was first run, a total of 16 experimentally solved 3D CCP module structures were available as templates. The experimentally determined 3D structures of the four DAF CCP modules were unavailable at the time and therefore not included. Subsequently, these served as a very useful means of validating the structural models (see below). A further five experimentally determined module structures, fH~5, β2GPI~1, β2GPI~3, C1s~2 and VCP~1, were not assigned to any of the nine clusters and therefore not used as templates.

The total of 135 modelled structures resulting from the first run included those of 63 RCA CCP modules (27 from CR1, nine from factor H, four from DAF, two from MCP, nine from C4BPα and β, 12 from CR2). All models are available at http://www.bru.ed.ac.uk/~dinesh/ccp-db.html. This database is updated as necessary, to reflect the growing number of available CCP module sequences and as more template structures are deposited. Fifty sequences, assigned to clusters D, E and I, were not modelled since no template structures were
available. While models for these modules could be obtained by modelling based on sequence similarity with the closest template identified using programs such as BLAST (Altschul et al., 1997) or MPSrch, they would be expected to be less reliable. We decided to refrain from such attempts also because I-cluster sequences stand out, compared (for example) with B- and A-cluster sequences which are hard to distinguish by visual inspection, in that they appear to lack the conserved or conservatively replaced hydrophobic residues in predicted β-strands 2 and 3. Similarly, cluster E members are characterized by short hypervariable loops and a lack of insertions between β-strands 5 and 6 or 7 and 8, using Wiles et al.’s convention for numbering the strands (Wiles et al., 1997) and D-cluster members have an extra pair of cysteines, allowing an additional putative disulfide bond (modelled in Norman et al., 1991).

The models generated in the current work may be compared with those reported previously from the other RCAs, namely C4BPα (Villotreux et al., 1998), MCP (Liszewski et al., 2000) and fH (Aslam and Perkins, 2001; Perkins and Goodship, 2002), for which no experimental structures are yet available. In the case of C4BPα, the r.m.s.d. values over Cα atoms range from 2.6 Å (C4BPα−5) to 3.2 Å (C4BPα−4) for MCP−3 and MCP−4 the values are 2.2 and 2.7 Å, respectively; for fH they range from 1.6 Å (fH−12) to 3.4 Å (fH−2). The new models are thus significantly different from previously published models, but how do they compare with experimentally determined structures?

**Validation of models against known structures**

The coordinates for the experimentally determined structures (Uhrinova et al., 2003; Williams et al., 2003; Lukacik et al., 2004) of the DAF CCP modules became available after the first run of the modelling procedure described above had been completed. This allowed a comparison of modelled and experimental structures. A high level of structural similarity—r.m.s.d.s (Cα) of 1.7, 2.0, 1.2 and 1.9 Å for DAF modules 1, 2, 3 and 4, respectively—was observed, supporting the modelling strategy used in this work (Figure 3a). Previously reported models of DAF, built when only fH−15 and fH−16 were available as templates (Kuttner-Kondo et al., 1996), exhibit a lower level of similarity with the crystal structure—r.m.s.d.s (Cα) of 2.1, 2.6, 3.3 and 2.5 Å for DAF modules 1, 2, 3 and 4, respectively. Models of DAF−2 and DAF−3 that were built (Kuttner-Kondo et al., 2001) after MCP−1 and MCP−2 also became available as templates, display r.m.s.d.s (Cα) of 2.0 and 1.6 Å compared with the equivalent modules of the crystal structure, i.e. were no more accurate than those obtained using the cluster-based modelling procedure described here. These data suggest that the current set of modelled structures are generally likely to be in the same range of accuracy or more accurate than previous ones. In addition, the advantages of an automated modelling procedure for straightforward updating of the database are evident.

The feasibility of using the new models in surface comparisons was examined by a comparison of electrostatic and lipophilic surfaces of modelled versus empirical CCP module structures. The two structures of DAF−3, for example, share very similar surface properties (Figure 3b). [The DAF−3 model was built using three templates—VCP−2 (Murthy et al., 2001), CR1−16 (Smith et al., 2002) and MCP−2 (Casasnovas et al., 1999)—with pairwise sequence identity percentages of 45, 38 and 34%, respectively.] The surface representation of our model is also in visibly better agreement with the experimentally determined structure than a previously published model of DAF−3 that was based

---

**Fig. 4.** Electrostatic surface clustering of CR1 modules 1−28. (a) Surface cluster diagram depicting electrostatic similarities amongst the N-terminal 28 modules of CR1 according to FIPSA and NMRClust. Labels are coloured according to sequence cluster. GRASP-derived images of electrostatic surfaces for selected modules are also shown (box) to validate the clustering. In each case, three views rotated by 120° around the y-axis are provided. (b) A more detailed comparison of surface electrostatics of CR1−2 and CR1−16. Three views, rotated by 120° about the y-axis, are shown for each module. The faces of CR1−2 and CR1−16 displayed in the left-hand frame display similarities to one another. In contrast, the faces displayed in the right-hand frame exhibit clear differences between modules. Those amino acid residues discussed in the text are highlighted.
on a template from a different cluster (Kuttner-Kondo et al., 1996). Hence this example illustrates that careful template choice could benefit subsequent application of the models involving surface comparison. Although we cannot be certain that all models generated by our procedure will be as accurate as this verifiable example, our strategy to select the best template and to update the CCP model database when new templates become available aims to ensure that the best possible surface representation is available to the user at any given time.

**Model of CR1~25 sheds light on haplotypic variants in malaria-exposed populations**

The database of modelled CCP structures is valuable in its own right. This is illustrated when the modelled structure of CR1~25 was examined in the context of single nucleotide polymorphisms (resulting in R1601G and K1590E) that occur with greatly increased frequencies in certain malaria-exposed populations in Africa (Moulds et al., 2000). Complement receptor type 1 is a receptor on the surface of uninfected erythrocytes for the *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEmp1) expressed on the surfaces of infected erythrocytes. This interaction promotes agglutination and erythrocyte rosetting, contributors to malarial pathogenesis (Rowe et al., 1997, 2000). Hence these polymorphisms might impart a degree of resistance to severe forms of malaria. The modelled structure (data not shown) reveals that the side chains of R1601 and K1590 are surface exposed and proximal to one another, implying that substitution with Gly and Glu (respectively) will result in drastic change in electrostatic properties in this region. In addition, this suggests further mutagenesis experiments for understanding the CR1–PfEmp1 interaction that would target surrounding surface residues.

**Surface electrostatic analysis of CR1**

Surface electrostatics play a key role in protein–ligand binding and protein–protein interactions. Electrostatic properties determine the relative orientation during molecular recognition, whereas charge–charge and hydrogen-bonding interactions contribute to binding specificity and affinity (for a review, see Honig and Nicholls, 1995). It has been demonstrated that where similar electrostatic properties are shared by a set of proteins, this may indicate similar behaviour and function (Ullman et al., 1997; Botti et al., 1998; Wade et al., 1998; Blomberg et al., 1999). There is plenty of evidence for the involvement of electrostatic interactions in the recognition by RCA proteins of C3b and C4b (Krych et al., 1998; Blom et al., 1999, 2000; Liszewski et al., 2000; Kuttner-Kondo et al., 2001). Surface electrostatic similarity amongst CCP modules was therefore examined using CR1 as an example.

The program PIPSA (Blomberg et al., 1999) was employed to compare electrostatic properties of the N-terminal 28 CR1 modules (25 modelled structures plus three experimentally determined ones). The PIPSA-generated cluster diagram (Figure 4a) was compared with manually grouped GRASP-generated electrostatic surface images of the CR1 modules. Reassuringly, most module associations made by PIPSA showed good agreement with the GRASP images. This implies that PIPSA could be applied to a larger set of CCP modules and used to compare module surfaces between proteins.

The majority of relationships in the PIPSA-generated cluster diagram based on electrostatic surfaces (Figure 4a) match up both with a sequence-based tree (not shown) and the cluster-assignments; e.g. CCP modules 4, 11, 18 and 25 (all in cluster F) are in the same surface-cluster. However, PIPSA also suggested associations based on electrostatic surfaces that are different to those based on sequence alone.

**Surface comparisons can identify modules important for function**

According to the sequence-based clustering, CR1~2 is most similar to CR1~9, CR1~16 and CR1~23, with all four modules belonging to cluster C. For example, CR1~2 has ~67% pairwise sequence identity with both CR1~9 and CR1~16. In the cluster diagram based on electrostatic surface-properties, however, CR1~2 is instead grouped with CR1~3, CR1~10 and CR1~17, which were all assigned to sequence cluster A (Figure 4a). Note that no significant sequence similarity was found through a pairwise BLAST comparison of CR1~2 and CR1~3. From inspection of Figure 4a, it is evident that the electrostatic surface of CR1~2 does indeed resemble more closely that of CR1~3 than it does those of CR1~9 or CR1~16 (see below). Thus a comparison of relationships based on surfaces versus those based on sequences readily identified a case where two modules (CR1~2 and CR1~9) had clearly diverged significantly in terms of their surface properties, but not in overall sequence. Given the close evolutionary relatedness of these modules as judged by their equivalent positions within the four LHRs in CR1 (Hourcade et al., 1990), this observation implies that some surface-exposed residues have been the subject of adaptive mutation while the remainder of the sequence has undergone predominantly neutral variation. Therefore, the modules affected are candidates for being functionally important ones, with their non-conserved residues representing possible sites of interaction with binding partners.

This notion is borne out by previously reported functional studies (Klickstein et al., 1988; Krych et al., 1991, 1994) showing that CR1~2 and CR1~9 (and the nearly identical CR1~16) contribute to two different multimodular binding sites, called functional sites 1 and 2, respectively. These sites have different, although related, functional profiles. CR1 binds to C3b and C4b, via its two functional sites 1 and 2 (Klickstein et al., 1988; Krych et al., 1991, 1994). Site 1 was shown to be composed of CCP modules 1–3; it primarily binds C4b, whereas C3b is only weakly bound and is the primary locus of convertase decay accelerating activity in CR1. Site 2—of which there are two copies, one formed by CCP modules 8–10 and the other by CCP modules 15–17—binds both C4b and C3b and is a key contributor to cofactor activity.

**Surface comparisons pinpoint functional sites**

While CR1~2 and CR1~9/CR1~16 do not cluster together on the basis of their electrostatic surfaces, the differences between them lie primarily on one face—that presented in the right-hand frame of Figure 4b. This face of CR1~16 is dominated by electronegative side chains while the equivalent face of CR1~2 displays significantly less positive charge and a substantial amount of negative charge. The remainder of the module surfaces on the other hand appear similar and those charged residues present are conserved or conservatively replaced.
(R64 is conservatively replaced with K964, D is conserved at positions 68 and 968 and R is conserved at 122 and 1022). Thus further inspection of a pair of modules that have diverged in terms of surface properties, but not in sequence, immediately suggests a binding face that could be explored using mutagenesis.

In fact, residues lying on the divergent faces of these two modules have, in previous work (Krych et al., 1991, 1998), already been substituted for one another. In CR1, two such individual substitutions—N1009D and K1016E (Figure 4b)—brought about in functional site 2 the loss of iC3-binding (iC3 is a form of C3 that, as a result of hydrolysis of the thioester bond, has a conformation and reactivity similar to that of C3b) and C4b-binding, as well as a loss of cofactor activity directed towards both these ligands. Strikingly, the reciprocal mutation D109N increased C3b-binding, while mutation E116K also conferred C3b-binding on site 1 and increased its cofactor activity with respect to C3b and C4b. These experimental results thus fully support the notion that the diversity in surface charge between these modules has occurred amongst functionally critical residues and has arisen through the need for the different sites to perform different functions. Therefore, at least in this example, the approach of multiple surface comparisons is remarkable in its ability to quickly pinpoint important regions. The extent to which such comparisons identify functional sites more generally thus warrants further investigation.

Acknowledgements

The authors thank Joan Moulds (Philadelphia) for useful discussions and Russ Hamilton (Edinburgh) for system administration support. D.C.S. acknowledges the Edinburgh Protein Interaction Centre (EPIC) for provision of a PhD studentship. This work was also funded by grants to P.N.B. from the Medical Research Council of the UK and the Wellcome Trust.

References

D.C. Soares et al.


Received January 18, 2005; revised May 6, 2005; accepted May 20, 2005

Edited by Michael Sternberg