BIOINFORMATICS

Vol. 15 no. 6 1999
Pages 501–509

Clustering of non-polar contacts in proteins

Finn Drabløs

SINTEF Unimed MR Centre, N-7465 Trondheim, Norway

Received on September 28, 1998; revised and accepted on March 20, 1999

Abstract

Motivation: *Hydrophobic or non-polar contacts in proteins are important for protein folding, protein stability and protein–protein interactions. In particular, in the interior of a protein, in the hydrophobic core, a large number of such contacts are found. The residues involved in these contacts often form a tightly packed cluster of atoms. It is useful for the understanding of protein structure to be able to identify and analyse such clusters.*

Results: *Tools for hierarchical cluster analysis of non-polar contacts in proteins are described. These tools allow for efficient identification of clusters of non-polar interactions in proteins, both internal clusters and clusters involved in protein–protein contacts. The non-polar contacts are represented by a dendrogram structure, which is a simple approach for flexible identification of clusters by visual inspection. The tools are demonstrated on the structure of crambin, the structure of the complex between human growth hormone and the human growth hormone binding protein, and a pair of lipase/esterase structures.*

Availability: *On request from the author.*

Contact: *finn.drablos@unimed.sintef.no*

Introduction

Hydrophobic or non-polar interactions in the interior of a protein represent a significant contribution towards the stability of the structure. It is difficult to get reliable experimental data on the relative importance of this contribution, but it has in most cases been assumed that hydrophobic interaction is the dominating force in maintaining a protein fold. Recent work has shown that hydrogen bonding probably plays an equally important role (Pace *et al*., 1996). Nevertheless, non-polar contacts are a very important factor in protein stability, and hydrophobic residues may play an important role in the formation of a 'folding nucleus' initiating the correct folding of a protein (Shakhnovich *et al*., 1996).

It has also been shown that non-polar interactions are important for protein–protein interactions (Jones and Thornton, 1996). There seems to be a difference between permanent and non-permanent contacts, as analysed by looking at homo-dimers versus hetero-dimers. The non-polar interactions seem to be more important for permanent contacts, whereas hydrogen bonding seems to be relatively important for non-permanent contacts. This difference can probably be explained by the requirement for the surfaces found in non-permanent contacts to interact with solvent as well as other protein surfaces.

The traditional point of view has been that the hydrophobic effect arises from the net changes in energy and entropy due to rearrangement of the local water structure as two hydrophobic species approach one another (Rigby *et al*., 1986). More recent studies indicate that favourable van der Waals interactions due to tight packing in the interior of the protein may make a substantial contribution to the hydrophobic effect (Pace *et al*., 1996). This makes it relevant to work on methods for identifying close contacts in proteins.

Several automatic and semi-automatic methods have been developed for analysing relationships between residues in a protein. The simplest approach is probably the distance matrix, as implemented in, for example, the X-PLOR program (Brünger, 1992). More detailed structural information can be gained from methods characterizing the protein environment at specific points (Bagley and Altman, 1995) or at the position of individual residues (Bordo, 1993; Gromiha and Selvaraj, 1997; Selvaraj and Gromiha, 1998). Such methods are useful in several applications, including fold recognition (Casari and Sippl, 1992; Vajda *et al*., 1997). However, the identification of a subset of residues forming a core structure is a slightly different problem. Cores of structurally conserved residues can be identified by comparison of related structures (Gerstein and Altman, 1995a,b; Schmidt *et al*., 1997). Several methods have been developed for finding cores in individual structures. Some of these are related to environment-describing methods, and may use contacts within an interaction radius (Heringa and Argos, 1991) or distances (Karpeisky and Ilyin, 1992; Karlin and Zhu, 1996; Zhu and Karlin, 1996) as input for some kind of grouping or cluster analysis. Compactness, based on solvent-accessible surface area, has also been used in a related approach (Zehfus, 1995, 1997). Cores have also been identified by a combination of properties, either by clustering (Swindells, 1995b) or by cutting of regions based on correspondence analysis (Tsai and Nussinov, 1997a,b). Identification of cores in proteins is related to identification of domains, and this has been used for finding protein domains (Swindells, 1995a). However, domains can also be identified directly, e.g. by finding sequence cut points giving minimal segment–seg-

ment contacts, followed by clustering of segments (Islam *et al*., 1995), or by cluster analysis of secondary structures (Sowdhamini and Blundell, 1995).

However, the existing methods for identification of cores are mainly designed for finding a unique set of clusters by a fully automatic approach. Although cluster analysis, at least in a simplified form, is used in many programs (Heringa and Argos, 1991; Islam *et al*., 1995; Sowdhamini and Blundell, 1995; Swindells, 1995b; Karlin and Zhu, 1996), these implementations do not utilize the potential for exploratory data analysis in statistical clustering when identifying the clusters.

This paper describes the implementation of a flexible method for investigating clusters or networks of residues, based on hierarchical clustering of pairwise contact areas. The clustering is basically standard hierarchical clustering as found in multivariate statistics and exploratory data analysis, and the output of the clustering process is represented as a dendrogram or tree structure. Representation of the data set as a dendrogram is useful for the identification of subclusters in the data set, possibly at several levels. The clusters may then be visualized using standard software for visualization of protein structures. This approach is user controlled, and it is not limited to predefined criteria for defining clusters. Rather, the user may test alternative criteria in a flexible way, using his or her knowledge about protein structure and interactions in order to identify interesting properties in an explorative manner. The explorative aspect is a unique and essential property of this tool. The examples described in this paper show that the software will identify important non-polar interactions in both protein interiors and in protein–protein interfaces.

System and methods

The system has been implemented as two separate programs, pdb_np_cont and pdb_np_clus, on an SGI workstation running IRIX 5.3. The pdb_np_cont program computes pairwise atom contact areas between non-polar atoms (see below) from structural protein data in a standard PDB coordinate file (Bernstein *et al*., 1977; Abola *et al*., 1987). This program is written in portable ANSI c and has been compiled with the IRIX/MIPS cc compiler (Version 3.19). The pdb_np_clus program reads the list of pairwise contact areas from pdb_np_cont, computes pairwise residue contact areas and clusters residues based on these contact areas. The clustering process is displayed as a dendrogram, and optionally commands for displaying selected clusters in a graphical representation of the three-dimensional structure may be generated. This program is written in awk and is compatible with both the nawk and the gawk (GNU awk) interpreter. The software is available from the author upon request.

Algorithm

Identification of pairwise atom contacts (pdb_np_cont)

Computation of pairwise atom contact areas is based on classification of points located on a sphere around each atom. This approach is inspired by the early work of Shrake and Rupley (1973) and the MS program (Connolly, 1983). It is also similar to the method used by Tsai *et al*. (1997). A predefined set of points is read into the program, and for each (non-polar) atom the points are translated and scaled so that they represent the defined interaction radius of the atom. The interaction radius of each atom is normally the van der Waals radius of the atom type plus the radius of a water molecule.

The classification of individual points is illustrated in Figure 1. The basic approach is to use a sphere of points corresponding to the interaction radius of a given atom. Then the closest interacting atom is identified for all points that are not buried by atoms belonging to the same residue as the atom we are looking at. Each point is then classified according to the nature of the closest interacting atom. A more precise description is given by the following pseudocode.

```
for each point {
    move point to interaction sphere of reference_atom;
    on_surface = TRUE;
    inside_same_residue = FALSE;
    inside_polar_atom = FALSE;
    for each neighbour_atom {
        if point is inside interaction radius of neighbour_atom
        then {
            on_surface = FALSE;
            if neighbour_atom belongs to same residue as
            reference_atom then {
                inside_same_residue = TRUE;
                goto done;
            } else if this is the closest contact so far then {
                remember this neighbour_atom;
                inside_polar_atom = 'this neighbour_atom is
                polar' (TRUE or FALSE);
            }
        }
    }
done:
    if not on_surface and not inside_same_residue and not
    inside_polar then {
        point represents non-polar contact between
        neighbour_atom and reference_atom;
    }
}
```

Two alternative definitions of 'closest contact' may be used. In Figure 1, the standard Euclidean distance is used, as it is easy to illustrate. However, this distance does not take

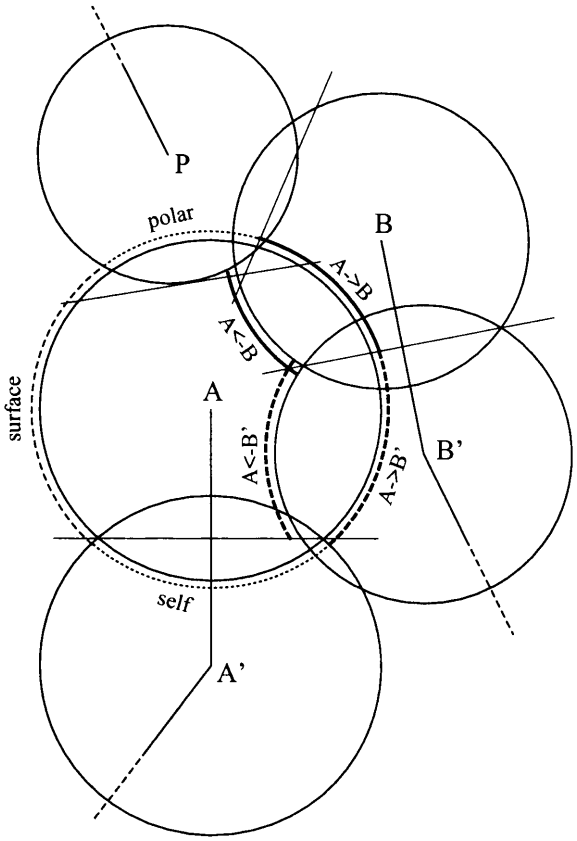

Fig. 1. Classification of regions of the interaction sphere around an atom. Non-polar atoms from two different residues (A and B) and a polar atom from a third residue (P) are in contact with each other. The classification for atom A is shown in detail, with surface region, region in contact with atom from the same residue (self), region in contact with polar atom, region in contact with B (A→B) and with B (A→B) indicated. The contacts of B and B with A (A←B and A←B) are also shown, and it is clear that the area for A→B may be different from the area for A←B. The thin lines at the sphere interfaces represent equidistant points with respect to corresponding sphere centres.

differences in van der Waals radius into account. An alternative distance measure where each Euclidean distance is weighted by the van der Waals radius of the interacting atom may therefore be used. The two distance measures will give different results only when atoms with significantly different radius are involved. In practice, this means that in the current implementation only interactions with oxygen will be significantly affected. This will be a general effect, affecting all atoms in contact with oxygen. The net effect on the final classification therefore tends to be small, and for the examples in this paper only the simpler (and slightly faster) Euclidean distance is used.

The non-polar contact between two atoms is thus represented by all points classified as being involved in this specific contact by the algorithm given above. This can be computed as a contact area by summation of the area represented by each point. This point area can easily be computed during the initial generation of the data points. By starting with a unit sphere and initial points located as a regular polyhedron on the sphere surface, new points can be generated by tessellation of the initial polyhedron, and the sphere area represented by individual data points can be computed. This approach makes the computation of contact area almost invariant to the relative orientation of the spheres, which is difficult to achieve when relying only on an even distribution of points on the sphere surface when computing the area.

The contact areas computed by this approach will not be symmetrical; the contact area of atom A with respect to atom B will not necessarily be the same as the contact area of atom B with respect to atom A. This can be seen in Figure 1. This difference is not necessarily a problem, and in the final clustering the contact area is made symmetrical by using the average value. However, this illustrates that although most definitions in use give comparable results, there is no obvious unique definition of contact area. Similar problems are also seen in other programs (Abagyan and Totrov, 1997).

There are alternative approaches to the computation of areas associated with atoms (Flower, 1997). However, the point-based approach is easy to implement, very flexible and it is easy to modify the classification to suite different needs.

Clustering of contact areas (pdb_np_clus)

The basic idea of the clustering stage is to group together residues based on contact areas in a stepwise (hierarchical) approach, in order to define groups or clusters of residues involved in non-polar interactions. The approach is inspired by the dendrograms (or tree diagrams) used for cluster analysis in traditional multivariate data analysis [see, for example, Everitt and Dunn (1983)]. However, in this case, there is a clear link between the structure of the dendrogram and the physical properties of the data set, as residues grouped together in the dendrogram are also in close (non-polar) contact in the protein.

Two approaches to clustering have been implemented in pdb_np_clus. The simplest approach is a standard single-linkage clustering, where at each step the residue pair with the largest contact area is joined together in the dendrogram.

The second approach is based on the fact that the interaction energy between two surfaces (and therefore the importance of this interaction) is the sum of several possibly small interactions between individual residues. It may, therefore, be more correct to use the sum of all contacts between clusters (total inter-cluster contact area), rather than individual atom–atom contacts. The output from this approach may be difficult to analyse. As new clusters form, the updated sum of contacts to other clusters will often increase compared to

the contact area of the newly formed cluster, and this will lead to inversions in the dendrogram. However, interpretation can be simplified by plotting the dendrogram versus the total sum of contact area (total intra-cluster contact area) for all clusters, rather than the more traditional approach using the clustering scale (the criterion for joining clusters at each step of the clustering process) as plot scale. The total intra-cluster contact area will be a monotonic increasing value during the clustering process, which will also be true for the pairwise contact area when single-linkage clustering is used.

Only contact areas for side-chain atoms are used for computing the pairwise residue contact areas. This reduces the potential problem of 'chaining' stretches of neighbouring residues together, thus making it easier to identify true clusters of residues, independent of sequence position. It has also been shown that distance measures based on side-chain atoms perform better than all-atom distance measures in some applications (Karlin *et al*., 1994). However, this also means that clusters separated by a β sheet will be classified as separate clusters. It may, in some cases, be relevant to merge such clusters (Swindells, 1995b). This is not done in the current implementation, based on the assumption that using separate clusters will give a better picture of how β sheets affect the structure of non-polar clusters.

Implementation

pdb_np_cont

In the current implementation, the sphere used for computation of contact area is represented by 512 data points, and the fraction of the sphere area is associated with each point. These data points are read in from an external file, and the user can therefore easily modify the number of data points. The data points used in this implementation were generated by a modified version of the sphere program, originally written by J.Leech (http://www.cs.unc.edu/~jon/sphere.html). The data points are scaled according to the sum of the van der Waals radius of the atom plus a water radius. The van der Waals radius is used as defined by Richmond and Richards (1978), and the water radius is 1.4 Å. All atoms are used for classification of contacts. However, only side-chain atoms, starting with Cα, are listed in the output and used by the pdb_np_clus program. All O and N atoms are regarded as polar, whereas C and S atoms are non-polar (Bowie *et al*., 1991). All hydrogens are ignored. These parameters are read in from external files and can easily be modified by the user.

pdb_np_clus

The output from pdb_np_cont can be read directly by pdb_np_clus. The contact area for each pair of contacts is recomputed as the average of A–B and B–A contact areas, as a consequence of the non-symmetrical nature of the algorithm described above. The default approach by the pdb_np_clus program is to carry out a single-linkage clustering using all contact areas and to draw a dendrogram based on the contact area at each step of the clustering process. This will generally make one very large cluster including all residues with non-polar contacts in the protein. In order to make interpretation and visualization easier, the clustering may be stopped when the contact area falls below a specified threshold, and all separate clusters with less than a specified number of members may be excluded from the dendrogram. This allows the user to focus on the most important clusters built from large contact areas. The dendrogram is plotted via postscript (EPS) commands to an external file.

In order to visualize a given set of non-polar clusters from the dendrogram onto the corresponding three-dimensional structure, a file of Biosym Command Language (BCL) commands may be generated by the program. These commands can be read by InsightII (Molecular Simulations Inc., San Diego, CA) and will then generate a colour-coded representation of the selected cluster. The colours that are assigned to clusters can be controlled by reading colour descriptions from an external file, and the dendrogram can be colour coded using the same colour scheme, which is very useful for analysing the data. By making BCL files corresponding to alternative cut-off levels in the dendrogram, the correspondence between the clustering process itself and the size and shape of the resulting clusters can be analysed in detail. Work is in progress to adapt this part of the output to other programs for molecular graphics, like Molscript (Kraulis, 1991). See Figure 3 (Crambin) for an example.

As described in the Algorithm section, it is also possible to use an alternative clustering technique, based on total inter-cluster contact area. One can specify cluster size and clustering cut-off for this approach in the same way as for single-linkage clustering. However, because of the inversions that are normally found in the standard plot mode, the clustering cut-off may be difficult to define. In these cases, it may be advantageous to use the alternative plot scale, base on total intra-cluster contact area for all clusters. This will be demonstrated in the first example (see below).

Discussion

Three examples will be discussed, in order to demonstrate how these tools may be used and what type of information one should expect to get. The first example, where crambin is analysed, is mainly used for demonstrating how some of the most important options in pdb_np_clus work. The second example, where the complex between the human growth hormone (hGH) and its binding protein (hGHbp) is studied, shows how these tools may be used to analyse and explain experimental data. In the third example, results from clustering on two different esterases are briefly described, showing

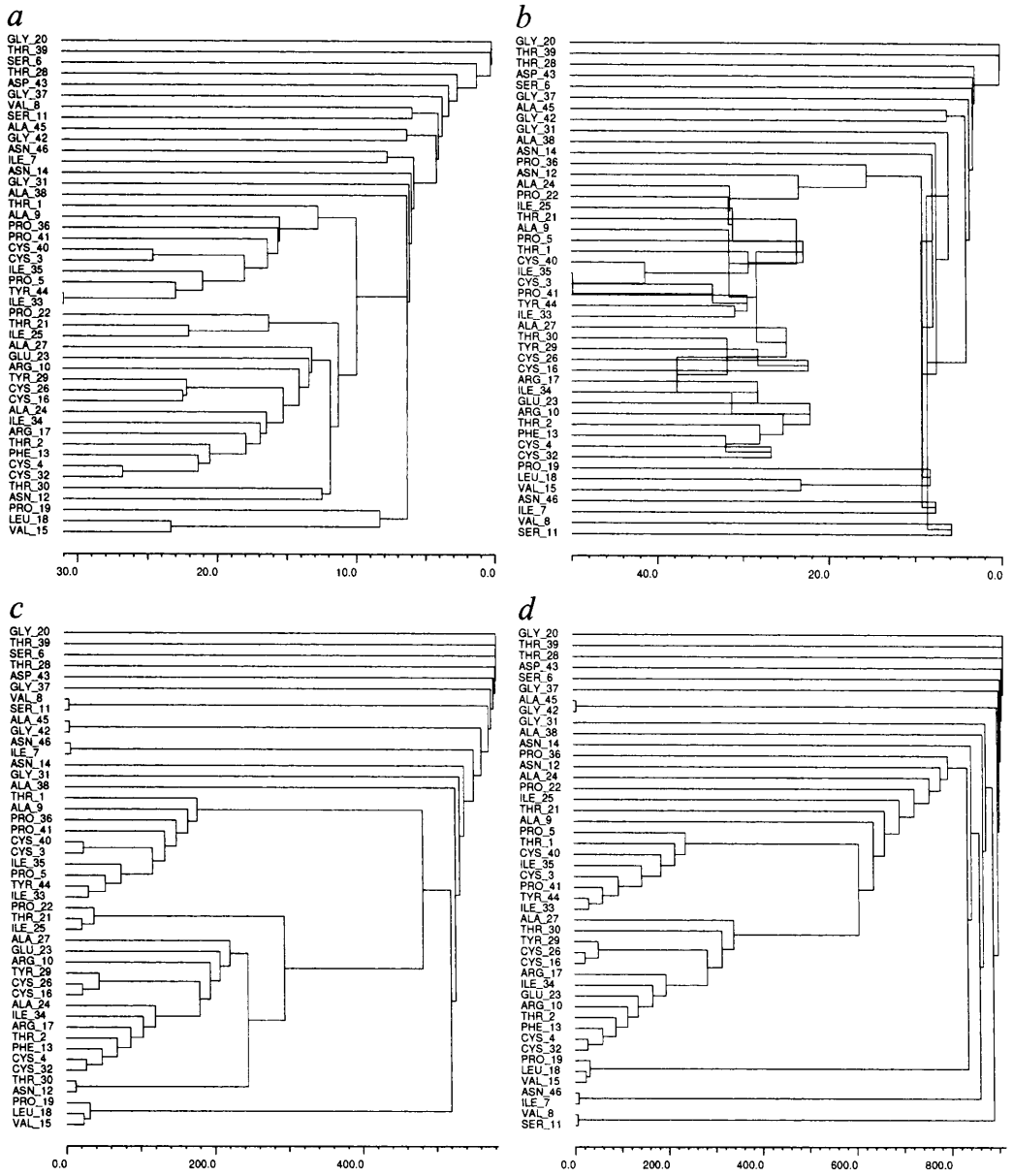

Fig. 2. Clustering of non-polar contacts in crambin using alternative approaches. (**a**) Default single-linkage clustering plotted versus the clustering scale. (**b**) Clustering based on total inter-cluster contact area plotted versus the clustering scale. (**c**) Single-linkage clustering plotted versus total intra-cluster contact area. (**d**) Clustering based on total inter-cluster contact area plotted versus total intra-cluster contact area.

that classification of contacts in related structures gives consistent results.

Crambin

Crambin is a hydrophobic 46 residue protein (Teeter *et al.*, 1981) with sequence similarity to thionins (Orrù *et al.*, 1997), a family of membrane-active toxins from plants. Crambin is found in the seeds of *Crambe abyssinica*, but the function of crambin in these seeds is still unknown.

For analysis of the crystal structure of crambin (Teeter, 1984), PDB reference code 1CRN was used. The four main approaches to clustering and dendrogram plotting are shown in Figure 2 (see the figure text for details). Two distinct clusters of residues can be seen, in particular when the dendrogram is drawn versus the total intra-cluster contact area (Figure 2c and d). These clusters can also be seen in the

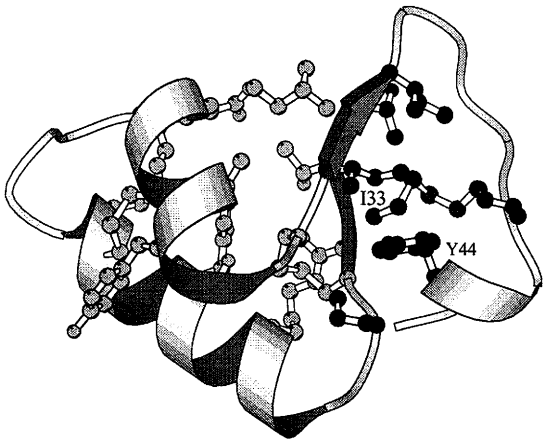

Fig. 3. The three-dimensional structure of crambin, showing the two clusters identified from Figure 2d displayed as ball-and-stick models in dark and light grey. The position of Ile-33 and Tyr-44 is indicated.

single-linkage dendrogram (Figure 2a), although they are less well separated from the rest of the dendrogram. The clustering based on the total inter-cluster contact area with the dendrogram drawn on the same scale (Figure 2b) is not very suitable for visual analysis.

The clusters formed at cut-off levels corresponding to comparable and well-defined cluster sizes (11.0 in Figure 2a, 400.0 in Figure 2a 2c and d) were analysed in detail. This showed that all residues found in the two main clusters in Figure 2d are also found in Figure 2a and c. The seven residues found in clusters in Figure 2a and c, but not in d, are all 'borderline cases' with relatively small contact areas. In general, clustering based on the total inter-cluster contact area seems to be well suited for generating tight and well-defined clusters.

The two main clusters from Figure 2d are visualized with Molscript (Kraulis, 1991) in Figure 3 as ball-and-stick models on a secondary structure representation of crambin. We see that the largest cluster links the two helices to the β-sheet region, whereas the smaller cluster links the β-sheet region to the C-terminal region. The single-linkage dendrogram is easier to analyse for important single-residue interactions. Inspection of Figure 2a shows that the largest single interaction is found between Ile-33 and Tyr-44. From Figure 3, we see that this interaction may be involved in keeping the C-terminal domain closely associated with the β-sheet region, and therefore these residues are probably important for the folding and stability of crambin in this region.

When the classification of residues into non-polar clusters is compared to a multiple alignment of the protein sequences of crambin, viscotoxins and thionins (Orrù *et al.*, 1997), it is seen that most conserved residues are found in the two main clusters identified in this analysis.

hGH/hGHbp

The growth hormone receptor is a trans-membrane protein involved in the regulation of processes related to growth. The extracellular part of the receptor, the growth hormone binding protein (GHbp), and in particular its interaction with growth hormone (GH), has been studied by a variety of approaches. The large amount of data generated on this system makes it a useful model for understanding interactions between 4-helix bundle growth factors in general and their receptors (Mott and Campbell, 1995). An essential step in growth hormone signalling is the formation of a 1:2 complex of GH:GHbp, where three inter-molecular contact surfaces can be defined: one GHbp–GHbp surface and two GH–GHbp surfaces. In this example, the non-polar properties of these contact surfaces will be analysed.

As discussed above, clustering based on the total inter-cluster contact area seems to be the best approach for the identification of robust, tight clusters. However, if we are looking at protein–protein interfaces, the non-polar residues at these interfaces will not necessarily form tight clusters, but more often loose networks of pairwise interactions. Such interactions are often not included in dendrograms based on total inter-cluster contact area, but can easily be identified in a single-linkage dendrogram.

The analysis is based on the crystal structure of human growth hormone and binding protein (DeVos *et al.*, 1992), PDB reference code 3HHR. Figure 4 shows a single-linkage dendrogram of the hGH/hGHbp complex. This complex consists of the growth hormone (A) and two molecules of binding protein (B and C). The complex is formed in a sequential manner (Cunningham *et al.*, 1991), so that the growth hormone first associates with one binding protein, followed by association to the second binding protein ($C + A + B \leftrightarrow C + AB \leftrightarrow CAB$). The dimerization of the receptor, which triggers the intracellular signal (Argetsinger and Carter-Su, 1996), is thus controlled by the growth hormone. The consequence of this is that the B–C interaction has to be weak, in order to avoid dimerization without the presence of growth hormone, whereas at least the A–B interaction has to be strong and specific, in order to bind the hormone in the correct orientation for subsequent dimerization. It has previously been shown that there is a significant electrostatic component in the A–B interaction (Cunningham and Wells, 1993), and the electrostatic interaction is probably important for initial orientation of the growth hormone. However, binding is not just a general electrostatic effect, as mutation of some of the charged residues in the binding interface had little effect on the on-rate of the hormone. This was the case for at least five residues in the growth hormone contact region, including K168 and K172.

The dendrogram in Figure 4 shows seven clusters. Most of the clusters are located within individual molecules: one in

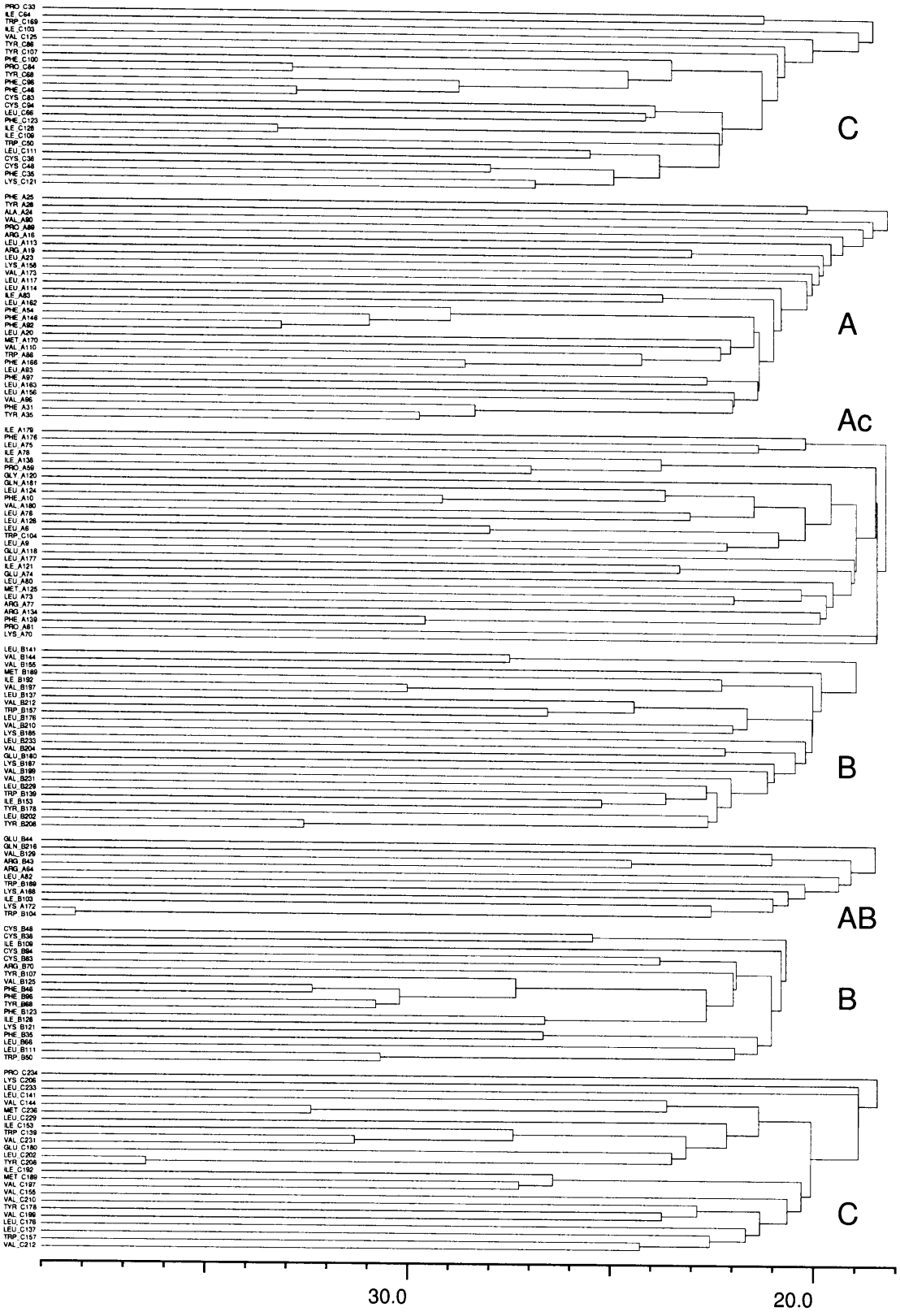

Fig. 4. A single-linkage dendrogram of the hGH/hGHbp complex, using a cut-off of 18.0 and removing clusters with less than eight members. The members of each cluster are indicated (A, B, C, Ac, AB).

A, two in B and two in C. These are structural clusters, and they are not important for inter-molecular interactions. One cluster in the dendrogram (Ac) is mainly from A, with only one residue from C, which is W104. One relatively small cluster (AB) has four residues from A (R64, L82, K168, K172) and seven residues from B (R43, E44, I103, W104, V129, W169, Q216). It has previously been shown that W104 and W169 in the binding protein have a very large influence on the binding free energy (Clackson and Wells, 1995). The current analysis confirms that they are important because of their non-polar interaction with residues from A, including K168 and K172. In particular, W104 has a very large contact area with K172, in fact the largest contact area seen in this complex.

Lipases/esterases

Lipases and esterases are water-soluble enzymes that catalyse the hydrolysis of ester bonds. A large number of related sequences and structures are known. In this example, an esterase from electric ray (*Torpedo californica*, PDB entry 2ACE) is compared to a lipase from yeast (*Candida rugosa*, PDB entry 1TRH). According to the FSSP database (Holm and Sander, 1996), these two structures can be aligned with an RMS deviation of 2.8 Å, although the sequence identity in this alignment is only 29%.

Single-linkage clustering gave a compact cluster in both structures (data not shown), consisting of 40 residues in the case of the esterase and 27 residues in the lipase. When these clusters are compared, using the alignment of the FSSP database, 21 residue positions are identified as common to both clusters, corresponding to 78% of the lipase cluster. These clusters are also found in the output from clustering based on total inter-cluster contact area, with 33 residues in the esterase and 29 residues in the lipase clusters. Here the number of common residue positions is similar, 22 positions or 76% of the lipase cluster. In clustering based on total contact area, a second compact cluster is found in each structure, consisting of 29 residues in the esterase and 25 residues in the lipase. In these clusters, 19 residue positions are common, corresponding to 76% of the lipase cluster. These clusters are not equally well defined in the single-linkage clustering, and this illustrates that clustering based on total contact area is the preferred method for identification of intra-molecular clusters. This example shows that clustering of related structures gives similar and consistent clusters.

The examples show that this new tool for the identification of clusters of non-polar interactions gives valuable information for understanding such interactions. By using the alternative approaches to clustering implemented in this software, the user can focus on different aspects of non-polar clusters, in particular inter-molecular complex-stabilizing clusters (or networks) and intra-molecular structural clusters, by using single-linkage clustering and total inter-cluster contact area clustering, respectively. The explorative approach to the identification of non-polar interactions found in the dendrogram representation of the clusters is unique to this software tool, and it is of great value for analysing the hierarchical nature of protein structure. It is an interesting question whether there is a correspondence between the hierarchy of non-polar interactions identified by this software and specific properties of the protein structure, e.g. protein folding. This has not yet been explored. However, others have previously shown that at least some folding processes seem to follow a hierarchical folding pathway [see, for example, Parker *et al.* (1996)].

Acknowledgements

The development of this tool has been supported by the European Commission (BIO2-CT94-3071) and the Norwegian Research Council (113733/420).

References

Abagyan,R.A. and Totrov,M.M. (1997) Contact area difference (CAD): A robust measure to evaluate accuracy of protein models. *J. Mol. Biol*., **268**, 678–685.

Abola,E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) Protein Data Bank. In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*. Data Commission of the International Union of Crystallography, Bonn, pp. 107–132.

Argetsinger,L.S. and Carter-Su,C. (1996) Mechanism of signaling by growth hormone receptor. *Physiol. Rev.*, **76**, 1089–1107.

Bagley,S.C. and Altman,R.B. (1995) Characterizing the microenvironment surrounding protein sites. *Protein Sci*., **4**, 622–635.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol*., **112**, 535–542.

Bordo,D. (1993) ENVIRON: a software package to compare protein three-dimensional structures with homologous sequences using local structural motifs. *Comput. Applic. Biosci*., **9**, 639–645.

Bowie,J.U., Lüthy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Brünger,A. (1992) *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR*. Yale University Press, New Haven, CT.

Casari,G. and Sippl,M.J. (1992) Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol*., **224**, 725–732.

Clackson,T. and Wells,J.A. (1995) A hot-spot of binding-energy in a hormone-receptor interface. *Science*, **267**, 383–386.

Connolly,M.L. (1983) Analytical molecular surface calculation. *J. Appl. Crystallogr*., **16**, 548–558.

Cunningham,B.C. and Wells,J.A. (1993) Comparison of a structural and a functional epitope. *J. Mol. Biol*., **234**, 554–563.

Cunningham,B.C., Ultsch,M., de Vos,A.M., Mulkerrin,M.G., Clauser,K.R. and Wells,J.A. (1991) Dimerization of the extracellular domain of the human growth hormone receptor by a single hormone molecule. *Science*, **254**, 821–825.

DeVos,A.M., Ultsch,M. and Kossiakoff,A.A. (1992) Human growth-hormone and extracellular domain of its receptor—crystal-structure of the complex. *Science*, **255**, 306–312.

Everitt,B.S. and Dunn,G. (1983) *Advanced Methods of Data Exploration and Modelling*. Heinemann Educational, London.

Flower,D.R. (1997) SERF: A program for accessible surface area calculations. *J. Mol. Graph. Model*., **15**, 238–244.

Gerstein,M. and Altman,R.B. (1995a) Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol*., **251**, 161–175.

Gerstein,M. and Altman,R.B. (1995b) Using a measure of structural variation to define a core for the globins. *Comput. Applic. Biosci*., **11**, 633–644.

Gromiha,M.M. and Selvaraj,S. (1997) Influence of medium and long range interactions in $(\alpha/\beta)_8$ barrel proteins. *J. Biol. Phys*., **23**, 209–217.

Heringa,J. and Argos,P. (1991) Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol*., **220**, 151–171.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.

Islam,S.A., Luo,J. and Sternberg,M.J. (1995) Identification and analysis of domains in proteins. *Protein Eng*., **8**, 513–525.

Jones,S. and Thornton,J.M. (1996) Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Karlin,S. and Zhu,Z.Y. (1996) Characterizations of diverse residue clusters in protein three- dimensional structures. *Proc. Natl Acad. Sci. USA*, **93**, 8344–8349.

Karlin,S., Zuker,M. and Brocchieri,L. (1994) Measuring residue associations in protein structures. Possible implications for protein folding. *J. Mol. Biol*., **239**, 227–248.

Karpeisky,M. and Ilyin,V.A. (1992) Analysis of non-polar regions in proteins. *J. Mol. Biol*., **224**, 629–638.

Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr*., **24**, 946–950.

Mott,H.R. and Campbell,I.D. (1995) 4-helix bundle growth-factors and their receptors—protein–protein interactions. *Curr. Opin. Struct. Biol*., **5**, 114–121.

Orrù,S., Scaloni,A., Giannattasio,M., Urech,K., Pucci,P. and Schaller,G. (1997) Amino acid sequence, S-S bridge arrangement and distribution in plant tissues of thionins from *Viscum album*. *Biol. Chem*., **378**, 989–996.

Pace,C.N., Shirley,B.A., McNutt,M. and Gajiwala,K. (1996) Forces contributing to the conformational stability of proteins. *FASEB J*., **10**, 75–83.

Parker,M.J., Sessions,R.B., Badcoe,I.G. and Clarke,A.R. (1996) The development of tertiary interactions during the folding of a large protein. *Fold. Des*., **1**, 145–156.

Richmond,T.J. and Richards,F.M. (1978) Packing of α-helices: Geometrical constraints and contact areas. *J. Mol. Biol*., **119**, 537–555.

Rigby,M., Smith,E.B., Wakeham,W.A. and Maitland,G.C. (1986) *The Forces Between Molecules*. Clarendon Press, Oxford.

Schmidt,R., Gerstein,M. and Altman,R.B. (1997) LPFC: an Internet library of protein family core structures. *Protein Sci.*, **6**, 246–248.

Selvaraj,S. and Gromiha,M.M. (1998) An analysis of the amino acid clustering pattern in $(\alpha/\beta)_8$ barrel proteins. *J. Protein Chem*., **17**, 407–415.

Shakhnovich,E., Abkevich,V. and Ptitsyn,O. (1996) Conserved residues and the mechanism of protein folding. *Nature*, **379**, 96–98.

Shrake,A. and Rupley,J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol*., **79**, 351–371.

Sowdhamini,R. and Blundell,T.L. (1995) An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci*., **4**, 506–520.

Swindells,M.B. (1995a) A procedure for detecting structural domains in proteins. *Protein Sci*., **4**, 103–112.

Swindells,M.B. (1995b) A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci*., **4**, 93–102.

Teeter,M.M. (1984) Water structure of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in crystals of crambin. *Proc. Natl Acad. Sci. USA*, **81**, 6014–6018.

Teeter,M.M., Mazer,J.A. and L'Italien,J.J. (1981) Primary structure of the hydrophobic plant protein crambin. *Biochemistry*, **20**, 5437–5443.

Tsai,C.J. and Nussinov,R. (1997a) Hydrophobic folding units at protein–protein interfaces: implications to protein folding and to protein–protein association. *Protein Sci*., **6**, 1426–1437.

Tsai,C.J. and Nussinov,R. (1997b) Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci*., **6**, 24–42.

Tsai,C.J., Xu,D. and Nussinov,R. (1997) Structural motifs at protein–protein interfaces: protein cores versus two-state and three-state model complexes. *Protein Sci*., **6**, 1793–1805.

Vajda,S., Sippl,M. and Novotny,J. (1997) Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol*., **7**, 222–228.

Zehfus,M.H. (1995) Automatic recognition of hydrophobic clusters and their correlation with protein folding units. *Protein Sci*., **4**, 1188–1202.

Zehfus,M.H. (1997) Identification of compact, hydrophobically stabilized domains and modules containing multiple peptide chains. *Protein Sci*., **6**, 1210–1219.

Zhu,Z.Y. and Karlin,S. (1996) Clusters of charged residues in protein three-dimensional structures. *Proc. Natl Acad. Sci. USA*, **93**, 8350–8355.