doi:10.1006/jmbi.2000.4092 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 302, 917–926

JMB

A Fast Method to Predict Protein Interaction Sites from Sequences

Xavier Gallet[1], Benoit Charloteaux[1], Annick Thomas[2] and Robert Brasseur[1]*

[1]*Centre de Biophysique Moléculaire Numérique, Faculté Agronomique, 5030 Gembloux Belgium*

[2]*INSERM Unité 410, Hôpital X. Bichat, 75018 Paris, France*

**Corresponding author*

A simple method for predicting residues involved in protein interaction sites is proposed. In the absence of any structural report, the procedure identifies linear stretches of sequences as ''receptor-binding domains'' (RBDs) by analysing hydrophobicity distribution. The sequences of two databases of non-homologous interaction sites eliciting various biological activities were tested; 59-80 % were detected as RBDs. A statistical analysis of amino acid frequencies was carried out in known interaction sites and in predicted RBDs. RBDs were predicted from the 80,000 sequences of the Swissprot database. In both cases, arginine is the most frequently occurring residue. The RBD procedure can also detect residues involved in specific interaction sites such as the DNA-binding (95 % detected) and Ca-binding domains (83 % detected). We report two recent analyses; from the prediction of RBDs in sequences to the experimental demonstration of the functional activities. The examples concern a retroviral Gag protein and a penicillin-binding protein. We support that this method is a quick way to predict protein interaction sites from sequences and is helpful for guiding experiments such as site-specific mutageneses, two-hybrid systems or the synthesis of inhibitors.

Keywords: receptor-binding domain; interaction site; hydrophobic moment; hydrophobicity; sequence

Introduction

Protein interaction sites are critical domains for selective recognition of molecules and for the formation of complexes. They are responsible for diverse important biological functions. Therefore, detection of interaction domains in sequences could help in identifying protein function. It could also help, for example, to validate functional hypotheses *via* the design of restricted fragments for two-hybrid assays (Vidal *et al*., 1996) or of specific mutageneses (Phizicky & Fields, 1995). Computational methods are of great interest in predicting protein interacting pairs, and thus to construct metabolic pathways or signalling cascades for recently sequenced genomes. The prediction of interaction sites should be a good starting point to help identify pharmacological targets and help drug design studies. Such analyses require the elaboration of docking procedures (Janin, 1995; Shoichet & Kuntz, 1996; Sternberg *et al*., 1998), the knowledge of protein and ligand structures (Bamborough & Cohen, 1996) and the consideration of conformational changes (Betts & Sternberg, 1999).

Several methods exist for predicting protein structure; they identify interaction domains by analysing the hydrophobicity, solvation, protrusion and the accessibility of residues (Young *et al*., 1994; Jones & Thornton, 1997a,b). Those approaches are interesting but cannot answer requests of the great number of biochemists with sequences, but no structural data. Indeed, despite the amount of protein structures already solved, the bank of structures is ridiculously small as compared to those of sequences.

To our knowledge, very few methods use sequences as their starting point. The algorithm by Kini & Evans (1995) supports that proline residues frequently occur near interaction sites. The frequency is 2.5 times higher than expected by random distribution. They suggest that ''proline-brackets'' encircle a large number of protein-

Abbreviations used: RBD, receptor-binding domain; NDV, Newcastle disease virsus.

E-mail address of the corresponding author: brasseur.r@fsagx.ac.be

0022-2836/00/040917–10 $35.00/0

protein interaction sites (Kini & Evans, 1996). Another method uses multiple sequence alignments and focuses on correlated mutations to detect protein interacting sites (Pazos *et al*., 1997). The hypothesis is that residues close to protein-protein interaction sites tend to mutate simultaneously during evolution. Therefore, from multiple sequence alignments, the authors detect the residues linking different protein domains and interacting in heterodimer complexes.

In a recent analysis, Marcotte *et al*. (1999) report that they can predict which proteins interact by analysing genome sequences. The hypothesis is that two proteins are interacting if, in another living organism, they are assembled as a single protein. The procedure is also very powerful to predict the functions of wide protein complexes if one can trace domain homologies. However, the procedure gives no information on the interacting amino acids *per se*.

Here, we test a fast and simple method to predict stretches of protein interaction sites from sequences in the absence of any structural report. Eisenberg *et al*. (1982) previously showed that plotting the mean alpha-helical hydrophobic moment $\langle\mu_H\rangle$ *versus* the mean hydrophobicity $\langle H\rangle$ allows us to classify protein fragments according to their location in the structures; either they are membrane segments, parts of globular domains or surface-seeking helices. The authors demonstrated that a high level of hydrophobicity, together with a low hydrophobic moment, support that the fragment is membranous, whereas residues from surface-seeking helices cover a wide diagonal area beginning at the upper left of the plot (Figure 1). The diagram was thus divided into four regions corresponding to globular, surface and membrane (monomeric and multimeric) domains, called G, S and M, respectively (Eisenberg *et al*., 1984; see Figure 1(a)). Here, a fifth domain, the ''receptor-binding domain'' (RBD) is investigated in which we detect some residues of protein interaction sites. The RBD method is described and is applied to different sequence databases. Results show that the plot drawn from the Eisenberg's method detects most of the experimentally known interaction sites. The effects of several parameters of the procedure were tested. The structures, the accessibility and the functional characterisation of predicted sites were also investigated on few 3D structures. The results obtained with the DNA-binding and the calcium-binding sequences and with the 3D structures, such as the ultrabithorax-extradenticle-DNA complex and the calcium-binding protein, demonstrate that our procedure can detect various types of interaction sites as long as they involve hydrophilic residues. Finally, we demonstrate that the RBD analysis could be valuable in identifying mutations. Two examples, in the Mason-Pfizer monkey virus Gag protein and in a penicillin-binding protein, are shown.

Results and Discussion

Apolipoprotein E and Newcastle disease virus fusion protein analysis

In the analysis of the apolipoprotein E sequence, De Loof *et al*. (1986) extended the concept previously suggested by Eisenberg by considering an

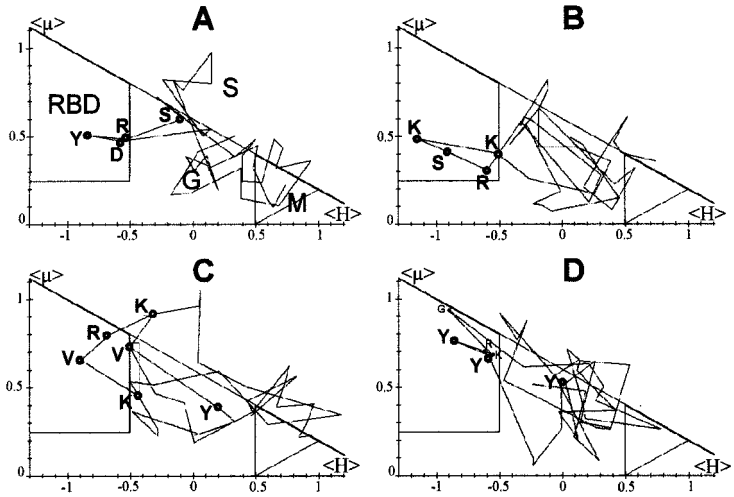


Figure 1. Plots of $\langle\mu_H\rangle$ *versus* $\langle H\rangle$ for four fragments of sequences from Table 1 ($N = 5$, $\delta = 100°$). Eisenberg *et al*. (1982) defined the G (globular), M (membrane) and S (surface) areas. The RBD area is the trapezium defined by $\langle H\rangle$ values lower than -0.5, $\langle\mu_H\rangle$ values greater than 0.25 and by the equation $\langle\mu_H\rangle = -0.4\langle H\rangle + 0.6$. Amino acid residues involved in known interaction sites are marked (○). Other residues lying in the RBD area are annotated (d). (a) SRYD binding site of the *Leishmanolysin* gp63 protein. (b) RSKK binding site of the human α1 follicle stimulating hormone. (c) YVKRVK binding site of the human band 3 protein. (d) Phosphorylated tyrosine residues in the chicken insulin receptor tyrosine kinase protein.

additional region of the hydrophobicity/hydrophobic moment plot that they called ''receptor-binding-domain'' (RBD). The RBD method is thus based on the calculation of the mean hydrophobic moment $\langle\mu_H\rangle$ and the mean hydrophobicity $\langle H\rangle$ of an N-residue window (N being odd) centred at the amino acid of interest. The δ angle is 100° to correspond to the calculation of the α-helical hydrophobic moment. The N-residue window is moved along the sequence. By using an 11-residue window, the authors predicted that two stretches of the apolipoprotein E sequence (Arg136-Ala160 and Leu214-Val236), exhibiting low mean hydrophobicity ($\langle H\rangle < -0.5$) and variable mean hydrophobic moments ($0.3 < \langle\mu_H\rangle < 0.8$) could be protein binding domains. The first stretch was already checked by mutant studies (Malhey *et al*., 1984): Malhey *et al*. (1984) had demonstrated that mutation R158C decreased the receptor binding of apolipoprotein E by 98%. De Loof *et al*. (1986) showed that change of arginine (R) to cysteine (C) shifted most residues off the RBD area in the Eisenberg's plot. Actually, the stretch is a binding site for the LDL apolipoprotein (B-E) receptor and part of the heparin-binding domain. The second stretch had no known function at the time the paper was written. It was later demonstrated that it is another part of the heparin-binding site (Cardin *et al*., 1986; Weisgraber *et al*., 1986).

The RBD procedure was also used to analyse the sequence of the Newcastle disease virus (NDV) fusion protein. Le *et al*. (1988) studied different virus strains, virulent and inactive. From the plot of $\langle\mu_H\rangle$ *versus* $\langle H\rangle$ with a seven-residue window, they demonstrated that a fragment of sequence located upstream the cleavage site is plotted in the RBD area of the Eisenberg's plot. The location is related to the virulence because the inactive virus strains have no residues in the RBD area. It was suggested that an increased level of hydrophobicity of the fragment prevents accessibility to the cleavage site and thus impairs the cleavage of the fusion protein for the non-virulent strains.

From a series of analyses that are not all reported here, we re-drew the RBD region of the Eisenberg's plot as a trapezoid area so that $\langle H\rangle < 0.5$, $\langle\mu_H\rangle > 0.25$ with $\langle\mu_H\rangle = -0.4\langle H\rangle + 0.6$.

Kini's and DIP databases analysis

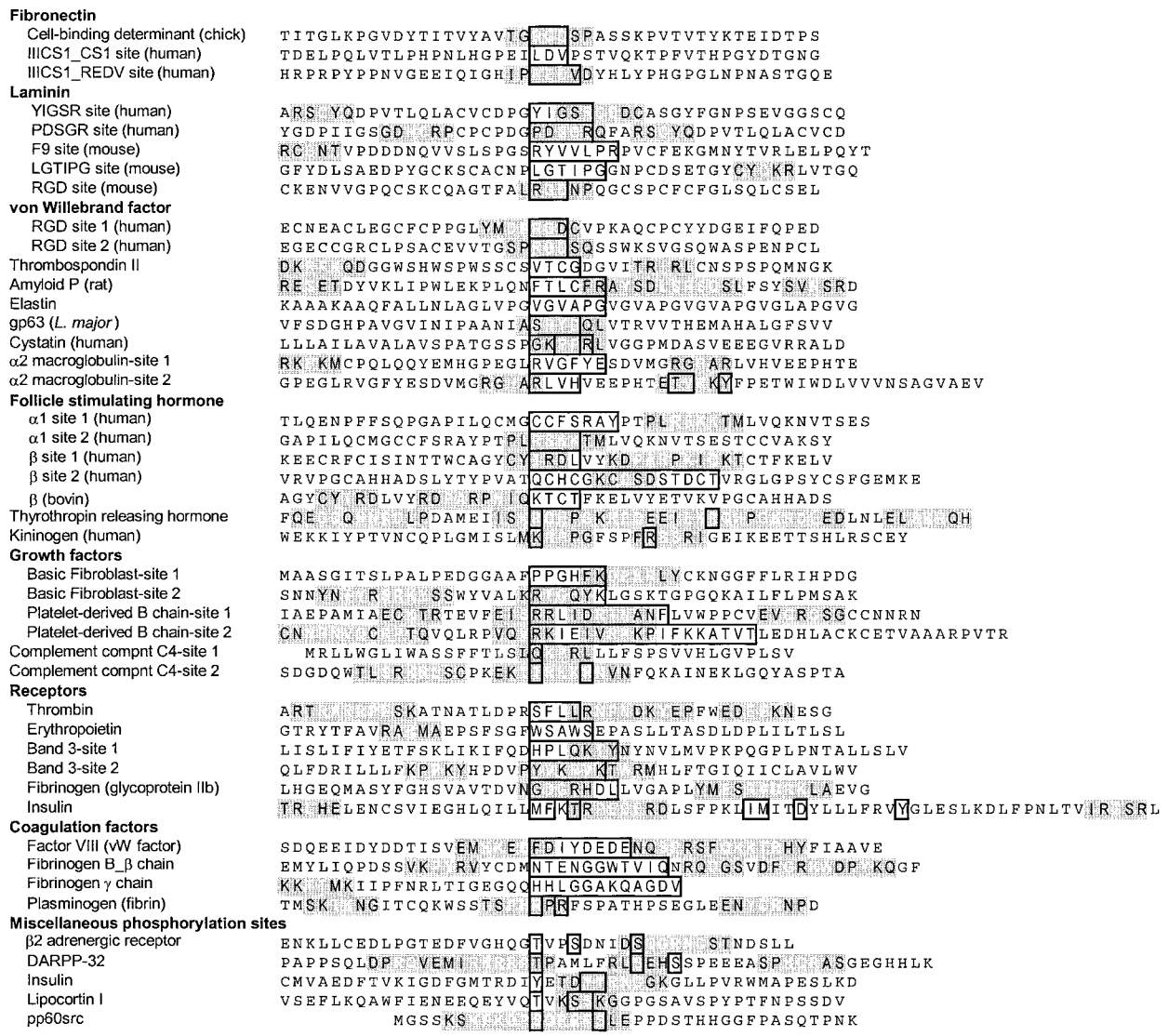
In previous studies, the RBD method was applied to two sequences (apolipoprotein E and NDV fusion protein). The α-helical hydrophobic moment ($\delta = 100^\circ$) was calculated but different windows were used (11 and 7, respectively). To generalise the procedure, we tested several databases of known interaction sites. The former was kindly provided by Kini (Kini & Evans, 1995) and contains about 1600 sequences. The second database is the DIP (Database of Interacting Proteins) with 1359 entries of protein-protein interactions. Those sites were experimentally identified by several laboratories and correspond to different biological activities. From Kini's database, 818 non-redundant sequences were used. The average length of these sequences is 14 amino acid residues. In the DIP, 244 sequences of interaction domains are listed. After discarding the redundant sequences and sequences larger than 100 residues, 136 fragments were kept, with an average length of 55 amino acid residues. Several parameters were tested; the window width (five, seven, nine and 11 residues were compared) and the δ angles (85°, 100° and 170° corresponding to β-turn, α-helix and β-sheet conformations, respectively). Table 1 lists the percentage of sequences where a ''RBD'' was detected. For Kini's database, as well as DIP database, the window width is crucial, while the δ angle has almost no influence. The best results are obtained with a window of five residues and an δ angle of 100°. In those conditions, interaction sites are predicted in 59.1% of Kini's database sequences and 80.1% of the DIP sequences. Increasing the window width significantly decreases the prediction efficiency while modifying the δ angle to 85° and 170° give similar percentages. A closer examination shows that the predicted residues are not strictly the same with the different δ angles. However, 54% (Kini's database) and 63% (DIP database) of the residues detected at 100° are also detected with δ equal to 85° and 170° (data not shown), especially when a short window is used. Differences often correspond to slight shifts along sequence so that, detected fragments are overlapping.

Table 1. Percentage of sequences in the databases

	window width (N)			
A	***5***	***7***	***9***	***11***
δ=85° (β-turn)	59.6	42.6	30.2	20
δ=100° (α-helix)	59.1	42.5	30.5	20.4
δ=170° (β-sheet)	56.4	40.4	26.2	16.2
B	***5***	***7***	***9***	***11***
δ=85° (β-turn)	78.6	66.9	55.8	45.5
δ=100° (α-helix)	80.1	64.7	57.3	47.8
δ=170° (β-sheet)	80.8	63.2	54.4	45.5
C	***5***	***7***	***9***	***11***
δ=85° (β-turn)	95.3	88.6	80.2	72.6
δ=100° (α-helix)	94.7	88.8	80.7	73.2
δ=170° (β-sheet)	95	88.2	79	72.5
D	***5***	***7***	***9***	***11***
δ=85° (β-turn)	84.4	51	32	25.8
δ=100° (α-helix)	82.7	52.2	33.7	20.1
δ=170° (β-sheet)	75.7	44.7	24.6	18.2

Percentage of sequences in the Kini's (A: 818 sequences), in the DIP (B: Database of Interacting Proteins; 136 sequences), in the ''DNA-interaction'' (C: 2298 sequences) and in the ''Ca-interaction'' (D: 527 sequences) databases in which RBDs were detected according to the window width (N) and the δ angle.

Table 2. Comparison of experimental and computed predictions of interaction sites in 45 fragments of proteins



Prediction of RBD in the database of 45 known interaction sites. The experimental interaction sites are boxed and their amino acid residues are in bold with a larger font than the rest of the sequence. The predicted RBD are in white and all the residues of the window (i.e. the residues involved in the calculation of the RBD, window width = 5) are shaded with grey.

45 sequences Kini's sub-database analysis

The residues directly in the interaction sites are known in some of Kini's database sequences. Some 45 of those sequences were selected and tested (Table 2). We predicted at least one interaction site in 98 % of them, but our prediction did not always match with the experimentally designated residues: 55 % of the residues experimentally involved in the interaction were detected by our method. For the SRYD site of the gp63 protein, the RSKK site of the human α1 follicle-stimulating hormone and the YVKRVK site of the human band 3 protein, all residues were detected (Figure 1). Other RBD stretches were close to the experimental binding domains as for the basic fibroblast growth factor and the bovine β follicle-stimulating hormone. For the fibronectin CS1 site and the elastin site, our procedure failed to detect any binding site. Those fragments contain mainly apolar amino acid residues; they are too hydrophobic to fit in the RBD criteria. For discontinuous interaction sites as phosphorylation sites, 58 % of the phosphorylable amino acid residues, serine, threonine and tyrosine were detected.

Procedure selectivity

The 80,000 sequences of the Swissprot database were screened using a five-residue window and δ equal to 100° (Figure 2(a)). The most frequent residues in RBD are arginine (19.6 %), lysine (12.5 %), glutamic acid (8.9 %), serine (7.0 %) and aspartic acid (6.9 %). When those frequencies are compared to the overall frequencies of these amino acids in the Swissprot, we conclude that the arginine is 3.8 times more frequent in RBD than a random distribution would predict. The scale of residue enrichment in RBD is as follows: Arg: 3.8 > Lys: 2.1 > Gln: 1.5 > Glu: 1.4 > Asp = Asn: 1.3 > His: 1.2 > Ser: 1.0 > Thr = Tyr: 0.9 > Pro = Cys: 0.8 > Gly = Met: 0.6 > Ala = Trp: 0.5 > Leu: 0.4 > Val = Phe: 0.3 > Ile: 0.2.

This highlights that the RBD procedure mainly detects charged and hydrophilic amino acid residues. Few hydrophobic residues are detected: their occurrences are lower in the RBD screening than in the Swissprot database.

To calibrate our results with respect to experimental data, we compared the residue composition of the 45 well-known interaction sites of Kini's database to the residue selection by the RBD procedure (Figure 2(a)). In both cases, arginine was the most frequent residue (Janin & Chothia, 1990). Arginine is 12.8 % in Kini's bank, as compared to 19.6 in the RBD screening. The major differences are found with glycine that is hardly detected in the RBD but is more frequent than random in the experimental sites.

Occurrence of amino acid residues in the complete Kini's database and in the DIP and the Swissprot databases were also compared to explore a larger set of data (Figure 2(b)). Even if not all residues are implicated in interactions, these two databases contain more charged residues such as arginine (R), lysine (K), aspartic (D) and glutamic (E) acids, and more proline (P), tyrosine (Y), cysteine (C) and methionine (M) residues than the Swissprot. By contrast, hydrophobic residues such as tryptophan (W), leucine (L), valine (V), phenylalanine (F) and isoleucine (I) are under-represented in these databases with respect to the Swissprot. In summary, our procedure points out charged residues that should be involved in interaction sites but underestimates the contribution of residues such as proline, cysteine, glycine and methionine (Figures 2 and 3).

Why is arginine more frequent in protein interaction sites than lysine or than negatively charged residues? The molecular hydrophobicity potentials (MHP) of amino acid side-chains (Brasseur, 1991) indicates that arginine has one of the widest radii of action, since its isopotential surface spreads farther than that of any other charged residue (Figure 3). The radius is comparable in asparagine and glutamine, but these should be involved in hydrogen bonds that require directional attack. Moreover, the charge of arginine is carried by one of the longest side-chains and thus, in a folded structure, the charge could be more accessible to a partner. This supports that arginine is the most frequent side-chain of interaction sites because it involves electrostatic interactions which require lower stereo-selectivity angle of attack than hydrogen bonds and because the charge should often be protruding at the protein surface.

Screening of specific interaction sites

We investigated the Swissprot databank to demonstrate that the RBD method is able to detect protein-DNA and protein-ion interaction

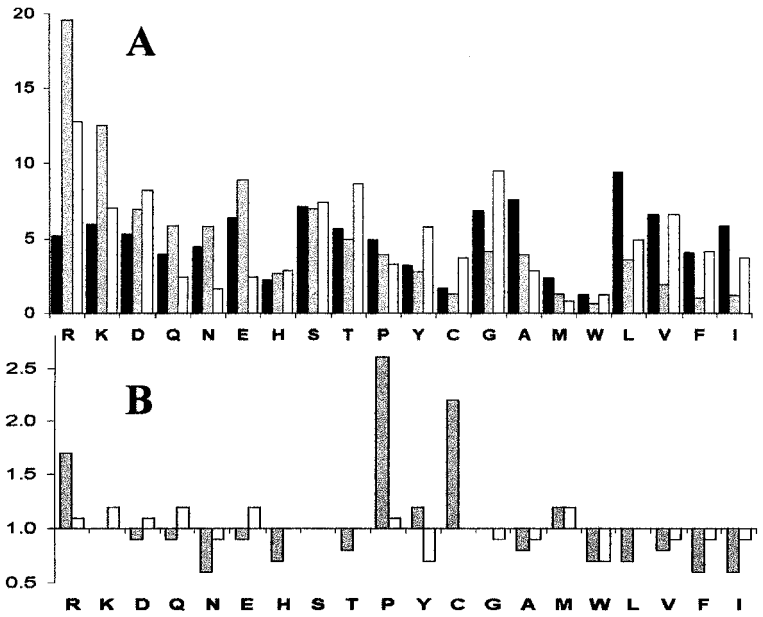


Figure 2. (a) Percentage of amino acid residues occurrence in the Swissprot database (black bars), in the RBD (grey bars) and in the 45 well-known protein interaction sites from the Kini's database (white bars). The residues are listed by increasing levels of hydrophobicity (from left to right) according to the Eisenberg's consensus hydrophobicity scale. RBD were detected with a five-residue window and an δ angle of 100°. (b) Percentage of amino acid residues occurrence in the sequences of the Kini's (grey bars) and the DIP (white bars) databases compared to percentage of amino acid residues occurrence in the Swissprot bank.

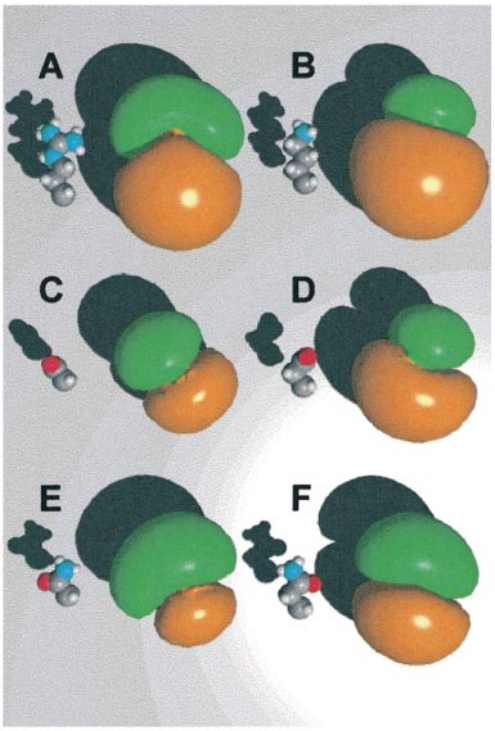


Figure 3. Distribution of MHP for the side-chains of arginine (a), lysine, (b), aspartic acid (c), glutamic acid (d), asparagine (e) and glutamine (f) residues. Atoms are displayed in CPK (Corey-Pauling-Koltun) representation. Orange surfaces are hydrophobic potentials and green, hydrophilic surfaces.

sites. Entries containing annotations such as ''DNA_BIND'' and ''CA_BIND'' were selected. They also provided running numbers of the segments of sequence that interact with DNA and calcium ions, respectively. Redundant sequences and fragments larger than 100 residues were removed. Finally, 2298 segments interacting with DNA, and 527 fragments interacting with calcium ions were selected. The corresponding databases were called ''DNA-interaction'' and ''Ca-interaction''. The average lengths of the sequences are 30 and 21 amino acid residues, respectively.

With a five-residue window and a δ value of 100°, 94.7% of the protein-DNA interacting sites and 82.7% of the protein-calcium ions binding sites are detected (Table 1).

Protein structures analysis

Features of known and predicted sites of few three-dimensional structures of the previous databases were analysed: the human fibronectin (Kini's database), the DNA-bound ultrabithorax-extradenticle complex (DNA-interaction database) and the calcium-binding protein (Ca-interaction database). Three RBD were detected in the sequence of the human fibronectin (PDB entry, 1fna) (Figure 4). The first one is the well-known RGD site, a domain for cell-attachment. The site is very accessible to the solvent and the molecular hydrophobic potential (MHP) surface, Brasseur, 1991) shows that it is both hydrophobic and hydrophilic (Figure 4(b)). The hydrophobic side corresponds to the alkyl side-chain of Arg73. The two other predicted RBDs, Tyr26-Tyr27 and Arg88 are also in solvent-accessible protein fragments. The Tyr26-Tyr27 RBD is two consecutive tyrosine residues and in the 3D structure, Tyr27 is buried in a pocket, its hydroxyl group pointing outside, 11 Å away from the RGD site. The last RBD (Arg88) is at the C-terminal end.

On the structure of the DNA-bound ultrabithorax-extradenticle complex (PDB entry 1b8i, Swissprot entries P02834 and P40427), the proteins

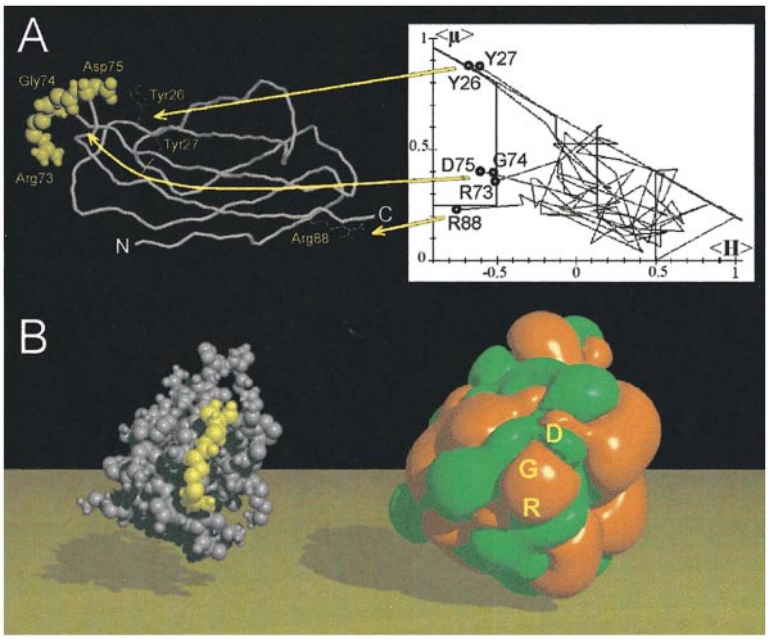


Figure 4. Mapping of RBD in the ribbon of the human fibronectin (PDB entry 1fna). (a) From the plot of $\langle \mu_H \rangle$ *versus* $\langle H \rangle$, three RBD are detected including the already known RGD site (shown in yellow CPK). The two other RBD (Tyr26-Tyr27 and Arg88) are displayed. (b) CPK model of fibronectin (the RGD site is in yellow) and plot of the MHP surface of the similarly oriented molecule. The MHP is calculated as explained in Materials and Methods. The RGD site has a hydrophilic and a hydrophobic surface. Orange surfaces are hydrophobic potentials and green, hydrophilic surfaces.

of the complex form homeodomains that bind opposite sides of the DNA (Passner *et al*., 1999). The amino acid residues interacting with the DNA were compared with those detected in the RBD analysis of the sequence (Figure 5): 16 residues (25.8 % of the sequence) of the ultrabithorax protein are in contact with the DNA; eight are detected by the RBD procedure. No RBD corresponds to the YPWM motif of ultrabithorax/extradenticle-interacting domain because of its low level of hydrophilicity. The two major interacting sites of the ultrabithorax protein, Arg7-Tyr10 and Arg45-Lys59 are detected: RBD detects the Gln8-Tyr13 and Arg54-Lys59 fragments. In addition, the RBD procedure detects a charged site (Thr29-Ile34) which is located at the N-end of helix α2 and is about 10 Å from the negatively charged oxygen atom of the DNA phosphate groups. For the extradenticle protein, among the 11 amino acid residues that bind to DNA (18.3 % of the sequence), four are detected in the RBD (Lys30, Arg52, Ile53 and Arg54).

In the calcium-binding protein (PDB entry 3icb, Swissprot entry P02633), each calcium-binding motif is made of five residues (Ala14, Glu17, Asp19, Gln22, Glu27 and Asp54, Asn56, Asp58, Glu60, Glu65, respectively). Using a five-residue window, three of those residues were detected: Glu27, Asp54 and Asn56. Besides this, three other amino acid residues were also detected: Pro3, Gly18, and Leu53. Except for Pro3, all the others are associated with the calcium-binding function. Leu53 is important for holding in place the hydrophobic core of the calcium-binding site, and Gly18 (separating Glu17 and Asp19 calcium-binding residues) is crucial for helix-loop-helix flexibility, i.e. for calcium affinity.

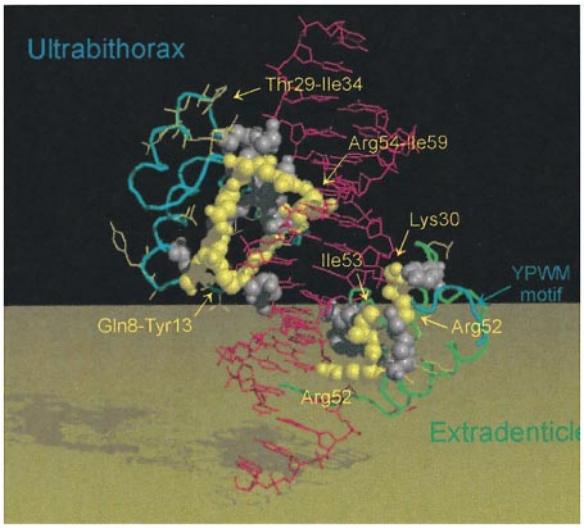


Figure 5. Mapping of known and predicted interaction sites in the 3D structure of the DNA-bound ultrabithorax-extradenticle complex (PDB entry 1b8i, Swissprot entries P02834 and P40427). The DNA is magenta and the ultrabithorax and extradenticle proteins are represented as a blue and a green ribbon, respectively. The YPWM motif is indicated. Residues involved in the protein-DNA interaction are shown in CPK representation and amino acid residues predicted as RBD are yellow.

Prediction of mutations

From predicting interaction sites, it seems it would be possible to predict point mutations; the RBD method could then be useful for predicting changes that would modify the interaction but not the structure. Three examples are described.

De Loof *et al*. (1986) previously reported the first example. The R158C shifted the fragment of sequence off the RBD area and inhibited apolipoprotein E binding activity.

Another example was recently derived from the conversion of the Mason-Pfizer monkey virus (M-PMV) morphogenesis. Some mutations are described and were studied *a posteriori* using the RBD prediction method. Functional inhibition is correlated to the displacement of the mutated residue out of the RBD trapezoid in the Eisenberg's plot. The M-PMV has a D-type morphogenesis: once assembled in the cytoplasm, its capsid migrates to the plasma membrane for budding. The capsid of other retroviruses, C-type retrovirus, assembles and buds simultaneously at the plasma membrane (Swanstrom & Wills, 1997). Rhee & Hunter (1990) described that the R55W mutation in the M-PMV matrix protein (the N-terminal domain of the Gag precursor) converts the D-type morphogenesis into a C-type. In our procedure, the fragment Arg55-Arg57 of the native virus is included in the RBD area. The R55W mutation displaces the fragment out of the RBD towards the surface domain (Figure 7). Our prediction suggests that the Arg55-Arg57 segment is a binding site implicated in the viral capsid assembly and that the R55W mutation causes its inactivation. By homology modelling, a three-dimensional structure of the R55W mutant was built using Modeller (Sali & Blundell, 1993) and the M-PMV matrix protein structure as template (Conte *et al*., 1997). Comparison of MHP for the native and the mutated structures shows that the patch corresponding to Arg55 is hydrophilic and becomes hydrophobic when the arginine (R) is changed to a tryptophan (W) residue in the mutant (Figure 7). Recently, Choi *et al*. (1999) identified in the M-PMV Gag protein an 18-residue fragment that could be crucial for the retroviral morphogenesis. This fragment includes Arg55 and the Arg55-Arg57 RBD. When the fragment is inserted into the MoMuLV (Moloney murine tumor virus) Gag protein, the C-type morphogenesis switches to a D-type. The authors propose that this stretch functions as a cytoplasmic targeting/retention signal peptide and is thus an interaction site as supported by our RBD detection.

The RBD procedure was finally applied to the class B penicillin-binding proteins (PBP). In that

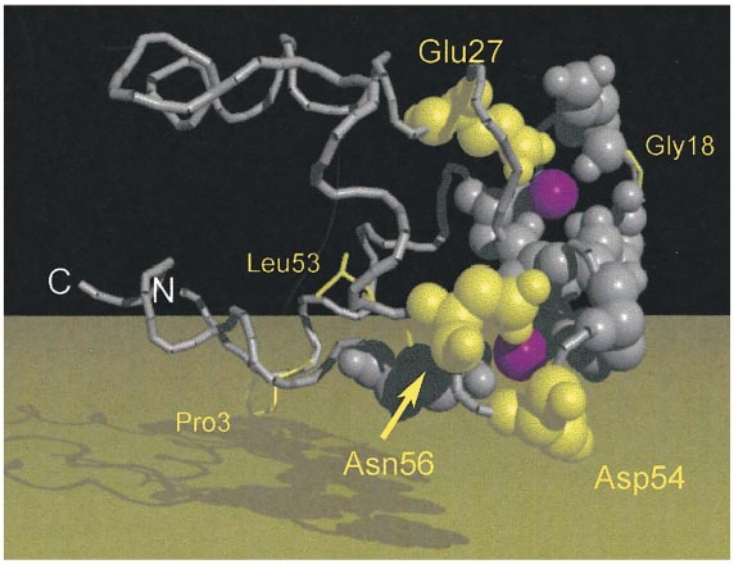


Figure 6. Ribbon of the calcium-binding protein structure (PDB entry 3icb, Swissprot entry P02633). Residues involved in the interaction with the calcium ions (magenta spheres) are displayed in CPK and amino acid residues predicted as RBD are yellow. N and C-terminal ends are indicated.

case, mutations were predicted *a priori* from the results of the RBD method on the PBP sequences. Those acyl serine transferases are involved in the assembly and metabolism of the bacterial cell wall peptidoglycan (Goffin & Ghuysen, 1998). Among this protein family, the Spn2x protein structure was recently solved (Pares *et al*., 1996). By homology modelling, we built a three-dimensional model of *Escherichia coli* PBP3. RBD were predicted and mapped in the 3D model. Several point mutations were proposed. The mutagenesis studies demonstrate that two of the ten predicted mutations modified the bacterial cell septation activity or the penicillin-binding affinity (Marrec-Fairley *et al*., 2000).

In conclusion, the RBD method is efficient as a first approach for localising putative interaction sites and proposing mutations from sequences. It is based on the analysis of sequence hydrophobicity and principally detects hydrophilic domains. The procedure could be combined with multiple sequence alignments to identify homologous binding sites or, in contrast, to elicit the absence of functional interaction domains. The RBD method is fast and easy and should be very useful for screening whole newly sequenced genomes, or is suitable

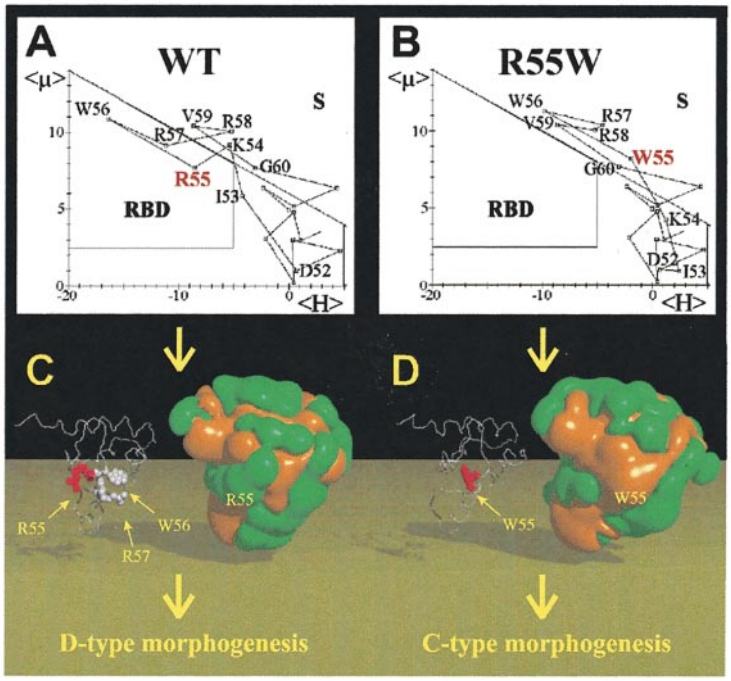


Figure 7. Analysis of the R55W mutation in the Mason-Pfizer monkey virus matrix protein. The native fragment Arg55-Arg57 should be an interaction site, since it is a RBD (a). The same fragment with R55W mutation is shifted out of the RBD (b). Isopotential surfaces of MHP for the native (c) and the mutated (d) M-PMV matrix protein structures. Orange surfaces are hydrophobic and green ones are hydrophilic.

for those who have sequences but no structural information and are looking for sequence-function relationships. Lastly, it can provide data for site-specific mutagenesis and two-hybrid system experiments.

Materials and Methods

The algorithm by Eisenberg is used to plot the mean hydrophobicity $\langle H_i \rangle$ *versus* the mean hydrophobic moment $\langle \mu_{Hi} \rangle$ (Eisenberg *et al.*, 1982) as follows:

$$\langle H_i \rangle = \frac{1}{N} \sum_{n=1}^{N} h_n \quad (1)$$

$$\langle \mu_{H_i} \rangle = \frac{1}{N} \left[\left(\sum_{n=1}^{N} h_n \sin(\delta n) \right)^2 + \left(\sum_{n=1}^{N} h_n \cos(\delta n) \right)^2 \right]^{1/2} \quad (2)$$

h_n is the hydrophobicity of the amino acid n according to the Eisenberg's consensus hydrophobicity scale (Eisenberg *et al.*, 1984). N is the number of residues in the window. The window is moved along the sequence and at each step $\langle H_i \rangle$,$\langle \mu_{Hi} \rangle$ values are assigned to the amino acid i in the centre of the window. δ is the gyration angle between two consecutive residues in the sequence: δ of 100° correspond to a α-helix, δ of 170° to a β-strand and δ of 85° to a β-turn.

The MHP visualises the hydrophobic/hydrophilic envelop of a molecule (Brasseur, 1991). MHP is plotted assuming that the hydrophobicity potential of an atom decreases exponentially with the distance so that:

$$\mathrm{MHP} = \sum_{i=1}^{P} E_{\mathrm{tr}_i} \mathrm{e}^{(r_i - d_i)}$$

r_i is the radius of atom i and d_i is the distance between the atom i and a point M, where the potential is calculated. P is the number of atoms in the molecule. Transfer energy E_{tr} for an atom i was calculated from the molecular transfer energies compiled by Tanford (1973). All E_{tr_i} are listed elsewhere (Brasseur, 1991). All M points corresponding to an isopotential value are joined to draw the isopotential hydrophobic and hydrophilic surfaces (MHP).

The group headed by R.M. Kini, Bioscience Centre at the National University of Singapore kindly provided the database of protein interaction sites. The Database of Interacting Proteins (DIP) is available at dip.doe-mbi.ucla.edu/. The Swissprot database (release 38; Bairoch & Apweiler, 2000) containing 80,000 sequences was used for the analyses. Three-dimensional structures of proteins were extracted from the Protein Data Bank (PDB) web site: www.rcsb.org/pdb/. Molecular visualisations were performed using WinMGM software (Rahman *et al.*, 1994) from Ab Initio Technology (Obernai, France).

Acknowledgements

The authors wish to thank J.M. Ghuysen and M. Nguyen-Distèche for their contribution and discussion during the analysis of PBP3. We are also grateful to A. Burny, F. Bex and S. Arnould for their constructive discussion about the M-PMV Gag protein and to M.R. Conte for kindly providing the atomic coordinates of the structure. We acknowledge R.M. Kini for the access to its database of known interaction sites. X.G. is supported by the Interuniversity Poles of Attraction Programme-Belgian State, Prime minister's Office- Federal Office for Scientific, Technical and Cultural Affairs contract no. P.4/03. R.B. is Research Director at the National Funds for Scientific Research of Belgium (FNRS). Requests for sequence analysis can be by E-mail to the corresponding author.

References

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.

Bamborough, P. & Cohen, F. E. (1996). Modeling protein-ligand complexes. *Curr. Opin. Struct. Biol.* **6**, 236-241.

Betts, M. J. & Sternberg, M. J. (1999). An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng.* **12**, 271-283.

Brasseur, R. (1991). Differentiation of lipid-associating helices by use of three-dimensional molecular hydrophobicity potential calculations. *J. Biol. Chem.* **266**, 16120-16127.

Cardin, A. D., Hirose, N., Blankenship, D. T., Jackson, R. L., Harmony, J. A., Sparrow, D. A. & Sparrow, J. T. (1986). Binding of a high reactive heparin to human apolipoprotein E: identification of two heparin-binding domains. *Biochem. Biophys. Res. Commun.* **134**, 783-789.

Choi, G., Park, S., Choi, B., Hong, S., Lee, J., Hunter, E. & Rhee, S. S. (1999). Identification of a cytoplasmic targeting/retention signal in a retroviral Gag polyprotein. *J. Virol.* **73**, 5431-5437.

Conte, M. R., Klikova, M., Hunter, E., Ruml, T. & Matthews, S. (1997). The three-dimensional solution structure of the matrix protein from the type D retrovirus, the Mason-Pfizer monkey virus, and implications for the morphology of retroviral assembly. *EMBO J.* **16**, 5819-5826.

De Loof, H., Rosseneu, M., Brasseur, R. & Ruysschaert, J. M. (1986). Use of hydrophobicity profiles to predict receptor binding domains on apolipoprotein E and the low density lipoprotein apolipoprotein B-E receptor. *Proc. Natl Acad. Sci. USA,* **83**, 2295-2299.

Eisenberg, D. ,Weiss R. M. & Terwilliger, T. C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature,* **299**, 371-374.

Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125-142.

Goffin, C. & Ghuysen, J. M. (1998). Multimodular penicillin-binding proteins: an enigmatic family of orthologs and paralogs. *Microbiol. Mol. Biol. Rev.* **62**, 1079-1093.

Janin, J. (1995). Protein-protein recognition. *Prog. Biophys. Mol. Biol.* **64**, 145-166.

Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 6027-6030.

Jones, S. & Thornton, J. M. (1997a). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121-132.

Jones, S. & Thornton, J. M. (1997b). Prediction of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 133-143.

Kini, R. M. & Evans, H. J. (1995). A hypothetical structural role for proline residues in the flanking segments of protein-protein interaction sites. *Biochem. Biophys. Res. Commun.* **212**, 1115-1124.

Kini, R. M. & Evans, H. J. (1996). Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site. *FEBS Letters,* **385**, 81-86.

Le, L., Brasseur, R., Wemers, C., Meulemans, G. & Burny, A. (1988). Fusion (F) protein gene of Newcastle disease virus: sequence and hydrophobicity comparative analysis between virulent and avirulent strains. *Virus Genes,* **1**, 333-350.

Mahley, R. W., Innerarity, T. L., Rall, S. C., Jr & Weisgraber, K. H. (1984). Plasma lipoproteins: apolipoprotein structure and function. *J. Lipid Res.* **25**, 1277-1294.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science,* **285**, 751-753.

Marrec-Fairley, M., Piette, A., Gallet, X., Brasseur, R., Hara, H., Fraipont, C., Ghuysen, J. M. & Nguyen-Distèche, M. (2000). Differential functionalities of amphiphilic peptide segments of the cell-septation penicillin-binding protein 3 of *Escherichia coli*. *Mol. Microbiol.* **370**, 1-15.

Pares, S., Mouz, N., Pétillot, Y., Hakenbeck, R. & Dideberg, O. (1996). X-ray structure of Streptococcus pneumoniae PBP2X, a primary penicillin target enzyme. *Nature Struct. Biol.* **3**, 284-289.

Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. (1999). Structure of a DNA-bound ultrabithorax-extradenticle homeodomain complex. *Nature,* **397**, 714-719.

Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523.

Phizicky, E. M. & Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59**, 94-123.

Rahman, M. & Brasseur, R. (1994). WinMGM: a fast CPK molecular graphics program for analyzing molecular structure. *J. Mol. Graphics,* **12**, 212-218.

Rhee, S. S. & Hunter, E. (1990). A single amino acid substitution within the matrix protein of a type D retrovirus converts its morphogenesis to that of a type C retrovirus. *Cell,* **63**, 77-86.

Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by saytisfaction spatial restraints. *J. Mol. Biol.* **234**, 779-815.

Shoichet, B. K. & Kuntz, I. D. (1996). Predicting the structure of protein complexes: a step in the right direction. *Chem. Biol.* **3**, 151-156.

Sternberg, M. J., Gabb, H. A. & Jackson, R. M. (1998). Predictive docking of protein-protein and protein-DNA complexes. *Curr. Opin. Struct. Biol.* **8**, 250-256.

Swanstrom, R. & Wills, J. W. (1997). Synthesis, assembly, and processing of viral proteins. In *Retroviruses* (Coffin, J. M., Hughes, S. H. & Varmus, H. E., eds), pp. 263, Cold Spring Harbor Laboratory Press, USA.

Tanford, C. (1973). *The Hydrophobic Effect: Formation of Micelles and Biological Membranes* (Tanford, C., ed.), pp. 1-217, John Wiley & Sons, Inc., New York, USA.

Vidal, M., Brachmann, R. K., Fattaey, A., Harlow, E. & Boeke, J. D. (1996). Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. *Proc. Natl Acad. Sci. USA,* **93**, 10315-10320.

Weisgraber, K. H., Rall, S. C., Jr, Mahley, R. W., Milne, R. W., Marcel, Y. L. & Sparrow, J. T. (1986). Human apolipoprotein E. Determination of the heparin binding sites of apolipoprotein E3. *J. Biol. Chem.* **261**, 2068-2076.

Young, L., Jernigan, R. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717-729.

Edited by B. Holland

(Received 14 April 2000; received in revised form 31 July 2000; accepted 31 July 2000)