J. Mol. Biol. (1997) **272**, 133–143

JMB

Prediction of Protein-Protein Interaction Sites using Patch Analysis

Susan Jones[1]* and Janet M. Thornton[1,2]

[1]*Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College Gower Street, London WC1E 6BT, England*

[2]*Department of Crystallography, Birkbeck College, Malet Street, London WC1 7HX, England*

**Corresponding author*

A method for defining and analysing a series of residue patches on the surface of protein structures is used to predict the location of protein-protein interaction sites. Each residue patch is analysed for six parameters; solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area. The method involves the calculation of a relative combined score that gives the probability of a surface patch forming protein-protein interactions. Predictions are made for the known structures of protomers from 28 homo-dimers, large protomers from 11 hetero-complexes, small protomers from 14 hetero-complexes, and antigens from six antibody-antigen complexes. The predictions are successful for 66% (39/59) of the structures and the remainder can usually be rationalized in terms of additional interaction sites.

Keywords: molecular recognition; protein-protein interaction; prediction; surface patches

Introduction

The reliable prediction of protein-protein interaction sites is an important goal in the field of molecular recognition. It is of direct relevance to the design of drugs for blocking or modifying protein-protein interactions. Predictions can be divided into two main areas. The first is the docking of two proteins of known structure; a problem which has been addressed extensively using shape complementarity (e.g. Greer & Bush 1978; Wodak & Janin, 1978; Kuntz *et al.*, 1982; Lee & Rose, 1985; Connolly, 1986; Jaing & Kim, 1991; Helmer-Citterich & Tramontano, 1994), chemical complementarity (e.g. Salemme, 1976; Warwicker 1989) and combinations of the two (e.g. Walls & Sternberg, 1992; Shoichet & Kuntz, 1993; Vakser & Aflalo, 1994). The second area of prediction, and the one addressed here, is the identification of putative interaction sites upon the surface of an isolated protein, known to be involved in protein-protein interactions, but where the structure of the partner or complex is not known.

It has been observed that protein-protein interaction sites in proteins have specific characteristics (e.g. Chothia & Janin, 1975; Argos, 1988; Janin *et al.*, 1988; Janin & Chothia, 1990; Jones & Thornton, 1995, 1996). In the accompanying paper (Jones & Thornton, 1997) we addressed the problem of comparing the observed interface with other similar sized patches on the protein surface using a series of parameters. It was concluded that it was possible to differentiate, to some degree, a protein interaction site from other similar patches on the surface of a protein. In the work presented here the use of multiple parameters for interface differentiation has been developed into a simple algorithm for the prediction of putative recognition sites for isolated proteins. Potentially this is a difficult problem, given that nothing is known about the partner protein. Therefore in this first attempt at prediction a relatively simple approach has been explored, to ascertain if prediction on this basis is possible. In this approach residue patches are defined on the surface of isolated proteins and analysed for a series of six parameters (solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area), with the aim of identifying those patches most likely to be involved in protein-protein interactions.

Results

Prediction of interface sites in homo-dimer proteins

The prediction algorithm, as described in Methods, was used to identify putative interface sites on the surface of isolated protomers from 28 non-homologous homo-dimers (see Table 1 in accompanying paper, Jones & Thornton, 1997). The

0022–2836/97/360133–11 $25.00/0/mb971233

interfaces of the homo-dimers were predicted by defining a combined score for each surface patch based on six parameters. The combined score was derived such that a surface patch that had a high solvation potential, a high residue interface propensity and was the most hydrophobic, protruding, accessible and planar scored the highest (see equation (3) in Methods).

As discussed in the accompanying paper the definition of a surface patch is a crude one, and with the use of an approximate relationship for the selection of patch size (see Methods), and the selection of an approximately circular patch, it was unlikely that one surface patch on a protomer would exactly match the residues in the observed interface. The use of the approximate relationship between the size of the protomer and the size of the interface resulted in patch sizes that ranged from 47% smaller to 81% larger than the observed interface size. The size of the patch used in the prediction will obviously influence how the observed interface is sampled. A patch that is defined to be significantly larger than the observed interface will be capable of sampling more of the observed interface than a patch defined to be significantly smaller. This was taken into account when the overlap between the predicted best patches and the observed interface was evaluated.

To evaluate the effectiveness of the combined scoring system using all six parameters two measures were calculated for each of the three patches with the highest combined scores in each protomer (Table 1). Firstly a percentage overlap value ($P1$) was calculated as defined in equation (1) in the accompanying paper (Jones & Thornton, 1997).

Then a relative overlap ($P2$) value was calculated as:

$$\text{relative overlap } (P2) = \frac{P1}{\text{maximum } P1} \qquad (1)$$

where $P1$ is defined in equation (1) in the accompanying paper and maximum $P1$ is patch with highest percentage overlap with observed interface.

Table 1. Results of the prediction algorithm for protomers from 28 homo-dimers

PDB code	No. patches[a]	Patch size[b]	% Overlap $P1$ of top three patches[c] 1st	2nd	3rd	Max $P1$[d] (%)	Relative % overlap of top three patches[e] 1st	2nd	3rd	% Random score[f]	No. diff patches[g] in top three	Rank order max $P1$[h]
1msb	88	28	100	84	74	100	100	84	74	13.6	1	1
1sdh	121	32	74	26	70	74	100	35	94	11.6	2	1
1ypi	174	43	85	52	48	85	100	62	56	10.3	1	1
2cts	304	59	47	46	46	47	100	98	96	22.4	1	1
2ts1	234	46	93	90	55	92	100	97	60	4.7	1	1
3grs	356	63	57	58	48	58	98	100	82	5.9	1	2
5adh	257	54	61	78	61	78	78	100	78	7.8	1	2
1pyp	219	47	46	46	85	85	55	81	100	6.8	2	3
1utg	67	21	41	38	44	44	94	88	100	45.5	1	3
2wrp	100	27	41	37	47	47	88	79	100	43.0	1	3
1cdt	57	19	60	60	67	93	64	64	71	17.5	2	5
4mdh	235	51	59	51	51	62	96	83	83	10.5	1	5
5hvp	81	25	54	54	54	56	96	96	96	28.4	1	5
2rve	177	43	70	68	70	87	81	77	80	9.6	1	7
1phh	276	56	35	35	59	79	44	44	75	6.2	1	8
2gn5	78	24	62	71	58	83	75	85	70	19.2	2	9
1pp2[i]	105	29	32	27	16	60	56	45	27	13.3	2	11
2sod	110	32	60	50	20	85	70	59	24	19.1	2	13
3gap[i]	178	39	13	13	10	58	23	23	18	11.2	1	19
2ssi[i]	93	27	19	15	4	73	26	21	5	15.1	1	45
2tsc[i]	193	58	0	4	0	58	0	6	0	15.0	1	28
2ccy[i]	109	29	50	40	30	90	56	44	33	7.3	1	34
3icd	298	58	56	47	52	59	95	80	89	18.8	1	36
1il8[i]	65	21	13	9	39	61	21	7	64	36.9	2	47
3sdp	151	36	18	40	74	81	22	49	91	10.6	2	70
2rus[i]	306	63	17	34	13	58	29	60	23	10.5	2	124
3enl[i]	273	59	0	0	16	82	0	0	19	5.5	1	144
3aat[i]	286	56	8	8	7	45	18	18	16	20.3	1	245

[a] Total number of patches on the surface of each protomer.
[b] The number of residues in a patch.
[c] Overlap value $P1$ for each of the three patches with the highest combined score.
[d] Overlap value $P1$ for the patch with the maximum overlap value with the observed interface.
[e] Relative overlap value for each of the three patches with the highest combined score.
[f] Random score which gave (as a percentage) the number of patches which had =70% overlap with the known interface.
[g] Number of different patches that the top three patches represent (patches were defined as different if they had an overlap of <50%).
[h] Rank order of the correct patch, i.e. that which overlaps most with the observed interface. A rank order of 1 denotes that the patch with the maximum overlap with the observed interface had the highest combined score of all surface patches. The homo-dimers are listed in increasing order of this ranking.
[i] Those protomers that were not predicted correctly, based on a relative percentage overlap cut-off of 70%.

By definition the surface patches are overlapping. To evaluate if the three patches with the highest combined scores overlapped, an overlap value between each pair of the top three patches in each protomer was calculated. If the overlap between any two patches in the set of three was $\geqslant 50\%$ then the two patches were counted as one patch (Table 1). In addition the rank order of the patch with the maximum overlap with the observed interface (Maximum *P*1), was calculated relative to the total number of patches on the surface of each protomer (Table 1). A rank order of 1 denoted that the patch with the maximum overlap with the observed interface had the highest combined score of all surface patches, i.e. the best possible prediction.

If the relative overlap *P*2 (equation (1)) was $\geqslant 70\%$ for any of the top three patches of a protomer the prediction was defined as correct. On this criterion 68% (19/28) of the homo-dimer interfaces were predicted correctly. Of these 19 correctly predicted interfaces, 16 had *P*1 values (the absolute overlap between predicted patches and the known interface) of $\geqslant 50\%$ in at least one of the three top scoring patches. Hence although the definition of a correct prediction is based on the relative overlap (*P*2), the absolute overlap (*P*1) in the correctly predicted cases is also high.

A random prediction score was also calculated (Table 1), which gave (as a percentage) the number of patches that had $\geqslant 70\%$ overlap with the known interface. This gave a value that could be used to evaluate the significance of the prediction. For example, a random prediction score of 80% would indicate that it would be possible to select a patch which overlapped the known interface (by $\geqslant 70\%$) 80% of the time just by chance. On average 15% of patches overlapped the known interface by $\geqslant 70\%$, but the random scores ranged from 4.7% to 45.5% for different proteins. The random prediction rate was not correlated to the rank order of the predictions: for example in 2tsl the random score was 4.7%, and the known interface is correctly predicted, but in 1il8, the random score was 36.9%, and yet the interface was not correctly predicted. Not surprisingly, the two structures (1utg and 2wrp) that had the largest random scores (45.5% and 43.0%, respectively) were predicted correctly. These very high random scores were caused by the patch size being 40 to 45% larger than the known interface.

The evaluation of the overlapping nature of the top three patches, with the highest combined score, revealed that in 19 of the 28 protomers the top three patches overlapped by $\geqslant 50\%$ and represented only a single discrete patch. Hence, in the majority of cases, the top 3 patches relate to the same area on the surface of the protein; and they do not represent alternative putative interface sites.

In any predictive algorithm of this nature the selection of criteria for the definition of correct and incorrect predictions is somewhat arbitrary. However with this method a correlation between the *P*1 overlap and the combined score was observed in the protomers with correctly predicted interfaces (e.g. Figure 1), indicating that the combined score does carry useful information for the selection of putative interface sites.

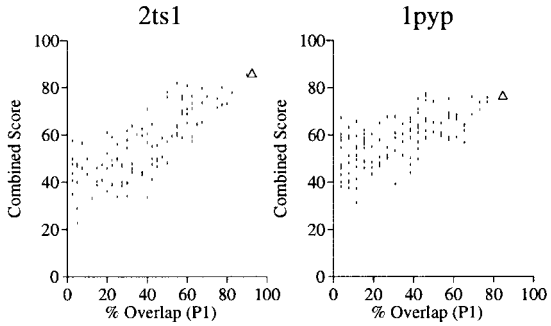

Figure 1. The relationships between the percentage overlap with the observed interface (*P*1) and the combined score for protomers from two homo-dimers with interfaces that were correctly predicted. (a) Tyrosyl-transfer/RNA synthetase (2ts1) (b) inorganic pyrophosphatase (1pyp). In each graph the black dots represent the calculated patches, and the open triangle the patch with the maximum *P*1 value (i.e. that which overlaps most with the observed interface).

Of the 28 protomers, there are nine structures (marked with **i** in Table 1) with interfaces that were never predicted (i.e. the relative overlaps for the top three patches for each protomer were $\geqslant 70\%$). The size of the patches used in each case could have been a significant factor in the failure of the prediction, hence the predictions were repeated using the size of the observed interface as the patch size. This resulted in the interfaces of four structures (1il8, 1pp2, 2ccy and 2rus) being correctly predicted. In the original predictions (when patch size was estimated using an approximate realtionship with the size of the protomer) patches for all four structrures were under-estimated. For example the estimated patch size for 1pp2 was 29 residues but the known interface contained 37 residues, similarly, the estimated patch of 2ccy was 20 residues but the known interface contained 29 residues. However patch size did not account for the failure of predictions for five other structures (2ssi, 2tsc, 3aat, 3enl, 3gap), as the predictions for these structures were still unsuccessful even when the size of the known interface was used as the size of the patch.

To understand why predictions for these structures failed, the surface patch predicted with the highest combined score was analysed, and its location compared with the observed interface patch and other known interaction sites on the surface. In the case of 2ssi (subtilisin inhibitor; Mitsui *et al* , 1979) the patch predicted as an interface site maps very closely to the interaction site observed between this structure and the enzyme subtilisin. Hence in this example the enzyme-inhibitor interface has been recognised in preference to the dimer

interface. The highest scoring patch in 2tsc (thymidylate synthase) did not include any residues from the observed dimer interface but mapped to a different location which approximately correlates with the site at which the enzyme binds a substrate and a co-factor analogue (Montfort *et al*. 1990). The patch predicted as the dimer interface in 3gap (gene activator protein) maps to a very small part of the dimer interface and to part of the location where the structure binds to DNA (Weber & Steitz, 1987). Some of the residues predicted as part of the dimer interface actually occur in the DNA recognition helix. Thus of five incorrect predictions, three could be rationalised, and revealed interesting interaction information.

The reason for the highest scoring patches in the remaining two structures is less clear. In 3aat (aspartate aminotransferase) the known interface is located on the large domain of this two-domain protein (Smith *et al*. 1989). The patch selected with the highest combined score mainly occupies a site on the small domain, at the junction of the large and small domain. The small number of residues which do overlap with the observed interface are those forming the enzyme's active site. Enolase (3enl) is a two domain structure with the C-terminal domain forming an αβ-barrel and the N-terminal domain a three-stranded meander (Stec & Lebioda, 1990). The observed interface involves both domains. The surface patch with the highest combined score was located only on the C-terminal domain, including two helices and connecting strands in the αβ-barrel which overlapped with neither the observed interface nor the active site.

Prediction of interface sites in hetero-complexes

The predictive algorithm was used to predict the interaction sites on the large protomer of 11 hetero-complexes and the small protomer from 14 hetero-complexes (see Table 1 in accompanying paper). Each data set contained a non-homologous set of proteins. The fact that these complexes represent the interaction between a large and a small component, meant that no valid relationship between the number of residues in a protomer and the number of residues in an interface could be made. Hence for the prediction the patch size was set to 26 residues for each large protomer and 16 for each small protomer. This was the mean number of residues involved in the observed interfaces in each data set. The results of the predictions for the 11 large protomers and the 14 small protomers are shown in Tables 2 and 3 respectively.

Prediction of large protomers

The interfaces of the large protomers were predicted by defining a combined score based on four parameters, where a surface patch that had high residue interface propensity and was non-protruding, accessible and planar, scored the highest (see equation (4) in Methods). The same criterion for a correct prediction was used for the large protomers as for the homo-dimers (if the relative overlap was $\geqslant$70% for any of the top three patches of an enzyme, the prediction was defined as correct). On this criterion seven of the 11 (68%) protomer interfaces were predicted correctly. Of these seven correctly predicted interfaces four had *P*1 values of $\geqslant$50% in at least one of the three top scoring patches. The rankings of the patch with the largest overlap with the observed interface, i.e. the ''correct'' patch, revealed that four protomers had the correct patch ranked in the top ten surface patches.

For all but one protomer the top three scoring patches relate to more than one site on the surface of each protein. For some of these examples the alternative sites can be explained by the presence of more than one interaction site on the surface of the protein. As mentioned in the accompanying paper both glycerol kinase and actin exist as oligomers and hence have more than one interaction site on the surface of the protomer. It is interesting that some of the predicted patches correlate with

Table 2. Results of the prediction algorithm for large protomers from 11 hetero-complexes.

PDB code	No. patches	% Overlap *P*1 of top three patches 1st	2nd	3rd	Max. *P*1 %	Relative % overlap of top three patches 1st	2nd	3rd	% Random score	N_o Different patches in top three	Rank order max *P*1
1cse E	172	0	0	52	67	0	0	78	4.7	2	4
1acb E	169	44	67	0	70	63	95	0	6.5	2	5
2pcb A	219	62	38	31	85	73	45	36	7.3	2	5
4cpa E[a]	204	44	0	48	73	59	0	65	6.4	2	9
2btf A	270	46	45	0	57	81	81	0	5.2	2	32
1fss A	334	39	10	26	55	71	18	47	4.5	2	34
1stf E	148	53	0	31	62	85	0	50	8.8	2	44
1smp E	322	41	42	42	58	71	71	71	6.2	1	57
1gla G[a]	311	12	0	0	94	13	0	0	4.5	2	63
1bgs A[a]	89	9	17	9	78	11	22	11	15.0	3	80
1udi E[a]	162	0	3	38	66	0	5	58	7.4	3	92

For a description of each column see footnotes caption to Table 1. The protomers are listed in order of the ranking in the right-hand column.

[a] See footnote [i] to Table 1.

Table 3. Results of the prediction algorithm for small protomers from 14 hetero-complexes

PDB code	No. patches	% Overlap *P*1 of top three patches 1st	2nd	3rd	Max. *P*1 (%)	Relative % overlap of top three patches 1st	2nd	3rd	% Random score	N^0 Different patches in top three	Rank order max *P*1
1udi I	70	13	50	50	50	27	100	100	18.0	2	2
2ptc I	51	69	15	85	85	82	18	100	27.0	1	3
1mct I	22	86	86	100	100	86	86	100	50.0	1	3
1bgs E	69	53	58	84	84	62	69	100	7.2	1	3
2sic I	91	0	0	64	93	0	0	70	9.9	2	4
1fss B[a]	55	10	5	10	85	12	6	12	14.0	1	5
1acb I	55	65	71	71	82	79	86	86	21.0	1	6
1tab I	30	7	57	71	92	8	62	77	30.0	2	8
1gla F	119	0	0	61	72	0	0	85	6.7	3	10
1stf I	77	60	70	65	75	80	93	87	20.0	1	12
2pcb B[a]	93	20	33	40	73	27	46	54	8.6	2	14
1cho I [a]	47	0	0	7	93	0	0	8	21.3	2	20
1smp I[a]	81	20	10	20	70	29	14	29	17.0	3	36
2btf P[a]	106	0	0	0	54	0	0	0	12.0	1	67

For a description of each column see footnotes to Table 1. The protomers are listed in order of the ranking in the right-hand column.
[a] See footnote [i] to Table 1.

contact sites between symmetry related molecules in the crystal. For example the α-chymotrypsin structure (1acb, chain E) has two alternative patches identified in the top three, one overlaps with the interface with eglin c and the other includes a loop (residues 122 to 125) which is involved in contacts with symmetry related molecules (Frigerio *et al*., 1992). The presence of crystal contacts in barnase could also influence the relative scoring of the known interface patch, which was not predicted correctly.

Prediction of small protomers

The interfaces of the small protomers were predicted by defining a combined score based on six parameters, where a surface patch that had a high solvation potential and residue interface propensity, and that was hydrophobic, protruding, accessible and planar, scored the highest (see equation (3) in Methods). Again the same criterion for a correct prediction was used for the small protomers as for the homo-dimers and on this basis nine of the 14 (64%) small protomer interfaces were predicted correctly. Of these nine corrrectly predicted interfaces eight had *P*1 values of $\geqslant 50\%$ in at least one of the three top scoring patches. The rankings of the patch with the largest overlap with the observed interface; i.e. the correct patch, revealed that in nine small protomers this patch ranked in the top ten surface patches.

As for the large protomers, the failure of some predictions and the presence of alternative sites in the top three scoring patches can be explained in some cases by the presence of alternative interaction sites. As previously described (see accompanying paper) profilin (2btf P) has two highly conserved hydrophobic patches on its surface that are not involved in the interaction with actin but are thought to have a regulatory role (Schutt *et al*., 1993). The presence of these patches must influence the relative scoring of the known interface patch, the position of which was not predicted. For another protein (ovomucoid third domain inhibitor (1cho, I)) the presence of domain interfaces must influence the relative scoring of the interface patch involved in the interaction with α-chymotrypisin. In the *Streptomyces* subtilisin inhibitor, the known interface with subtilisin BPN is predicted correctly but an alternative patch is also identified. The *Streptomyces* subtilisin inhibitor is a dimer (Takeuchi *et al*., 1991) and the alternative patch which is selected as one of the top scoring patches actually forms part of the dimer interface. For glucose specific factor III (1glaF) the interface patch involved in the interaction with glycerol kinase is identified, but an alternative patch is also selected as one of the top three scoring patches. This alternative site is comprised of two overlapping patches centred on residues 1 and 9. It is known that glucose specific factor III forms contacts with two different glycerol kinase tetramers in the crystal and the residues involved in these additional contacts have been identified (Hurley *et al*., 1993). It was very interesting to find that residues contained within the alternative site (identified in the patch prediction), overlapped by 80% those residues involved in the crystal contacts.

Prediction of interface sites in antigens

The predictive algorithm was used to predict the interaction sites on antigens involved in six antibody-antigen complexes (see Table 1 in accompanying paper) and the results are shown in Table 4. For the prediction the patch size was set to 20, the mean number of residues involved in the observed interfaces in the data set. The combined score was based on five parameters, where a surface patch that had a low solvation potential and was hydrophilic, protruding, accessible and planar scored the highest (see equation (5) in Methods). The same criterion for a correct prediction was used for the antigens as for the homo-dimers and on this basis

Table 4. Results of the prediction algorithm for antigens from 6 antibody-antigen complexes

PDB code and chain	No. patches	% Overlap *P*1 of top three patches 1st	2nd	3rd	Max. *P*1 (%)	Relative % overlap of top three patches 1st	2nd	3rd	% Random score	Nº Different patches in top three	Rank order max *P*1
1jhl Y	96	59	76	29	76	78	100	38	13.1	1	2
1fdl Y	99	59	41	41	82	72	50	50	10.1	1	8
2hfl Y[a]	97	0	0	0	71	0	0	0	11.2	1	8
1jel Y	68	31	75	12	88	35	85	14	11.4	3	12
3hfmY	96	50	64	4	73	69	88	6	8.3	2	16
1nca Y[a]	240	10	17	3	59	17	29	5	5.4	2	94

For a description of each column see footnotes to Table 1. The protomers are listed in order of the ranking in the right-hand column.
[a] See footnote [i] to Table 1.

four of the six antigen interface sites were predicted, and the correct patch ranked in the top ten surface patches for three antigens. Of the four correctly predicted interfaces all had *P*1 values of $\geqslant 50\%$ in at least one of the three top scoring patches.

Four of the antigens (1jhl, 1fdl, 2hfl, and 3hfm) are lysozymes, which have pairwise sequence identities between 92% and 100%. The antibody binding site is different for each antigen, but for three of the antigens (1jhl, 1fdl and 3hfm), some regions of the binding sites overlap (Figure 2). In Figure 2 it can be clearly seen that for each of the antigenic lysozymes the predicted interface patch is essentially the same in each case (the top scoring patch is one centred on glycine 22 for 1fdl, 1jhl and 3hfm and one centred on tyrosine 20 for 2hfl). This predicted patch contains many of the residues that are common to two or more antibody binding sites on the antigenic lysozymes. Hence the predictions were correct for three of the four antigens, those with antibody binding sites which overlapped (1fdl, 3hfm, 1jhl; Table 4). The fourth lysozyme structure, 2hfl, has an antibody binding site that does not share residues with the other three antibody binding sites. The top patch selected in the prediction was essentially the same as for the other three lysozyme structures, and thus did not overlap with the known antibody binding site on the 2hfl lysozyme structure. For this structure the alternative antibody binding site has been selected on the surface of lysozyme, in preference to the site specific to the HYHEL-5 Fab.

For the influenza virus neuraminidase antigen the prediction was unsuccessful, with two alternative patches identified, neither of which overlapped significantly with the known antibody binding site (Table 4). It is known that neuraminidase is a tetramer (Tulip *et al.*, 1992) and it was thought possible that the presence of interface sites involved in interactions with subunits within the tetrameric structure could influence the relative scoring of the patch which constituted the antibody binding site. To find if this was the case, the prediction was repeated on the tetrameric structure of the influenza virus neuraminidase antigen. However, although the top scoring patch was different to that when the predictions were conducted on the isolated protomer, it was still not the correct one. In fact the best patch (the patch with the

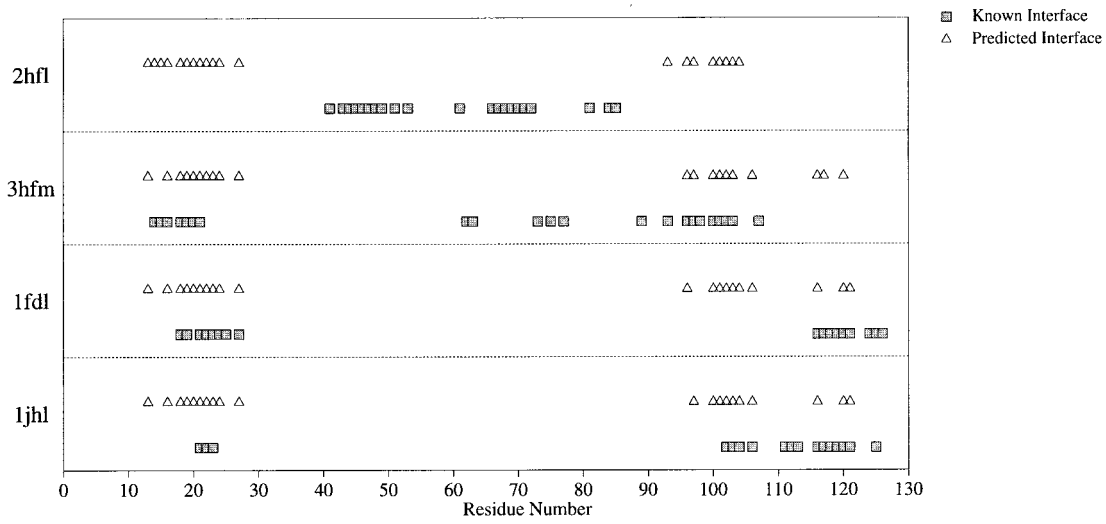

Figure 2. The location of residues in predicted and known interfaces on lysozyme from four antibody-antigen complexes. Each marker on the graph indicates the location of a residue, where the residue number is shown on the *x*-axis.

largest overlap with the known interface site) dropped from ranking at 94th place to 95th place. Hence although the presence of alternative interaction sites on the protomer did change the relative scoring of the surface patches, it did not account for the failure to predict the antibody binding site.

In the two examples described above, where the predictions failed, the definition of approximately circular surface patches resulted in poor sampling of the known interface. The surface patches with maximum overlap with the known interface contained only 75% of the known interface residues. This clearly indicates the crude nature of the surface patch definition, and the poor sampling of the known interface could explain the failure of the predictions. In addition the antigen binding sites were predicted using only five parameters (see equation (5) in Methods). The interface residue propensities were not used as this parameter was not found to contribute discriminating information for the antigen data set (see Jones & Thornton, 1997). This gives an indication that the antigen binding sites are different from those observed in homo-dimers and other hetero-complexes, and that more information on these interfaces is required before they can be predicted successfully.

Discussion

Patch predictions were made for 59 complexes and 66% of the predictions were defined as correct. It was found that in some cases the predictions were unsuccessful because the size of patch used was either too large or too small. In addition some unsuccessful predictions could be attributed to the presence of alternative interaction sites on the surface of the proteins. This also explained why, in some cases where the known interface was predicted, alternative patches were also identified as potential interaction sites. Some of these alternative sites could be attributed to interactions with subunits in oligomeric structures (in the case of the hetero-complexes), interactions with other molecules such as protein inhibitors or DNA (in the case of the homo-dimers) or to contacts between different molecules in a crystal. Hence some, but not all, unsuccessful predictions and alternative interface patches could be rationalised.

The predictive algorithm is based on the definition and comparison of surface patches at the C^{α} atom level. The nature of this definition has advantages and disadvantages. Like all docking and prediction algorithms there is a balance to be reached between the accuracy of the method and the time taken. The algorithm described is simplistic and was only intended as a first attempt to explore the chances of success in this complex problem. The definition of the surface patches at the C^{α} atom level meant a patch was never defined that contained all the observed interface atoms and no others. To achieve this the patches would need to be defined at the atom level and the definition allow for discontinuous patches (i.e. allowing for gaps and irregularly shaped patches). However this would result in a combinatorial explosion, with thousands of patches being defined. Restricting the patch definition to contiguous surface patches defined at the residue level, reduces the combinatorial problem, but at the expense of the accuracy of the predictions. However, the prediction algorithm is relatively fast; the prediction of the interaction site on HIV protease (PDB code 5hvp) takes 36 seconds on an SGI-Challenge with R4400 CPU running at 150 MHz.

By definition the surface patches were overlapping and this caused problems in the evaluation of the predictions. In many homo-dimer structures the three patches with the highest combined score overlapped each other by $\geqslant 50\%$, and represented a single region on the surface of each protomer. One method of overcoming the problem of overlapping patches, would be to assign the combined score for a patch centred on a single residue to that central residue. Hence each residue would be assigned a combined score that describes its local environment upon the surface of the protomer. Putative interface sites could then be selected to comprise those residues with the highest combined scores. In the current method the six parameters are weighted equally and are relative rather than absolute. It is likely that the predictions could be improved if the parameters were weighted according to their relative importance in the interactions. Future work will involve the calculation of the optimum weights for the parameters using neural networks. A further refinement to the method would be to include additional physical and chemical parameters for interface prediction. The recent work of Lichtarge *et al*. (1996) on the analysis of multiple sequences provides additional important information that could be used to improve the patch analysis predictions.

The method described here is similar to that used by Young *et al*. (1994) in their analysis of the hydrophobicity of surface residue clusters in proteins. However, whilst the surface patches described here are more simple that those calculated by Young *et al*. (1994), the current method moves a stage further by analysing surface patches for multiple parameters. The approach described here is useful for identifying candidate interface residues, which can be mutated experimentally, and tested for their effect on complex formation. As was observed, the nature of the interface can vary and a ''perfect'' prediction would be an unrealistic expectation, unless the structure of the partner is known and full docking can be pursued. The method described provides a rapid means to identify possible interaction sites as a guide to future experiments. For example comparative patch analysis between non-homologous proteins could provide information useful for designing species specific antibodies against proteins. Further refinements to extend the ''patch analysis'' to ligands other than

proteins (e.g. nucleic acids and carbohydrates) are in progress.

Methods

Definition of a surface patch

A patch was defined as described in the accompanying paper, with a central surface accessible residue and n nearest neighbours, where n was defined as a variable. The choice of the size of the patch (n) was crucial to the prediction. It has been observed that the size of an interface region is approximately correlated to the size of the protomer (Jones & Thornton, 1995). For the homo-dimer predictions, this correlation was calculated in terms of the number of residues in the protomer (NR_p) and the number of residues in the observed interface region (Nr_i; Figure 3). A regression line, of the form $y = ax^b$, fitted to the data of 28 protomers gave the equation:

$$NR_i = 1.9NR_p^{0.6} \quad (2)$$

with a correlation coefficient (r) of 0.7.

The prediction algorithm

For each isolated protein all surface patches were generated and the six parameters calculated for each patch. The predictive procedure involved three stages: scoring of patches for individual parameters, calculation of a combined score from multiple parameters and the selection of best patches.

Individual parameter score

The definition of the six parameters (solvation potential, interface residue propensity, hydrophobicity, planarity, protrusion and accessible surface area) have been described in the accompanying paper. For an individual parameter there was a range of values over all surface patches, and hence the patches could be scored by their relative ranking to all other surface patches in a single protomer. For each parameter the range was calculated for a given protein, and then divided into 100 separate ranges. Thus each patch parameter value was normalised and assigned a score of 1 to 100. The lowest parameter value was assigned a score of 1, and the highest parameter value a score of 100. Thus each patch had six individual parameter scores assigned; for example a patch could have a score of 60 for solvation potential, 20 for residue interface propensity, 80 for hydrophobicity, 1 for planarity, 100 for protrusion and 90 for ASA. This approach weights all six parameters equally, and is relative, rather than absolute.

Combined parameter score

A combined score was calculated which incorporated the individual scores of a patch for all six parameters. The combined score gave a probability (on a scale of 1 to 100) of any one patch (P_j) forming protein-protein interactions. The combined score is on a scale of 1 to 100, where 1 denotes a very low (poor) probability of forming a putative interaction site, and 100 a very high (good) probability of forming a putative interaction site. As previously discussed interfaces in different types of protein-protein interactions have different properties, and this is reflected in the variable definitions of the combined score. Three definitions are derived here, one for each of the data sets;

(a) Homo-dimers and small protomers from hetero-complexes:

$$\text{combined score } P_j = \frac{S_{\text{sp}} + S_{\text{rp}} + S_{\text{hy}} + S_{\text{pi}} + S_{\text{asa}} + S_{\text{pl}}}{N_{\text{par}}} \quad (3)$$

Thus we are searching for patches that have a high solvation potential and residue interface propensity

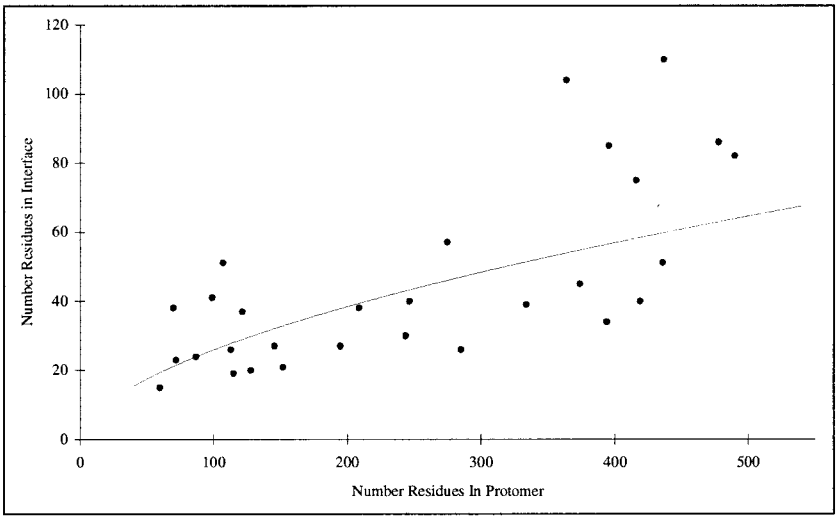

Figure 3. The relationship between the number of residues in the interface and the number of residues in the protomer for the data set of 28 homo-dimers. A regression line of $y = 1.92x^{0.56}$ has been fitted.

and that are hydrophobic, protruding, accessible and planar.

(b) Large protomers from hetero-complexes:

$$\text{combined score } P_j = \frac{S_{\text{rp}} + (100 - S_{\text{pi}}) + S_{\text{asa}} + S_{\text{pl}}}{N_{\text{par}}} \tag{4}$$

Thus we are searching for patches that have a high residue interface propensity and that are non-protruding, accessible and planar.

(c) Antigens:

$$\text{combined score } P_j = \frac{(1 - S_{\text{sp}}) + (1 - S_{\text{hy}}) + S_{\text{pi}} + S_{\text{asa}} + S_{\text{pl}}}{N_{\text{par}}} \tag{5}$$

Thus we are searching for patches that have a low solvation potential and that are hydrophilic, protruding, accessible and planar. Where, S_{sp} is score of patch P_j in the solvation potential distribution; S_{rp} is score of patch P_j in the residue interface

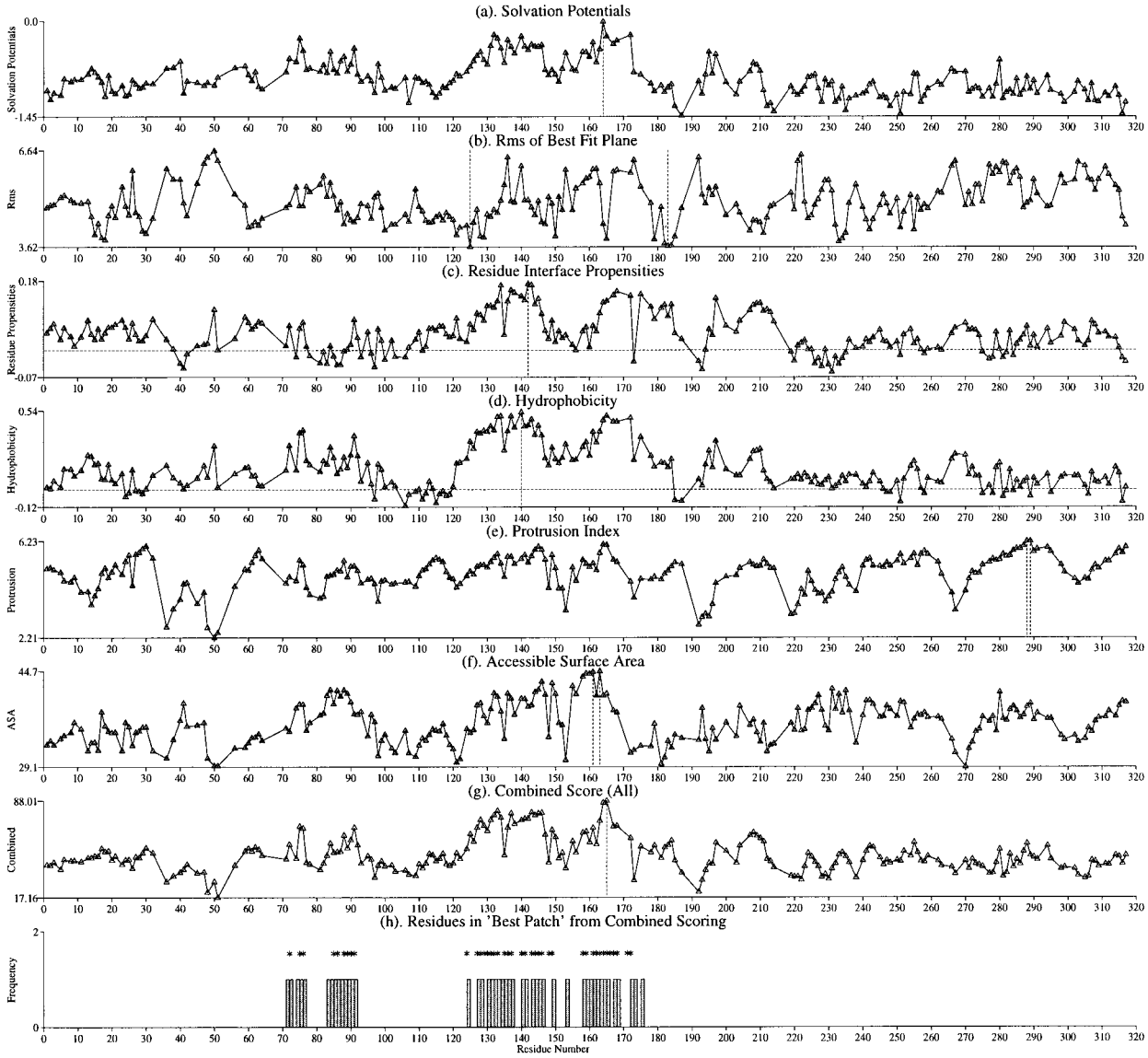

Figure 4. Each of the parameters used for a prediction can be displayed as a profile, and the combined profiles for up to six parameters can be created. The patch profiles of tyrosyl-transfer/RNA synthetase (2ts1) are shown for (a) solvation potentials, (b) rms of the best fit plane, (c) residue interface propensities, (d) hydrophobicity, (e) protrusion index. (f) accessible surface area (ASA), (g) combined score from all six parameters (a to f). The frequency of occurrence of residues in the top scoring patches are indicated as a histogram in the final profile (h). The * on the histogram indicate those residues in the known interface. On each profile the residue numbers are indicated on the *x*-axis and the value of the parameter on the *y*-axis. The dotted vertical lines on each profile indicate the residue number at the centre of the top scoring patch for each parameter.

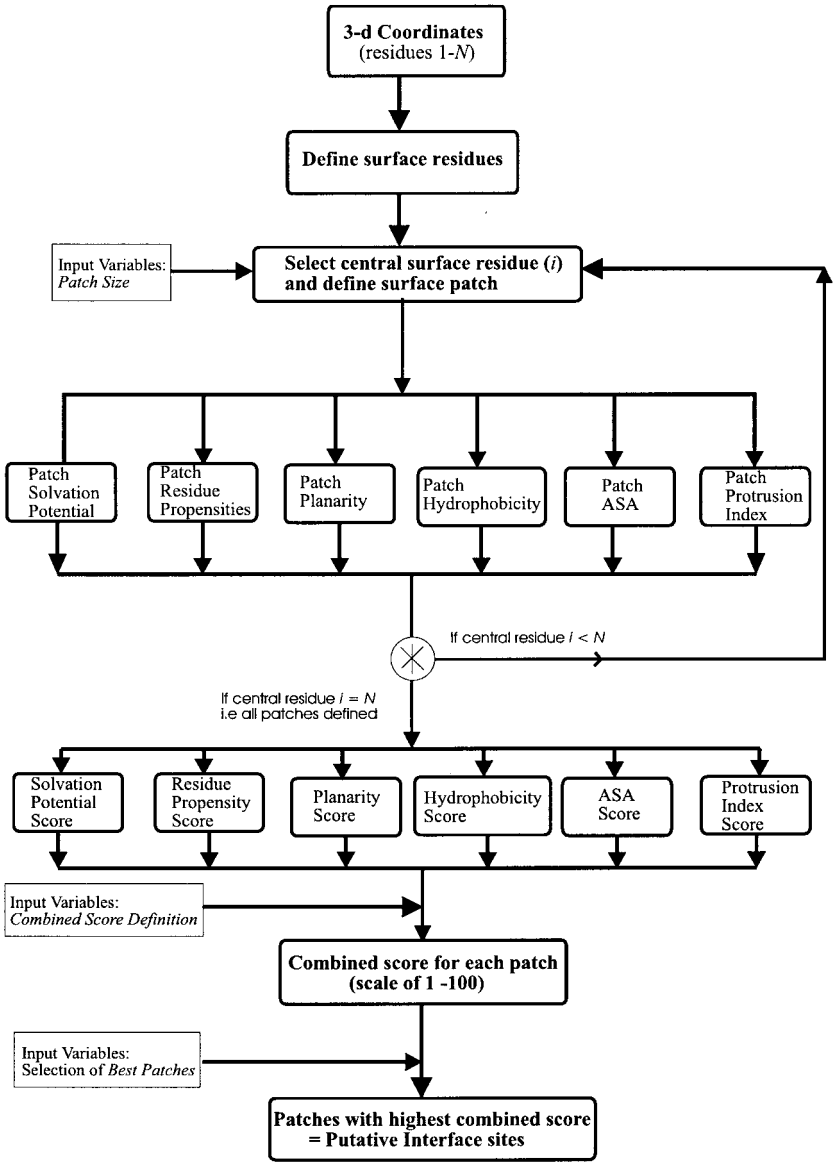

Figure 5. Flow diagram summarizing the events in the prediction algorithm, designed to predict putative interaction sites on the surfaces of isolated proteins.

propensity distribution; S_{hy} is score of patch P_j in the hydrophobicity distribution; S_{pi} is score of patch P_j in protrusion index distribution; S_{asa} is score of the patch P_j in the accessible surface area distribution; S_{pl} is score of patch P_j in the planarity distribution; N_{par} is number of parameters.

Best surface patches

The algorithm places each patch in descending order of its combined score and the first n number of patches can be selected as best patches. In the evaluation of all the interface predictions the three patches with the highest combined scores were selected as best patches. Each of the six parameters can be used to create a profile and the residues comprising the best patch indicated in a histogram (e.g. Figure 4).

The flow diagram shown in Figure 5 summarises the main elements of the predictive algorithm.

Acknowledgements

S. J. was funded by a BBSRC studentship, sponsored by Zeneca Pharmaceuticals. We thank D. Tims for useful discussions.

References

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.

Chothia, C. & Janin, J. (1975). Principles of protein–protein recognition. *Nature*, **256**, 705–708.

Connolly, M. L. (1986). Shape Complementarity at the hemoglobin A1B1 subunit interface. *Biopolymers*, **25**, 1229–1247.

Frigerio, F., Coda, A., Pugliese, L., Lionetti, C., Menegatti, E., Amiconi, G., Schnebli, H. P., Ascenzi, P. A. & Bolognesi, M. (1992). Crystal and molecular structure of the bovine α-chymotrypsin-eglin *c* complex at 2.0 Å resolution. *J. Mol. Biol.* **225**, 107–123.

Greer, J. & Bush, B. L. (1978). Macromolecular shape and surface maps by solvent exclusion. *Proc. Natl Acad. Sci. USA,* **75**, 303–307.

Helmer-Citterich, M. & Tramontano, A. (1994). Puzzle: A new method for automated protein docking based on surface shape complementarity. *J. Mol. Biol.* **235**, 1021–1031.

Hurley, J. H., Faber, H. R., Worthylake, D. & Meadow, N. D. (1993). Structure of the regulatory complex of *Escherichia coli* III GLC with glycerol kinase. *Science,* **259**, 673–677.

Jaing, F. & Kim, S. (1991). ''Soft docking'': Matching of molecular surface cubes. *J. Mol. Biol.* **219**, 79–102.

Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 16027–16030.

Janin, J., Miller, S. & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204**, 155–164.

Jones, S. & Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63**, 31–65.

Jones, S. & Thornton, J. M. (1996). The principles of protein-protein interactions. *Proc. Natl Acad. Sci., USA,* **93**, 13–20.

Jones, S. & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol* **272**, 132–143.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, E. (1982). A geometric approach to macromolecular-ligand interactions. *J. Mol. Biol.* **161**, 269–288.

Lee, R. H. & Rose, G. D. (1985). Molecular recognition. I. Automatic identification of topographic surface features. *Biopolymers,* **24**, 1613–1627.

Lichtarge, O., Bourne, H. A. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.

Mitsui, Y., Satow, Y., Watanabe, Y., Hirono, S. & Iitaka, Y. (1979). Crystal structures of *Streptomyces* subtilisin inhibitor and its complex with subtilisin BPN′. *Nature,* **277**, 447–452.

Montfort, W. R., Perry, K. M., Fauman, E. B., Finer-Moore, J. S., Maley, G. F., Hardy, L., Maley, F. & Stroud, R. M. (1990). Structure, multiple site binding, and segmental accommodation in thymidylate synthase on binding dUMP and anti-folate. *Biochemistry,* **29**, 6964–6976.

Salemme, F. R. (1976). An hypothetical structure for an intermolecular electron transfer complex of cytochromes *c* and b_5. *J. Mol. Biol.* **102**, 563–568.

Schutt, C. E., Myslik, J. C., Rozycki, M. D., Goonesekere, N. C. W. & Lindberg, U. (1993). The structure of crystalline profilin-β-actin. *Nature,* **365**, 810–816.

Shoichet, B. K. & Kuntz, I. D. (1993). Matching chemistry and shape in molecular docking. *Protein Eng.* **6**, 723–732.

Smith, D. L., Almo, S. C., Toney, M. D. & Ringe, D. (1989). 2.8 Å resolution crystal structure of an active-site mutant of aspartate aminotransferase from *Escherichia coli*. *Biochemistry,* **28**, 8161–8167.

Stec, B. & Lebioda, L. (1990). Refined structure of yeast apo-enolase at 2.25 Å resolution. *J. Mol. Biol.* **211**, 235–248.

Takeuchi, Y., Satow, Y., Nakamura, K. T. & Mitsui, Y. (1991). Refined structure of the complex of subtilisin BPN′ and *Streptomyces* subtilisin inhibitor at 1.8 Å resolution. *J. Mol. Biol.* **221**, 309–325.

Tulip, W. R., Varghese, J. N., Laver, W. G., Webster, R. G. & Colman, P. M. (1992). Refined crystal structure of the influenza virus N9 neuraminidase-NC41 Fab complex. *J. Mol. Biol.* **227**, 122–148.

Vakser, I. A. & Aflalo, C. (1994). Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins: Struct. Funct. Genet.* **20**, 320–329.

Walls, P. H. & Sternberg, M. J. E. (1992). New algorithm to model protein-protein recognition based on surface complementarity: applications to antibody-antigen docking. *J. Mol. Biol.* **228**, 277–297.

Warwicker, J. (1989). Investigating protein-protein interaction surfaces using a reduced stereochemical and electrostatic model. *J. Mol. Biol.* **206**, 381–395.

Weber, I. T. & Steitz, T. A. (1987). Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Å resolution. *J. Mol. Biol.* **198**, 311–326.

Wodak, S. J. & Janin, J. (1978). Computer analysis of protein-protein interaction. *J. Mol. Biol.* **124**, 323–342.

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 G. von Heijne

(Received 14 April 1997; received in revised form 17 June 1997; accepted 26 June 1997)