This is a preprint of an article accepted for publication in
*Proteins, Structure, Function, and Genetic*s

Evaluation of Protein Docking Predictions Using *Hex* 3.1 in CAPRI Rounds 1–2

David W. Ritchie

Department of Computing Science,
King's College, University of Aberdeen, Aberdeen AB24 3UE, U.K.

dritchie@csd.abdn.ac.uk

Tel. (+44)(–0)1224 272282, Fax (+44)(–0)1224 273422

This article describes and reviews our efforts using *Hex* 3.1 to predict the docking modes of the seven target protein-protein complexes presented in the CAPRI (Critical Assessment of Predicted Interactions) blind docking trial. For each target, the structure of at least one of the docking partners was given in its unbound form, and several of the targets involved large multimeric structures (e.g. *Lactobacillus* HPr kinase, hemagglutinin, bovine rotavirus VP6). Here, we describe several enhancements to our original spherical polar Fourier docking correlation algorithm. For example, a novel surface sphere smothering algorithm is introduced to generate multiple local coordinate systems around the surface of a large receptor molecule which may be used to define a small number of initial ligand docking orientations distributed over the receptor surface. High resolution spherical polar docking correlations are performed over the resulting receptor surface patches, and candidate docking solutions are refined using a novel soft molecular mechanics energy minimisation procedure. Overall, this approach identified two good solutions at rank 5 or less for two of the seven CAPRI complexes. Subsequent analysis of our results shows that *Hex* 3.1 is able to place good solutions within a list of 20 or less for four of the seven targets. This demonstrates that useful *in silico* protein-protein docking predictions can now be made with increasing confidence, even for very large macromolecular complexes.

Keywords: blind docking trial; protein shape; shape complementarity; Fourier correlation; fast Fourier transform; spherical harmonics; OPLS potentials.

Abbreviations: FFT: Fast Fourier Transform; HPr: Histidine-containing Phosphocarrier Protein; OPLS: Optimised Potentials for Liquid Simulations; PC: Personal Computer; PDB: Protein Data Bank; RMS: Root Mean Squared; TCR: T-Cell Receptor; 1D: One-Dimensional; 3D: Three-Dimensional.

Introduction

If we are to understand how proteins function at the molecular level, it is necessary to develop good computational models of how large biomolecules might interact. However, early efforts to predict the association, or docking, of globular protein domains quickly showed that this was by no means a trivial task [1, 2]. Although some progress has been made towards developing improved docking algorithms, and several successes have been described, a general solution to the so-called docking problem seems to remain elusive [3]. Nonetheless, with recent improvements in protein expression and X-ray crystallography techniques, the number, size and diversity of proteins whose interactions we wish to model have never been greater. Thus there is an increasing need to develop accurate and fast docking algorithms.

In order to assess and compare current protein docking algorithms, and to stimulate further developments in the field, the CAPRI (Critical Assessment of Predicted Interactions; *http://capri.ebi.ac.uk*) blind docking experiment [4] was launched. During the summer of 2001 and early in 2002 seven target structures were presented to the docking community. Several of the targets involved large enzymes or viral surface proteins, similar to the antibody/hemagglutinin complex [5] which proved difficult to model in the CASP2 docking section [6, 7]. This article describes and reviews our efforts using *Hex* 3.1 to predict the docking modes of the seven CAPRI target complexes. We describe several enhancements to our original spherical polar Fourier correlation algorithm [8], and we assess the usefulness of these enhancements in light of the revealed complex structures.

Our basic approach to the docking problem is to represent the steric shape, electrostatic potential, and charge density of each protein as expansions of spherical polar Fourier basis functions [8]. However, unlike conventional three-dimensional (3D) fast Fourier transform (FFT) docking approaches [9, 10, 11] which accelerate translational correlations, our approach uniquely favours rotational searches, although translational correlations may also be calculated. Here we describe how the rotational correlations at the heart of our algorithm may be accelerated by implementing the innermost loop of a docking search as a one-dimensional (1D) FFT. We also introduce several further enhancements to our algorithm. For example, very large complexes may now be modelled by performing multiple local dockings over a small set of surface patches on the larger of the docking partners. Candidate docking solutions may be refined using a "soft" molecular mechanics energy minimisation procedure, and the list of docking solutions may be clustered to help identify distinct orientations and reduce the number of "false-positives". Additionally, protein surface shapes are now calculated using a new "marching tetragons" algorithm [12] to contour atomic Gaussian density representations [13] of each protein. This treats re-entrant surface regions more reliably than our former dot surface sampling scheme [8], and allows improved graphical visualisation of results.

To try to honour the spirit of a blind trial, and to test our algorithm as thoroughly as possible, we elected to use only knowledge of the hypervariable loops in the antibody and T-cell receptor (TCR) problems (targets 2-6, and target 7, respectively). However, because we expected the first target, a novel bacterial HPr kinase (HPrK) [14] in complex with HPr, to be difficult to solve we used knowledge of key residues in each subunit to try to select a feasible docking orientation. Overall, our approach produced two close

solutions for two of the seven targets (target 3: antibody HC63/hemagglutinin; target 6: antibody AMB9/α-amylase). Subsequent analysis of our results shows that we are able to place good solutions within a list of 20 or less for four of the seven targets.

Methods

Gaussian Density Representation of Protein Shape

As an enhancement to our original shape-sampling algorithm [8], we now use a Gaussian density representation of protein shape [13]:

$$\rho_i(\underline{r}) = \alpha \exp\{-\beta(r/r_i)^2\}, \tag{1}$$

where $\rho_i(\underline{r})$ is the density function for atom i, r_i is its van der Waals (VDW) radius, and α and β are adjustable parameters. Following Grant & Gallardo [13], we use $\alpha = 2.70$ and $\beta = 2.3442$. For any given atom type, the density at a distance r_i from the atom is $\rho_i(r_i) = 2.7 \exp\{-2.3442\} = 0.259$. Hence a good estimate of the VDW surface of a protein may be calculated by summing the atom density contributions at each node in a 3D grid, and by contouring the grid using a density threshold of 0.259. Similarly, the solvent accessible surface (SAS) may be calculated and contoured using enlarged atom radii. We define the surface skin as the volume bounded by the SAS and VDW surfaces. This skin volume is central to our model of protein shape complementarity [8]. Contouring is performed using an adaptation of Guéziec and Hummel's tetrahedral decomposition algorithm [12], which we call "marching tetragons". Compared to the "marching cubes" algorithm [15], tetrahedral contouring has the advantages that there are significantly fewer ways for a surface to cut a tetrahedron than a cube, and the resulting surface triangles are implicitly oriented in a consistent sense. However, additional processing is required to remove thin edges [12].

Fourier Expansions and Coordinate Operations

The use of orthonormal spherical polar basis functions to represent protein shape and electrostatic properties has been described previously [8], so we give only a brief summary here. A Fourier expansion to order N of some property $A(\underline{r})$ in spherical polar coordinates $\underline{r} = (r, \theta, \phi)$ may be written as

$$A(\underline{r}) = \sum_{nlm}^{N} a_{nlm} R_{nl}(r) y_{lm}(\theta, \phi) \; ; \quad N \geq n > l \geq |m| \geq 0, \tag{2}$$

where a_{nlm} is an expansion coefficient, $R_{nl}(r)$ represents either an harmonic oscillator or a Coulomb-type radial function, and $y_{lm}(\theta, \phi)$ is a real spherical harmonic. Because our basis functions are entirely real, it is straightforward to reconstruct protein shape and electrostatic properties from the expansion coefficients. For example, Figure 1 shows the Fv fragment of the MCV antibody (CAPRI target 2, M.C. Vaney & F.A. Rey, unpublished results) reconstructed at different expansion orders, N. The low order $N = 16$

expansion encodes considerable global shape information, whilst individual atoms are clearly discernible with high order $N = 25$ and $N = 30$ expansions. However, it is worth noting the reduction in detail for those atoms furthest from the origin, even for high order expansions.

[Figure 1 about here.]

One of the advantages of a Fourier-based approach is that the correlation between a pair of functions (i.e. their overlap as a function of coordinate transformations) can be calculated easily. For example, it can be shown that rotational and translational coordinate operations on spherical polar Fourier expansions may be represented as:

$$\hat{R}(\alpha,\beta,\gamma)A(\underline{r}) = \sum_{nlm}^{N} a'_{nlm} R_{nl}(r) y_{lm}(\theta,\phi) \tag{3}$$

and

$$\hat{T}_z(R)A(\underline{r}) = \sum_{nlm}^{N} a''_{nlm} R_{nl}(r) y_{lm}(\theta,\phi), \tag{4}$$

where the rotated and translated expansion coefficients are respectively given by [16]

$$a'_{nlm} = \sum_{m'=-l}^{l} R^{(l)}_{mm'}(\alpha,\beta,\gamma) a_{nlm'} \tag{5}$$

and (manuscript in preparation)

$$a''_{nlm} = \sum_{n'l'}^{N} T^{(|m|)}_{nl,n'l'}(R) a_{n'l'm}. \tag{6}$$

Hence during a rigid body docking search, it is convenient to represent steric and electrostatic complementarity as overlap integrals between corresponding pairs of 3D functions. For example, if both molecules are initially located at the origin, the correlation S_{AB} between any pair of functions $A(\underline{r})$ and $B(\underline{r})$ for molecules A and B, respectively, is calculated as

$$S_{AB}(R,\beta_1,\gamma_1,\alpha_2,\beta_2,\gamma_2) = \int [\hat{T}_z(-R)\hat{R}(0,\beta_1,\gamma_1)A(\underline{r})]\times [\hat{R}(\alpha_2,\beta_2,\gamma_2)B(\underline{r})]\mathrm{d}V. \tag{7}$$

Conceptually, at each trial intermolecular separation, each protein is incrementally rotated using rotation angles (β,γ) generated from icosahedral tessellations of the sphere [8], and a search over the twist angle, α_2 is performed. For a given partial orientation $(R,\beta_1,\gamma_1,\beta_2,\gamma_2)$, the correlation in α_2 may be expressed as a 1D Fourier series (see Eq 27 in ref. [8]), which we have recently implemented as an FFT. Compared to our earlier implementation, this gives a speed-up of 30% or 50% when using 64 or 128 steps in α_2, respectively.

Docking Very Large Molecules

By itself, our spherical polar approach is unsuitable for docking very large molecules because our radial basis functions fall off rapidly beyond about 30Å from the chosen origin. Hence molecular shapes larger than this are represented poorly. Nonetheless, it is not necessary to rely on a single coordinate origin. We have developed an automatic method of generating multiple local coordinate systems with which to define initial ligand docking orientations about a large receptor. The four main steps of this algorithm are illustrated in Figure 3. First, the smaller ligand molecule is (optionally) oriented along the negative z axis to face the receptor, if knowledge of the ligand binding surface is available. Second, a low resolution (L=5) spherical harmonic surface [16] is calculated for the receptor. The surface is discretised by projecting it onto an icosahedral tessellation of the sphere, as shown in Figure 3(B). At each triangular facet of the surface, a normal vector is calculated and a 15Å radius sphere is centred on each outward normal, tangential to the surface. This smothers the surface with spheres. In the third step, the surface spheres are culled by iteratively identifying and striking out that sphere which has the greatest volume overlap with its neighbours. This procedure is repeated until no overlap volume exceeds $5Å^3$. This yields a fairly even distribution of the surviving spheres over the surface of the receptor. Finally, each surviving sphere (normal vector) is used to define a local intermolecular axis for docking, with the initial ligand/axis orientation being transferred onto the outward normal, and a local coordinate origin for the receptor being defined at an equal distance along the inward normal.

Initially, each tessellation triangle is associated with the chain identifier (ID) of the atom nearest to that triangle's centre. Thus surface spheres (normals) may be associated with chain IDs. If the receptor is composed of symmetry-related chains, surface spheres may be restricted to a selected group of chains. This helps avoid the expense of over-sampling symmetry related orientations during the docking search. Figure 3(C) shows the result of applying the surface spheres algorithm to only the C chain of the VP6 trimer, and Figure 3(D) shows the trimer covered with 23 generated MCV Fv starting orientations.

Soft Molecular Mechanics Refinement

Although our correlation approach implicitly provides a "soft" docking scheme, we wished to incorporate an additional, more sensitive, scoring function to try to reduce the number of false positives. Ideally, this function should reliably identify the correct orientation when docking bound subunits, yet still be able to accommodate small conformational changes when docking unbound structures. Hence "soft" Lennard-Jones (12-6) and hydrogen bond (12-10) potential functions were constructed from the OPLS (Optimised Potentials for Liquid Simulations) parameter set [17] in such a way as to retain the long range nature and minima of the original 12-6 and 12-10 forms whilst dramatically reducing the short-range repulsive behaviour. For example, each 12-6 potential of the form

$$E(r) = \frac{A}{r^{12}} - \frac{B}{r^6} \tag{8}$$

was fitted to a three-term expansion, $E_{LJ}(r)$:

$$E_{LJ}(r) = \sum_{n=1}^{3} e_n R_{n0}(\rho);\ \rho = r^2/(r_0/2)^2, \tag{9}$$

where $R_{nl}(\rho)$ are harmonic oscillator basis functions, ρ is a scaled distance, $r_0 = (A/B)^{1/6}$ is the zero-crossing point of the 12-6 potential, and e_n are the expansion coefficients. The coefficients, e_n, were determined by least-squares using six sample values of the target function. The first sample point, $E_{LJ}(0) = -33E(r_{\rm eq})$, where $r_{\rm eq} = (2A/B)^{1/6}$ is the location of the minimum of the 12-6 potential, limits the repulsive contribution to 33 times the well depth. The second sample point, $E_{LJ}(r_0 - 0.5) = 0$, moves the zero crossing point approximately 0.5Å closer to the origin in order to allow moderate atomic contacts to occur before being penalised by the repulsive contribution. The remaining sample points, $E_{LJ}(kr_{\rm eq}) = E(kr_{\rm eq})$ (with k = 1,2,4,8), serve as guide points for fitting the attractive part of the curve. Figure 2 shows a comparison of this softened potential with the original 12-6 form, using the OPLS parameters for a pair of ALA C_β atoms. Softened 12-10 potential functions, $E_{HB}(r)$, are calculated for all hydrogen-bonding atom types in a similar manner.

Following a correlation search, the first few hundred orientations are rigidly energy-minimised using soft OPLS energies calculated for all pairs of atoms i and j across the protein-protein interface (within a distance threshold of 10Å):

$$E_{OPLS} = \sum_{i\in A}\sum_{j\in B} E_{LJ}(r_{ij}) + E_{HB}(r_{ij}) + \frac{q_i q_j}{(4\overline{r}_{ij})\overline{r}_{ij}}, \tag{10}$$

where $4\overline{r}_{ij}$ is a distance-dependent dielectric [18], and $\overline{r}_{ij} = \max(r_{ij}, 1)$ avoids producing electrostatic spikes at sterically forbidden close contacts. The final docking score for each orientation is taken as the sum of the OPLS and shape-based correlation energies at the minimised orientation:

$$E_{TOTAL}(R, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2) = E_{SHAPE} + E_{OPLS}. \tag{11}$$

Clustering Solutions

Because our macromolecular surface sphere smothering procedure tends to over-sample the orientational search space, all low energy solutions are clustered in order to identify distinct orientations. The clustering algorithm first orders the docking solutions by energy, and allocates the lowest energy solution as the "seed" member of the first cluster. The list of remaining solutions is then scanned for unallocated entries, and any orientations for which the ligand C_α atoms fall within 2Å RMS of the corresponding atoms in the seed member are allocated to the current cluster. The list is then re-scanned for the next unallocated low energy solution which becomes the next cluster's seed, and the procedure is repeated until all solutions have been allocated to a cluster.

Even when it is not necessary to use multiple ligand starting orientations, this clustering algorithm provides a useful way to reduce the number of "false-positives" generated

by a docking search. For example, in an exhaustive search such as ours, many similar but nonetheless distinct orientations may be found, and these would tend to "push a good solution down the list" if clustering were not used. Clustering is also useful when using energy minimisation because multiple docking solutions can coalesce to a single minimised orientation.

[Figure 3 about here.]

Results

CAPRI Targets 1–7

In each of the seven CAPRI targets at least one of the docking partners was too big to be represented with reasonable accuracy by spherical polar Fourier expansions centred on a single coordinate origin. Hence the sphere smothering algorithm (Methods) was used to define multiple initial ligand docking orientations for all calculations except for targets 1 (HPrK/HPr) and 7 (SpeA/TCR-14.3.D). For each target, an initial low order $N = 16$ shape complementarity scan was performed using 492 and 642 icosahedral tessellation angular samples for the receptor and ligand rotations (giving rotational increments in (β_1, γ_1) and (β_2, γ_2) of approximately 10° and 8.5°, respectively), 64 twist angle increments, and 53 intermolecular separations in steps of ±0.75Å from the starting orientation(s). For all targets, the search was constrained using angular cutoffs for β_1 and/or β_2. The best 10,000 orientations from the $N = 16$ scan phase were then refined using combined high order shape and electrostatic correlations at $N = 30$, and using 128 steps for the twist angle search. The best 500 orientations from each $N = 30$ correlation were energy-minimised and clustered to give a final list of distinct docking orientations. The specific protocols used and the results obtained are summarised in Table I, and are described in further detail below.

Target 1: *Lactobacillus* HPr Kinase / *B. Subtilis* HPr

Although the kinase structure in target 1 is hexameric, each individual subunit was considered to be sufficiently small (265 residues) to be represented with reasonable accuracy using a single Fourier expansion about a coordinate origin placed near the C domain P-loop phosphate binding site. However, because the provided HPrK structure had an unresolved loop (residues 240-253), we first modelled this loop as an α-helix into the 1JB1 structure [14] using Modeller 4.0 [19]. The model-built monomer was then fitted to each domain in the hexamer, and all seleno-methionine residues were changed to methionines. The VAL-142:C_β atom of the C domain was chosen as the "receptor" origin, and the coordinate origin of the "ligand" HPr [20] was taken as the all-atom centre of mass of the HPr A chain. The HPr B chain was discarded. The expected phosphate binding residue HPr HIS-15 [21] was initially positioned close to the residues of the HPrK P-loop, and docking was performed with a receptor cutoff angle of 45°. Hence the HPr A domain was docked primarily onto the HPrK C domain, whilst retaining the remaining HPrK domains

in the final force-field calculation.

There is evidence that HPr SER-46 is strongly implicated in phosphate transfer [21], and we expected to see docking solutions in which SER-46 was physically close to the the HPrK P-loop. However, visual inspection of our low energy docking solutions showed only one orientation (rank 11 after clustering) with a close approach of both a serine residue (SER-12) and the phosphate binding residue HIS-15 to the HPrK P-loop, hence this orientation was selected (incorrectly) as the only feasible solution predicted by our algorithm. However, the revealed crystallographic solution [22] shows that the HPr HIS-15 residue is located at a substantial distance from the kinase P-loop, and that SER-46 is significantly closer to the P-loop than is SER-12 in our prediction.

Target 7: *Streptococcal* Exotoxin A1 / TCR 14.3.D

The superantigen *Streptococcal* pyrogenic exotoxin A1 (SpeA) [23] has a highly similar fold to the *Staphylococcus aureus* enterotoxin SEC3, despite a very low sequence identity. The structure of a complex between SEC3 and the TCR 14.3.D has been solved [24], and we assumed SpeA might bind to the TCR [25] in a similar manner to SEC3. Hence the SpeA moiety was manually positioned near the complementarity determining region (CDR) loops CDR1 and CDR2 of the TCR Vβ domain, TYR-48:C_α was selected as the Vβ coordinate origin, and docking was performed by tumbling the SpeA ligand over the TCR, with a receptor cutoff angle of $\beta_2 = 45°$. Thus the docking search ranged over the entire SpeA surface, but was largely constrained to the CDR loops of the Vβ domain. After clustering, visual inspection of the first 100 orientations showed no solutions which resembled the known TCR/SEC3 binding mode [24], hence the five lowest energy orientations were submitted as our predictions for this target.

Subsequent comparison of our predictions with the revealed structure of the complex [26] (PDB code 1L0Y) showed that we failed to identify practically any of the SpeA interface residues. In fact, the SpeA-TCR binding mode is highly reminiscent of the SEC3-TCR complex (PDB code 1JCK), despite several SpeA/SEC3 residue mutations at the TCR-binding surface.

Targets 2–6: Antibody / Large-Antigen Complexes

Each of the remaining targets called for the docking of an antibody to a large antigen. Hence we used the surface sphere smothering algorithm to generate multiple local coordinate systems around the antigen "receptor". In each case, the CDR loops of the smaller antibody "ligand" were manually centred about the negative z axis before generating multiple initial docking orientations of the ligand over the receptor surface. For targets 2 and 3 (antibody MCV/bovine rotavirus VP6, and antibody HC63/hemagglutinin) only the Fv fragment of the antibody was used in the docking calculation. For the remaining targets 4–6, each of which involved docking a different camelid antibody heavy chain variable domain (VHH) [27] onto porcine α-amylase [28], each VHH domain was initially oriented with its centre of mass at the origin and with the C_α of CDR1 residue 30 positioned near the negative z axis.

The surface sphere smothering algorithm was applied to the C chain of the rotavirus

VP6 trimer [29] and to the A and B chains of the hemagglutinin (HA) trimer [30], but ranged over the entire α-amylase surface. This generated from 21 (α-amylase) to 26 (HA) starting orientations, and each system was docked using angular search cutoffs of $\beta_1 = \beta_2 = 45°$, giving around 5–6$\times 10^8$ relative orientations in which at least one of the CDRs faced the antigen, but generally excluded trial orientations not involving the CDR loops. The best 500 orientations from each starting orientations were energy-minimised and clustered into a single list of solutions. For the VP6 and HA dockings, the structure of an intact antibody (PDB code 1IGT [31]) was superposed onto the first few docked Fv orientations, and those orientations which were judged to have an infeasible Fc take-off direction relative to the membrane-proximal region of the antigen were discarded. The lowest energy members of the first 5 surviving clusters were submitted to CAPRI. The 5 solutions submitted for each VHH/α-amylase complex were selected using only the calculated energies without further filtering.

This procedure successfully found a reasonably good solution for one of the two antibody binding modes in the large HA/HC63 complex [32] (rank 4, 43/63 correct residue contacts, with an Fv C_α RMS of 7.43Å, docking to the HA A and C domains), and a very good solution for the somewhat smaller AMD9/α-amylase complex [27] (rank 5, 53/65 correct residue contacts, with a VHH C_α RMS of 2.16Å). These docking orientations are illustrated in Figure 4. The revealed structures of the two unsuccessfully docked antibody/α-amylase complexes both have novel VHH binding modes, in which the majority of the antigen-binding residues are composed of framework and CDR3 residues [27]. However, the chosen angular cutoffs were too tight for the "side-on" binding modes of AMB7 and AMD10 to appear in the search space.

[Figure 4 about here.]

Retrospective Re-Docking

Although it is pleasing that good predictions were found for two of the seven targets, we wished to investigate the extent to which the various components in our scoring scheme helped or hindered our predictions. Hence each docking run was recalculated in the presence of the revealed complex structure, using correlations at both $N = 30$ and $N = 25$ [8], and both with and without the electrostatic correlation and soft OPLS energy minimisation procedure. For each target, the coordinates of the known complex were initially superposed onto the undocked receptor structure and the ligand was randomly positioned close to, but not coincident with, its position in the complex. This allowed the RMS deviation between each docked ligand orientation and the correct orientation in the complex to be calculated easily. All calculations used the same search parameters as above, except that the ligand cutoff angle β_2 was increased to 90° for targets 4 and 5 in order to include the unusual VHH binding orientations in the search space.

Table II shows the results of these retrospective docking calculations. This table shows that in the majority of cases using correlations to $N = 30$ gives superior predictions than using softer $N = 25$ correlations. It also shows that our electrostatic correlation is beneficial in some cases (targets 3–6) but not in others, notably target 7. Similarly, considering the $N = 30$ calculations, our OPLS refinement procedure improves the rank and/or RMS

of some solutions (targets 3, 5, 6, & 7) but worsens the remainder. It is interesting, and indeed ironic, to note that shape-only correlations at $N = 30$ identified rank-1 solutions for targets 1 and 6. Nonetheless, the combined $N = 30$ steric plus electrostatic plus OPLS scoring scheme gives a significantly improved average rank compared to shape-only correlations. It is worth noting that the best orientation obtained retrospectively for target 6 is superior to our original CAPRI submission. This suggests that despite the small search increments used here, there remains some dependence in our calculations on the starting orientation, and hence using yet smaller search increments could be beneficial.

Overall, Table II shows that our current algorithm could not have correctly predicted more than two low RMS orientations for the seven CAPRI targets within the limit of five submissions permitted by the CAPRI assessors. On the other hand, the final column of ranks in this table shows that when using steric plus electrostatic correlations to $N = 30$ in conjunction with soft OPLS energy minimisation, a good docking solution is ranked within the top 20 for four of the seven CAPRI targets.

Discussion

We have presented several enhancements to our original spherical polar Fourier docking correlation algorithm. In particular, the new surface sphere smothering algorithm allows a large receptor surface to be divided into a feasible number of smaller surface regions over which a ligand may be docked in a series of high resolution angular docking searches. There is essentially no limit to the size of proteins which may be docked with this approach, although the current implementation allows multiple local coordinate systems to be defined for only one (the larger) of the two docking partners. This approach allowed the five large antibody/antigen targets to be docked almost fully automatically. In these cases, it was expedient to use knowledge of the antibody CDRs and to arrange that these loops always faced the antigen during the docking search. This was easily achieved using two simple constraints on the angular degrees of freedom. However, our chosen angular constraints were too tight for the novel framework/CDR3 binding modes of the ABD10 and AMB7 VHH domains to be included in the search space around the α-amylase surface (targets 4 and 5).

The speed-up achieved by implementing the innermost cycle of our correlation search as a 1D FFT allowed finer angular search increments and higher order correlations to be used compared to our previous study [8]. Table II shows that using $N = 30$ correlations improves the average rank of good docking orientations compared to the softer $N = 25$ correlations used formerly by approximately a factor of 2. However, our retrospective docking results for target 6 indicate that the use of yet finer angular search increments would be desirable. This is practicable because despite the exhaustive nature of our search algorithm, the initial $N = 16$ scan of the search space is very fast (calculating up to 800,000 orientations/second on a dual processor 800MHz Pentium III Xeon PC), and total execution times are quite reasonable (Table I). Therefore, there is considerable scope to trade speed for increased accuracy.

Consistent with our earlier results for enzyme/inhibitor and antibody/antigen complexes [8], the benefit to be gained by including electrostatic correlations in the docking

score seems variable and unpredictable. We do not have a satisfactory explanation for this except to note that the $1/r$ form of an electrostatic potential has a much weaker distance dependence than a 12-6 or step-like "steric potential". Therefore, our current scoring function may be giving too much weight to this much weaker discriminant of complementarity. On the other hand, our soft OPLS refinement scheme seems promising. By design, the use of softened OPLS potentials significantly improves the energy and rank when re-docking the bound subunits of known complexes (data not shown), yet compared to shape-only correlations it still improves the rank and/or RMS deviations in five of the seven unbound docking problems studied here. Furthermore, the best average rank of good solutions is achieved when using shape-only $N = 30$ correlations followed by our soft OPLS refinement scheme (Table II). However, even a softened Lennard-Jones potential energy function is likely to be quite sensitive to small conformational changes in the binding site residues. Hence our soft potentials would probably not perform well if the conformational changes are much larger than those observed here (column 2 of Table II). Although there is scope to optimise the soft potential parameters used here, we expect it will be necessary to incorporate an explicit model of conformational flexibility if significantly more accurate models of protein-protein interactions are to be achieved.

Following the Round 1 docking runs (targets 1–3), considerable time was spent attempting to assess visually the feasibility of the docked solutions. However, we failed to recognise a good solution for target 1, and the best calculated solution for target 2 had too poor a rank to be considered. Visual inspection did not improve the satisfactory orientation (rank 4) submitted for target 3. In Round 2 (targets 4–7), the "ligand" CDR loops were initially oriented towards the corresponding antigen, but the final submissions for these targets were essentially selected automatically. Hence it is gratifying that a further good docking orientation was identified for target 6.

Conclusions

We have described several enhancements to our docking program, *Hex*. A novel surface sphere smothering algorithm allows our original approach to be extended to dock arbitrary-sized macromolecules such as large viral surface proteins. Using high order ($N = 30$) spherical polar Fourier correlations in conjunction with a soft molecular mechanics potential function often improves the rank obtained for low RMS docking orientations, even when docking unbound subunits. There remains scope to optimise the soft potential parameters, but we believe it will be necessary to incorporate an explicit model of conformational flexibility if we are to calculate significantly more accurate models of protein-protein interactions. Nonetheless, *Hex* 3.1 successfully identified good docking orientations for two of the seven target complexes presented in the blind CAPRI docking experiment, and subsequent analysis of our results shows that our algorithm can place a good solution within the top 20 orientations for four of the seven targets. This demonstrates that useful *in silico* protein-protein docking predictions can now be made with increasing confidence, even for very large macromolecular complexes. *Hex* 3.1 is available at *http://www.biochem.abdn.ac.uk/hex/*.

Acknowledgements

We are grateful to Russell Hamilton for assistance with the surface sphere smothering algorithm, Guillaume Valadon for assistance with surface rendering, and Graham Kemp for useful discussions. Much of *Hex* was developed during projects funded by the BBSRC.

References

1. S. J. Wodak and J. Janin. Computer analysis of protein-protein interaction. *J. Mol. Biol.*, 124:323–342, 1978.

2. M. L. Connolly. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. *Biopolymers*, 25:1229–1247, 1986.

3. I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct. Func. Genet.*, 47:409–443, 2002.

4. J. Janin. Welcome to CAPRI: A critical assessment of predicted interactions. *Proteins: Struct. Func. Genet.*, 47(3):257, 2002.

5. D. Fleury, R. S. Daniels, J. J. Skehel, M. Knossow, and T. Bizebard. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins: Struct. Func. Genet.*, 40:572–578, 2000.

6. J. S. Dixon. Evaluation of the CASP2 docking section. *Proteins: Struct. Func. Genet.*, Suppl. 1:198–204, 1997.

7. I. Vakser. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins: Struct. Func. Genet.*, Suppl. 1:226–230, 1997.

8. D. W. Ritchie and G. J. L. Kemp. Protein docking using spherical polar Fourier correlations. *Proteins: Struct. Func. Genet.*, 39(2):178–194, 2000.

9. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, and C. Aflalo. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci.*, 89:2195–2199, 1992.

10. H. A. Gabb, R. M. Jackson, and M. J. E. Sternberg. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, 272(1):106–120, 1997.

11. R. Chen and Z. Weng. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins: Struct. Func. Genet.*, 47:281–294, 2002.

12. A. Guéziec and R. Hummel. Exploiting triangulated surface extraction using tetrahedral decomposition. *IEEE Trans. Vis. Comp. Graph.*, 1(4):328–342, 1995.

13. J. A. Grant and B. T. Pickup. A Gaussian description of molecular shape. *J. Phys. Chem.*, 99:3503–3510, 1995.

14. S. Fieulaine, S. Morera, S. Poncet, V. Monedero, V. Gueguen-Chaignon, A. Galinier, J. Janin, J. Deutscher, and S. Nessler. X-ray structure of HPr kinase: a bacterial protein kinase with a P-loop nucleotide-binding domain. *EMBO J.*, 20(15):3917–3927, 2001.

15. W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *Computer Graphics*, 21(4):163–169, 1987.

16. D. W. Ritchie and G. J. L. Kemp. Fast computation, rotation and comparison of low resolution spherical harmonic molecular surfaces. *J. Comp. Chem.*, 20(4):383–395, 1999.

17. W. L. Jorgensen and J. Tirado-Rives. The OPLS potential functions for proteins. Energy minimisations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1671, 1988.

18. R. W. Pickersgill. A rapid method of calculating charge-charge interaction energies in proteins. *Protein Eng.*, 2:247–248, 1988.

19. A. Šali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial constraints. *J. Mol. Biol.*, 234:779–815, 1993.

20. D. I. Liao and O. Herzberg. Refined structures of the active Ser83→Cys and impaired Ser46→Asp histidine-containing phosphocarrier proteins. *Structure*, 2:1203–1216, 1994.

21. O. Herzberg, P. Reddy, S. Sutrina, M. H. Saier, J. Reizer, and G. Kapadia. Structure of the histidine-containing phosphocarrier protein HPr from *bacillus subtilis* at 2.0-Å resolution. *Proc. Natl. Acad. Sci.*, 89:2499–2503, 1992.

22. S. Fieulaine, S. Morera, S. Poncet, I. Mijakovic, A. Galinier, J. Janin, J. Deutscher, and S. Nessler. X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr. *Proc. Natl. Acad. Sci.*, 99(21):13437–13441, 2002.

23. A. C. Papageorgiou, C. M. Collins, D. M. Gutman, J. B. Kline, S. M. O'Brien, H. S. Tranter, and K. R. Acharya. Structural basis for the recognition of superantigen streptococcal pyrogenic exotoxin A (SpeA1) by MHC class II molecules and T-cell receptors. *EMBO J.*, 18:9–21, 1995.

24. B. A. Fields, E. L. Malchiodi, H. Li, X. Ysern, C. V. Stauffacher, P. M. Schlievert, K. Karjalainen, and R. A. Mariuzza. Crystal structure of a T-cell receptor β-chain complexed with a superantigen. *Nature*, 384:188–192, 1996.

25. G. A. Bentley, G. Boulot, K. Karjalainen, and R. A. Mariuzza. Crystal structure of the beta chain of a T cell antigen receptor. *Science*, 267:1984–1987, 1995.

26. E. J. Sundberg, H. Li, A. S. Liera, J. K. McCormick, J. Torma, P. M. Schlievert, K. Karjalainen, and R. A. Mariuzza. Structures of two streptococcal superantigens bound to TCR β chains reveal diversity in the architecture of T cell signaling complexes. *Structure*, 10:687–699, 2002.

27. A. Desmyter, S. Spinelli, F. Payan, M. Lauwereys, L. Wyns, S. Muyldermans, and C. Cambillau. Three camelid VHH domains in complex with porcine pancreatic α-amylase. *J. Biol. Chem.*, 277(26):23645–23650, 2002.

28. M. Machius, L. Vértesy, R. Huber, and G. Wiegand. Carbohydrate and protein-based inhibitors of porcine pancreatic α-amylase; structure analysis and comparison of their binding characteristics. *J. Mol. Biol.*, 260:409–421, 1996.

29. M. Mathieu, I. Petitpas, J. Navaza, J. Lepault, E. Kohli, P. Pothier, B. V. Venkataram Prasad, J. Cohen, and F. A. Rey. Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. *EMBO J.*, 20(7):1485–1497, 2001.

30. N. K. Sauter, J. E. Hanson, G. D. Glick, J. H. Brown, R. L. Crowthers, S. J. Park, J. J. Skehel, and D. C. Wiley. Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: Analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography. *Biochemsitry*, 31:9609–9621, 1992.

31. L. J. Harris, S. B. Larson, K. W. Hasel, J. Day, A. Greenwood, and A. McPherson. The three-dimensional structure of an intact monoclonal antibody for canine lymphoma. *Nature*, 360:369–372, 1992.

32. C. Barbey-Martin, B. Gigant, T. Bizebard, L. J. Calder, S. A. Wharton, J. J. Skehel, and M. Knossow. An antibody that prevents the hemagglutinin low pH fusogenic transition. *Virology*, 294:70–74, 2002.

Table I: *Hex* 3.1 Blind Docking Results for CAPRI Targets 1–7

Target	Description	Constraints[a]	Samples[b]	Rank[c]	Contacts[d]	RMS[e]	Time/h[f]
1	HPrK/HPr	$\beta_1 \leq 45°$	1.5×10^8	-/1	2/56	-	0.2
2	VP6/MCV	$\beta_1, \beta_2 \leq 45°$	5.6×10^8	-/5	0/52	-	5.3
3	Hemagglutinin/HC63	$\beta_1, \beta_2 \leq 45°$	6.3×10^8	4/5	43/63	7.43	6.5
4	α-Amylase/AMD10	$\beta_1, \beta_2 \leq 45°$	5.6×10^8	-/5	0/58	-	4.7
5	α-Amylase/AMB7	$\beta_1, \beta_2 \leq 45°$	5.6×10^8	-/5	0/64	-	4.7
6	α-Amylase/AMD9	$\beta_1, \beta_2 \leq 45°$	5.6×10^8	5/5	53/65	2.16	4.7
7	SpeA/14.3.D	$\beta_1 \leq 45°$	1.5×10^8	3/5	2/37	-	0.25

[a]The angular constraints applied during the docking search. β_1 and β_2 refer to the latitudinal rotation angles for the receptor and ligand, respectively.

[b]The number of orientations evaluated in the $N = 16$ scan of the search space.

[c]The rank of the best prediction and the total number of predictions submitted to CAPRI. A maximum of 5 submissions per target was permitted. A hyphen denotes no submission within 10Å RMS of the complex.

[d]The fraction of correct residue contacts in the predicted docking orientation.

[e]The C_α RMS deviation of the best docked ligand orientation with respect to the revealed complex structure, following least-squares superposition of the docked structure onto the complex using all receptor C_α atoms. The RMS deviation for target 3 was calculated using only the Fv fragment of HC63.

[f]Total docking time in hours on a dual processor 800MHz Pentium III Xeon PC.

Table II: *Hex* 3.1 Retrospective Docking Results for CAPRI Targets 1–7. Listed are the rank and RMS deviation of the first docked ligand orientation found within 10Å RMS of the orientation of the revealed complex, calculated using different combinations of scoring functions. Each docking calculation used a random initial ligand orientation and the same search cutoff angles as used in the blind trial submissions, with the exception of targets 4 and 5 for which the ligand cutoff angle was increased to $\beta_2 = 90°$ in order to include the unusual VHH binding modes in the search space. A hyphen indicates no solution found within the top 5,000 orientations.

			Shape Only[a]		+EL[b]		+MM[c]		+EL+MM[d]	
Target	R/L RMS[e]	N[f]	Rank	RMS	Rank	RMS	Rank	RMS	Rank	RMS
1	2.39/0.57	30	1	4.06	11	4.06	33	4.18	19	4.18
2	0.62/0	30	37	3.05	41	3.05	123	2.90	119	2.90
3-AC[g]	1.13/0	30	112	9.99	18	2.75	3	6.95	3	6.95
3-CE	1.13/0	30	62	8.91	39	8.91	18	8.43	15	8.43
4	0.41/0	30	63	3.05	10	3.05	125	8.74	106	8.74
5	0.41/0	30	826	2.74	118	2.74	18	2.49	20	2.49
6	0.41/0	30	1	0.75	1	0.75	1	0.64	1	0.64
7	1.09/0.49	30	122	5.37	295	9.22	39	5.60	118	8.83
Average Rank:			175		76		51		57	
1	2.39/0.57	25	6	4.06	104	4.06	13	9.16	16	7.41
2	0.62/0	25	452	3.02	715	3.02	359	2.85	333	2.85
3-AC	1.13/0	25	245	2.75	27	2.75	54	3.38	43	3.38
3-CE	1.13/0	25	544	8.20	85	1.80	129	9.96	88	9.96
4	0.41/0	25	70	3.69	13	3.79	19	1.50	20	1.50
5	0.41/0	25	185	2.72	29	2.67	169	2.55	5	2.49
6	0.41/0	25	40	0.75	8	0.75	4	0.64	4	0.64
7	1.09/0.49	25	234	5.80	-	-	64	5.89	-	-
Average Rank:[h]			257		163		124		92	

[a]Shape-only docking correlations.

[b]Shape plus electrostatic correlations.

[c]Shape-only correlations followed by soft molecular mechanics minimisation of the best 500 orientations from each starting orientation.

[d]Shape plus electrostatic correlations followed by soft molecular mechanics minimisation of the best 500 orientations from each starting orientation.

[e]R/L RMS denotes the C_α RMS deviation between the unbound and complexed structures of the receptor (R) and ligand (L), tabulated as R/L.

[f]The order, N, of the high resolution shape and electrostatic correlations.

[g]There are two antibody/HA binding modes in target 3. One antibody binds across the AC domains (labelled 3-AC) and the other spans the CE domains (3-CE).

[h]Average Rank calculated using results only for targets 1–6 at $N = 25$.

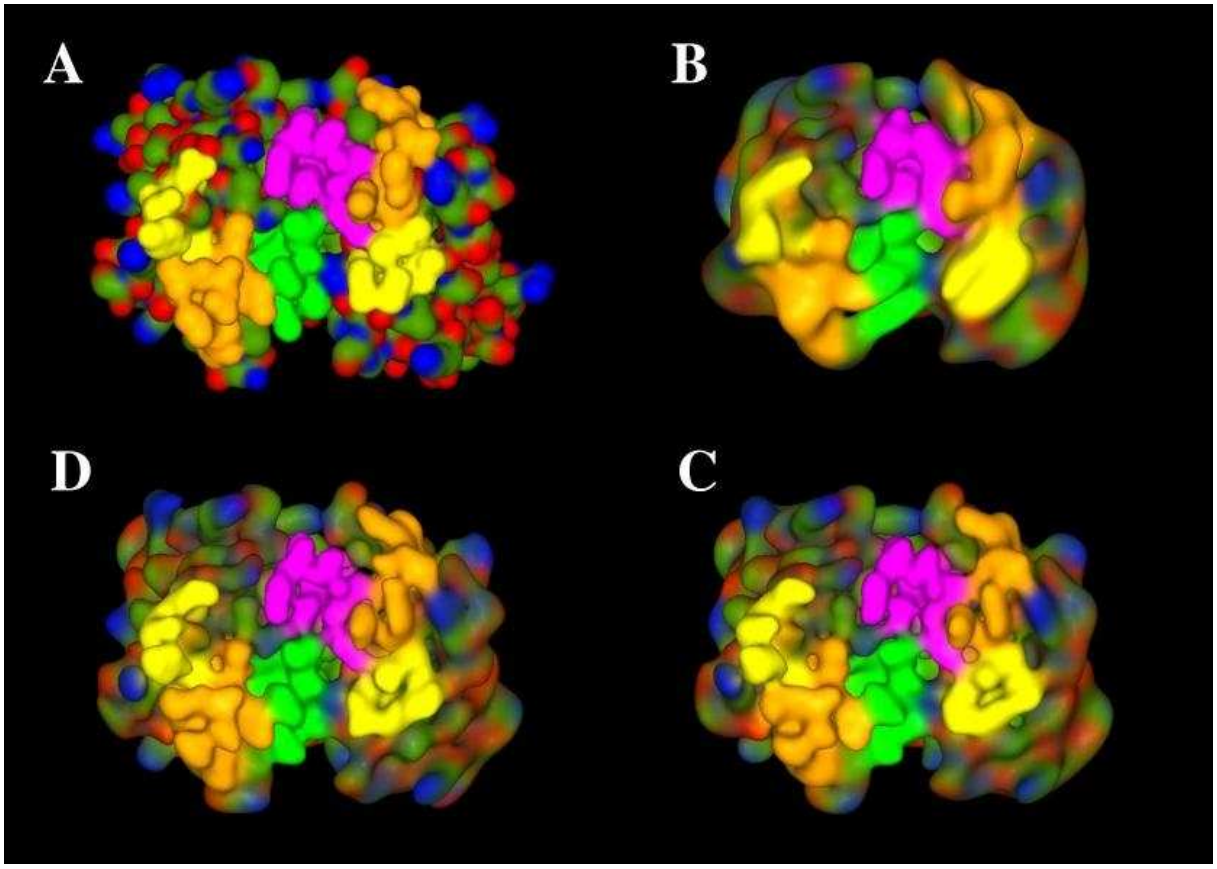

Figure 1: Spherical polar Fourier shape reconstruction of the MCV Fv fragment, looking directly at the CDR loops. (A) The molecular surface contoured from atomic Gaussians. (B, C, D) The spherical polar Fourier molecular surface reconstructed from expansions to order $N = 16$, 25, and 30, respectively, contoured with a steric density threshold of 0.25. The coordinate origin for the spherical polar expansion is taken as the centre of mass of the Fv fragment. In each image, the VH domain is on the left and the VL domain is on the right. The CDR loop atoms are coloured as H1: gold; H2: yellow; H3: green; L1: yellow; L2: gold; L3: magenta. All other atoms are coloured by atom type; carbon: green; nitrogen: blue; oxygen: red. Atom colours are assigned to surface vertices using a Gaussian weighting rule. Hence surface colours become smeared in regions of low shape resolution. Each surface was contoured using the marching tetragons algorithm (Methods).

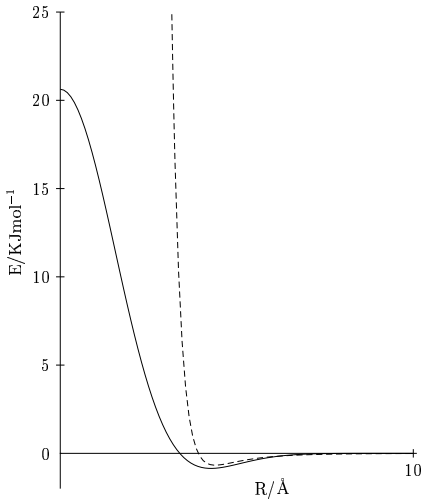

Figure 2: Comparison of a softened Lennard-Jones potential (solid line) with the original OPLS 12-6 form (dashed line). The softened potential is derived from the 12-6 potential by least squares fitting to a three term harmonic oscillator expansion (Eq 9).

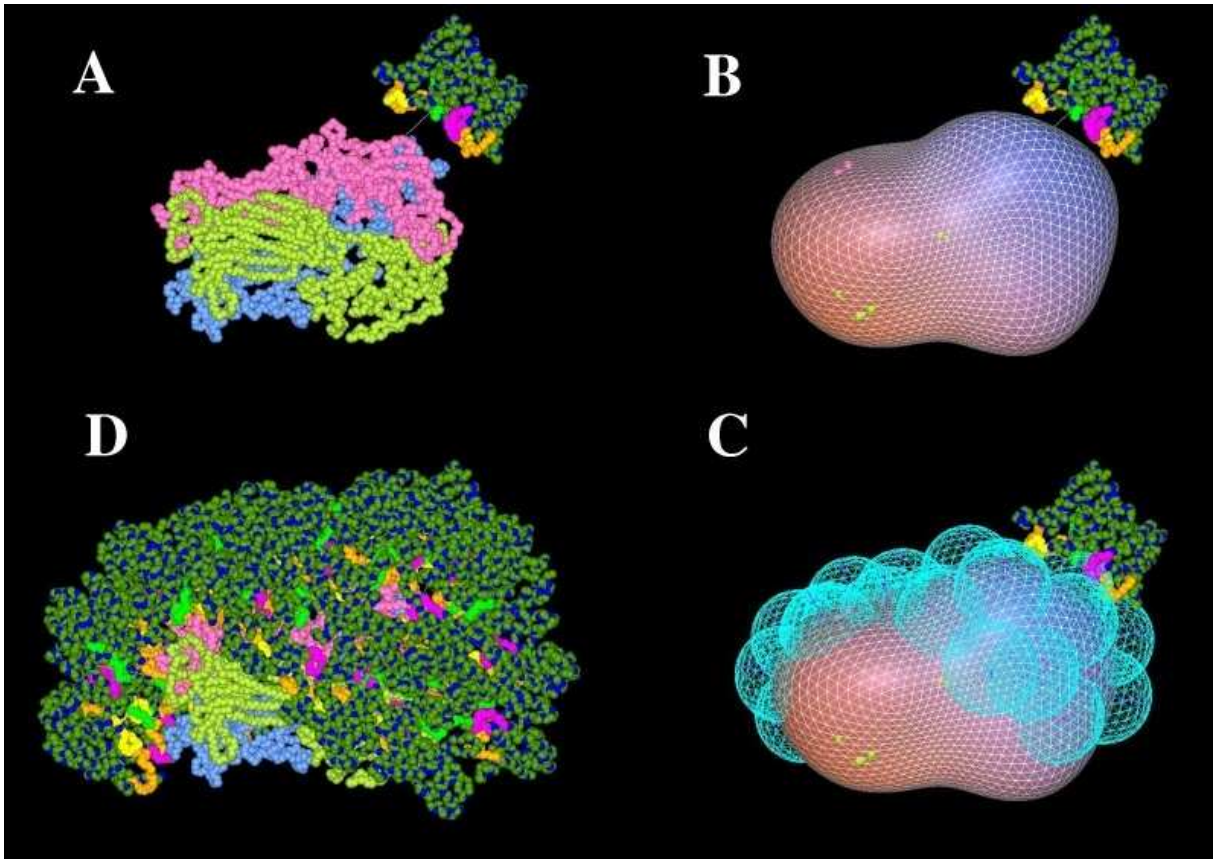

Figure 3: The four stages of the macromolecular surface sampling algorithm, illustrated schematically for the antibody MCV/VP6 complex (CAPRI target 2). (A) The CDR loops of the MCV Fv fragment (the "ligand") are initially oriented to face the VP6 trimer (the "receptor"). The VP6 chains are coloured as A: blue; B: yellow; C: pink. The CDR loops of the much smaller antibody are coloured as in Figure 1. (B) A low resolution (L=5) spherical harmonic surface is calculated for the receptor (2,252 surface triangles). (C) The spherical harmonic receptor surface after applying the sphere smothering algorithm to the C chain of the receptor. (D) Multiple initial docking orientations for the ligand are generated from the sphere centres. This example shows 23 MCV Fv fragments distributed over the VP6 C chain.

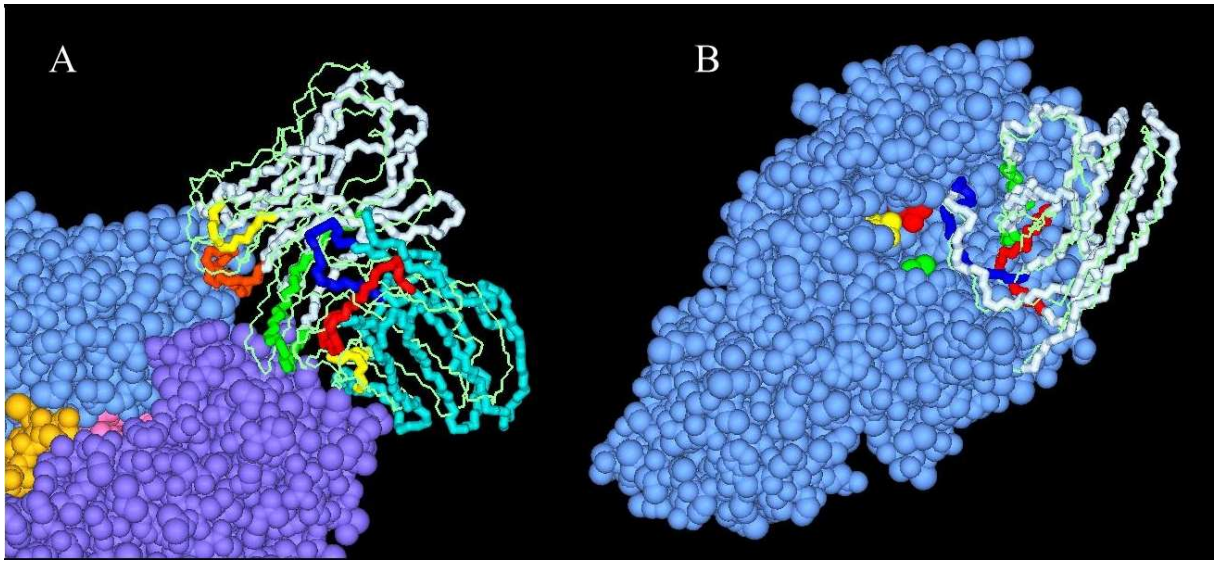

Figure 4: (A) The best docking solution obtained for the HA/HC63 complex (target 3, rank 4, 43/63 correct residue contacts, Fv RMS 7.43Å). The HC63 Fv fragment is coloured as VH: white; VL: cyan; H1: orange; H2: yellow; H3: green; L1: red; L2: yellow; L3: blue. The crystallographic orientation of the Fv is shown in light green. The HA chains are coloured as A: light blue; C: pink; E: dark blue; F: orange. (B) The best docking solution obtained for the camelid antibody AMB9/α-amylase complex (target 6, rank 5, 53/65 correct residue contacts, VHH RMS 2.16Å). The α-amylase is in blue and the AMB9 VHH domain is in white. The active-site amylase residues are coloured ASP-197: red; GLU-233: yellow; ASP-300: green. The VHH CDR loops are coloured as CRD1: red; CDR2: green; CDR3: blue. The crystallographic orientation of the VHH is shown in light green.