

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

David W. Ritchie*

Department of Computing Science, King's College, University of Aberdeen, Aberdeen, United Kingdom

ABSTRACT: 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 by using a novel soft molecular mechanics energy minimization 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 for four of the seven targets. This finding shows that useful *in silico* protein–protein docking predictions can now be made with increasing confidence, even for very large macromolecular complexes. *Proteins* 2003;52:98–106.

© 2003 Wiley-Liss, Inc.

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

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 toward developing improved docking algorithms and several successes have been described, a general solution to the so-called docking problems 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 want to model have never been greater. Thus, there is an increasing need to develop accurate and fast docking algorithms.

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 experiment⁴ 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,^{9–11} which accelerate translational correlations, our approach uniquely favors 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 modeled 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 by using a “soft” molecular

Abbreviations: FFT: fast Fourier transform; HPr: histidine-containing phosphocarrier protein; OPLS: optimized potentials for liquid simulations; PC: personal computer; PDB: Protein Data Bank; RMS: root-mean-square; TCR: T-cell receptor; 1D: one-dimensional; 3D: three-dimensional.

*Correspondence to: David W. Ritchie, Department of Computing Science, King's College, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom. E-mail: dritchier@csd.abdn.ac.uk

Received 8 October 2002; Accepted 19 December 2002

mechanics energy minimization procedure, and the list of docking solutions may be clustered to help identify distinct orientations and reduce the number of “false-positives.” In addition, protein surface shapes are now calculated by using a new “marching tetragons” algorithm¹² to contour atomic Gaussian density representations¹³ of each protein. This treats reentrant surface regions more reliably than our former dot surface sampling scheme⁸ and allows improved graphical visualization of results.

To try to honor 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 for four of the seven targets.

MATERIALS AND 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

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

where $\rho_i(r)$ 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 and Pickup,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 by 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 by using an adaptation of the tetrahedral decomposition algorithm by Guézic and Hummel,¹² which we call “marching tetragons.” Compared with 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 sum-

mary here. A Fourier expansion to order N of some property $A(r)$ in spherical polar coordinates $r = (r, \theta, \phi)$ may be written as

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

where a_{nlm} is an expansion coefficient, $R_{nl}(r)$ represents either a 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, and F. A. Rey, unpublished results) reconstructed at different expansion orders, N . The low-order $N = 16$ expansion encodes considerable global shape information, whereas 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 farthest from the origin, even for high-order expansions.

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:

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

and

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

where the rotated and translated expansion coefficients are, respectively, given by (see Ref. 16)

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

and (manuscript in preparation)

$$a''_{nlm} = \sum_{n'l'}^N T_{nl,n'l'}^{(|m|)}(R) a_{n'l'm} \quad (6)$$

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 $A(r)$ and $B(r)$ for molecules A and B, respectively, is calculated as

$$S_{AB}(R, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2) = \int [T_z(-R)\hat{R}(0, \beta_1, \gamma_1)A(r)] \\ \times [\hat{R}(\alpha_2, \beta_2, \gamma_2)B(r)]dV \quad (7)$$

Conceptually, at each trial intermolecular separation, each protein is incrementally rotated by 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 , β_1 , γ_1 , β_2 , γ_2), the correlation in α_2 may be expressed as a 1D Fourier series (see Eq. 27 in Ref. 8), which we recently implemented as an FFT. Compared to our earlier implementation, this gives a speedup 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 30 Å 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 2. 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 surface¹⁶ is calculated for the receptor. The surface is discretized by projecting it onto an icosahedral tessellation of the sphere, as shown in Figure 2(B). At each triangular facet of the surface, a normal vector is calculated, and a 15 Å radius sphere is centered 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 the sphere that has the greatest volume overlap with its neighbors. This procedure is repeated until no overlap volume exceeds 5 Å³. 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 center. 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 2(C) shows the result of applying the surface spheres algorithm to only the C chain of the VP6 trimer, and Figure 2(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 wanted 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 (Optimized Potentials for Liquid Simulations) parameter set¹⁷ in such a way as to retain the long-range nature and minima of the original 12-6 and 12-10 forms while dramatically reducing the short-range repulsive behavior. For example, each 12-6 potential of the form

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

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

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

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, $E_{LJ}(0) = -33E(r_{eq})$, where $r_{eq} = (2A/B)^{1/6}$ 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 0.5 Å closer to the origin to allow moderate atomic contacts to occur before being penalized by the repulsive contribution. The remaining sample points, $E_{LJ}(kr_{eq}) = E(kr_{eq})$ (with $k = 1, 2, 4, 8$), serve as guide points for fitting the attractive part of the curve. Figure 3 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 minimized by using soft OPLS energies calculated for all pairs of atoms i and j across the protein–protein interface (within a distance threshold of 10 Å):

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

where $4\bar{r}_{ij}$ is a distance-dependent dielectric,¹⁸ and $\bar{r}_{ij} = \max(r_{ij}, 1)$ 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 minimized orientation:

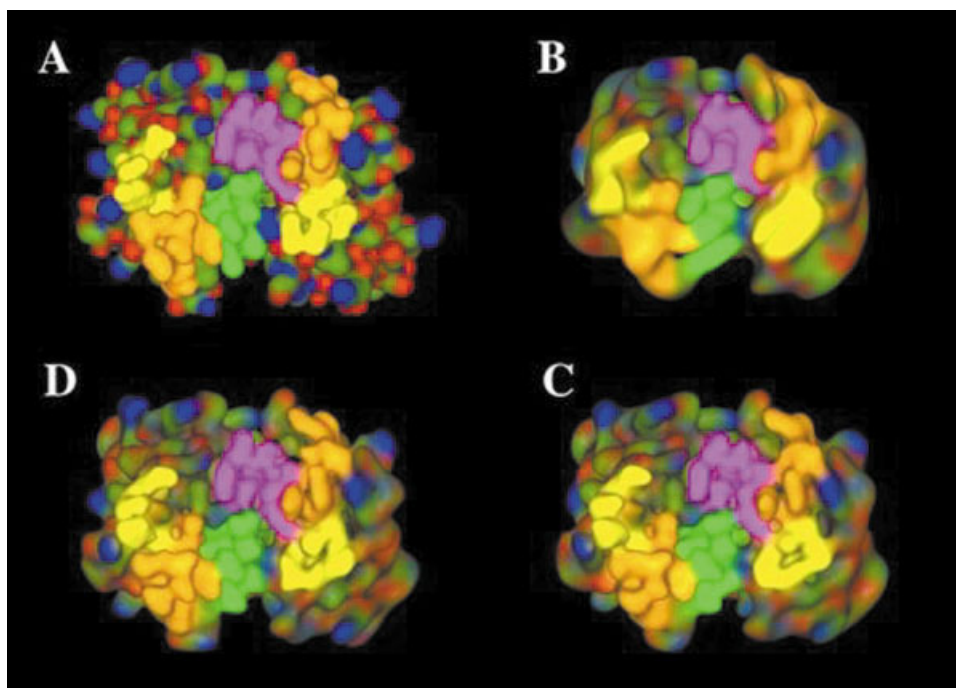


Fig. 1. Spherical polar Fourier shape reconstruction of the MCV Fv fragment, looking directly at the CDR loops. **A:** The molecular surface contoured from atomic Gaussians. **B, C, D:** The spherical polar Fourier molecular surface reconstructed from expansions to order $N = 16, 25$, and 30 , respectively,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

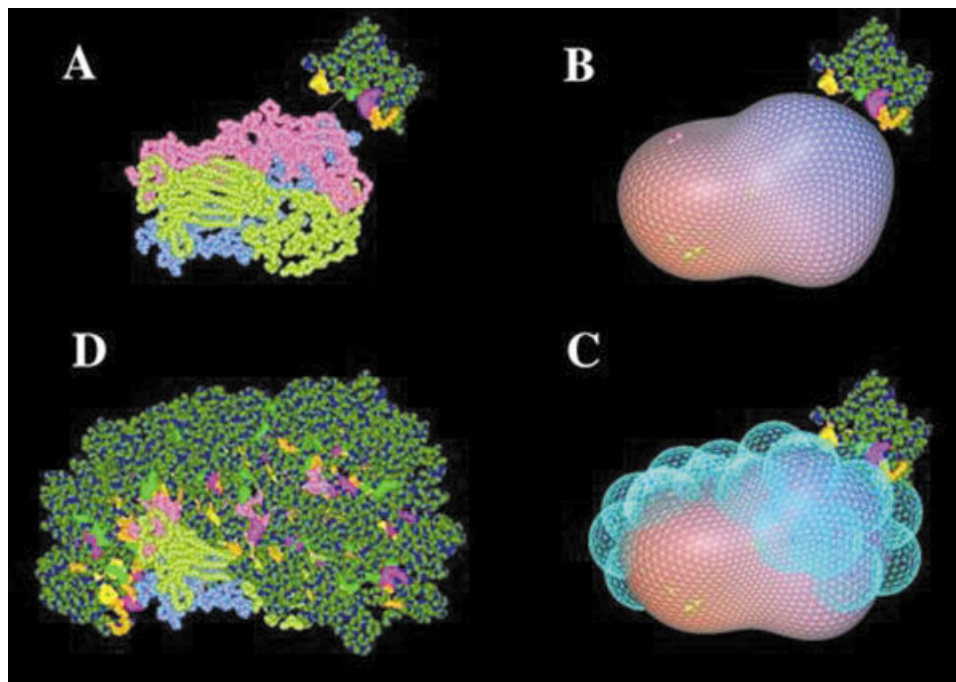


Fig. 2. The four stages of the macromolecular surface sampling algorithm, illustrated schematically for the antibody MCV-VP6 complex (CAPRI target 2). **A:** The CDR loops of the MCV Fv fragment (the “ligand”) are initially oriented to face the VP6 trimer (the “receptor”). The VP6 chains are colored as follows: A: blue; B: yellow; C: pink. The CDR loops of the much smaller antibody are colored as in Figure 1. **B:** A low-resolution ($L = 5$) spherical harmonic surface is calculated for the receptor (2252 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 centers. This example shows 23 MCV Fv fragments distributed over the VP6 C chain.

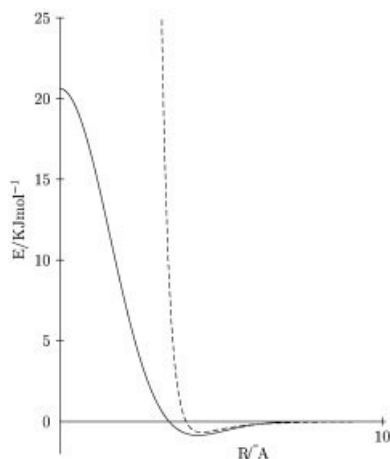


Fig. 3. Comparison of a softened Lennard-Jones potential (solid line) with the original OPLS 12-6 form (dashed line). The softened potential is derived from the 12-6 potential by least-squares fitting to a three-term harmonic oscillator expansion (Eq. 9).

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

Clustering Solutions

Because our macromolecular surface sphere smothering procedure tends to over-sample the orientational search space, all low-energy solutions are clustered 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 2 Å RMS of the corresponding atoms in the seed member are allocated to the current cluster. The list is then rescanned 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 minimization because multiple docking solutions can coalesce to a single minimized orientation.

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 centered on a single coordinate origin. Hence, the sphere smothering algorithm (see Materials and 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 by 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 ± 0.75 Å from the starting orientation(s). For all targets, the search was constrained by using angular cutoffs for β_1 and/or β_2 . The best 10,000 orientations from the $N = 16$ scan phase were then refined by 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 minimized and clustered to give a final list of distinct docking orientations. The specific protocols used and the results obtained are summarized 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 modeled this loop as an α -helix into the 1JB1 structure¹⁴ by using Modeller 4.0.¹⁹ The model-built monomer was then fitted to each domain in the hexamer, and all selenomethionine 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” HPr²⁰ was taken as the all-atom center of mass of the HPr A chain. The HPr B chain was discarded. The expected phosphate-binding residue HPr HIS-15²¹ 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, while 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 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 solution²² 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 SER-12 is 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

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	$\beta_1 \leq 45^\circ$	1.5×10^8	—/1	2/56	—	0.2
2	VP6/MCV	$\beta_1, \beta_2 \leq 45^\circ$	5.6×10^8	—/5	0/52	—	5.3
3	Hemagglutinin/HC63	$\beta_1, \beta_2 \leq 45^\circ$	6.3×10^8	4/5	43/63	7.43	6.5
4	α -Amylase/AMD10	$\beta_1, \beta_2 \leq 45^\circ$	5.6×10^8	—/5	0/58	—	4.7
5	α -Amylase/AMB7	$\beta_1, \beta_2 \leq 45^\circ$	5.6×10^8	—/5	0/64	—	4.7
6	α -Amylase/AMD9	$\beta_1, \beta_2 \leq 45^\circ$	5.6×10^8	5/5	53/65	2.16	4.7
7	SpeA/14.3.D	$\beta_1 \leq 45^\circ$	1.5×10^8	3/5	2/37	—	0.25

^aThe angular constraints applied during the docking search. β_1 and β_2 refer to the latitudinal rotation angles for the receptor and ligand, respectively.

^bThe number of orientations evaluated in the $N = 16$ scan of the search space.

^cThe rank of the best prediction and the total number of predictions submitted to CAPRI. A maximum of five submissions per target was permitted. A dash denotes no submission within 10 Å RMS of the complex.

^dThe fraction of correct residue contacts in the predicted docking orientation.

^eThe C_α RMSD between the best docked ligand orientation and the revealed complex structure, following least-squares superposition of the docked structure onto the complex using all receptor C_α atoms. The RMSD for target 3 was calculated by using only the Fv fragment of HC63.

^fTotal docking time in hours on a dual processor 800 MHz Pentium III Xeon PC.

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 TCR²⁵ 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 V β domain, TYR-48: C_α was selected as the V β coordinate origin, and docking was performed by tumbling the SpeA ligand over the TCR, with a receptor cutoff angle of $\beta_2 = 45^\circ$. Thus, the docking search ranged over the entire SpeA surface but was largely constrained to the CDR loops of the V β domain. After clustering, visual inspection of the first 100 orientations showed no solutions that 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 complex²⁶ (PDB code 1LOY) 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 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 centered 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 center 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 trimer²⁹ and to the A and B chains of the hemagglutinin (HA) trimer³⁰ but ranged over the entire α -amylase surface. This generated from 21 (α -amylase) to 26 (HA) starting orientations, and each system was docked by using angular search cutoffs of $\beta_1 = \beta_2 = 45^\circ$, giving around $5\text{--}6 \times 10^8$ 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 orientation were energy minimized and clustered into a single list of solutions. For the VP6 and HA dockings, the structure of an intact antibody (PDB code 1IGT³¹) was superposed onto the first few docked Fv orientations, and those orientations that were judged to have an infeasible Fc takeoff direction relative to the membrane-proximal region of the antigen was discarded. The lowest energy members of the first five surviving clusters were submitted to CAPRI. The five solutions submitted for each VHH- α -amylase complex were selected by 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 complex³² (rank 4, 43 of 63 correct residue contacts, with an Fv C_α RMS of 7.43 Å, docking to the HA A and C domains), and a very good solution for the somewhat smaller AMD9- α -amylase complex²⁷ (rank 5, 53 of 65 correct residue contacts, with a VHH C_α RMS of 2.16 Å). 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 most 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.

Retrospective Redocking

Although it is pleasing that good predictions were found for two of the seven targets, we wanted to investigate the

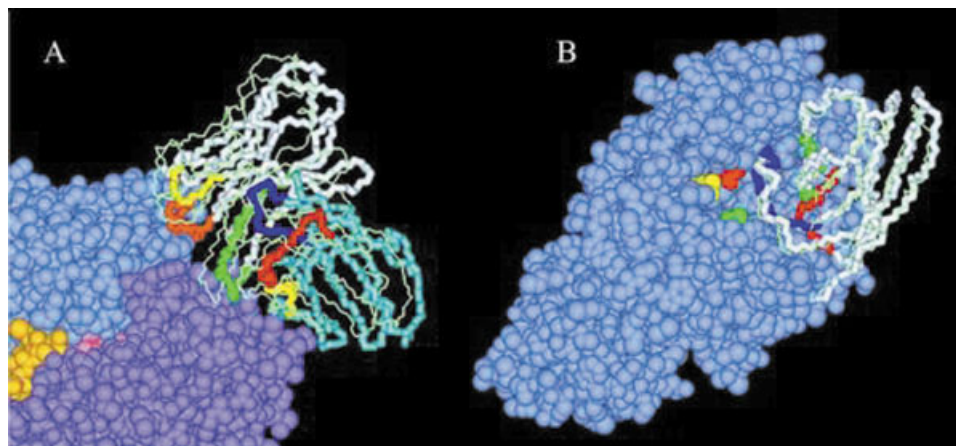


Fig. 4. **A:** The best docking solution obtained for the HA-HC63 complex (target 3, rank 4, 43 of 63 correct residue contacts, Fv RMS 7.43 Å). The HC63 Fv fragment is colored as follows: VH: white; VL: cyan; H1: orange; H2: yellow; H3: green; L1: red; L2: yellow; L3: blue. The crystallographic orientation of the Fv is shown in light green. The HA chains are colored as follows: 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 of 65 correct residue contacts, VHH RMS 2.16 Å). The α -amylase is in blue and the AMB9 VHH domain is in white. The active-site amylase residues are colored as follows: ASP-197: red; GLU-233: yellow; ASP-300: green. The VHH CDR loops are colored as follows: CDR1: red; CDR2: green; CDR3: blue. The crystallographic orientation of the VHH is shown in light green.

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 minimization 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 RMSD 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 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 most 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, and 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 finding 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 minimization, 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 that 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 constrain these loops to face the antigen during the docking search. This was easily achieved by 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).

TABLE II. *Hex* 3.1 Retrospective Docking Results for CAPRI Targets 1–7[†]

Target	R/L RMS ^e	N ^f	Shape only ^a		+EL ^b		+MM ^c		+EL + MM ^d	
			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	

[†]Listed are the rank and RMSD of the first docked ligand orientation found within 10 Å RMS of the orientation of the revealed complex, calculated by 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 $\beta_2 = 90^\circ$ to include the unusual VHH-binding modes in the search space. A dash indicates no solution found within the top 5000 orientations.

^aShape-only docking correlations.

^bShape plus electrostatic correlations.

^cShape-only correlations followed by soft molecular mechanics minimization of the best 500 orientations from each starting orientation.

^dShape plus electrostatic correlations followed by soft molecular mechanics minimization of the best 500 orientations from each starting orientation.

^eR/L RMS denotes the C_α RMSD between the unbound and complexed structures of the receptor (R) and ligand (L), tabulated as R/L.

^fThe order, N, of the high-resolution shape and electrostatic correlations.

^gThere are two antibody-HA binding modes in target 3. One antibody binds across the AC domains (labeled 3-AC), and the other spans the CE domains (3-CE).

^hAverage rank calculated by using results only for targets 1–6 at $N = 25$.

The speedup 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 than in 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/s on a dual processor 800 MHz 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 steplike “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 redocking the bound subunits of known complexes (data not shown); yet compared to shape-only correlations it still improves the rank and/or RMSD 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 optimize 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 recognize 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 toward 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 large 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 optimize 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 finding shows 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/>.

ACKNOWLEDGMENTS

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

- Wodak SJ, Janin J. Computer analysis of protein-protein interaction. *J Mol Biol* 1978;124:323–342.
- Connolly ML. Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface. *Biopolymers* 1986;25:1229–1247.
- Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* 2002;47:409–443.
- Janin J. Welcome to CAPRI: a critical assessment of predicted interactions. *Proteins* 2002;47:257.
- Fleury D, Daniels RS, Skehel JJ, Knossow M, Bizebard T. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins* 2000;40:572–578.
- Dixon JS. Evaluation of the CASP2 docking section. *Proteins* 1997;Suppl 1:198–204.
- Vakser I. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins* 1997;Suppl. 1:226–230.
- Ritchie DW, Kemp GJL. Protein docking using spherical polar Fourier correlations. *Proteins* 2000;39:178–194.
- Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci USA* 1992;89:2195–2199.
- Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J Mol Biol* 1997;272:106–120.
- Chen R, Weng Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins* 2002;47:281–294.
- Guézic A, Hummel R. Exploiting triangulated surface extraction using tetrahedral decomposition. *IEEE Trans Vis Comp Graph* 1995;1:328–342.
- Grant JA, Pickup BT. A Gaussian description of molecular shape. *J Phys Chem* 1995;99:3503–3510.
- Fioulaine S, Morera S, Poncet S, Monedero V, Gueguen-Chaignon V, Galinier A, Janin J, Deutscher J, Nessler S. X-ray structure of HPr kinase: a bacterial protein kinase with a P-loop nucleotide-binding domain. *EMBO J* 2001;20:3917–3927.
- Lorenson WE, Cline HE. Marching cubes: a high resolution 3D surface construction algorithm. *Computer Graphics* 1987;21:163–169.
- Ritchie DW, Kemp GJL. Fast computation, rotation and comparison of low resolution spherical harmonic molecular surfaces. *J Comp Chem* 1999;20:383–395.
- Jorgensen WL, Tirado-Rives J. The OPLS potential functions for proteins: energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc* 1988;110:1657–1671.
- Pickersgill RW. A rapid method of calculating charge-charge interaction energies in proteins. *Protein Eng* 1988;2:247–248.
- Şali A, Blundell TL. Comparative protein modelling by satisfaction of spatial constraints. *J Mol Biol* 1993;234:779–815.
- Liao DI, Herzberg O. Refined structures of the active Ser83→Cys and impaired Ser46→Asp histidine-containing phosphocarrier proteins. *Structure* 1994;2:1203–1216.
- Herzberg O, Reddy P, Sutrina S, Saier MH, Reizer J, Kapadia G. Structure of the histidine-containing phosphocarrier protein HPr from *bacillus subtilis* at 2.0-Å resolution. *Proc Natl Acad Sci USA* 1992;89:2499–2503.
- Fioulaine S, Morera S, Poncet S, Mijakovic I, Galinier A, Janin J, Deutscher J, Nessler S. X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr. *Proc Natl Acad Sci USA* 2002;99:13437–13441.
- Papageorgiou AC, Collins CM, Gutman DM, Kline JB, O'Brien SM, Tranter HS, Acharya KR. Structural basis for the recognition of superantigen streptococcal pyrogenic exotoxin A (SpeA1) by MHC class II molecules and T-cell receptors. *EMBO J* 1995;18:9–21.
- Fields BA, Malchiodi EL, Li H, Ysern X, Stauffacher CV, Schlievert PM, Karjalainen K, Mariuzza RA. Crystal structure of a T-cell receptor β -chain complexed with a superantigen. *Nature* 1996;384:188–192.
- Bentley GA, Boulot G, Karjalainen K, Mariuzza RA. Crystal structure of the beta chain of a T cell antigen receptor. *Science* 1995;267:1984–1987.
- Sundberg EJ, Li H, Liera AS, McCormick JK, Torma J, Schlievert PM, Karjalainen K, Mariuzza RA. Structures of two streptococcal superantigens bound to TCR β chains reveal diversity in the architecture of T cell signaling complexes. *Structure* 2002;10:687–699.
- Desmyter A, Spinelli S, Payan F, Lauwereys M, Wyns L, Muyldermans S, Cambillau C. Three camelid VHH domains in complex with porcine pancreatic α -amylase. *J Biol Chem* 2002;277:23645–23650.
- Machius M, Vértessy L, Huber R, Wiegand G. Carbohydrate and protein-based inhibitors of porcine pancreatic α -amylase; structure analysis and comparison of their binding characteristics. *J Mol Biol* 1996;260:409–421.
- Mathieu M, Petitpas I, Nayaza J, Lepault J, Kohli E, Pothier P, Venkataram Prasad BV, Cohen J, Rey FA. Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. *EMBO J* 2001;20:1485–1497.
- Sauter NK, Hanson JE, Glick GD, Brown JH, Crowthers RL, Park SJ, Skehel JJ, Wiley DC. 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. *Biochemistry* 1992;31:9609–9621.
- Harris LJ, Larson SB, Hasel KW, Day J, Greenwood A, McPherson A. The three-dimensional structure of an intact monoclonal antibody for canine lymphoma. *Nature* 1992;360:369–372.
- Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, Knossow M. An antibody that prevents the hemagglutinin low pH fusogenic transition. *Virology* 2002;294:70–74.