Journal of Computer-Aided Molecular Design **18:** 55–70, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

Mapping protein pockets through their potential small-molecule binding volumes: QSCD applied to biological protein structures

Keith Mason, Nehal M. Patel, Aric Ledel, Ciamac C. Moallemi & Edward A. Wintner*
NeoGenesis Pharmaceuticals, Inc., 840 Memorial Dr., Cambridge, MA 02139, USA

Received 11 December 2003; accepted in revised form 2 February 2004

Key words: cavity detection, molecular recognition, protein diversity, protein pockets, QSCD, quantized surface complementarity, small molecule binding volumes

Summary

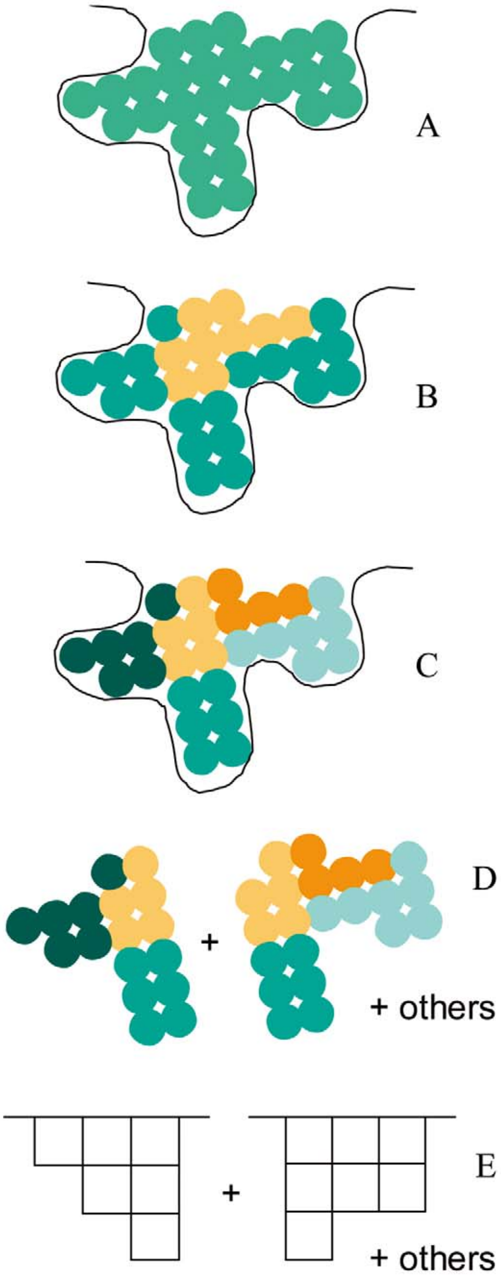
Previously we demonstrated a method, Quantized Surface Complementarity Diversity (QSCD), of defining molecular diversity by measuring shape and functional complementarity of molecules to a basis set of theoretical target surfaces [Wintner E.A. and Moallemi C.C., J. Med. Chem., 43 (2000) 1993]. In this paper we demonstrate a method of mapping actual protein pockets to the same basis set of theoretical target surfaces, thereby allowing categorization of protein pockets by their properties of shape and functionality. The key step in the mapping is a 'dissection' algorithm that breaks any protein pocket into a set of potential small molecule binding volumes. It is these binding volumes that are mapped to the basis set of theoretical target surfaces, thus measuring a protein pocket not as a single surface but as a collection of molecular recognition environments.

Introduction

With each passing year, both in academia and the pharmaceutical industry, an ever increasing number of small molecules is screened against an ever increasing set of protein targets. The role of molecular diversity models is to classify these small molecules in a manner that will rapidly match a given protein with a concentrated set of likely binding compounds. Current diversity models are based on a wide variety of chemical property measurement techniques, including physicochemical properties, 2D substructure fingerprints, and 3D pharmacophores; several excellent reviews of the field have been authored by Yvonne C. Martin [2, 3] and Hans Matter [4]. This work, in which we attempt to use a single theoretical framework to encompass the diversity of both small molecules and protein pockets, is an extension of our own diversity model based on 3D shape and functionality [1].

The importance of modeling protein diversity cannot be overstated at a time when the majority of human proteins have yet to enter a pharmaceutical screening campaign. If one can group proteins using a measure that is relevant to their ability to bind small molecules, then one can match focused sets of screening compounds to multiple targets at once. Not only is this an obvious economy in terms of time and chemistry; it also suggests new areas of discovery as soon as potent structures have been developed for one protein within a structural class, thus allowing so-called 'privileged structures' to be applied to novel targets. Classification of proteins using sequence and secondary structure is a well established field [5, 6]. Less so is the general characterization of un-aligned proteins by their 3D structures (for 3D analysis of pre-aligned proteins, see [7]). To date, most general 3D protein characterization algorithms compare the position and connectivity of secondary structural elements, classifying proteins based on similar 3D conglomerations of alpha-helices and/or beta sheets, often termed 'folds' [6, 8–10]. Another promising method is to search for specific 3D arrangements of three or more residues, such as a catalytic triad [11–13]. In this paper, based on our assumption that the properties of molecular recognition are determining factors in the binding of a

*To whom correspondence should be addressed. E-mail: wintner@neogenesis.com



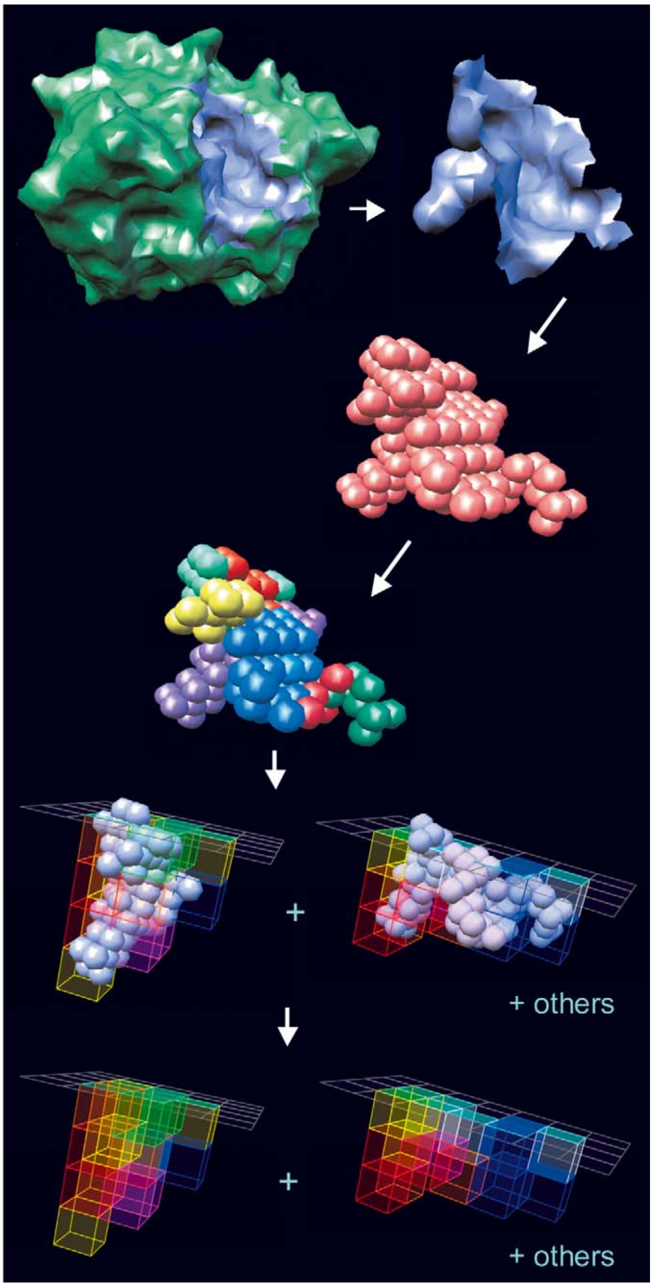


Figure 1. A schematic representation of the algorithm used to dissect a protein pocket into quantized binding volumes. (A) Filling a protein pocket with a lattice of balls. (B) Determination of 'surface balls' (blue) and 'interior balls' (yellow). (C) Partitioning of surface balls into surface partitions and interior balls into interior partitions. (D) Conglomeration of partitions into potential binding volumes. (E) Quantization of binding volumes.

Figure 2. Screenshots of the dissection of a protein pocket into quantized binding volumes. The pocket shown is the active site of protein structure 1dre (dihydrofolate reductase).

ligand to a protein pocket, we theorized that a comparison of the shape and functionality defined by the concave surface of protein pockets should also result in a pharmaceutically relevant model of protein diversity.

The diversity model of our previous work [1] was based on the cubic 'quantization' of small molecules: each conformation of a compound was represented as a set of contiguous 4.24 Å cubes that encompassed the VDW volume of the molecule's atoms. Each of the cubes of the quantized molecule was given one of seven functionalities based on the atoms it contained: negatively charged, positively charged, H-bond donor, H-bond acceptor, H-bond donor/acceptor, polarizable, or hydrophobic. Once in a cubically quantized format, the small-molecules were then mapped to a basis set of theoretical target surfaces: a complete enumeration under a given set of rules of all possible shapes containing 6–14 4.24 Å cubes. For each molecule, this resulted in a collection of all theoretical target surfaces to which any of the molecule's quantized conformations were complementary. The process, which we named QSCD (for Quantized Surface Complementarity Diversity), was shown to be a valid method of classifying small molecules; when small molecules were grouped according to the similarity of their QSCD mappings, these groupings correlated well with the actual biological activity of the molecules.

An obvious extension of the QSCD approach was to map actual protein pockets to the same basis set of theoretical target surfaces, thus bringing both proteins and small molecules into the same frame of reference. Carrying this concept through to a practical conclusion proved more difficult than expected, however. The problem which arose was one of size: most protein pockets are large compared to the ligands that they bind. In this paper we detail one solution to the problem, presenting a method of 'dissecting' protein pockets into their potential small molecule binding volumes. We then map these binding volumes to our basis set of theoretical target surfaces, and we validate the use of the QSCD approach to classify proteins through the shape and functionality of their pockets.

Methods

A basis set of theoretical target surfaces is created as previously described [1], but expanded from 6–14 to 5–25 cubes. Each theoretical surface is formed by successively carving contiguous cubic units out of an initially flat surface. The cubic units represent 'negative space' that a potential ligand could occupy. Each cube is 4.24 Å on a side and may have any one of eight functionalities: negatively charged, positively charged, H-bond donor, H-bond acceptor, H-bond donor/acceptor, polarizable, hydrophobic, and 'average' (slightly hydrophobic). The basis set thus approximates all possible binding pockets of volume between 380 and 1900 Å^3, and creates a total set of approximately 1568 trillion theoretical surfaces.

To map actual protein pockets onto the basis set of theoretical target surfaces, the given protein pockets must be quantized into the cubic framework of the basis set. At first glance, one could simply quantize the entire negative space of a protein pocket. However, many concave areas of a protein are quite large and contain within them multiple, often overlapping, small molecule binding pockets. Thus, in order to map protein pockets onto a basis set of theoretical target surfaces such that the mapping contains information relevant to small molecule binding, it was necessary to develop an algorithm that dissected a potentially large protein pocket into a set of constituent 'binding volumes'. The set of binding volumes that results from dissection of a given protein pocket must both be able to characterize the protein pocket in comparison to other protein pockets, and must also be complementary to the small molecules that bind to the protein pocket. As verified in the results and discussion section, these criteria are met by the algorithm detailed below.

Protein pocket dissection

The algorithm used to dissect a protein pocket into binding volumes is given in steps 1–9 below: (A schematic dissection is shown in Figure 1, and screenshots of the process for the MMDB protein 1dre are shown in Figure 2.)

(1) A simulated H_2O contact surface is generated for a protein structure of interest (see Experimental for details).

(2) Concave pocket surfaces are found on the protein by using an in-house slicing algorithm (see Experimental). Other pocket detection algorithms [14–17] may be used as well.

(3) All concavities at least 400 Å^3 in volume are filled with a lattice of balls on the same axis. Figure 1A shows this process schematically; details may be found in the experimental section. A lattice is used that mimics the average VDW distribution of a small

molecule: experimentally, a 1.65 Å cubic lattice of balls with a radius of 1.3 Å was found to give good results. (This correlates well with a typical C-C bond distance of 1.5 Å and a typical VDW radius for H of 1.2 Å.)

(4) All sets of contiguous balls are grouped as a 'pocket volume'. The total volume of each protein pocket is thus defined by a set of contiguous balls in a cubic lattice.

(5) All balls of a given pocket volume whose radii are within 0.7 Å of a VDW radius of a protein atom are deemed to be 'surface balls'. The remainder are deemed to be 'interior balls' (see Figure 1B). Both surface and interior balls are then ranked according to the number of balls adjacent to them in the lattice, where the maximum number of lattice neighbors that a ball can have is 26. Balls with least neighbors are ranked highest, with equally ranked balls sub-prioritized at random.

(6) Starting with the surface ball of highest rank (fewest neighbors), a contiguous set of balls called a 'surface partition' is created. (The results of a partitioning are diagrammed schematically in Figure 1C.) First, all surface balls within 1.5× of the lattice space distance are added to the partition. Next, all surface balls are added to the partition that (a) are within 1.5× of the lattice space distance to any of the partition's current balls, and (b) have no more than 20 neighboring balls of any type. Finally, all surface balls are again added to the partition that (a) are within 1.5× of the lattice space distance to any of the partition's current balls, and (b) have no more than 20 neighboring balls of any type. Once the first surface partition is created, the next is created starting from the highest ranked surface ball of all remaining surface balls that have not been added to any partition.

(7) Starting with the interior ball of highest rank, all interior balls within 1.95× of the lattice space distance are added to create an 'interior partition'. (The results of a partitioning are diagrammed schematically in Figure 1C.) Once the first interior partition is created, the next is created starting from the highest ranked interior ball of all remaining interior balls.

(8) Interior partitions not adjacent to any other interior partitions are grafted onto the smallest adjacent surface partition. Surface partitions that have 5 or fewer balls are grafted to the smallest adjacent surface partition. If there are more than 15 surface partitions, the smallest surface partition is grafted to the smallest contiguous surface partition. The process is repeated until the maximum of 15 surface partitions is met. This limit prevents a combinatorial explosion when partitions of a very large protein pocket are conglomerated in step 9 below.

(9) Primary binding volumes are created by conglomerating contiguous surface partitions into all possible permutations of at least three surface partitions. A secondary set of binding volumes is created by adding all contiguous interior partitions to each primary binding volume. A tertiary set of binding volumes is created by adding all contiguous interior and surface partitions to each secondary binding volume. The final set is the sum of all *unique* binding volumes from the three subclasses above, less all binding volumes that exceed 120 surface balls. Sample binding volumes are shown schematically in Figure 1D.

Quantization of binding volumes

For each binding volume, as defined by its total set of contiguous balls, a quantized cubic representation (Figure 1E) of the pocket's shape and functionality is created using the algorithm below. The process is analogous to our previous algorithm for the quantization of molecules [1], with specific changes as noted:

(1) Cubes are drawn around the binding volume's balls (instead of around the VDW radii of a molecule's atoms), thus yielding sets of contiguous 4.24 Å cubes that represent the shape of a given binding volume.

(2) The alignment of the cube grid is initially set by the principle axes of rotation of the ball centers. As in our molecular quantizations, the grid alignment is then shifted slightly through all directions and rotations, and the quantization is saved that: (a) has the fewest cubes, and (b) is closest to the principal alignment.

(3) The 'top' or 'opening' of the quantized binding volume is designated as that direction (of six possible) which contains the most cube face normals that contact no protein surface.

(4) The functionality of each cube is assigned by assessing the protein atoms that surround the cube and determining the dominant molecular environment as selected from the following list: negatively charged, positively charged, H-bond donor, H-bond acceptor, H-bond donor/acceptor, polarizable, hydrophobic, and 'average' (slightly hydrophobic).

Mapping of binding volumes

The quantized representations of all binding volumes for a given protein pocket are mapped against the basis set of theoretical target surfaces by the following algorithm:

(1) Orient the 'top' of the quantization with its normal to the +Z axis.
(2) If the top layer of cubes contains fewer than 3 cubes, remove the top layer until this condition is met.
(3) If the total number of cubes in the quantization is greater than 25, remove the top layer until this condition is met.
(4) If the total number of cubes is less than 8, delete the quantization, otherwise the quantization of the binding volume now maps directly to a single member of the basis set of QSCD theoretical target surfaces.

Training

Dissection parameters were optimized with the following three training structures chosen from the MMDB database of protein crystal sructures: 1dre [18], 1hck [19], and 1qkt [20]. Each structure has a co-crystallized ligand, which was removed before the pocket was subjected to the pocket-finding and dissection algorithms.

Dissection parameters were optimized in a visually guided progression following the criteria below:
(1) Distinct topological regions in a given protein pocket, such as a 'hole' into which an amino acid residue might bind, should be exactly encompassed by a single partition (see Figure 1C).
(2) At least one predicted binding volume (one union of the partitions) should closely encompass the VDW volume of the known binding ligand in its co-crystallized conformation.
(3) We set parameters to obtain the fewest possible number of predicted binding volumes while maintaining the above conditions.

Results and discussion

For a given protein pocket, the above dissection process results in a set of potential small molecule binding volumes. This data is then used to create a list of all theoretical target surfaces that match a binding volume of the protein pocket. The list of quantized surfaces thus created is a numerical representation of the protein pocket in terms of the shape and functionality of the binding volumes it contains.

Model validation

To test the validity of analyzing protein pockets by mapping their dissected binding volumes to a basis set of theoretical target surfaces, we asked the following

Table 1. Listing of the 30 protein structures used in the pocket comparison experiment of Table 2.

PDB code	Protein name	Resolution (Å)	Ref
Kinases			
1a9u	MAP kinase p38	2.5	21
1b6c	Ser/Thr-kinase of TGF-beta receptor type I	2.6	22
1csn	Casein kinase 1	2	23
1fvt	Cyclin-Dependent Kinase 2 (Cdk2)	2.2	24
1gag	Insulin receptor kinase	2.7	25
1ia8	Human cell cycle checkpoint kinase chk1	1.7	26
1jnk	C-Jun N-terminal kinase (jnk3S)	2.3	27
1phk	Phosphorylase kinase	2.2	28
1qpj	Lymphocyte-specific kinase (Lck)	2.2	29
1stc	Camp-dependent protein kinase (capk)	2.3	30
2src	Human tyrosine protein kinase C-Src	1.5	31
Other proteins			
1alo	Aldehyde oxidoreductase	2	32
1ap8	Translation initiation factor eif4e	NMR	33
1b87	Aminoglycoside n6′-acetyltransferase type 1	2.7	34
1cou	Nematode anticoagulant protein c2	NMR	35
1csj	Hepatitis c virus RNA polymerase	2.8	36
1d8c	Malate synthase g	2	37
1dik	Pyruvate Phosphate Phosphotransferase	2.3	38
1dlc	Insecticidal delta-endotoxin cryiiia (bt13)	2.5	39
1ex1	Beta-d-glucan exohydrolase isoenzyme	2.2	40
1gcb	Gal6, Bleomycin hydrolase DNA-binding protease	2.2	41
1gnd	Guanine nucleotide dissociation inhibitor	1.8	42
1i78	Outer membrane protease ompt	2.6	43
1i9b	Acetylcholine binding protein (achbp)	2.7	44
1kuh	Zinc protease	1.6	45
1ndo	Naphthalene 1,2-dioxygenase	2.3	46
1p32	Mitochondrial matrix protein sf2p32	2.3	47
1ppn	Papain	1.6	48
1prh	Prostaglandin h2 synthase-1 (cyclooxygenase I)	3.5	49
1ryt	Rubrerythrin	2.1	50

Table 2. QSCD scores of each pairing of 11 different active site kinase pockets and 19 non-kinase protein pockets. Scoring is based on the number of matching quantized surface shapes that have at least six functional cubes identical in type and location. Scores are colored to show value according to the key.

Pockets at the Active Site of Kinase Proteins — **Other Pockets Selected Randomly from the PDB**

	1a9u	1b6c	1csn	1fvt	1gag	1ia8	1jnk	1phk	1qpj	1stc	2src	1alo	1ap8	1b87	1cou	1csj	1d8c	1dik	1dlc	1ex1	1gcb	1gnd	1i78	1i9b	1kuh	1ndo	1p32	1ppn	1prh	1ryt
1a9u	--	5.860	0.021	0.000	0.159	0.071	0.009	0.294	0.000	0.113	4.265	0.008	0.040	0.003	0.000	0	0.000	0.017	0.003	0.000	0.030	0.000	0.002	0	0.000	0	0.000	0.000	0.000	0.003
1b6c	5.860	--	0.167	10.25	5.314	21.97	1.335	2.414	0.163	4.900	62.46	0.016	0.087	0.044	0.006	0	0.013	0.004	0.243	0	0.089	0.016	0.084	0	0	0.001	0.005	0.001	0.070	0
1csn	0.021	0.167	--	0.143	0.186	0.327	0.060	0.015	0.161	0.308	0.701	0	0.028	0.026	0	0	0.001	0.001	0	0.004	0	0	0	0	0	0	0	0.049	0	0
1fvt	0.000	10.25	0.143	--	0.068	0.004	0.005	0.014	0.288	0.037	0.355	0	0	0.003	0	0	0	0	0.003	0	0	0	0	0	0	0	0	0.005	0.001	0
1gag	0.159	5.314	0.186	0.068	--	0.308	0.626	0.975	0.021	0.014	8.530	0.001	0.016	0.015	0	0	0.001	0	0	0	0.001	0	0.002	0	0	0.002	0	0.003	0.010	0
1ia8	0.071	21.97	0.327	0.004	0.308	--	0	0.107	0.004	0.073	0.205	0	0	0.022	0	0	0	0	0	0	0	0.002	0.020	0	0	0	0	0	0	0.024
1jnk	0.009	1.335	0.060	0.005	0.626	0	--	0.015	0.006	0.005	0.013	0.002	0.004	0	0	0	0	0	0	0	0	0	0	0	0	0	0.003	0	0	0
1phk	0.294	2.414	0.015	0.014	0.975	0.107	0.015	--	0.047	0.837	1.533	0	0.001	0.038	0	0	0.001	0	0	0	0	0	0	0	0	0	0	0.002	0.008	0.012
1qpj	0.000	0.163	0.161	0.288	0.021	0.004	0.006	0.047	--	0.127	0.013	0	0	0.080	0	0	0.001	0	0	0	0	0	0.022	0	0	0	0.003	0.005	0	0
1stc	0.113	4.900	0.308	0.037	0.014	0.073	0.005	0.837	0.127	--	2.771	0.000	0.015	0.035	0	0	0	0.000	0.015	0	0.000	0	0	0	0	0	0.001	0.005	0.004	0.001
2src	4.265	62.46	0.701	0.355	8.530	0.205	0.013	1.533	0.013	2.771	--	0.012	0.025	0.069	0	0	0	0.000	0.005	0.000	0.005	0.002	0.002	0	0	0	0.013	0.006	0.005	0.002
1alo	0.008	0.016	0	0	0.001	0	0.002	0	0	0.000	0.012	--	0.012	0.001	0	0	0.001	0	0	0	0	0	0	0	0	0	0.001	0	0.001	0.002
1ap8	0.040	0.087	0.028	0	0.016	0	0.004	0.001	0	0.015	0.025	0.012	--	0.059	0	0	0.001	0	0	0.003	0	0	0	0	0	0	0	0	0.001	0
1b87	0.003	0.044	0.026	0.003	0.015	0.022	0	0.038	0.080	0.035	0.069	0.001	0.059	--	0.003	0	0.008	0.007	0	0	0	0	0	0	0	0	0.002	0.003	0.001	0
1cou	0.000	0.006	0	0	0	0	0	0	0	0	0	0	0	0.003	--	0	0	0	0.003	0	0	0.002	0	0	0	0	0.022	0	0.012	0
1csj	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--	0	0	0	0	0	0	0	0	0	0	0	0	0.002	0
1d8c	0.000	0.013	0.001	0	0.001	0	0	0.001	0.001	0	0	0.001	0.001	0.008	0	0	--	0	0	0	0	0	0	0	0	0	0	0	0	0
1dik	0.017	0.004	0.001	0	0	0	0	0	0	0.000	0.000	0	0	0.007	0	0	0	--	0	0	0	0.001	0	0	0	0	0	0	0	0
1dlc	0.003	0.243	0	0.003	0	0	0	0	0	0.015	0.005	0	0	0	0.003	0	0	0	--	0	0	0	0	0	0	0	0	0.004	0.001	0.003
1ex1	0.000	0	0.004	0	0	0	0	0	0	0	0.000	0	0.003	0	0	0	0	0	0	--	0	0	0	0	0	0	0	0	0	0
1gcb	0.030	0.089	0	0	0.001	0	0	0	0	0.000	0.005	0	0	0	0	0	0	0	0	0	--	0	0.001	0	0	0	0	0	0.001	0
1gnd	0.000	0.016	0	0	0	0.002	0	0	0	0	0.002	0	0	0	0.002	0	0	0.001	0	0	0	--	0	0	0	0	0	0	0.001	0
1i78	0.002	0.084	0	0	0.002	0.020	0	0	0.022	0	0.002	0	0	0	0	0	0	0	0	0	0.001	0	--	0	0	0	0	0	0	0.002
1i9b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--	0	0	0	0	0	0
1kuh	0.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--	0	0	0.009	0	0
1ndo	0	0.001	0	0	0.002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	--	0.002	0	0	0
1p32	0.000	0.005	0	0	0	0	0.003	0	0.003	0.001	0.013	0.001	0	0.002	0.022	0	0	0	0	0	0	0	0	0	0	0.002	--	0.023	0	0
1ppn	0.000	0.001	0.049	0.005	0.003	0	0	0.002	0.005	0.005	0.006	0	0	0.003	0	0	0	0	0.004	0	0	0	0	0	0.009	0	0.023	--	0	0
1prh	0.000	0.070	0	0.001	0.010	0	0	0.008	0	0.004	0.005	0.001	0.001	0.001	0.012	0.002	0	0	0.001	0	0.001	0.001	0	0	0	0	0	0	--	0
1ryt	0.003	0	0	0	0	0.024	0	0.012	0	0.001	0.002	0.002	0	0	0	0	0	0	0.003	0	0	0	0.002	0	0	0	0	0	0	--

Score Coloring Key:

0	to	0.01
0.01	to	0.05
0.05	to	0.1
0.1		and higher

Score Summary:

Type	Total	#	Avg Score
Kin-Kin	277.3	110	2.521
Kin-Oth	2.85	418	0.007
Oth-Oth	0.39	342	0.001
Non-Kin	3.24	760	0.004

Signal to Noise:
592

question: Can the mapping of the shapes and functionalities of potential binding volumes classify proteins known to have similar pockets as similar, and proteins known to have different pockets as different?

To answer this question, we chose a set of 11 different active site kinase pockets and 19 non-kinase protein pockets. Table 1 lists the 30 proteins used in the experiment. The 11 kinases were selected from the MMDB database and any co-crystallyzed ligands were removed. The non-kinase pockets were chosen by selecting 19 random proteins from the MMDB database, verifying that they were not known kinases, removing any ligands, and then selecting one random pocket each from these 19 proteins (minimum pocket volume of 400 Å^3). All 30 pockets were dissected using the algorithm detailed above, and their resulting binding volumes were quantized as described above. The quantized representations of all binding volumes for a given protein pocket were then mapped against our basis set of theoretical target surfaces. This yielded for each kinase a list of quantized surfaces of which its pocket was comprised. All 30 sets of quantized surfaces were then compared pairwise to each other in a matrix format.

For each pairing of the 30 kinase and non-kinase protein pockets, the following information was calculated:

P1Size: # Protein 1 quantized surface shapes (Protein 1 being always the greater number)

P2Size: # Protein 2 quantized surface shapes (Protein 2 being always the lesser number)

Isect: # Matching quantized surface shapes (This is the Intersection of Proteins 1 and 2)

FIsect: # Matching quantized surface shapes that have at least 6 functional cubes identical in type and location. (This is the Functional Intersection of Proteins 1 and 2)

Score = FIsect ˆ3 / ((P1Size + P2Size) / 2)

In the scoring method above, FIsect is cubed to place a greater importance on the differences in FIsect as compared to the differences in average surface shapes when comparing pairs of protein pockets. Thus, in this scoring system, greatest weight is being placed on matching potential binding volumes, with a lesser regard to the size of the pockets being compared. If FIsect is not cubed when comparing pairs of protein pockets, differences regarding matching potential binding volumes are swamped by the great fluctuation in pocket sizes.

The resulting score of each pairing of the 30 protein pockets is shown in the matrix of Table 2 and is broken down in the listing of Table 3. As we had hoped, the matrix clearly shows that the method in question categorizes kinase pockets as being more similar to each other than to random protein pockets. Thus, our analysis of protein pockets by mapping their dissected binding volumes to a basis set of theoretical target surfaces appears to have captured some portion of the molecular recognition properties that give proteins their various biological functions. Overall, the ‘signal’ of kinase-kinase pocket comparisons is well above the expected ‘noise’ of random pocket comparisons (S/N > 500). To confirm that the method is not artifactual, and is truly categorizing the proteins based on the similar 3D structure of their binding pockets, we examined many of the matches between pairwise comparisons in-depth. In particular, we found that when we aligned the matching quantized surfaces of any high scoring kinase pair, their peptide backbones were brought into a very close structural fit. This overlay was good even in cases such as 1b6c versus 2src (Figure 3A–F), where there are gross differences in the proteins’ structures outside of the kinase active sites.

In a further analysis of the comparison matrix, several cases were found of quantized surface shapes common to more than two kinases. One such quantized surface shape, shown in Figure 4A, was common to the five kinases 1a9u, 1jnk, 1src, 1stc, and 1gag. Rewardingly, the functionality conserved within this quantized surface shape corresponds to two key elements of ATP binding in kinases: A positively charged and polar phosphate binding region, and the critical hydrogen-bond donating region that binds the N1 of adenine [51]. These five proteins happen to have three different co-crystallized ligand types: three are bound with ATP mimics, one is bound with Staurosporine, and one is bound with inhibitor Sb203580. As shown in Figure 4B–D, when the ligands are re-introduced in their co-crystallized frame of reference and their proteins aligned by their common quantized surface, the key H-bond acceptor of all three ligand types is correctly oriented within the conserved quantization. The H-bond acceptor (green) is in each case positioned within the conserved H-bond donating cube (orange).

The comparison matrix of Table 2 shows a number of ‘false positives’; Figure 5 illustrates the highest scoring of these erroneously matched protein pockets, 1b6c and 1dlc. Figure 5A shows the kinase pocket of 1b6c (TGF-beta receptor type I), and Figure 5C shows the pocket of 1dlc (insecticidal delta-endotoxin) which has no kinase activity. Figure 5B shows the spatial overlap of the pockets as aligned by their match-

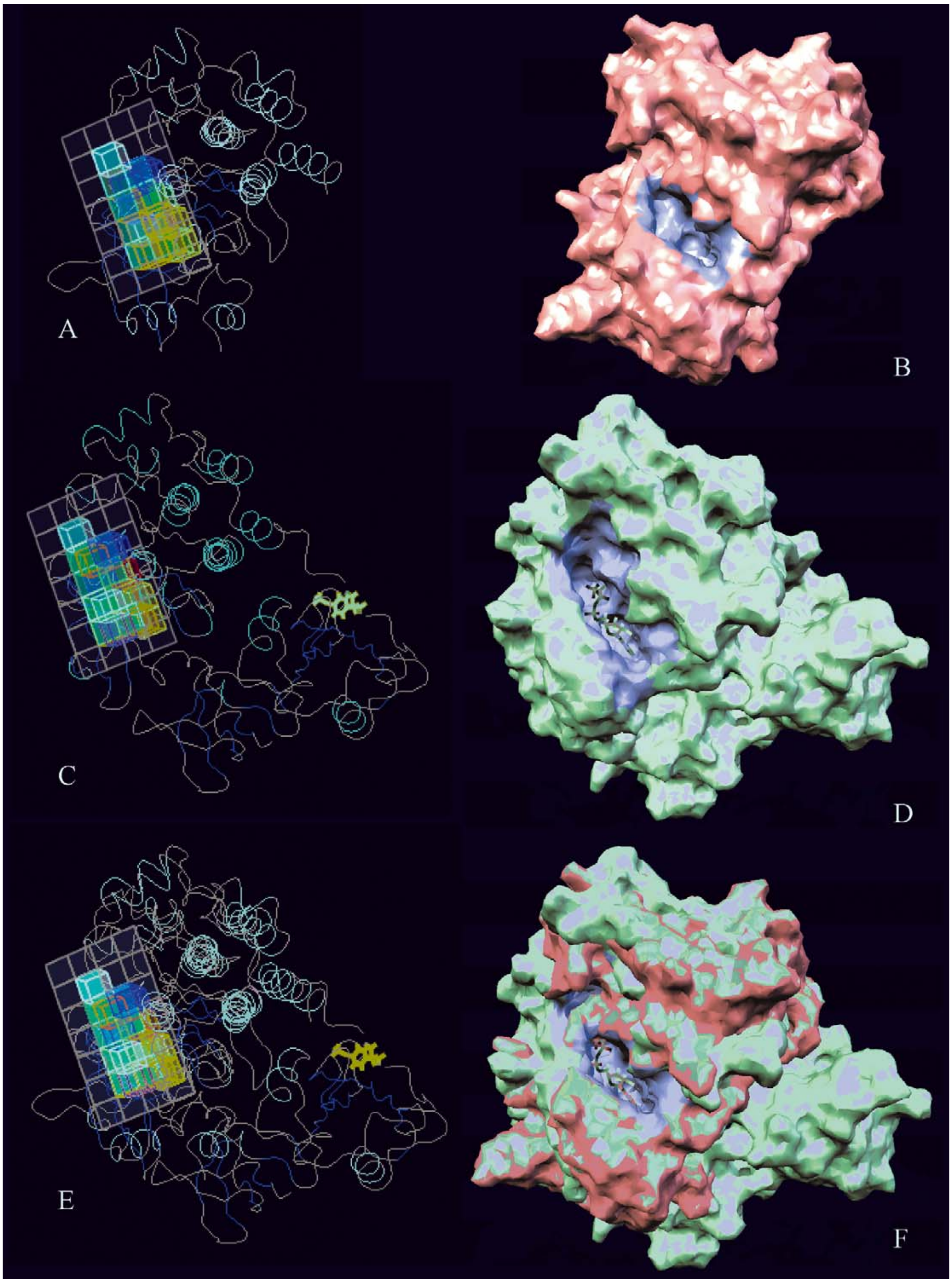
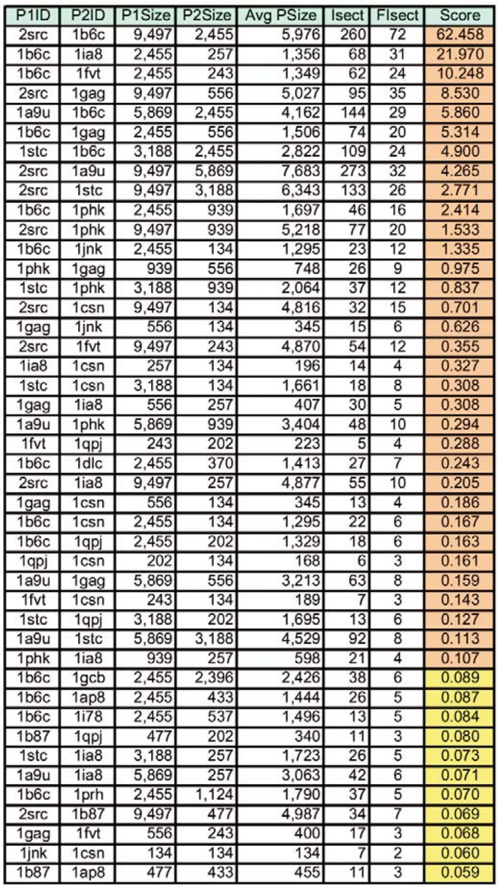
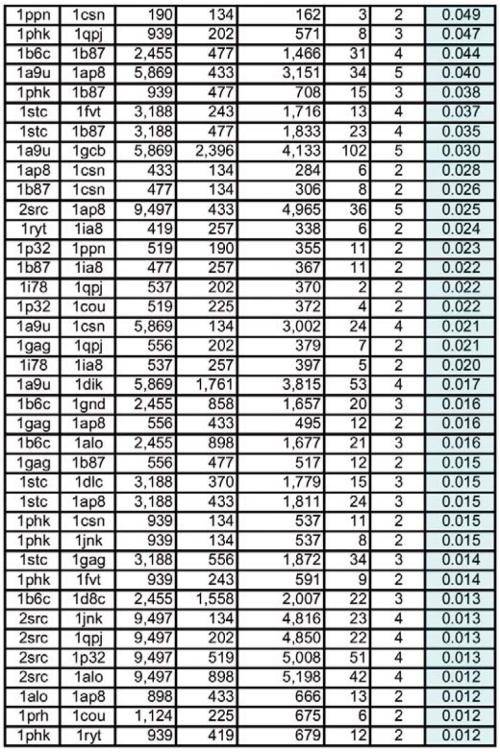


Figure 3. Pairing of kinases 1b6c and 2src as aligned by QSCD (top score in Tables 2 and 3). (A) Backbone of kinase 1b6c and its quantized surface. (B) Surface and pocket of kinase 1b6c. (C) Backbone of kinase 2src and its quantized surface. (D) Surface, pocket, and ligand of kinase 2src. (E) Alignment of matching quantized surfaces brings kinase backbones into good overlay. (F) Alignment of matching quantized surfaces brings catalytic pockets into good overlay.

Table 3. Details of the top pairings of 11 different active site kinase pockets and 19 non-kinase protein pockets as shown in Table 2. PSize = number of quantized surface shapes for a given pocket P, Isect = number of matching quantized surface shapes, FIsect = number of matching quantized surface shapes that have at least six functional cubes identical in type and location, Score = FIsect ˆ3 / ((P1Size + P2Size) / 2).





P1ID	P2ID	P1Size	P2Size	Avg PSize	Isect	FIsect	Score
2src	1b6c	9,497	2,455	5,976	260	72	62.458
1b6c	1ia8	2,455	257	1,356	68	31	21.970
1b6c	1fvt	2,455	243	1,349	62	24	10.248
2src	1gag	9,497	556	5,027	95	35	8.530
1a9u	1b6c	5,869	2,455	4,162	144	29	5.860
1b6c	1gag	2,455	556	1,506	74	20	5.314
1stc	1b6c	3,188	2,455	2,822	109	24	4.900
2src	1a9u	9,497	5,869	7,683	273	32	4.265
2src	1stc	9,497	3,188	6,343	133	26	2.771
1b6c	1phk	2,455	939	1,697	46	16	2.414
2src	1phk	9,497	939	5,218	77	20	1.533
1b6c	1jnk	2,455	134	1,295	23	12	1.335
1phk	1gag	939	556	748	26	9	0.975
1stc	1phk	3,188	939	2,064	37	12	0.837
2src	1csn	9,497	134	4,816	32	15	0.701
1gag	1jnk	556	134	345	15	6	0.626
2src	1fvt	9,497	243	4,870	54	12	0.355
1ia8	1csn	257	134	196	14	4	0.327
1stc	1csn	3,188	134	1,661	18	8	0.308
1gag	1ia8	556	257	407	30	5	0.308
1a9u	1phk	5,869	939	3,404	48	10	0.294
1fvt	1qpj	243	202	223	5	4	0.288
1b6c	1dlc	2,455	370	1,413	27	7	0.243
2src	1ia8	9,497	257	4,877	55	10	0.205
1gag	1csn	556	134	345	13	4	0.186
1b6c	1csn	2,455	134	1,295	22	6	0.167
1b6c	1qpj	2,455	202	1,329	18	6	0.163
1qpj	1csn	202	134	168	6	3	0.161
1a9u	1gag	5,869	556	3,213	63	8	0.159
1fvt	1csn	243	134	189	7	3	0.143
1stc	1qpj	3,188	202	1,695	13	6	0.127
1a9u	1stc	5,869	3,188	4,529	92	8	0.113
1phk	1ia8	939	257	598	21	4	0.107
1b6c	1gcb	2,455	2,396	2,426	38	6	0.089
1b6c	1ap8	2,455	433	1,444	26	5	0.087
1b6c	1i78	2,455	537	1,496	13	5	0.084
1b87	1qpj	477	202	340	11	3	0.080
1stc	1ia8	3,188	257	1,723	26	5	0.073
1a9u	1ia8	5,869	257	3,063	42	6	0.071
1b6c	1prh	2,455	1,124	1,790	37	5	0.070
2src	1b87	9,497	477	4,987	34	7	0.069
1gag	1fvt	556	243	400	17	3	0.068
1jnk	1csn	134	134	134	7	2	0.060
1b87	1ap8	477	433	455	11	3	0.059
1ppn	1csn	190	134	162	3	2	0.049
1phk	1qpj	939	202	571	8	3	0.047
1b6c	1b87	2,455	477	1,466	31	4	0.044
1a9u	1ap8	5,869	433	3,151	34	5	0.040
1phk	1b87	939	477	708	15	3	0.038
1stc	1fvt	3,188	243	1,716	13	4	0.037
1stc	1b87	3,188	477	1,833	23	4	0.035
1a9u	1gcb	5,869	2,396	4,133	102	5	0.030
1ap8	1csn	433	134	284	6	2	0.028
1b87	1csn	477	134	306	8	2	0.026
2src	1ap8	9,497	433	4,965	36	5	0.025
1ryt	1ia8	419	257	338	6	2	0.024
1p32	1ppn	519	190	355	11	2	0.023
1b87	1ia8	477	257	367	11	2	0.022
1i78	1qpj	537	202	370	2	2	0.022
1p32	1cou	519	225	372	4	2	0.022
1a9u	1csn	5,869	134	3,002	24	4	0.021
1gag	1qpj	556	202	379	7	2	0.021
1i78	1ia8	537	257	397	5	2	0.020
1a9u	1dik	5,869	1,761	3,815	53	4	0.017
1b6c	1gnd	2,455	858	1,657	20	3	0.016
1gag	1ap8	556	433	495	12	2	0.016
1b6c	1alo	2,455	898	1,677	21	3	0.016
1gag	1b87	556	477	517	12	2	0.015
1stc	1dlc	3,188	370	1,779	15	3	0.015
1stc	1ap8	3,188	433	1,811	24	3	0.015
1phk	1csn	939	134	537	11	2	0.015
1phk	1jnk	939	134	537	8	2	0.015
1stc	1gag	3,188	556	1,872	34	3	0.014
1phk	1fvt	939	243	591	9	2	0.014
1b6c	1d8c	2,455	1,558	2,007	22	3	0.013
2src	1jnk	9,497	134	4,816	23	4	0.013
2src	1qpj	9,497	202	4,850	22	4	0.013
2src	1p32	9,497	519	5,008	51	4	0.013
2src	1alo	9,497	898	5,198	42	4	0.012
1alo	1ap8	898	433	666	13	2	0.012
1prh	1cou	1,124	225	675	6	2	0.012
1phk	1ryt	939	419	679	12	2	0.012

ing quantizations, and Figure 5D shows the matching quantization with functional overlaps denoted by color. While this paring of proteins must be classified as a false positive based on their known activities, there *is* a similarity of shape and functionality between these two pockets that is recognized by the dissection-quantization algorithm at 4 Å resolution. The value of such a categorization lies in its potential application to lead discovery; one could make the case, for example, that both targets should be screened against similar combinatorial libraries.

Comparison to PID matrix

In total, the results of the kinase/non-kinase experiment confirm the validity of analyzing protein pockets by mapping their dissected binding volumes to a basis set of theoretical target surfaces. For purposes of comparison, the same experiment was performed using established percent identity (PID) software, which aligns proteins by sequence and secondary structure and compares them by the percent similarity of their residue types at each amino acid position (see details in Experimental). The results of this experiment are given in Table 4 as a matrix of percent identities between protein pairings.

Analysis of the results shown in Tables 2 and 4 reveals that the QSCD dissection and mapping method compares favorably with the sequence percent identity method. While the PID algorithm has a lower signal to noise ratio, both methods clearly group kinases as being more similar to each other than they are to random proteins. The methods are also complementary: the 1D/2D PID algorithm is good at scoring kinase 1jnk against 8 of 10 other kinases despite a differentiating peptide loop across 1jnk's ATP phosphate binding region, whereas the latter active site variation in 1jnk gives the protein a low QSCD score against 7 of 10

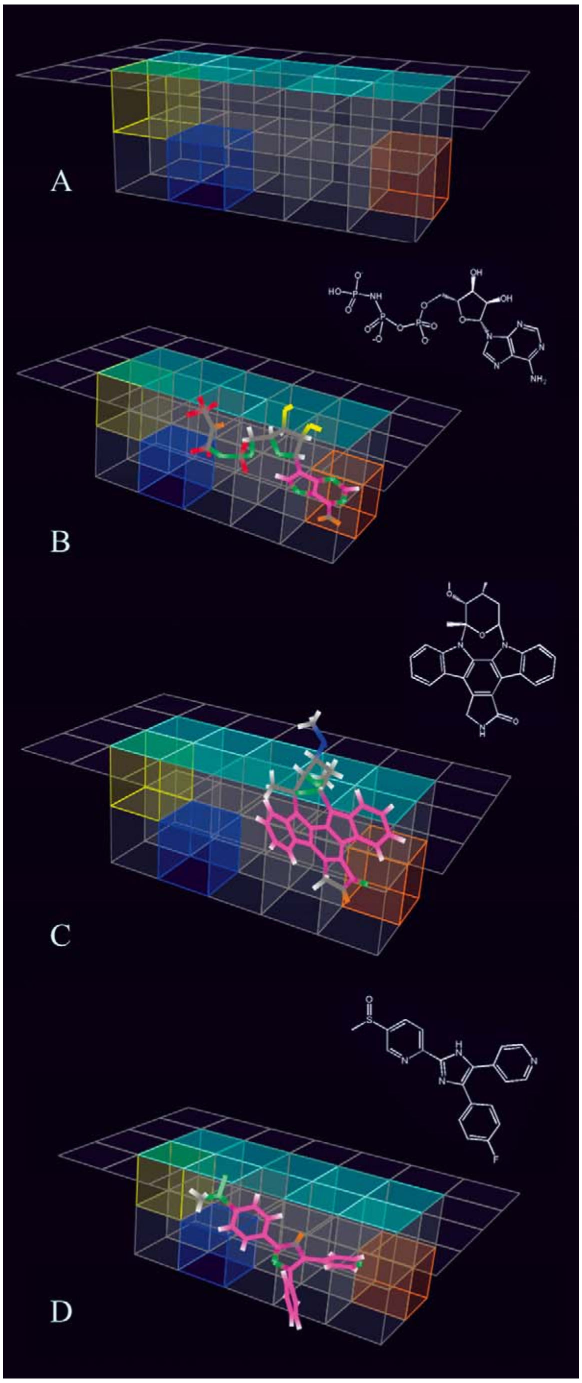


Figure 4. (A) A quantized surface shape common to the five kinases 1a9u, 1jnk, 1src, 1stc, and 1gag, with conserved functionality denoted in color. Orange = hydrogen-bond donating region, blue = positively charged region, yellow = hydrogen-bond donor/acceptor region. (B,C,D) Co-crystallized ATP mimic, Staurosporine, and inhibitor Sb203580 in the same frame of reference as the quantized surface. Ligand coloring: red = negatively charged, blue = positively charged, orange = H-bond donor, green = H-bond acceptor, yellow = H-bond donor/acceptor, magenta = polarizable, white = hydrophobic.

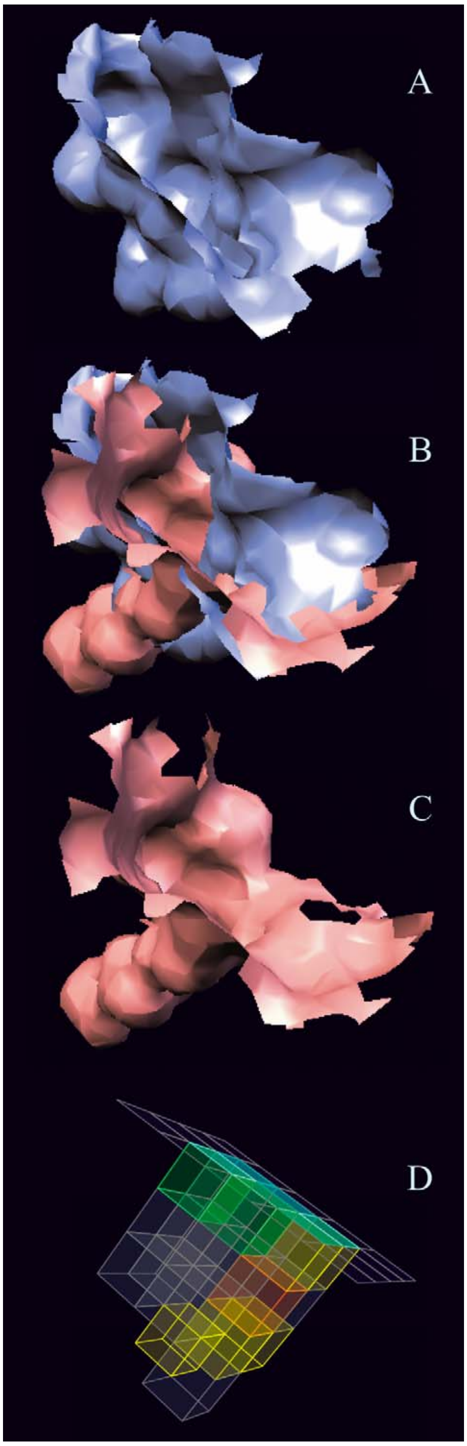


Figure 5. (A) The kinase pocket of 1b6c (TGF-beta receptor type I). (B) The spatial overlap of the pockets of 1b6c and 1dlc as aligned by their matching quantized surfaces. (C) The pocket of 1dlc (Insecticidal delta-endotoxin). (D) The matching quantization and its functional overlaps: green = HB-acceptor; orange = HB-donor; yellow = HB-donor/acceptor.

Table 4. PID scores of each pairing of 11 different active site kinase pockets and 19 non-kinase protein pockets. Scoring is based on percent sequence identities between protein pairings. Scores are colored to show value according to the key.

	Kinase Proteins											Other Proteins Selected Randomly from the PDB																		
	1a9u	1b6c	1csn	1fvt	1gag	1ia8	1jnk	1phk	1qpj	1stc	2src	1alo	1ap8	1b87	1cou	1csj	1d8c	1dik	1dlc	1ex1	1gcb	1gnd	1i78	1i9b	1kuh	1ndo	1p32	1ppn	1prh	1ryt
1a9u	--	0.23	0.14	0.39	0.22	0.23	0.52	0.28	0.25	0.24	0.19	0.03	0.03	0.07	0.03	0.01	0.03	0.04	0.03	0.03	0.05	0.03	0.07	0.04	0.08	0.11	0.14	0.03	0.04	0.04
1b6c	0.23	--	0.21	0.28	0.36	0.33	0.23	0.23	0.33	0.20	0.25	0.07	0.05	0.12	0.07	0.02	0.03	0.02	0.04	0.03	0.05	0.08	0.04	0.02	0.07	0.05	0.06	0.04	0.02	0.08
1csn	0.14	0.21	--	0.26	0.24	0.21	0.15	0.19	0.24	0.19	0.15	0.03	0.05	0.06	0.04	0.03	0.03	0.03	0.06	0.05	0.04	0.06	0.09	0.04	0.02	0.04	0.06	0.04	0.11	0.03
1fvt	0.39	0.28	0.26	--	0.29	0.27	0.43	0.38	0.31	0.23	0.21	0.02	0.10	0.05	0.03	0.03	0.03	0.04	0.03	0.02	0.15	0.05	0.02	0.05	0.04	0.08	0.06	0.03	0.04	0.12
1gag	0.22	0.35	0.24	0.29	--	0.24	0.27	0.24	0.41	0.24	0.31	0.03	0.04	0.07	0.06	0.02	0.02	0.04	0.03	0.06	0.03	0.02	0.03	0.08	0.05	0.03	0.08	0.03	0.03	0.04
1ia8	0.25	0.33	0.21	0.27	0.24	--	0.21	0.28	0.23	0.29	0.20	0.02	0.02	0.03	0.05	0.02	0.02	0.06	0.05	0.06	0.05	0.11	0.07	0.03	0.10	0.11	0.05	0.03	0.02	0.05
1jnk	0.52	0.23	0.14	0.43	0.27	0.21	--	0.32	0.24	0.21	0.18	0.03	0.03	0.12	0.05	0.05	0.03	0.02	0.02	0.01	0.16	0.12	0.04	0.06	0.04	0.07	0.04	0.03	0.08	0.02
1phk	0.29	0.23	0.19	0.38	0.24	0.28	0.32	--	0.19	0.31	0.17	0.03	0.02	0.03	0.04	0.03	0.03	0.01	0.04	0.04	0.05	0.05	0.04	0.04	0.09	0.06	0.04	0.02	0.03	0.03
1qpj	0.24	0.33	0.25	0.31	0.41	0.23	0.24	0.19	--	0.21	0.50	0.03	0.04	0.03	0.03	0.05	0.08	0.07	0.04	0.06	0.08	0.04	0.06	0.04	0.04	0.03	0.03	0.14	0.06	0.04
1stc	0.24	0.20	0.19	0.23	0.22	0.29	0.21	0.31	0.21	--	0.17	0.05	0.02	0.05	0.02	0.02	0.02	0.02	0.04	0.02	0.07	0.12	0.02	0.03	0.03	0.03	0.04	0.07	0.02	0.06
2src	0.19	0.26	0.15	0.21	0.31	0.20	0.18	0.18	0.50	0.17	--	0.04	0.03	0.10	0.01	0.01	0.03	0.07	0.04	0.03	0.03	0.04	0.09	0.09	0.04	0.04	0.03	0.07	0.07	0.04
1alo	0.03	0.07	0.03	0.02	0.03	0.02	0.03	0.03	0.03	0.05	0.04	--	0.04	0.02	0.02	0.02	0.01	0.04	0.02	0.04	0.04	0.05	0.03	0.03	0.04	0.03	0.02	0.02	0.03	0.03
1ap8	0.03	0.05	0.05	0.10	0.04	0.02	0.03	0.02	0.04	0.02	0.03	0.04	--	0.05	0.03	0.07	0.05	0.02	0.04	0.02	0.06	0.07	0.04	0.08	0.04	0.04	0.13	0.03	0.03	0.09
1b87	0.07	0.12	0.06	0.05	0.07	0.03	0.12	0.03	0.03	0.05	0.10	0.02	0.05	--	0.05	0.06	0.02	0.08	0.06	0.10	0.05	0.01	0.10	0.07	0.07	0.06	0.08	0.03	0.05	0.08
1cou	0.03	0.07	0.04	0.03	0.06	0.05	0.05	0.04	0.03	0.02	0.01	0.02	0.03	0.05	--	0.02	0.02	0.01	0.03	0.02	0.02	0.05	0.06	0.09	0.04	0.03	0.05	0.07	0.04	0.03
1csj	0.01	0.02	0.03	0.03	0.02	0.02	0.05	0.03	0.05	0.02	0.01	0.02	0.07	0.06	0.02	--	0.04	0.04	0.01	0.05	0.07	0.03	0.07	0.08	0.03	0.02	0.03	0.03	0.04	0.02
1d8c	0.03	0.03	0.03	0.03	0.02	0.02	0.03	0.03	0.08	0.02	0.03	0.01	0.05	0.02	0.02	0.04	--	0.03	0.03	0.03	0.02	0.03	0.03	0.02	0.03	0.02	0.05	0.08	0.03	0.02
1dik	0.04	0.02	0.03	0.04	0.04	0.06	0.02	0.01	0.07	0.02	0.07	0.04	0.02	0.08	0.01	0.04	0.03	--	0.02	0.06	0.04	0.01	0.03	0.04	0.04	0.04	0.06	0.02	0.02	0.09
1dlc	0.03	0.04	0.06	0.03	0.03	0.05	0.03	0.04	0.04	0.04	0.04	0.02	0.04	0.06	0.03	0.01	0.03	0.04	--	0.06	0.06	0.04	0.13	0.02	0.03	0.08	0.08	0.02	0.09	0.02
1ex1	0.03	0.03	0.05	0.02	0.06	0.06	0.01	0.04	0.06	0.02	0.03	0.04	0.02	0.10	0.02	0.02	0.04	0.06	0.06	--	0.07	0.02	0.04	0.02	0.06	0.05	0.04	0.07	0.04	0.04
1gcb	0.05	0.05	0.04	0.15	0.03	0.05	0.16	0.05	0.08	0.07	0.03	0.04	0.06	0.05	0.02	0.07	0.02	0.04	0.06	0.07	--	0.01	0.07	0.04	0.03	0.01	0.05	0.04	0.01	0.07
1gnd	0.03	0.08	0.06	0.05	0.02	0.11	0.12	0.05	0.04	0.12	0.04	0.05	0.07	0.01	0.05	0.03	0.03	0.01	0.04	0.02	0.01	--	0.02	0.06	0.04	0.04	0.03	0.05	0.14	0.04
1i78	0.07	0.04	0.09	0.02	0.03	0.07	0.03	0.04	0.06	0.02	0.09	0.03	0.04	0.10	0.06	0.07	0.03	0.03	0.13	0.04	0.07	0.02	--	0.08	0.08	0.09	0.02	0.07	0.12	0.05
1i9b	0.04	0.02	0.04	0.05	0.08	0.03	0.06	0.04	0.04	0.03	0.09	0.03	0.08	0.07	0.09	0.08	0.02	0.04	0.02	0.02	0.04	0.06	0.08	--	0.08	0.06	0.04	0.04	0.03	0.05
1kuh	0.08	0.07	0.02	0.04	0.05	0.10	0.02	0.08	0.04	0.03	0.04	0.04	0.04	0.07	0.04	0.03	0.03	0.04	0.03	0.06	0.03	0.04	0.08	0.08	--	0.08	0.05	0.11	0.04	0.05
1ndo	0.11	0.05	0.04	0.08	0.03	0.11	0.07	0.06	0.03	0.01	0.04	0.03	0.04	0.06	0.03	0.02	0.02	0.05	0.08	0.05	0.01	0.04	0.09	0.06	0.08	--	0.13	0.04	0.04	0.03
1p32	0.14	0.06	0.06	0.06	0.08	0.05	0.04	0.04	0.03	0.04	0.03	0.02	0.13	0.08	0.05	0.02	0.05	0.06	0.08	0.04	0.05	0.03	0.02	0.04	0.05	0.12	--	0.04	0.04	0.11
1ppn	0.03	0.04	0.04	0.03	0.03	0.08	0.03	0.02	0.14	0.07	0.07	0.02	0.03	0.03	0.07	0.03	0.08	0.02	0.02	0.07	0.04	0.05	0.07	0.04	0.11	0.04	0.04	--	0.05	0.21
1prh	0.04	0.06	0.11	0.04	0.03	0.02	0.08	0.03	0.06	0.02	0.07	0.03	0.03	0.05	0.04	0.04	0.03	0.02	0.09	0.04	0.01	0.14	0.12	0.03	0.04	0.04	0.04	0.05	--	0.05
1ryt	0.04	0.08	0.03	0.12	0.04	0.05	0.02	0.02	0.04	0.06	0.04	0.03	0.09	0.08	0.03	0.02	0.02	0.08	0.02	0.03	0.07	0.05	0.05	0.05	0.05	0.03	0.11	0.21	0.05	--

Score Coloring Key:

0	to	0.1
0.1	to	0.15
0.15	to	0.2
0.2	and higher	

Score Summary:

Type	Total	#	Avg Score
Kin-Kin	28.67	110	0.261
Kin-Oth	19.88	418	0.048
Oth-Oth	16.16	342	0.047
Non-Kin	36.04	760	0.047

Signal to Noise:
5.50

other kinases. Conversely, 2src is well recognized as a kinase by the 3D dissection algorithm because of its characteristic pocket, but the PID algorithm is not as good at scoring 2src against other kinases because 2src contains a large additional domain. In short, the QSCD method has the benefit of looking at 3D data, and will thus have a better chance to see pockets that are similar in shape and functionality but different in sequence and/or secondary structure. The PID method, looking only at 1D/2D data, will recognize conserved motifs even in the presence of structural variation at the active site.

Relevance of 'Potential Binding Volumes'

Having established that the process of dissecting protein pockets into their potential binding volumes is a valid method of comparing protein surfaces, we wished to confirm that those binding volumes do in fact encompass the binding modes of known ligands. Thus, we tested the process on 20 co-crystal structures, including a range of varied protein types and several structures of the same protein type with varying ligands. We included our three training proteins as a benchmark. Ligands were removed from all 20 co-crystal structures and their relevant protein pockets were dissected as above. The resulting potential binding volumes were mapped to quantized theoretical surfaces as detailed in the Methods section. Meanwhile, as described in the Introduction and in our previous communication [1], the removed ligands were quantized and mapped by complementarity to the same basis set of theoretical surfaces. The quantized surfaces of each protein's binding volumes were then compared to those of their respective small molecule ligands. If the set of cubes for a ligand either (a) exactly matched the set of cubes of a binding volume, or (b) was completely contained within the set of cubes of a binding volume with the binding volume having no more than three unfilled cubes, then the cube-aligned orientation of the quantized ligand was analyzed with respect to the known orientation of the ligand within the given co-crystal structure. The predicted orientations were said to be valid if they had an RMS of less than 4.0 Å compared to their actual bound ligand orientations, an RMS of 4 Å being the intrinsic margin of error for comparisons using 4 Å cubes. If this were a docking study, 4.0 Å would be a generous RMS allowance; however, the aim of this experiment, as noted below, is not to create a docking algorithm, but rather to validate the proposed method of generating potential binding volumes.

As shown in Table 5, in all cases at least one quantized binding volume produced by the dissection algorithm aligned the ligand to its actual binding mode within an RMS of 4.0 Å. In most cases, due to our above allowance of three unfilled cubes, multiple quantized binding volumes properly encompassed the quantized ligand. It should be noted that this procedure is *not* proposed as a good method of molecular 'docking' of ligands to proteins on an individual scale; the experiment was designed only to show that the potential binding volumes predicted by our dissection algorithm *included* the actual binding volumes of known ligands. The dissection algorithm may, however, be used as a first-pass filter where protein–ligand analysis is concerned; if no quantized conformations of a molecule fit any of a pocket's quantized binding volumes, the molecule is unlikely to have the necessary shape for binding in the given pocket. While such information does not predict binding, it may allow a large library of molecules to be narrowed prior to screening against a target or target family.

Conclusions

This paper extends the method of Quantized Surface Complementarity Diversity, or QSCD, to the analysis of actual protein pockets. By dissecting protein pockets into their potential binding volumes and mapping the results to a single basis set of theoretical target surfaces, proteins may be compared in three dimensions on a large scale. Efforts are currently underway to use the algorithms detailed herein to map the pockets of all available protein crystals structures, thus allowing proteins to be categorized based purely on the shape and functionality of their pockets.

Furthermore, we have demonstrated that the potential binding volumes which result from the dissection algorithm successfully encompass the binding modes of known small-molecule ligands. This suggests that by mapping both small-molecules and actual protein pockets to the same diversity space on a large scale, we may be able to develop QSCD as a first measure of 'biorelevance': the likelihood of a small molecule to have properties which will allow it to bind a given set of natural proteins. This is a measure based on properties of molecular recognition, and thus should be very complementary to measures such as molecular weight, Log P, and number of rotatable bonds, which predict

Table 5. Quantizations of potential binding volumes for 20 proteins compared to quantizations of respective known ligands. The matching procedure is detailed in the text. Predicted ligand orientations are said to be valid if they had an RMS of less than 4.0 Å compared to the actual bound ligand orientations. Starred proteins were used in training runs and are thus expected to have valid matches.

PDB code of co-crystal structure	Ref.	Protein	Ligand	Number of quantized binding volumes	Number of valid matches to ligand quantizations at 4.0 Å RMS
1cbs	52	cellular retinoic acid binding protein	all trans retinoic acid	1,701	59
1srj	53	streptavidin	2-((4′-hydroxynaphthyl)-azo)benzoate	366	37
2qwk	54	influenza virus neuraminidase	4-(N-acetylamino)-5-amino-3-(1 ethylpropoxy)-1-cyclohexene-1-carboxylic acid	2,192	29
11bd	55	nuclear hormone receptor	9-cis retinoic acid	1,515	28
1cqe	56	cyclooxygenase-1	flurbiprofen	14,925	27
1epb	57	epididymal retinoic acid binding protein	all trans retinoic acid	2,692	17
1dre*	18	dihydrofolate reductase	methotrexate	505	17
1rud	58	rhinovirus 14 coat protein	5-(7 (5-hydro-4-methyl-2-oxazolyl)phenoxy)heptyl)-3-methyl isoxazole	2,401	12
1tlp	59	thermolysin (metalloprotease)	phosphoramidon	5,347	12
3aig	60	adamalysin II (metalloprotease)	peptidomimetic inhibitor	123	9
1yet	61	HSP-90 (chaperone protein)	geldanamycin	8,200	8
1bx6	62	cAMP-dependent protein kinase	balanol	4,311	7
1qkt*	20	estrogen receptor	estradiol	753	6
1dyr	63	dihydrofolate reductase	trimethoprim	335	6
1b9s	64	influenza virus neuraminidase	4-(N-acetylamino)-3-[N-(2-ethylbutanoylamino)]benzoic acid	1,223	5
1dwd	65	thrombin (serine protease)	N-(2-naphthyl-sulfonyl-glycyl)-aipha-(para-benzamidyl)-alanyl-piperidine	4,783	4
1ajv	66	HIV-1 aspartyl protease	cyclic sulfamide inhibitor	27	3
2src	31	protein kinase C-Src	phosphoaminophosphonic acid adenylate ester	9,497	2
1hck*	19	cyclin-dependent kinase 2	ATP	2,621	2
2wea	67	penicillopepsin (aspartyl-protease)	macrocyclic phosphonate inhibitor	3,165	1

pharmaceutical relevance based on a molecule's bulk physical properties.

Experimental

Proteins were downloaded from NCBI's Molecular Modeling DataBase (MMDB, (http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml). Sybyl software version 6.8 (Tripos Inc., St. Louis, MO) on an R10000 Silicon Graphics workstation was used to remove water molecules and other non-covalent artifacts of co-crystallization. Proteins were saved in mol2 format as output by Sybyl. Any desired co-crystallized ligands were saved as separate mol2 files. All in-house software was developed for Intel-based workstations using the JAVA programming language (JDK 1.4.1) and the Java3D graphics API (version 1.3).

Surface generation

We used proteins in a mol2 format and generated simulated H_2O contact surfaces by rolling a spherical probe of radius 1.8 Å over the VDW surface of the protein. The set of points at which the probe generated tangential contacts to the VDW surface of the protein was taken as the simulated H_2O contact surface.

Pocket determination

Our internal 'slicing algorithm' used to detect concave pockets on a protein surfaces is as follows:
(1) Create a set of 1024 tessellated axes radiating from the center of the protein.
(2) For each axis, at every 1.5 Å interval along the axis, slice the protein with a plane that is perpendicular to the axis.
(3) If the intersection of the slicing plane and the protein surface defines a closed loop A which is further surrounded by another closed loop B, then loop A defines on the protein surface the edge of either a concave pocket or a tunnel.
(4) If the intersection of the slicing plane and the protein surface defines two closed loops C and D, neither of which is enclosed by any other loop, then attempt to join C and D with a pair of lines j and k, each of which have one endpoint on C, one endpoint on D, are not greater than 10 Å in length, and have their centers maximally distant from one another.
(5) If j and k exist under the specifications in step 4, test to see if a plane perpendicular to the slicing plane and containing j has an intersection with the protein surface to give a curve j′ that runs from the start of j to the end of j, and test to see if the area enclosed by j and j′ is less than 50 Å^2. Further test to see if a plane perpendicular to the slicing plane and containing k has an intersection with the protein surface to give a curve k′ that runs from the start of k to the end of k, and test to see if the area enclosed by k and k′ is less than 50 Å^2.
(6) If j and k pass the test of step 5, then curves j′ and k′ together with the inner portions of loops C and D define on the protein surface the edge of either a concave pocket or a tunnel.
(7) All concave pockets from steps 3 and 6 are allowed if their volume is between 400 and 2300 Å^3.
(8) All tunnels from steps 3 and 6 are 'capped' at their far ends by a planar surface that is as far from the slicing plane as possible without exceeding an area of 80 Å^2. If no such cap exists, then the tunnel is discarded.
(9) All capped tunnels from step 8 are allowed as concave pockets if their volume is between 400 and 2300 Å^3.

Filling of pockets with balls prior to dissection

Concave pocket surfaces of a protein were filled with a 1.65 Å cubic lattice of balls with a radius of 1.3 Å. The lattice was oriented using the origin and principal axes of the set of all atom centers of the given protein. Balls were discarded if their centers did not lie within the volume bounded by the concave pocket surface and the opening plane of the concave pocket surface. Balls were also discarded if their radii protruded more than 0.7 Å beyond the van Der Waals radius of any protein atom.

PID calculation

To create the percent identity (PID) matrix of Table 4, protein alignments and the number of matching residues in the aligned sequences were generated using 'ssearch33'. This program is an implementation of the Smith–Waterman sequence alignment algorithm [68] and is distributed as part of the well established 'FASTA' alignment package [69]. The PID scores in Table 4 were calculated as follows for each pair of proteins:
PID = #MatchingResiduesInAlignedSequences / ((ProteinLengthA + ProteinLengthB) / 2)

The value of #MatchingResiduesInAlignedSequences in the formula above was calculated by multiplying the 'FASTA' output, which is the percentage of matching residues in the aligned segments of the two sequences, by the number of residues that are aligned. Note that the matrix of Table 4 is not perfectly symmetrical, as the protein alignments of 'fasta' are based on the seed sequence. Thus, the alignment of A with B may differ slightly from the alignment of B with A.

References

1. Wintner, E.A. and Moallemi, C.C., J. Med. Chem., 43 (2000) 1993.
2. Martin, Y.C., J. Comb. Chem., 3 (2001) 1.
3. Bures, M.G. and Martin, Y.C., Curr. Opin. Chem. Biol., 2 (1998) 376.
4. Matter, H., Modern Methods Drug Discov., 93 (2003) 125.
5. Rehm, B.H.A., Appl. Microbiol. Biotechnol., 57 (2001) 579.
6. Xu, D., Xu, Y. and Uberbacher, E.C., Curr. Protein Pept. Sci., 1 (2000) 1.
7. Naumann, T. and Matter, H., J. Med. Chem., 45 (2002) 2366.
8. Swindells, M.B., Orengo, C.A., Jones, D.T., Hutchinson, E.G. and Thornton, J.M., Bioessays, 11 (1998) 884.
9. Hegyi, H. and Gerstein, M., J. Mol. Biol., 288 (1999) 147.
10. Holm, L. and Sander, C., Nucleic Acids Res., 26 (1998) 68.
11. Dawe, J.H., Porter, C.T., Thornton, J.M. and Tabor, A.B., Proteins, 52 (2003) 427.
12. Wallace, A.C., Laskowski, R.A. and Thornton, J.M., Protein Sci., 5 (1996) 1001.
13. Fetrow, J.S. and Skolnick, J., J. Mol. Biol., 281 (1998) 949.
14. Brady, G.P. Jr and Stouten, P.F., J. Comput.-Aided Mol. Des., 14 (2000) 383.
15. Liang, J., Edelsbrunner, H. and Woodward, C., Protein Sci., 7 (1998) 1884.
16. Ruppert, J., Welch, W. and Jain, A.N., Protein Sci., 6 (1997) 524.
17. Peters, K.P., Faulk, J. and Frommel, C., J. Mol. Biol., 256 (1996) 201.
18. Sawaya, M.R. and Kraut, J., Biochemistry, 36 (1997) 586.
19. Schulze-Gahmen, U., De Bondt, H.L. and Kim, S.H., J. Med. Chem., 39 (1996) 4540.
20. Ruff, M., Gangloff, M., Eiler, S., Duclaud, S., Wurtz, J.M. and Dino, M., to be published.
21. Wang, Z., Canagarajah, B.J., Boehm, J.C., Kassisa, S., Cobb, M.H., Young, P.R., Abdel-Meguid, S., Adams, J.L. and Goldsmith, E.J., Structure, 6 (1998) 1117.
22. Huse, M., Chen, Y.G., Massague, J. and Kuriyan, J., Cell, 96 (1999) 425.
23. Xu, R.M., Carmel, G., Sweet, R.M., Kuret, J. and Cheng, X., Embo J., 14 (1995) 1015.
24. Davis, S.T., Benson, B.G., Bramson, H.N., Chapman, D.E., Dickerson, S.H., Dold, K.M., Eberwein, D.J., Edelstein, M., Frye, S.V., Gampe Jr., R.T., Griffin, R.J., Harris, P.A., Hassell, A.M., Holmes, W.D., Hunter, R.N., Knick, V.B., Lackey, K., Lovejoy, B., Luzzio, M. J., Murray, D., Parker, P., Rocque, W.J. and Shewch, L., Science, 291 (2001) 134.
25. Parang, K., Till, J.H., Ablooglu, A.J., Kohanski, R.A., Hubbard, S.R. and Cole, P.A., Nat. Struct. Biol., 8 (2001) 37.
26. Chen, P., Luo, C., Deng, Y., Ryan, K., Register, J., Margosiak, S., Tempczyk-Russell, A., Nguyen, B., Myers, P., Lundgren, K., Chen Kan, C.-C. and O'Connor, P.M., Cell (Cambridge, MA), 100 (2000) 681.
27. Xie, X., Gu, Y., Fox, T., Coll, J.T., Fleming, M.A., Markland, W., Caron, P.R., Wilson, K.P. and Su, M.S., Structure, 6 (1998) 983.
28. Owen, D.J., Noble, M.E., Garman, E.F., Papageorgiou, A.C. and Johnson, L.N., Structure, 3 (1995) 467.
29. Zhu, X., Kim, J.L., Rose, P.E., Stover, D.R. and Toledo, L.M., Structure London, 7 (1999) 651.
30. Prade, L., Engh, R.A., Girod, A., Kinzel, V., Huber, R. and Bossemeyer, D., Structure, 5 (1997) 1627.
31. Xu, W., Doshi, A., Lei, M., Eck, M.J. and Harrison, S.C., Mol. Cell, 3 (1999) 629.
32. Romao M.J., Archer M., Moura I., Moura J.J., LeGall J., Engh R., Schneider M., Hof P. and Huber R., Science, 270 (1995) 1170.
33. Matsuo, H., Li, H., McGuire, A.M., Fletcher, C.M., Gingras, A.C., Sonenberg, N. and Wagner, G., Nat. Struct. Biol., 4 (1997) 717.
34. Wybenga-Groot, L.E., Draker, K., Wright, G.D. and Berghuis, A.M., Structure London, 7 (1999) 497.
35. Duggan, B.M., Dyson, H.J. and Wright, P.E., Eur. J. Biochem., 265 (1999) 539.
36. Bressanelli, S., Tomei, L., Roussel, A., Incitti, I., Vitale, R.L., Mathieu, M., De Francesco, R. and Rey, F.A., Proc. Natl. Acad. Sci. USA, 96 (1999) 13034.
37. Howard, B.R., Endrizzi, J.A. and Remington, S.J., Biochemistry, 39 (2000) 3156.
38. Herzberg, O., Chen, C.C., Kapadia, G., McGuire, M., Carroll, L.J., Noh, S.J. and Dunaway-Mariano, D., Proc. Natl. Acad. Sci. USA, 93 (1996) 2652.
39. Li, J.D., Carroll, J. and Ellar, D.J., Nature, 353 (1991) 815.
40. Varghese, J.N., Hrmova, M. and Fincher, G.B., Structure London, 7 (1999) 179.
41. Joshua-Tor, L., Xu, H.E., Johnston, S.A. and Rees, D.C., Science, 269 (1995) 945.
42. Schalk, I., Zeng, K., Wu, S.K., Stura, E.A., Matteson, J., Huang, M., Tandon, A., Wilson, I.A. and Balch, W.E., Nature, 381 (1996) 42.
43. Vandeputte-Rutten, L., Kramer, R.A., Kroon, J., Dekker, N., Egmond, M.R. and Gros, P., Embo J., 20 (2001) 5033.
44. Brejc, K., Van Dijk, W.J., Klaassen, R.V., Schuurmans, M., Van Der Oost, J., Smit, A.B. and Sixma, T.K., Nature, 411 (2001) 269.
45. Kurisu, G., Kinoshita, T., Sugimoto, A., Nagara, A., Kai, Y., Kasai, N. and Harada, S., J. Biochem. Tokyo, 121 (1997) 304.
46. Kauppi, B., Lee, K., Carredano, E., Parales, R.E., Gibson, D.T., Eklund, H. and Ramaswamy, S., Structure, 6 (1998) 571.
47. Jiang, J., Zhang, Y., Krainer, A.R. and Xu, R.M., Proc. Natl. Acad. Sci. USA, 96 (1999) 3572.
48. Harris, G.W., Pickersgill R.W., Howlin, B. and Moss D.S., Acta Crystallogr. B, 48 (1992) 67.
49. Picot, D., Loll, P.J. and Garavito, R.M., Nature, 367 (1994) 243.
50. deMare, F., Kurtz Jr., D.M. and Nordlund, P., Nat. Struct. Biol, 3 (1996) 539.
51. Dumas, J., Exp. Opin. Ther. Patents, 11 (2001) 405.
52. Kleywegt, G.J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K. and Jones, T.A., Structure, 2 (1994) 1241.
53. Weber, P.C., Pantoliano, M.W., Simons, D.M. and Salemme, F.R., J. Am. Chem. Soc., 116 (1994) 2717.
54. Varghese, J.N., Smith, P.W., Sollis, S.L., Blick, T.J., Sahasrabudhe, A., McKimm-Breschkin, J.L. and Colman, P.M., Structure, 6 (1998) 735.
55. Bourguet, W., Ruff, M., Chambon, P., Gronemeyer, H. and Moras, D., Nature, 375 (1995) 377.

56. Picot, D., Loll, P.J. and Garavito, R.M., Nature, 367 (1994) 243.
57. Newcomer, M.E., Structure, 1 (1993) 7.
58. Hadfield, A.T., Oliveira, M.A., Kim, K.H., Minor, I., Kremer, M.J., Heinz, B.A., Shepard, D., Pevear, D. C., Rueckert, R.R. and Rossmann, M.G., J. Mol. Biol., 253 (1995) 61.
59. Tronrud, D.E., Monzingo, A.F. and Matthews, B.W., Eur. J. Biochem., 157 (1986) 261.
60. Gomis-Ruth, F.X., Meyer, E.F., Kress, L.F. and Politi, V., Protein Sci., 7 (1998) 283.
61. Stebbins, C.E., Russo, A.A., Schneider, C., Rosen, N., Hartl, F.U. and Pavletich, N.P., Cell, 89 (1997) 239.
62. Narayana, N., Diller, T.C., Koide, K., Bunnage, M.E., Nicolaou, K.C., Brunton, L.L., Xuong, N.H., Ten Eyck, L.F. and Taylor, S.S., Biochemistry, 38 (1999) 2367.
63. Champness, J.N., Achari, A., Ballantine, S.P., Bryant, P.K., Delves, C.J. and Stammers, D.K., Structure, 2 (1994) 915.
64. Finley, J.B., Atigadda, V.R., Duarte, F., Zhao, J.J., Brouillette, W.J., Air, G.M. and Luo, M., J. Mol. Biol., 293 (1999) 1107.
65. Banner, D.W. and Hadvary, P., J. Biol. Chem., 266 (1991) 20085.
66. Backbro, K., Lowgren, S., Osterlund, K., Atepo, J., Unge, T., Hulten, J., Bonham, N.M., Schaal, W., Karlen, A. and Hallberg, A., J. Med. Chem., 40 (1997) 898.
67. Ding, J., Fraser, M.E., Meyer, J.H., Bartlett, P.A. and James, M.N.G., to be published.
68. Smith, T.F. and Waterman, M.S., J. Mol. Biol., 147 (1981) 195.
69. Pearson, W.R., Genomics, 11 (1991) 635 (ftp://ftp.virginia.edu/pub/fasta).