J Mol Model (2009) 15:489–498
DOI 10.1007/s00894-008-0424-7

ORIGINAL PAPER

Vicinity analysis: a methodology for the identification of similar protein active sites

A. McGready · A. Stevens · M. Lipkin · B. D. Hudson · D. C. Whitley · M. G. Ford

Received: 13 March 2008 / Accepted: 17 November 2008 / Published online: 16 December 2008

Abstract Vicinity analysis (VA) is a new methodology developed to identify similarities between protein binding sites based on their three-dimensional structure and the chemical similarity of matching residues. The major objective is to enable searching of the Protein Data Bank (PDB) for similar sub-pockets, especially in proteins from different structural and biochemical series. Inspection of the ligands bound in these pockets should allow ligand functionality to be identified, thus suggesting novel monomers for use in library synthesis. VA has been developed initially using the ATP binding site in kinases, an important class of protein targets involved in cell signalling and growth regulation. This paper defines the VA procedure and describes matches to the phosphate binding sub-pocket of cyclin-dependent protein kinase 2 that were found by searching a small test database that has also been used to parameterise the methodology.

Key Words Clique detection · *In silico* protein screening · Protein binding site similarity · Protein kinase · Receptor pocket similarity

Introduction

The wealth of experimental crystallographic data now available [1] enables the development of data mining techniques that can contribute to the goal of knowledge-based drug design. Given a known active site, it is now possible to search the public database for proteins having active sites that are similar, in terms of their three-dimensional (3D) chemical structure, to the target site. Active sites discovered this way may be only loosely connected, in a biochemical, functional or phylogenetic sense, to the target site and will therefore not have been identified using traditional homology-based searching techniques such as BLAST [2].

The lack of two-dimensional (2D) structural or functional homology has not prevented the development of a number of successful 3D homology models. Locally conserved 3D structures are important despite the lack of homology in secondary sequence or phylogenetics. Current computational approaches to this problem are well known, e.g. the CATH [3] database of observed protein structures, and the CavBase [4] method of searching for similarity amongst active sites.

The conservation of "sub-pockets" has important consequences for drug design and, in particular, for the construction of targeted combinatorial libraries. In the combinatorial approach, a chemically tractable "scaffold" is "adorned" with "flora" from the chemical universe. These flora, in effect simple monomers that are amenable to combinatorial chemistry, are selected either on the basis of chemical diversity or are based on fragments known to be of interest to the relevant pharmacology. In this latter approach, small molecular weight compounds are screened at high concentration to establish potential binding to a site. Experimental, X-ray or NMR methods are then used to confirm the binding and establish the fragment as a potential piece of flora to add into a combinatorial design.

Medicinal chemists are interested in improving the process of focussed drug design. This is often facilitated by in silico approaches such as automated docking.

A. McGready · B. D. Hudson (✉) · D. C. Whitley · M. G. Ford
Centre for Molecular Design, Institute of Biomedical and Biomolecular Sciences, University of Portsmouth,
King Henry Building, King Henry I St.,
Portsmouth PO1 2DY, UK
e-mail: brian.hudson@port.ac.uk

A. Stevens · M. Lipkin
Biofocus DPI,
Chesterford Research Park,
Saffron Walden, Essex CB10 1XL, UK

Although in silico methods are useful for docking molecules and fragments, they are less efficient at identifying which are the most appropriate alignments to select. The strength of the vicinity analysis (VA) approach is that it can help the synthetic chemist to make that decision with improved reliability. The growing knowledge base provided by the Protein Data Bank (PDB) enables a data mining approach. The identification of sub-pockets in proteins having similar tertiary structure and chemically similar residues, despite being functionally and phylogenetically different, would enable a computational approach to the rational design of potential generic monomers by the study of ligands that are observed experimentally to bind in "similar" sites. The PDB contains ca. 49,000 protein structures. A significant proportion of these also contain bound small molecule inhibitors. This represents a potentially highly valuable source of (evidence-based) ligand-binding interaction data. Computational approaches to this problem would therefore appear to be very relevant.

There are various computational approaches available for assessing the similarity between proteins. One approach is to compare the structures of whole proteins similar to a 3D version of BLAST. A good example is the TOP system [5]. This describes each protein secondary structure element as a pair of points, which are systematically overlaid to give the best match. Another approach is to look for similarities between smaller parts of the proteins. LFMpro [6] uses the distance field to backbone atoms to identify local features specific to a particular protein family, distinguishing that family from a training set of proteins in other families. However, such methods look at the overall similarity of the protein structures and are of most use in protein categorisation problems.

Of more interest to the medicinal chemist are methods that attempt to match protein binding sites or pockets. The FEATURE package [7] searches databases for similar arrangements of physicochemical property spaces defined from a training set of known sites of interest. The features are defined by using a supervised learning algorithm on this training set of known sites. An example is the identification of new calcium binding sites from a training set of known sites and sites known not to bind calcium.

When the binding site is known, detailed comparisons of the site with other proteins in a database can be attempted. Relibase [8] achieves this by performing matches of the Cα positions of the target structure with those of the database. The aim is to identify novel ligands based on the similarity of their binding sites to the target. The ASSAM system [9] searches for patterns of amino acids, defined as a set of vectors between the backbone and the functional side chain, using the Ullman subgraph isomorphism algorithm [10]. This geometric method has been applied to search for the serine protease catalytic triad. SPASM [11] is another method that identifies similar motifs by aligning the Cα positions and the centres of gravity of the residues using an exhaustive search.

Several methods for identifying and characterising protein binding sites are based on graphs where nodes represent certain features and edges are labelled by inter-feature distances. Pairs of such feature graphs are compared by combining them into a secondary, correspondence graph in such a way that a maximal set of matching features is represented in the correspondence graph by a clique: a set of nodes where each node is connected to every other node. For example, Samudrala [12] describes a system where the interaction energy between the sidechain and backbone of each residue, plus up to four of its neighbours, are used to form the nodes in an undirected graph. Clique analysis is then used to determine structural similarities. The SURFCOMP program [13] also uses clique analysis, this time with the pockets defined as collections of critical points on molecular surfaces rather than as simple atom positions or other geometrical constructs.

By recommending VA for this task, we are proposing a "direct-evidence" based approach that uses experimental data as a source to identify which small molecule fragments are most likely to bind into a pocket or sub-site. The results of VA can therefore be used as a basis for guiding a focussed drug design campaign. The key is to find similar micro environments in a binding site and using this to 'map' ligand fragments into a target cavity.

VA also provides a technique for finding similar small molecule binding sub-sites based on similarity of proximate residues. Residue similarity can be estimated using any appropriate descriptor metric, e.g. phylogenetic or chemical similarity. The VA algorithm identifies sub-site matches between the target query and a data set of crystal structures using a clique detection algorithm. Development of this technique has been facilitated using the ATP binding site in the kinase family of proteins, an important class of protein targets involved in cell signalling and growth regulation. This example was chosen because it provides a test set with known geometries for each sub-site.

The aim of this work is to develop a technology to mine PDB ligand-binding data that can be readily exploited in rational drug design programmes. The approach identifies similarities between ligand-binding sites based on the 3D coordinates of the residues and the degree of chemical similarity between matched residues.

Methods and theory

The presence of an active site within a PDB entry is indicated by the inclusion of a HET record for a ligand. HET records corresponding to metals, salts or modified

amino acids are omitted, and each of those remaining is considered a bona fide ligand, defining an active site formed by the set of residues sufficiently close to the ligand. A residue is considered to be within the active site if the shortest distance between the centres of atoms in the residue and the ligand is less than the sum of their atomic radii plus an interaction tolerance I_{tol}. The atomic radii used were based on the values in the CHARMM [14] united atom force field.

To reduce the number of distance calculations required, each active site was identified in two stages, using a modification of the LPC procedure of Sobolev et al. [15]. In the first stage, each residue was treated as a sphere based at the centre of gravity, *RC*, of the residue's atom centres, with radius *RR* equal to the maximum distance from *RC* to a residue atom centre. Similarly, the ligand was treated as a sphere based at its centre of gravity, *LC*, with radius *LR*. Letting *MR* denote the maximum atomic radius, residues with *RC* further than $RR + LR + 2MR + I_{tol}$ from *LC* were discarded as they cannot fall within the active site. In order not to artificially reject residues close to, but not directly in contact with, the ligand I_{tol} was set to 4.5 Å. The procedure is illustrated in Fig. 1. In the second stage, the remaining residues were processed atom-by-atom, checking whether the distances between any pair of residue and ligand atom centres fell within the prescribed range.

Vicinity analysis takes pairs of binding sites *S* and *T* identified as above and attempts to identify sub-pockets that can be matched in a one-to-one fashion so that they have similar 3D geometry, in terms of the relative positions and orientations of matching residues, and so that matching residues have similar chemical properties. Geometric similarity is based on the inter-residue distance, *d*, defined by the Euclidean distance between the C_α positions. A measure of chemical similarity between residues was obtained by Hellberg et al. [16] from the principal components of a matrix of 29 physicochemical properties for the amino acids (including molecular weight, several partition coefficients, NMR chemical shifts and high performance liquid chromatography retention times). The first three principal components z_1, z_2 and z_3 are broadly associated with hydrophilicity, bulk and electronic properties, respectively. The chemical similarity employed by VA is based on the inter-residue distance δ measured in the 3D chemical property space defined by z_1, z_2 and z_3.

Similar sub-pockets are found by locating the cliques of a specially constructed correspondence graph [17]. This graph *G* has a vertex for each pair of residues (*s*, *t*), where

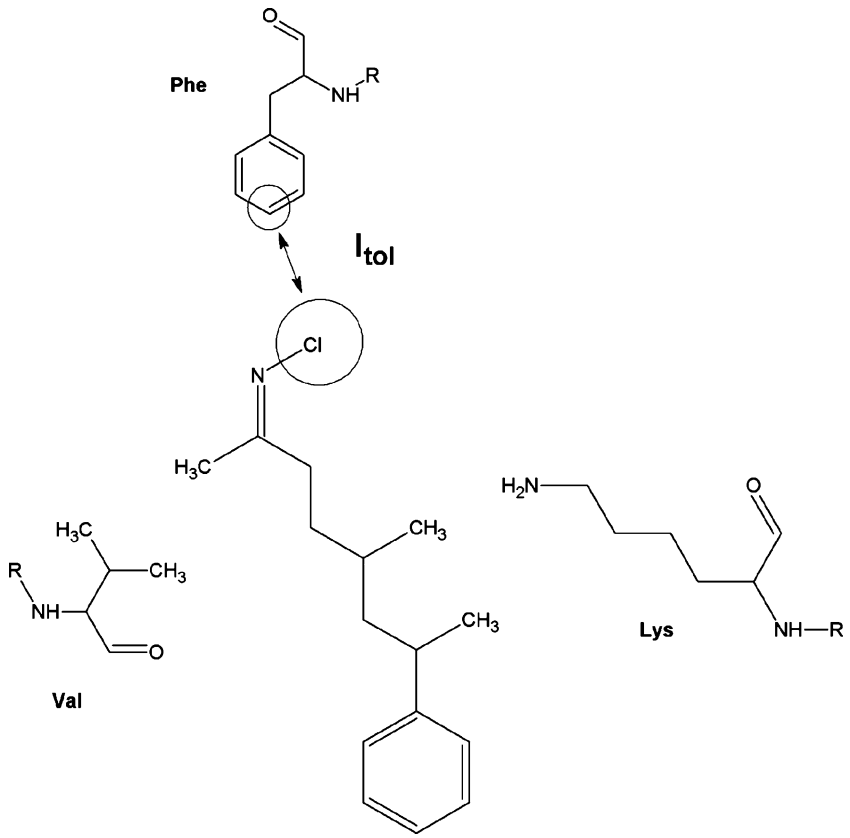

Fig. 1 Identification of potential binding sites

$s \in S$ and $t \in T$, and two vertices representing pairs (s, t) and (s', t') are joined by an edge if

$$|d(s,s') - d(t,t')| < D_{tol} \text{ and } \max\{\delta(s,t), \delta(s',t')\} < C_{tol} \quad (1)$$

where D_{tol} and C_{tol} are preset tolerances. A clique of G is a complete subgraph of G ; that is, a subgraph in which each vertex is connected to every other vertex [18]. Cliques of G represent sets of matching residue pairs that satisfy collectively both the geometric and chemical constraints in Eq. 1. The Bron-Kerbosch algorithm [19] was used to find the maximal cliques of G.

Several measures are used to filter and rank the VA results, which may include large numbers of matches. The size of a clique is the simplest indication of match quality, and cliques below a minimum size C_{min} are discarded. To establish the geometric quality of matches, for each clique the two matching sub-pockets are aligned by a least-squares rigid-body fit of the corresponding C_α positions and the resulting root mean square distance (RMSD) calculated. Matches with RMSD greater than some value R_{max} are discarded. An additional option is to measure the relative orientations of matched residues in their respective sub-pockets. For each matching pair of residues, the angle θ between the aligned $C_\alpha \rightarrow C_\beta$ vectors is found. If θ is greater than some preset bound θ_{max} the offending vertex is removed from the clique and the reduced clique is re-assessed (for glycine this step is omitted). The mean chemical distance (MCD), the mean of the distances $\delta(s, t)$ between matching residues in the chemical property space, provides an indication of the chemical quality of the match. The filters applied in the present work were C_{min}=4, R_{max}= 1.5 Å, and the matches were ranked by clique size, MCD and RMSD. In the following examples the θ_{max} tolerance was not used.

Results and discussion

In order to test the algorithms, a small set of 98 proteins containing 128 pockets with bound ligands was selected from the PDB. All the proteins in the test database are derived from X-ray structures and were chosen to include members of several different superfamilies. The distribution of binding pockets comprises 40 kinases (the primary superfamily of this study), 19 transferases, 19 proteases, 42 hydrolases and 8 others (including nucleases, aldolases, etc). The test query comprises a sub-site consisting of eight residues from the ATP binding pocket of cyclin-dependent protein kinase 2 (CDK2; PDB code 1AQ1). The query sub-site consists of eight residues: GLY13, TYR15, VAL18, LYS33, LYS129, GLN131, ALA144 and ASP145. Figure 2 shows the query sub-pocket with a bound ligand, staurosporine, depicted using SYBYL [20]. This set of residues comprises a readily identifiable micro-environment for binding the triphosphate group of ATP in CDK2.

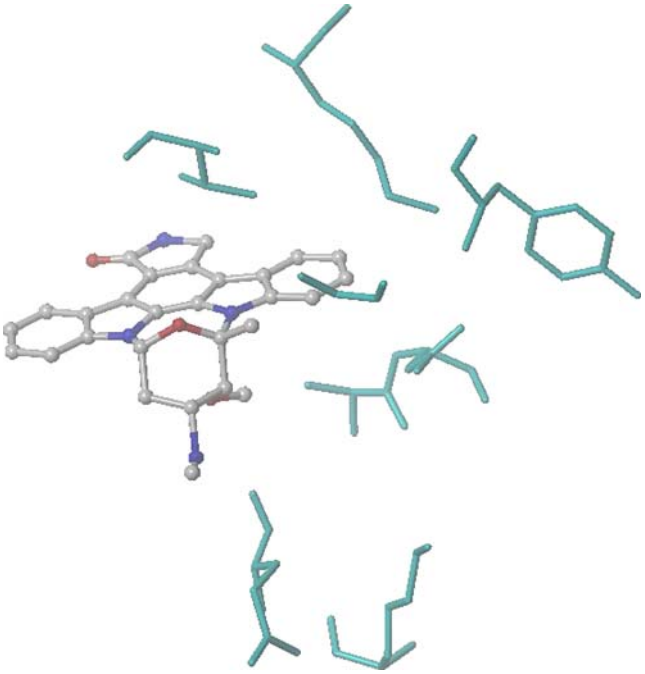

Fig. 2 Cyclin-dependent protein kinase 2 (CDK2) test query with staurosporine (1AQ1)

Table 1 shows the matching residues between the test query and c-AMP-dependent protein kinase (PDB ref 1ATP). This represents a different type of kinase from CDK2 from which the query was derived. VA has identified a clique of size seven with an RMSD of 1.09 Å between the matching residues, and it can clearly be seen that most of the residues match not only in terms of the C_α positions but also in terms of chemical similarity. In fact, five of the seven residues in the clique are identical, and the other two are conservative replacements (Ala144–Thr183 and Gln131–Glu170). Furthermore, the values for the $C_\alpha \rightarrow C_\beta$ angles are all below 30°. The overlay of the query with the matching residues in 1ATP is shown in Fig. 3.

A more interesting example is shown by the match between the test query and catechol O-methyl transferase (PDB ref 1JR4). This gave a matching clique of size five with an RMSD of 1.4 Å. The matching residues are shown in Table 2. Four of the residues are exact matches and the other match is a phylogenetically conservative replacement (Gln131–Arg146). Figure 4 shows the overlay of this pocket with the target query. It is of interest that the ligand involved in 1JR4 is similar to ATP in that it also possesses an adenine ribonucleoside, but in this case it is attached to a 2,3-dihydroxy-5-nitrophenylamido moiety. This is illustrated in Fig. 5, where the circles indicate the possible

Table 1 A clique of size seven matching the phosphate sub-pocket with c-AMP dependent protein kinase (1ATP)

Phosphate sub-site	1ATP	$C_\alpha \rightarrow C_\beta$ angle	Chemical similarity
ASP 145	ASP 184	18.6	0.000
ALA 144	THR 183	20.5	1.753
GLY 13	GLY 52	0.0	0.000
VAL 18	VAL 57	5.4	0.000
LYS 129	LYS 168	9.9	0.000
GLN 131	GLU 170	4.7	1.405
LYS 33	LYS 72	1.2	0.000

bioisosteric replacement of the triphosphate group. Thus, VA has shown its utility in identifying ATP-like binding sites in proteins other than the query, and also its potential for identifying putative binding sites, and hence bioisosteric ligands, derived from ligand matches in other classes of proteins.

A series of experiments was performed matching the CDK2 phosphate binding pocket against the test database to identify the parameters D_{tol} and C_{tol} that best discriminate between the kinase and non-kinase subsets. In this context, a binding pocket in the database is predicted to belong to a kinase if a clique is found matching that pocket to the CDK2 pocket. This procedure relies on the assumption that the phosphate-binding sub-pocket is present in most of the kinase set of proteins and absent in most of the non-kinase proteins. The VA program was run for a 2D grid of D_{tol} values from 1.0 Å to 4.0 Å in steps of 0.1 Å, and C_{tol} values from 2.0 to 6.0 in steps of 0.1, and a confusion matrix, showing the numbers of true and false, positive and negative results, was generated for each (D_{tol}, C_{tol}) pair. For each value of D_{tol} the area under the receiver operating characteristic (ROC) curve, obtained by plotting the sensitivity (no. of kinase hits / total no. of kinases) against 1-specificity (no. non-kinase misses/total no. non-kinases) for each value of C_{tol}, was calculated. The optimal value of D_{tol} was chosen to maximise this area. The optimal value of C_{tol} was chosen to correspond to the point in the D_{tol} optimal ROC curve closest to the upper left corner (x=0, y=1) of the plot.

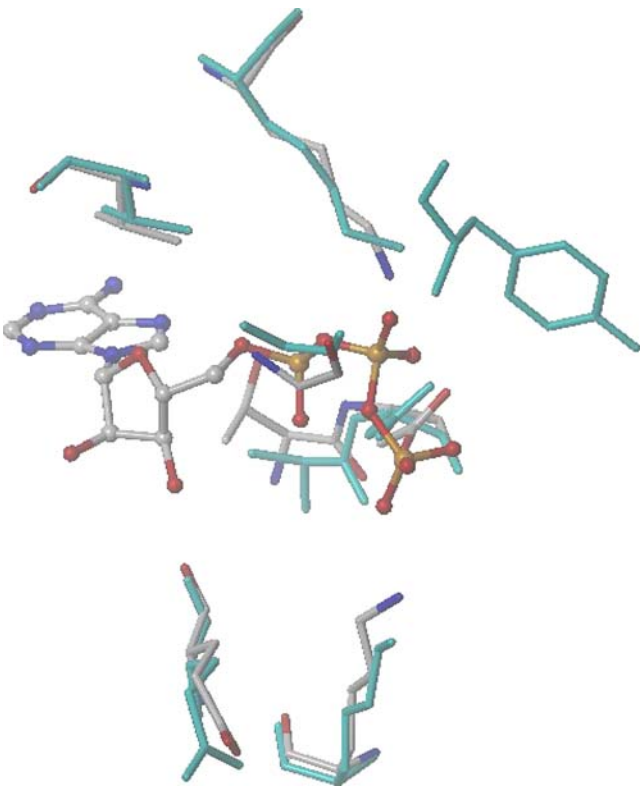

Fig. 3 Match of test query with 1ATP

ROC curves for several values of D_{tol} are shown in Fig. 6. The areas under the ROC curves for D_{tol} in the range from 1.0 Å to 4.0 Å, obtained by matching the CDK2 kinase phosphate binding pocket against the test database, are plotted in Fig. 7. The largest ROC area (0.87) occurs when D_{tol}=2.8. The point on the ROC curve for D_{tol}=2.8 closest to the upper left corner corresponds to a value of C_{tol}=3.4. Hence the optimal parameters were chosen to be D_{tol}=2.8 and C_{tol}=3.4.

To further test the utility of VA, the CDK2 test query was matched against a diverse database of 1,000 active sites extracted from X-ray crystal structures, selected at random from the PDB and using the parameters derived above. Figure 8 shows the best match of the CDK2 query against this expanded dataset (PDB code 1E1X). The match is a clique of size six with a C_α RMSD of 0.92 Å. This structure is that of CDK2 co-crystallised with a different ligand (NU6027). It is encouraging that VA has identified this match.

Figure 9 shows a match with hydroxysteroid dehydrogenase complexed with NADPH+ (PDB code 1EQU) comprising a five-residue clique with an RMSD of 1.36 Å and a mean chemical distance of 0.8. Despite the low chemical similarity of the match, the ligand is similar to ATP in that it contains adenine and phosphate moieties, although these are not visible in Fig. 9 as they lie some distance from the binding site. This match gives further confidence in the validity of the methodology.

Figures 10 and 11 show the matches found with 1JG1 and 1LEV. Figure 10 is 1JG1, which is protein-L-isoaspartate(D-aspartate) O-methyltransferase(1JG1) complexed with

Table 2 A clique of size five matching the phosphate sub-pocket with catechol O-methyltransferase (1JR4)

Phosphate sub-site	1JR4	$C_\alpha \rightarrow C_\beta$ angle	Chemical similarity
ASP 145	ASP 141	126.1	0.000
LYS 129	LYS 144	107.7	0.000
TYR 15	TYR 68	47.1	0.000
GLY 13	GLY 66	0.0	0.000
GLN 131	ARG 146	46.7	3.121

S-adenosyl-L-homocysteine. The clique size was five and the RMSD was 1.25 Å. This represents a very dissimilar protein, albeit with a similar ligand. Figure 11 is fructose-1,6-biphosphatase (1LEV) complexed with 3-(2-carboxyethyl)-4,6-dichloro-1H-indole-2-carboxylic acid. This has a clique size of six with an RMSD of 1.33 Å. This contains similar residue matches but a very dissimilar ligand.

The above examples demonstrate the utility of the VA approach. An advantage of the method is that other similarity measures and associated tolerances can be easily incorporated. This includes geometrical as well as chemical tolerances. As an example of this, the crystallographic B factors have been included as an extra tolerance. It is well known that the atomic positions in crystal structures have a degree of fuzziness due to a combination of experimental error and thermal fluctuations in the structure. This would suggest that the distance tolerances for atoms with large B factors can be relaxed compared to those with lower B factors. These B factor tolerances are incorporated by using a value of D_{tol} that varies with the B factors of the associated atoms.

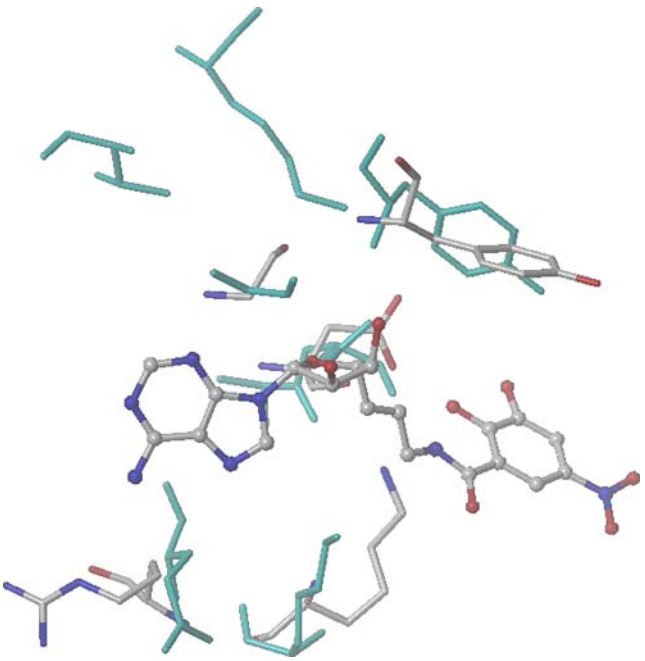

Fig. 4 Match of test query with 1JR4

Preliminary experiments suggest that this improves the results, giving more, better quality matches. For example, incorporation of a B factor tolerance within the CDK2 query identifies two additional high ranking hits (2PHK clique size seven, RMSD 1.046 Å, and 1O6K clique size seven, RMSD 1.246 Å). These matches are not identified using the static tolerances and are at positions three and four in the new ranking. The top two matches remain the same. In addition, this loosening of the tolerances identified an extra kinase from the test database at the expense of an increase in the number of non-kinase matches (from 11 to 24).

Comparison with other methods

In order to validate the algorithm, the above results were compared with those obtained using the standard primary sequence alignment routines in BLAST and PSI-BLAST, as well as with results obtained using the 3D algorithm SPASM.

For the simple case of matching of the 1AQ1 pocket with the other kinases in the dataset, the sequence alignment results are similar to those obtained with VA. Using the 3D chemical similarity matrix with the optimal tolerances of D_{tol}=2.8 and C_{tol}=3.4 described above, VA

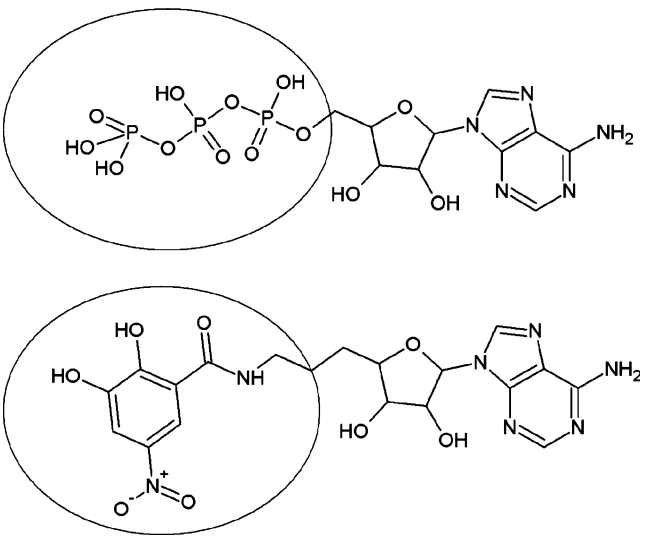

Fig. 5 Putative phosphate bioisosteric replacement in 1JR4

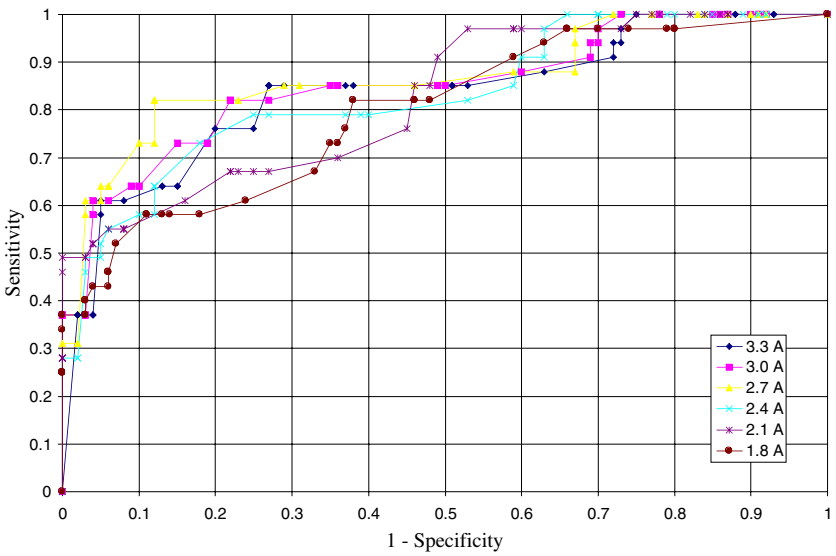

Fig. 6 Receiver operating characteristic (ROC) curves for different values of D_{tol}

identified 29 of the 40 kinase sites as being close matches. The standard BLAST analysis identified only four of these structures, based on the sequence alignment of the whole proteins. PSI-BLAST gave much more similar results to those of VA, identifying 19 of the kinases. Most of the high ranking hits identified by VA can also be identified by a simple PSI-BLAST analysis. There were, however, four of the VA kinases that were not identified by PSI-BLAST, whilst PSI-BLAST identified one kinase that was not identified by VA. This is not surprising since the kinase set was chosen as a well defined problem with which to calibrate and validate the methodology. Across the set, the sequence similarity of the non-identical kinases to CDK2 is between 52% and 64% and it is therefore to be expected that sequence alignment tools will perform well in this case.

Of more interest is to compare the performance of VA and PSI-BLAST in identifying similar pockets in non-kinase proteins, since this is the major rationale for the development of the VA methodology. In these experiments, a subset of the CDK2 sequence (residues 15–147) was used for the BLAST alignments. This sequence covers the range of the residues in the CDK2 ATP binding pocket, and was compared with the full sequence of the non-kinases in the dataset. Neither BLAST nor PSI-BLAST was able to find any significant sequence similarities in these structures. However, VA identified several non-kinase pockets as matches to the CDK2 test query and the five highest ranked of these are listed in Table 3. Three of these proteins have ligands closely similar to ATP (AMP in the case of 1FRP, 2′-deoxyadenosine-5′-diphosphate for 1G4A, and 1JR4, discussed previously). Table 3 also shows the results of a structural comparison of the proteins containing the non-kinase matches and the test query obtained using the DaliLite program [21]. These consist of the Z-score, the number of residues matched by DaliLite, the RMSD between matched alpha-carbons and the percentage se-

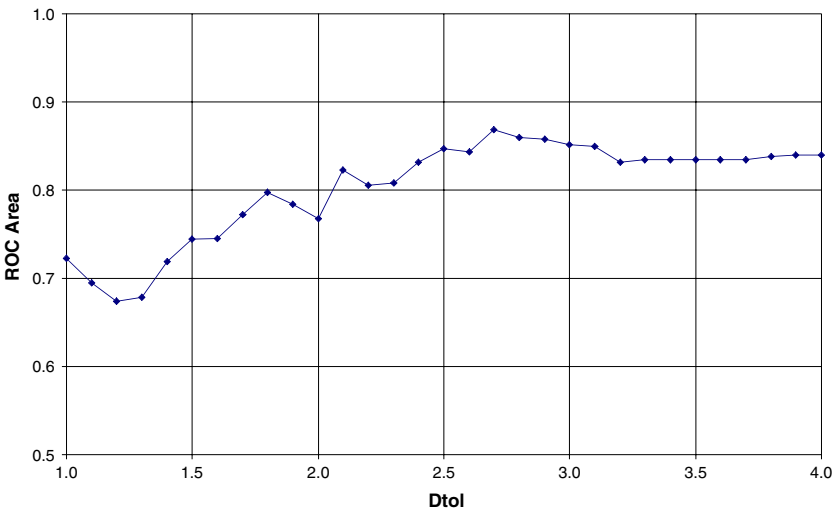

Fig. 7 Area under ROC curve against D_{tol}

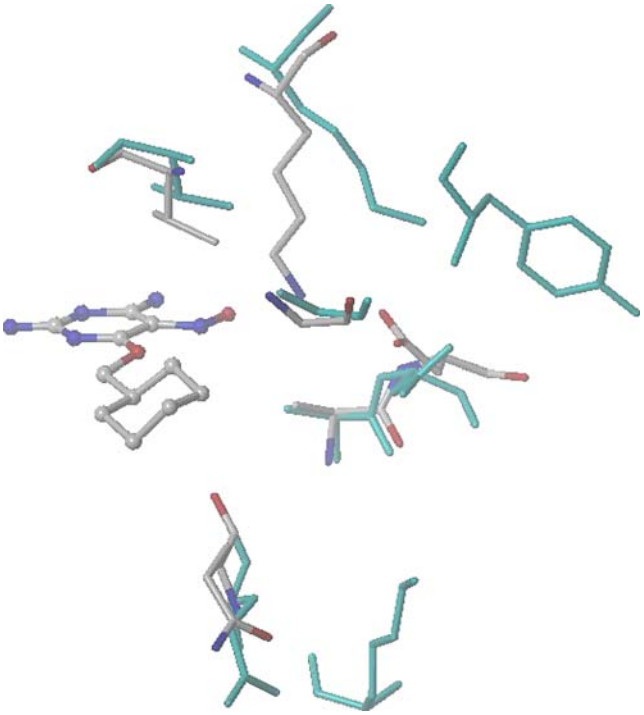

Fig. 8 Match of test query with 1E1X

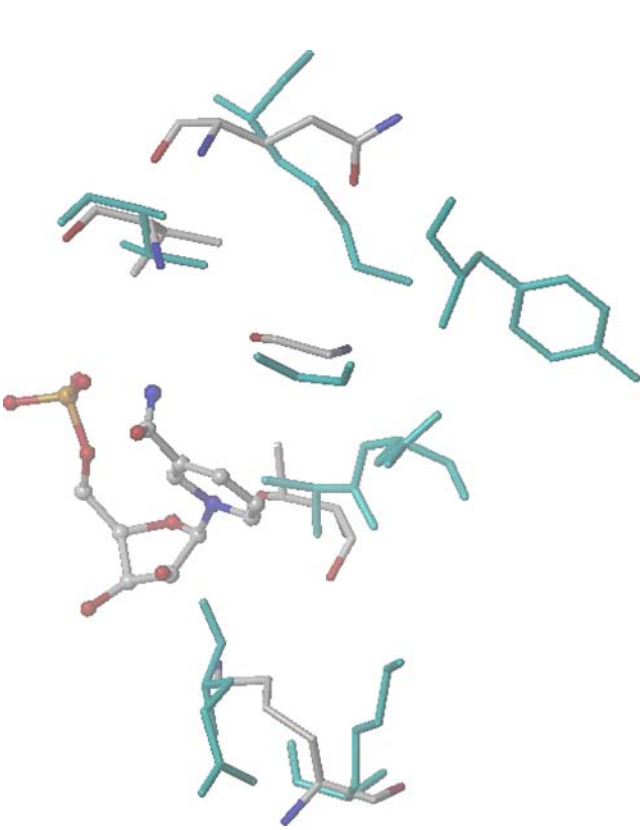

Fig. 9 Match of test query with 1EQU

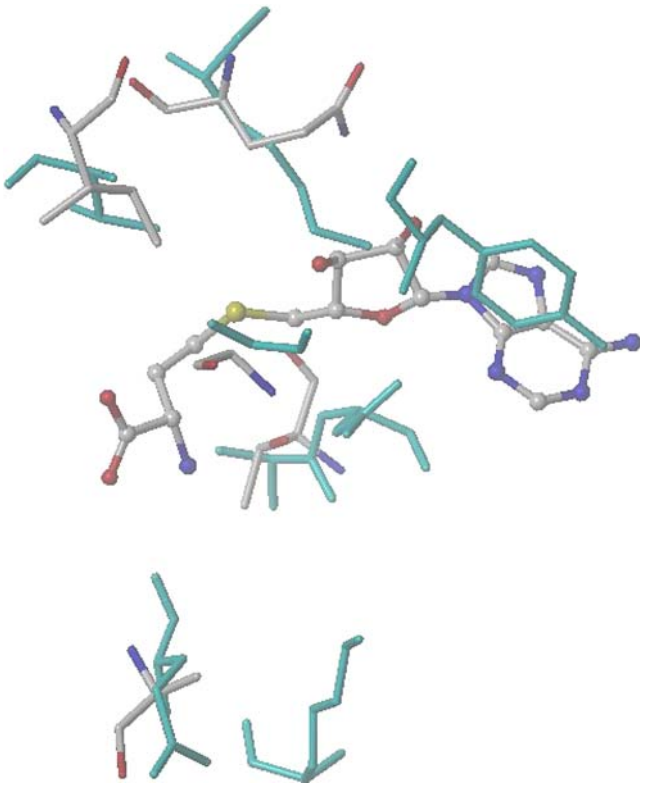

Fig. 10 Match of test query with 1JG1

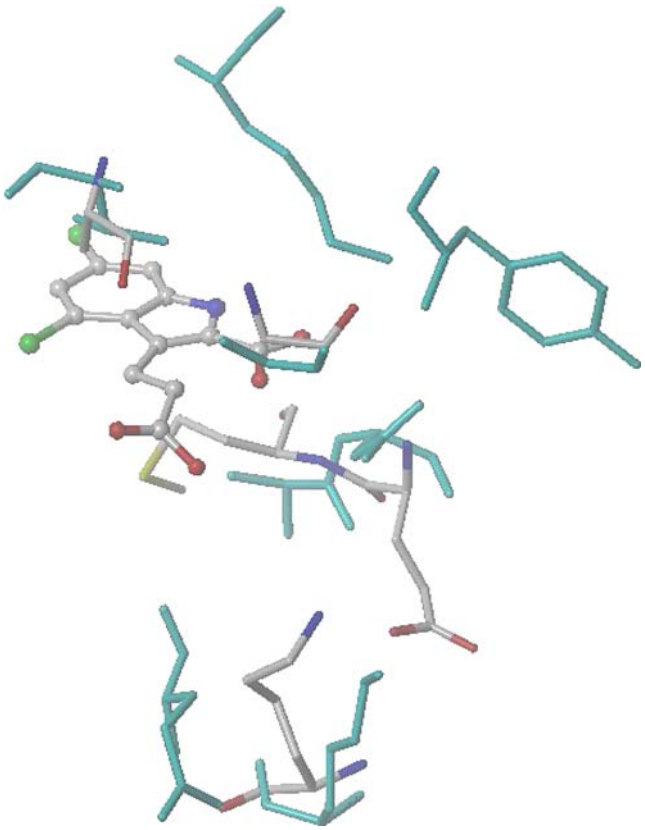

Fig. 11 Match of test query with 1LEV

Table 3 Non-kinase hits identified by vicinity analysis (VA) as matches to the 1AQ1 query pocket, with details of the DALI structural comparison. *PDB* Protein Data Bank, *RMSD* root mean square distance

PDB code	Protein	Type	Z-score	Aligned residues	RMSD	Sequence identity (%)
1FRP	Fructose-1,6-biphosphatase	Hydrolase	0.7	64	4.8	2
1JR4	Catechol O-methyl transferase	Transferase	1.2	40	3.6	5
1A2T	Staphylococcal nuclease	Nuclease	0.2	65	10.1	3
1G4A	ATP-dependent HSLV protease	Protease	0.1	67	4.3	7
1A16	Aminopeptidase P	Peptidase	0.6	63	7.1	8

quence similarity between the matched residues. From these results it is clear that the non-kinase matches found by VA belong to proteins that are not homologous to CDK2. The DaliLite matches have low sequence identity, large RMSD and Z-scores below the value of 2.0 required for a significant match.

Since the sequences are identical, when the same analysis is performed using the geometry of the 1E1V site (from the CDK2/ATP structure), once again no hits are found using PSI-BLAST. However, additional matches in non-kinase proteins were identified by VA; the highest ranked of these are listed in Table 4. The optimal DaliLite matches between these proteins and the 1E1V query, reported in the last four columns of Table 4, show that these VA matches also come from proteins that are structurally dissimilar to the query. In this case, the Z-scores for 1KNY lie above the 2.0 value required for statistical significance, but well below the value of 8.0 required for a probable homologous match.

The 3D method most closely related to VA is SPASM. This uses a depth-first search to find matches to the whole query pocket, rather than the maximal clique procedure employed by VA that enables the identification of matching sub-pockets. Both programs allow matching either solely on the C_α positions or with an additional constraint on residue orientation, through the use of the residue centres of gravity in SPASM and the $C_\alpha - C_\beta$ vectors in VA. The default residue replacements allowed by SPASM are based on a phylogenetic similarity matrix (BLOSUM-45), but if this is over-ridden to use the replacements determined by the VA C_{tol} parameter, then the hits found by SPASM for a fixed subset Q′ of the query are essentially the same as the subset of VA cliques containing Q′. To obtain results equivalent to those of VA, the SPASM program would have to be applied to all possible subsets of the query, down to the minimum clique size. VA therefore provides a much more extensive selection of hits in a single run than SPASM thus allowing parameter optimisations of the type presented here, which would be extremely labourious with SPASM.

Conclusions

Vicinity analysis can be used to match ligand-binding sites between proteins with similar active sites. It has been exemplified using the phosphate subsite of a known kinase and has identified matches with both other kinases as well as non-kinases from a diverse database. The method has been successfully applied to the problem of matching a test kinase sub-site with other kinases from a diverse database of proteins. Identification of these matches means the associated ligand fragments can be used in a focussed drug design campaign.

Vicinity analysis is capable of detecting areas of similarity across different protein superfamilies. The software has the ability to analyse any protein structure for areas of similarity. Because the phosphate pocket was derived from CDK2 (1AQ1), these areas of similarity can be used to suggest potential drug fragments that are specific targets for kinases. In the case of 1JR4, the area of similarity, i.e. the five matching residues, contain a drug fragment in the terminal ligand group that may act as a phosphate isostere. This illustrates the capability of the tool.

The approach could also be used to suggest molecular fragments for any given protein, and so reduce the time required to design novel ligands and, ultimately, drugs that fit it. Identification of the correct maximal clique is, of course, critical to the power of the approach. The matches

Table 4 Non-kinase hits identified by VA as matches to the 1E1V query pocket, with details of the DALI structural comparison

PDB code	Protein	Z-score	Aligned residues	RMSD	Sequence identity (%)
4FUA	L-Fuculose-1-phosphate aldolase	0.5	54	3.4	9
1A3L	Exo-dielsalderase antibody complex	0.9	42	3.5	14
1KNY	Kanamycin nucleotidyl transferase	3.4	96	4.8	8

identified with VA, together with knowledge of the associated ligand, can be used to develop novel fragments for drug design.

Acknowledgements We would like to thank Drs V.S. Rose and C.J. Harris for helpful comments and support for this work. A.M.G. acknowledges financial support from Biofocus DPI.

References

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissag H, Shindyalov IN, Bourne P (2000) Nucleic Acids Res 125:235–242. doi:10.1093/nar/28.1.235
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389–3402. doi:10.1093/nar/25.17.3389
3. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) Structure 5:1093–1108. doi:10.1016/S0969-2126(97)00260-8
4. Schmitt S, Kuhn D, Klebe G (2002) J Mol Biol 323:387–406. doi:10.1016/S0022-2836(02)00811-2
5. Guoguang L (2000) J Appl Cryst 1:176–183
6. Sacan A, Oztuck O, Feratosmanoglu H, Wang Y (2007) Bioinformatics 23:709–716. doi:10.1093/bioinformatics/btl685
7. Wei L, Altman RB (1998) Pac Symp Biocomput 3:497–508
8. Hendlich M, Bergner A, Gunther J, Klebe G (2003) J Mol Biol 326:607–620. doi:10.1016/S0022-2836(02)01408-0
9. Spriggs VR, Artymiuk PJ, Willett P (2003) J Chem Inf Comput Sci 43:412–421. doi:10.1021/ci0255984
10. Ullman J (1976) J ACM 23:31–42. doi:10.1145/321921.321925
11. Kleywegt GJ (1999) J Mol Biol 285:1887–1897. doi:10.1006/jmbi.1998.2393
12. Samudrala R, Moult J (1998) J Mol Biol 279:287–302. doi:10.1006/jmbi.1998.1689
13. Hofbauer C, Lohninger H, Aszódi A (2004) J Chem Inf Comput Sci 44:837–847. doi:10.1021/ci0342371
14. Brooks BR, Bruccoleri RE, Olafson DJ, States DJ, Swaminathan S, Karplus M (1983) J Comput Chem 4:187–217. doi:10.1002/jcc.540040211
15. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M (1999) Bioinformatics 15:327–332. doi:10.1093/bioinformatics/15.4.327
16. Helberg S, Sjostrom M, Skagerberg B, Wold S (1987) J Med Chem 30:1126–1135. doi:10.1021/jm00390a003
17. Gardiner EJ, Willet P (2000) J Chem Inf Comput Sci 40:273–279. doi:10.1021/ci990262o
18. Harary F (1994) Graph theory. Addison-Wesley, Reading, MA
19. Bron C, Kerbosch J (1973) Commun ACM 16:575–577. doi:10.1145/362342.362367
20. SYBYL, Tripos Associates, St Louis, MO
21. Holm L, Park J (2000) Bioinformatics 16:566–567. doi:10.1093/bioinformatics/16.6.566