ELSEVIER

Journal of Molecular Graphics and Modelling 21 (2002) 71–79

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

A simple method for visualizing the differences between related receptor sites

Robert P. Sheridan [a,*], M. Katharine Holloway [b], Georgia McGaughey [b], Ralph T. Mosley [a], Suresh B. Singh [a]

[a] *Department of Molecular Systems, RY50SW-100 Merck Research Laboratories, Rahway, NJ 07065, USA*
[b] *Department of Molecular Systems, WP53F-301 Merck Research Laboratories, West Point, PA 19486, USA*

Abstract

Pastor and Cruciani [J. Med. Chem. 38 (1995) 4637] and Kastenholz et al. [J. Med. Chem. 43 (2000) 3033] pioneered methods for comparing related receptors, with the ultimate goal of designing selective ligands. Such methods start with a reasonable superposition of high-resolution three-dimensional (3D) structures of the receptors. Next, molecular field maps are calculated for each receptor. Then the maps are analyzed to determine which map features are correlated with a particular subset of receptors. We present a method FLOGTV, based on the trend vector paradigm [J. Chem. Inf. Comput. Sci. 25 (1985) 64] to perform the analysis. This is mathematically simpler than the GRID/CPCA method of Kastenholz et al. and allows for the simultaneous comparison of many receptor structures. Also, the trend vector paradigm provides a method of selecting isopotential contours that are well above "noise". We demonstrate the method on four examples: HIV proteases versus two-domain acid proteases, thrombin versus trypsin and factor Xa, bacterial dihydrofolate reductases (DHFRs) versus vertebrate DHFRs, and P38 versus ERK protein kinases.

Keywords: Trend vector; FLOG; Homologs; Specificity

1. Introduction

One of the desirable properties of a drug is its selectivity, that is, its ability to bind to only one of a number of related receptors. The selectivity may be between, for instance, related enzymes in bacterial and mammalian systems, or between related enzymes in the same species. Given the atomic-level three-dimensional (3D) structures of the receptors, one may explain selectivity or design selective inhibitors. Since such atomic-level structures are almost always X-ray crystal structures of enzymes, we will hereafter speak of "enzymes", with the understanding that the methodology can apply to any kind of receptor for which there is a high-resolution structure. A series of papers [1,2] has suggested a method by which one can visualize the differences between related enzymes. The more recent methodology GRID/CPCA, as described by Kastenholz et al. [2] has the following steps:

1. Find a reasonable superposition of the enzyme structures. The more structures one has, including more than one determination of the same enzyme, the less chance that one can be misled by the idiosyncrasies of a single structure.
2. Calculate a series (typically a dozen) of molecular interaction maps around the binding site, each map being defined by a "probe atom". The interaction map for a probe is a 3D grid with a number at each *x*, *y*, *z* location representing the energy of interaction of the probe with the enzyme at that location. The program GRID [3] is a popular way of generating such fields.
3. The molecular interaction maps are assembled into a large matrix, where each row of the matrix represents an enzyme, and each column is a grid point for a particular probe. For *n* probes and *k* grid points, there will be *nk* columns.
4. A principal components analysis (PCA) is done to find the linear combination of grid values that best explains the variance between the enzymes.
5. The loadings of vectors best separating enzymes of interest are projected as contour plots. This allows the user to visualize which parts of the binding site favor specific probes for specific sets of enzyme.

Kastenholz et al. [2] have pointed out that, since the range of interaction energies may vary from probe to probe, it is very useful to rescale the individual molecular interaction maps before assembling them into a large matrix.

* Corresponding author. Tel.: +1-732-594-3859; fax: +1-732-594-4224.
E-mail address: sheridan@merck.com (R.P. Sheridan).

1093-3263/02/$ – see front matter
PII: S1093-3263(02)00122-5

Interpretation is aided by setting all repulsive interaction energies to zero before the PCA. Also, it is useful to ignore grid points more than a certain distance from the binding site, as these may vary a great deal between related receptors, but are not relevant for binding.

In this paper, we present a mathematically simpler method of visualizing the differences between receptors based on the trend vector formulation. This method does not require rescaling, and provides a natural method of assigning isopotential contours above "noise".

2. Methods

2.1. Superposition

All methods of this type start with a reasonable superposition between the various enzymes. Since for most selectivity problems the enzymes are presumed to be related, finding some superposition is not hard. In practice superposition is usually done by overlaying subsets of atoms. Some possibilities are:

1. The alpha-carbons of the enzymes, especially around the binding site.
2. Selected atoms in the enzymes that are known to interact with ligands.
3. Selected atoms in the ligands.

Here we use alpha-carbon superposition because, given that there are many alpha-carbons distributed over the entire binding site, the superpositions are less sensitive to idiosyncrasies of particular crystal structures, for instance unique chi values of sidechains. Also, it is not necessary that certain sidechains be conserved over all enzymes, or that a co-crystallized ligand be present with every enzyme. In our experience, once the alpha-carbons are overlaid, the enzyme atoms relevant for binding and/or ligands are almost always overlaid as well.

One way to overlay alpha-carbons is to specify which pairs of alpha-carbons are equivalent by sequence alignment, and we have done that in one of our examples. However, we found that SQ, [4] our tool for finding superpositions of drug-like molecules, can automatically generate very credible superpositions of alpha-carbon traces without the user having to specify the equivalence between residues. It does this in two steps: (1) Selecting sets of alpha-carbons with similar through-space distances to make an initial orientation. (2) Refining the orientation of the two traces so that there is maximum volume overlap of Gaussians centered on the alpha-carbons. Thus, SQ uses a similar scoring function to that recently published by Maggiora et al. [5]. In SQ a number of possible superpositions are generated. In almost all cases the highest scoring superposition, using the default SQ parameters, is the one that superimposes conserved 3D features. Of course, for work of this type any reasonable method of alpha-carbon superposition can be used.

2.2. Generation of maps

We are familiar with FLOG [6] maps, which were originally developed for the docking of databases of drug-like molecules onto a receptor. FLOG maps are conceptually similar to those from GRID, but do not require atomic charges and allow for certain ambiguities in the placement of polar hydrogens. By default, there are five probe types in FLOG: donor/cation, acceptor/anion, polar (i.e. donor and/or acceptor), hydrophobic, and other. "Other" atoms feel only van der Waals interactions. The idea that, for instance, donors and cations should be treated as equivalent, comes from the observation [7] that intra-protein interactions seem well-predicted by hydrogen bonding, but not by electrostatics. For the purposes of this work, it is usually sufficient to examine only the donor/cation, acceptor/anion, and hydrophobic maps because the two other maps provide information that is redundant.

Typically a FLOG map is constructed as a box 5 Å around a ligand. The FLOG interactions are all short-range, so it is necessary to include only those enzyme atoms within 5 Å of a map point to capture all the relevant interactions. For the purpose of visual inspection, a 0.6 Å spacing of map points is sufficient. The FLOG convention is that more positive numbers mean more attractive interactions. Although we use FLOG maps exclusively in this paper, the methodology applies equally well to any type of interaction field map.

2.3. Trend vector

The trend vector [8,9] was first formulated by Carhart et al. [8] to relate some biological activity to the presence or absence of topological descriptors in large sets of drug-like molecules. Effectively we are asking: What is the difference in descriptor space between the more active and the less active molecules? It is trivially extended to comparing maps:

$$T(x, y, z, j) = \frac{1}{W_{\text{tot}}} \sum_{\text{i}} W(i) M(x, y, z, j, i)\, A(i)$$

The index i goes over enzymes. $W(i)$ is the weight on enzyme i, W_{tot} the sum of weights, $M(x, y, z, j, i)$ the map value for probe j against enzyme i at location x, y, z and $A(i)$ is the "activity" of enzyme i. The values of $A(i)$ have been normalized so that the weighted mean activity of all the enzymes is zero, and the weighted S.D. is 1. For the purposes of this work, we will set the activity of an enzyme before normalization to 1.0 if it is desirable to selectively inhibit it, and to 0.0 if it is desirable not to inhibit it. Thus, T will capture the difference in map space between "desirable" and "undesirable" enzymes. $W(i)$ can be used to set the importance of a particular enzyme. In this paper, we will set all weights to 1.0.

We concur with Kastenholz et al. [2] that interpretability is greatly helped by ignoring repulsive potentials and potentials far away from the binding site. In our case, we set $M(x, y, z, j, i) = 0$, if $M(x, y, z, j, i) < 0$. Also, we

set $M(x, y, z, j, i) = 0$ if x, y, z is more distant than a user-specified cut-off, usually 5 Å, from a reference ligand or ligands.

Trend vector analysis has been used to find correlations between the biological activity and the presence of topological descriptors. Any given descriptor can have a non-zero correlation strictly due to chance, and it is crucial to determine what level of correlation is seen in a set for which there is no real relationship between activities and descriptors, i.e. the level of "noise". The same applies to descriptors in receptor map space. Also, here we deal with at least two additional sources of noise having to do with the 3D nature of the descriptors: variations in atomic coordinates between crystal determinations (even for the same enzyme), and imperfect superpositions. The usual approach in the trend vector paradigm is to generate "null hypothesis" sets by scrambling the activities (i.e. randomly reassigning them among the molecules), renormalizing the activities, and recalculating T. We prefer this Monte Carlo approach to any analytical approach because it implicitly accounts for the size of the dataset, the number of actives, problem-specific errors in superposition, etc. Typically, several such scrambled sets are generated.

2.4. Contouring

T is already in the form of a set of maps, and can be contoured directly. Scaling of different probes is implicitly handled by the fact that the map for each probe j is independent and can be contoured at its own appropriate level.

We would like to pick isocontour levels low enough that we can detect subtle differences between enzymes, but high enough that the differences are very unlikely to be due to chance. One way of doing this is to compare the real T with T's generated from scrambled activities. The distribution of the values in T is centered around zero and much narrower than Gaussian, not surprising because a significant fraction of the map points is set to exactly zero. There are an appreciable number of values far from zero. Fig. 1 shows a typical example. The distribution for scrambled sets tend to be much more symmetric around zero and narrower, as might be expected. One measure of the spread around zero is the S.D. σ. Our heuristic is to choose the isopotential for each probe at $\pm c\sigma_{\text{mean}}$, where σ_{mean} is the average σ over 10 or more scrambled sets. We have had good results using $c > 12$, and here we use $c = 15$. There are hardly any map values from the scrambled T's $> 15\sigma_{\text{mean}}$, but many map values from the real T. Thus, we can be confident that the isopotential surfaces reflect real differences.

2.5. Setup of problems

Enzyme structures were taken from the protein data base, [10] superimposed onto a reference structure using alpha-carbons. Ligands, if present, were removed from the active-sites, and all waters were deleted. Enzyme atoms were assigned default ionization states based on the residue type (Arg, Lys cations; Asp, Glu anions). FLOG maps were generated such that their sides extended 5 Å around the union of all the superimposed ligands. For the calculation of T,

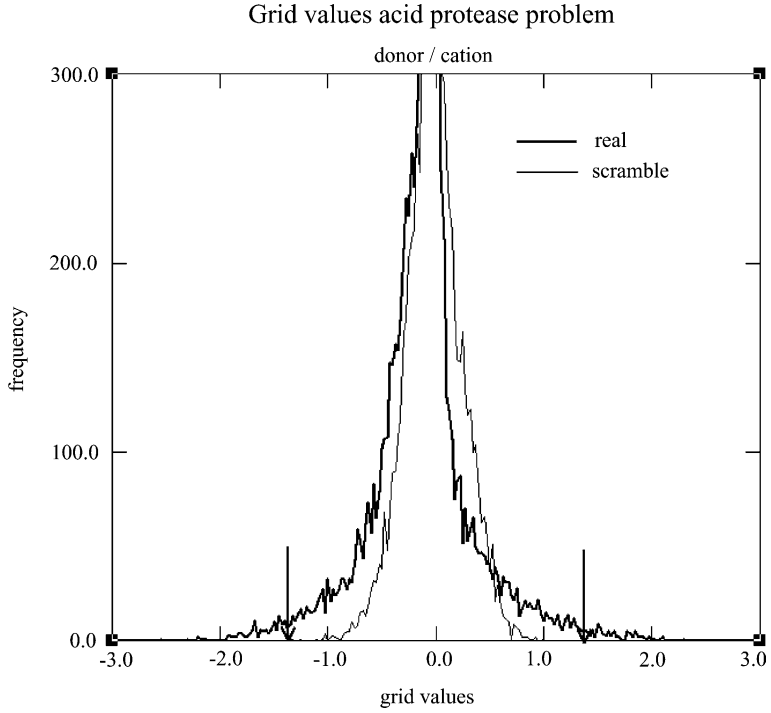

Fig. 1. The distribution of values for the donor/cation map of T for the acid protease problem compared to the distribution from T calculated from scrambled data. The arrows indicate $\pm 15\sigma_{\text{mean}}$ for 10 sets of scrambled data.

map points were set to zero where x, y, z was more than 5 Å away from the atoms of the ligand associated with the reference structure.

3. Results

3.1. Acid proteases

In this problem, we examine the difference between HIV protease, a homodimer, and two-domain acid proteases. The enzymes are listed in Table 1. HIV protease is similar to

Table 1
Enzymes for acid protease problem

PDB code	Description	Source	Activity
1HSH[a]	Acid protease	HIV-1	1
4PHV	Acid protease	HIV-1	1
2BPZ	Acid protease	HIV-1	1
1AJX	Acid protease	HIV-1	1
1PSO	Pepsin	Human	0
1HRN	Renin	Human	0
6APR	Rhizopuspepsin	*Rhizopus*	0
1APV	Penicillopepsin	*Penicillium*	0
1E5O	Endothiapepsin	*Endothia*	0

[a] Reference structure for superposition.

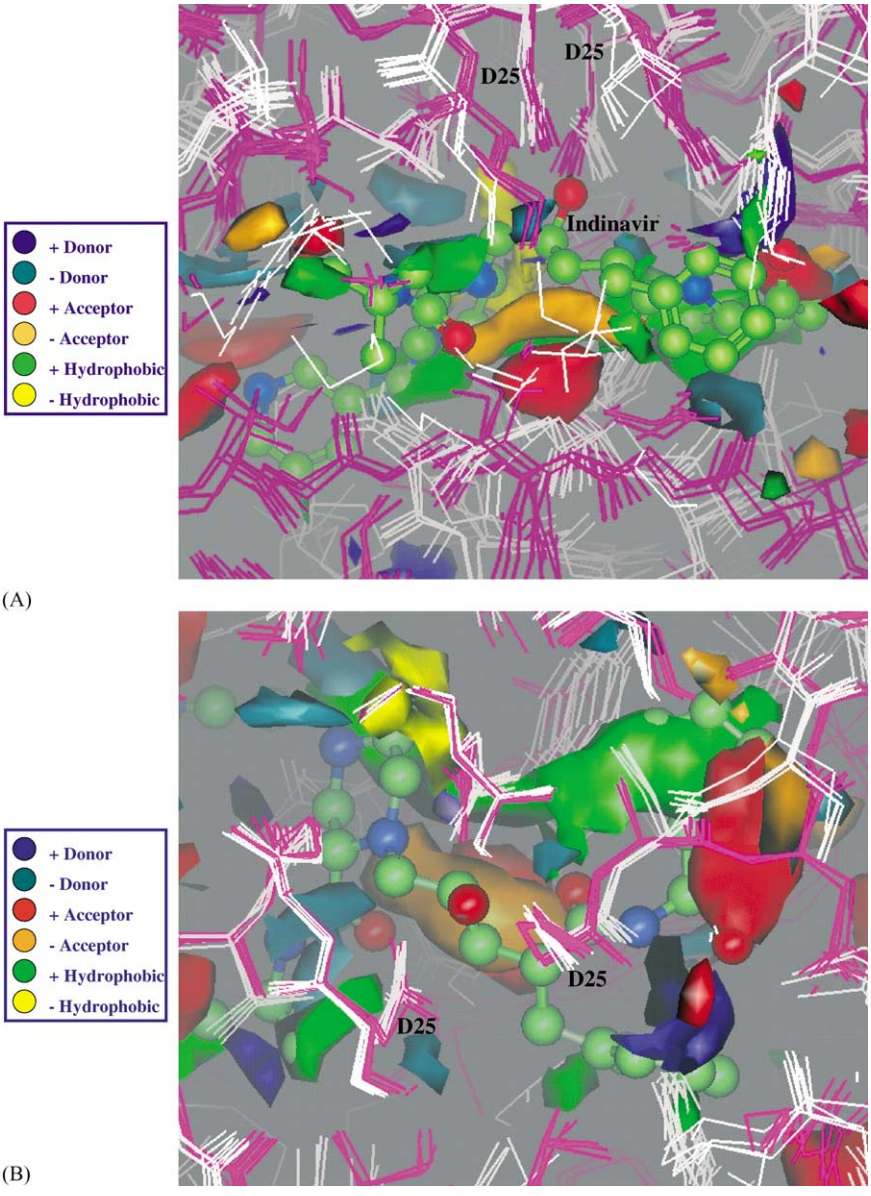

Fig. 2. *T* maps for the acid protease problem. Enzymes with activity 1 (HIV proteases) are shown in purple wire, those with activity 0 (two-domain acid proteases) are shown in white wire. The inhibitor indinavir is shown as ball and stick. Positive contours are where the specified probe would selectively bind to HIV proteases: donor/cation, blue; acceptor/anion, red; hydrophobic, green. Negative contours are where the probe would selectively bind to two-domain proteases: donor/cation, blue–gray; acceptor/anion, orange; hydrophobic, yellow. (A) The view where the catalytic aspartates are at the top of the figure, and the flaps are at the bottom. (B) The view where the catalytic aspartates are toward the reader.

two-domain proteases in the region of the active-site aspartates, but different elsewhere, so the superpositions were done such that corresponding alpha-carbons of the other enzymes were superimposed onto the alpha-carbons A24–A28 and B24–B28 of 1HSH.

Fig. 2A shows T contoured at $c = 15$ for donor/cation, acceptor/anion and hydrophobic maps. The ligand for 1HSH (indinavir) is shown as ball and stick. The orientation is with the catalytic aspartates up, and the flaps down. The upper part of all enzymes are very similar, as indicated by a lack of difference contours. The most compact feature in the lower portion is a red sphere of positive acceptor contour. This is where the conserved water would be found in HIV proteases. Above that is an orange crescent of negative acceptor contour that touches the amide carbonyls of indinavir. This difference is due to backbone nitrogens from the flap being closer to the inhibitor in the two-domain acid proteases, whereas the flaps in HIV protease are far enough away to allow a water molecule between. Indinavir does not place any atom in the acceptor portion, but known cyclourea-based HIV-specific inhibitors do, e.g. the ligand for 1AJX. The hydrophobic contours, better seen in Fig. 2B, are due to the presence of a few residues closing off parts of the active-site. Ile-84 in HIV protease closes off part of the active-site that is open to two-domain proteases, hence the yellow negative hydrophobic contour. Thr-75 and Tyr-77 (1PSO numbering) close off parts available to HIV protease, hence, the green positive hydrophobic contours. The red contour on the right is due to the backbone nitrogens of Asp-A29 to Asp-A30 which approach closer to the binding site in HIV protease than in the two-domain acid proteases.

Table 2
Enzymes for thrombin problem

PDB code	Description	Source	Activity
1DWD[a]	Alpha-thrombin	Human	1
1DWB	Alpha-thrombin	Human	1
1DWC	Alpha-thrombin	Human	1
1QUR	Alpha-thrombin	Human	1
7KME	Alpha-thrombin	Human	1
1FAX	Coagulation factor Xa	Human	0
1HCG	Coagulation factor Xa	Human	0
1XKB	Coagulation factor Xa	Human	0
1PPC	Trypsin	Bovine	0
1MTS	Trypsin	Bovine	0
1MTU	Trypsin	Bovine	0
1MTV	Trypsin	Bovine	0
1MTW	Trypsin	Bovine	0

[a] Reference structure for superposition.

3.2. Thrombin

In the second problem (Table 2), we examine the difference between thrombin and two other serine proteases, trypsin and factor Xa. Enzymes were superimposed onto 1DWD by SQ using all alpha-carbons within 12 Å of NAPAP, the inhibitor co-crystallized with 1DWD.

Fig. 3 shows the T maps. The binding site can be divided into three sections. The D site is to the left and contains the naphthyl portion of NAPAP, the P site is on the top and contains the piperidine, and the S1 pocket is to the right and binds the amidine of NAPAP. In the D site, the blue contour is due to backbone carbonyls provided by a loop containing residues 96–98 in thrombin. This region is blocked in trypsin and factor Xa because the corresponding loop approaches

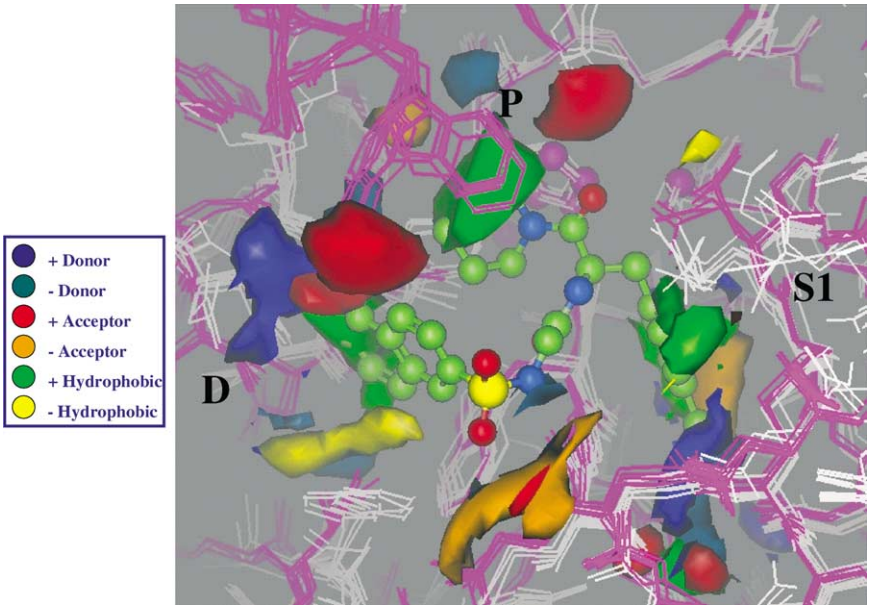

Fig. 3. The T maps for the thrombin problem. Thrombin structures are shown as purple wire, trypsin and factor Xa as white wire. The inhibitor NAPAP is shown as ball and stick. The contour convention is as in Fig. 2.

closer in those enzymes. The green contour is a hydrophobic area in thrombin that is blocked by the same loop in the other two enzymes. The red contour is due to hydrogen bonding from Tyr-60A in thrombin, which does not have a counterpart in the other enzymes. The yellow contour is a hydrophobic region that is blocked by Ile-174 in thrombin.

In the P site the green contour is a hydrophobic site provided by Tyr-60A and Trp-60D in thrombin. There are no residues from trypsin and factor Xa in this region. The red contour is provided by hydrogen bonding from Lys-60F in thrombin. This region is blocked off by trypsin and factor Xa.

In the S1 site, the orange contour is from hydrogen bonding due to Ser-190 from trypsin. The blue and red contours are due to the backbone differences between the loop containing Gly-216 to Gly-219 in thrombin and the corresponding loop in trypsin and factor Xa. Another orange contour is due to hydrogen bonding from Ser-217 in trypsin. These are similar to the features noted by Kastenholz et al. [2] in their treatment of a similar problem.

3.3. Dihydrofolate reductase

In the third problem (Table 3) we examine the difference between bacterial dihydrofolate reductases (DHFRs) and their vertebrate counterparts. The enzymes were superimposed onto 3DFR using all alpha-carbons within 12 Å of methotrexate, the ligand from 3DFR.

There are few differences in the *T* maps seen in Fig. 4. The yellow contour nearest the pteridine ring of methotrexate (center) is a hydrophobic region that is open in chicken and human DHFR. However, the loop containing His-18 and Leu-19 (3DFR numbering) in bacterial DHFRs approaches more closely and blocks this region. On the top of the figure, the blue region is due to hydrogen bonding from the ribose oxygens of bacterial DHFRs. The corresponding oxygens in chicken and human DHFR approach closer and block off this region. The blue–gray region below it is due to hydrogen bonding from Glu-21 in chicken and human DHFR (1BOZ numbering). The red region at the alpha-carboxylate

Table 3
Enzymes for bacterial dihydrofolate problem

PDB code	Description	Source	Activity
3DFR[a]	Dihydrofolate reductase	*Lactobacillus*	1
1AOE	Dihydrofolate reductase	*Candida*	1
1CD2	Dihydrofolate reductase	*Pneumocystis*	1
3CD2	Dihydrofolate reductase	*Pneumocystis*	1
1DYR	Dihydrofolate reductase	*Pneumocystis*	1
1D8R	Dihydrofolate reductase	*Pneumocystis*	1
1DAJ	Dihydrofolate reductase	*Pneumocystis*	1
1RX1	Dihydrofolate reductase	*E. coli*	1
7DFR	Dihydrofolate reductase	*E. coli*	1
1RA2	Dihydrofolate reductase	*E. coli*	1
1RA3	Dihydrofolate reductase	*E. coli*	1
1DG5	Dihydrofolate reductase	*Mycobacterium*	1
1DG7	Dihydrofolate reductase	*Mycobacterium*	1
1DG8	Dihydrofolate reductase	*Mycobacterium*	1
1DR1	Dihydrofolate reductase	Chicken	0
8DFR	Dihydrofolate reductase	Chicken	0
1BOZ	Dihydrofolate reductase	Human	0
1HFP	Dihydrofolate reductase	Human	0
1HFR	Dihydrofolate reductase	Human	0

[a] Reference structure for superposition.

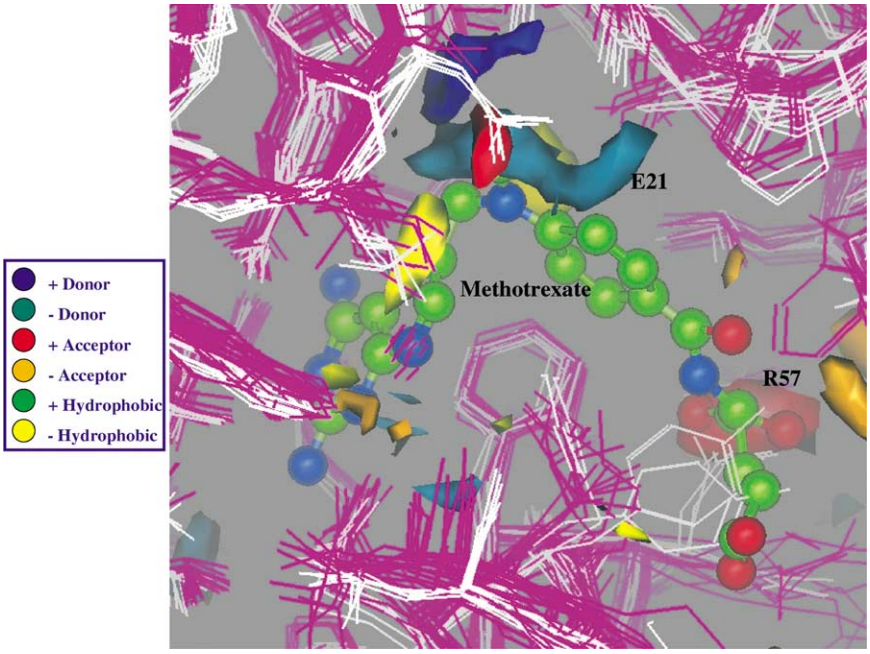

Fig. 4. The *T* maps for the bacterial DHFR problem. Bacterial DHFRs are shown as purple wire, vertebrate DHFRs as white wire. The inhibitor methotrexate is shown as ball and stick. The contour convention is as in Fig. 2.

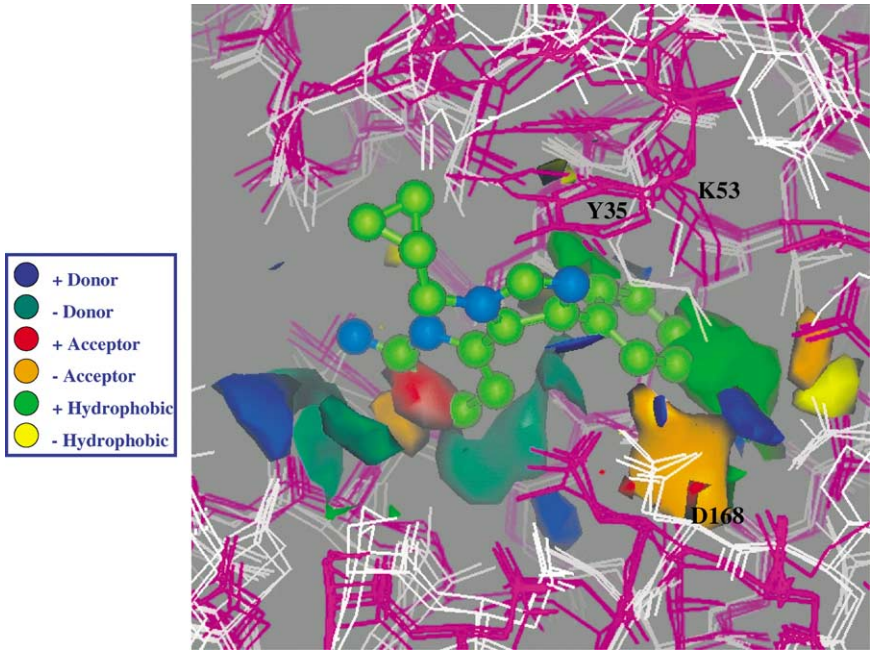

Fig. 5. The *T* maps for the P38 vs. ERK protein kinase problem. P38 kinases are shown as purple wire, ERK kinases as white wire. The inhibitor SB-218655 is shown as ball and stick. The contour convention is as in Fig. 2.

of methotrexate (right) is due to hydrogen bonding from Arg-57, which is conserved in all DHFRs. However, in three of the vertebrate DHFRs (1BOZ, 1DR1, 8DFR) the sidechain of Gln-35 is blocking off this region. This is a reflection of the fact that none of the ligands of vertebrate DHFR reach as far as the alpha-carboxylate of methotrexate and the sidechain of Gln-35 can occupy this space. If we move Gln-35 to a position more consistent with homologous sidechains, the red contour disappears.

3.4. Protein kinases

In the last example, we show the differences between P38 and ERK protein kinases (Table 4). The enzymes were superimposed by alpha-carbons onto 1BMK within 12 Å of the inhibitor for 1BMK, SB-218655. The *T* maps are shown in Fig. 5. The enzymes are oriented so that the conserved Lys-53 (1BMK numbering) is at the top and Tyr-35, the residue at the tip of the Gly-rich loop, is toward the front.

Table 4
Enzymes for P38 versus ERK protein kinase problem

PDB code	Description	Source	Activity
1BLK[a]	Map kinase P38	Human	1
1BL6	Map kinase P38	Human	1
1BL7	Map kinase P38	Human	1
1A9U	Map kinase P38	Human	1
1ERK	Extracellular regulated kinase 2	Rat	0
3ERK	Extracellular regulated kinase 2	Rat	0
4ERK	Extracellular regulated kinase 2	Rat	0

[a] Reference structure for superposition.

The contours on the left are due to differences in the conformation of the loop containing the hinge residues 108–112. These are relatively small differences. The large blue–gray arc of negative donor contour near the center is due to two sources. One source is the carbonyl oxygen of His-107, the conserved H-bond acceptor for N-6 of adenine. Its location is very similar in both types of enzyme. However, the region is blocked by the sidechain of Met-109 in the P38 enzymes. The second source is the sidechain carbonyl of Gln-106 in ERK. The corresponding residue in P38 is Thr.

On the right is a large green positive hydrophobic contour. This is a hydrophobic region that is blocked off by Gln-106 and the gamma-carbon of Ile-104 in the ERK enzymes. This space is occupied by a fluorine in the ligand SB-218655. Also on the right is a large orange negative acceptor contour. This is due to hydrogen bonding from the sidechain amide nitrogen in Gln-106. This region is blocked off by Ile-84 in the P38 enzymes. Interestingly, none of the ligands co-crystallized with these enzymes occupies that region, but there are known P38 inhibitors that put polar substituents or polarizable atoms such as halogens in that location [11].

4. Discussion

We have presented a method FLOGTV that uses the trend vector paradigm to visualize the differences in molecular field maps between closely related enzymes superimposed in a reasonable way. The differences appear to be consistent with a visual inspection of the atomic-resolution enzyme structures. It should be noted in this context that methods

like FLOGTV and GRID/CPCA are very useful because differences between enzymes are very hard to perceive from looking at the structures alone. For instance, in the acid protease problem, it would not be obvious that the complex differences in the flap structure between HIV and two-domain proteases would resolve into a few localized changes in the interaction maps.

It should be emphasized that all methods that compare molecular field maps, regardless of the method used to analyze the maps, depend on having a reasonable superposition and the results depend on what superposition is chosen. This is universally recognized as an issue for small molecules in CoMFA [12] and applies just as much to enzymes for GRID/CPCA and FLOGTV. Superimposing related enzymes is less arbitrary than superimposing small molecules in the sense that there is usually an obvious global correspondence between structures suggested by sequence homology. However, since hydrogen bond minima and similar features in molecular field maps are only about 1 Å wide, differences in coordinate placement of 1 Å become important, and at that level of detail there is an unavoidable sensitivity to arbitrary choices on how the superposition is to be made. This can only partly be overcome by considering more structures. To give a highly simplified example, say we have two enzymes A and B, with two very similar pockets 1 and 2. The distance between A1 and A2 is 2 Å smaller than the distance between B1 and B2. We can superimpose A1 onto B1 exactly. In that case, we will see no difference contours at pocket 1, but will see many differences at pocket 2. On the other hand, we can superimpose A2 onto B2, in which case the differences will all be at pocket 1. If we do our best to simultaneously superimpose A1 onto B1 and A2 onto B2, we will see differences in both pockets. It becomes a matter of judgment which is the best frame of reference for a particular problem. Additional complications sometimes arise with sidechain positions, as in our DHFR example. It is a matter of judgment whether to keep the structures as given or to "correct" the sidechain positions to make them more consistent among the enzymes.

GRID/CPCA and FLOGTV approach the same goal in different ways. GRID/CPCA suppresses the "noise" by finding the principal components in map descriptor space, the idea being that the most variable features among all the enzymes are most likely to capture a real difference. Differences between sets of enzymes are embodied in the vectors pointing from one set to another in the reduced space defined by the first two principal components. The loadings of these vectors onto the original maps generate a new set of maps that can be contoured. In contrast, FLOGTV finds the differences between the specified enzymes as a vector in the original descriptor space. Noise is suppressed by finding contour levels well above chance correlations.

Much of the appeal of FLOGTV has to do with its simplicity. We avoid constructing huge matrices and having to extract principal components from them. T is a simple sum. This makes the computations fast enough for us to analyze tens of enzymes in a few minutes on a modest workstation. T is already a set of maps that can be directly contoured. The probes remain unmixed, so scaling is not necessary. Our use of FLOG maps rather than GRID maps has the additional advantage that there are many fewer probe types.

Another appealing feature of trend vectors is that, in correlating all descriptors with activity, we avoid a limit of principal component analysis that is a potential problem for any type of QSAR. Taking the first few principal components selects for descriptors that account for the most variability between all molecules. Since the activity is ignored at that step, it is not guaranteed that the variable that best distinguishes actives from inactives will be included in the first principal components.

One of the most useful ideas from the trend vector paradigm is the ability to define the level of "noise" by scrambling the activities. This is of great utility when there are at least a moderate number of enzyme structures being compared, as there are in this work. However, as the number of structures becomes very small, say <5, it may be impossible to pick the signal from the noise, and some other method of deriving sensible contours must be used.

Mathematically, the trend vector is proportional to the first component in a partial-least-squares (PLS) fit [9]. While we can in principle extend FLOGTV with PLS, we have not done so for two reasons. First, PLS would require scaling of the maps for different probes, something that we prefer to avoid. Second, the number of enzymes in a typical problem is usually not large enough such that a second PLS component would be statistically justified.

The use of the trend vector to analyze molecular field maps has precedent. A method called SOMFA has been proposed by Robinson et al. [13] to analyze fields surrounding small molecules (i.e. CoMFA fields). Although not called by that name, SOMFA is mathematically identical to the original trend vector formulation [8]. In that formulation "activities" are mean-centered, as opposed to normalized as in later formulations [9] and this work. Robinson et al. recognized the simplicity of the approach relative to the PLS method, which is standard for CoMFA analyses.

Acknowledgements

FLOGTV was written with the in-house modeling system MIX. The authors thank the MIX team for their tireless efforts.

References

[1] M. Pastor, G. Cruciani, A novel strategy for improving ligand selectivity in receptor-based drug design, J. Med. Chem. 38 (1995) 4637–4647.

[2] M.A. Kastenholz, M. Pastor, G. Cruciani, E.E.J. Haaksma, T. Fox, GRID/CPCA: a new computational tool to design selective ligands, J. Med. Chem. 43 (2000) 3033–3044.

[3] P.J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, J. Med. Chem. 28 (1985) 849–857.
[4] M.D. Miller, R.P. Sheridan, S.K. Kearsley, SQ: a program for rapidly producing pharmacophorically relevant molecular superpositions, J. Med. Chem. 42 (1999) 1505–1514.
[5] G.M. Maggiora, D.C. Rohrer, J. Mestres, Comparing protein structures: a Gaussian approach to the three-dimensional structural similarity of proteins, J. Mol. Graph. Model. 19 (2001) 168–178.
[6] M.D. Miller, S.K. Kearsley, D.J. Underwood, R.P. Sheridan, FLOG: a system to select quasi-flexible ligands complementary to a receptor of known three-dimensional structure, J. Comput.-Aided Mol. Des. 8 (1994) 153–174.
[7] R.S. Bohacek, C. McMartin, Definition and display of steric, hydrophobic, and hydrogen-bonding properties of ligand binding sites in proteins using Lee and Richards accessible surface: validation of a high-resolution graphical tool for drug design, J. Med. Chem. 35 (1992) 1671–1684.
[8] R.E. Carhart, D.H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure–activity studies: definition and application, J. Chem. Inf. Comput. Sci. 25 (1985) 64–73.
[9] R.P. Sheridan, R.B. Nachbar, B.L. Bush, Extending the trend vector: the trend matrix and sample-based partial-least-squares, J. Comput.-Aided Mol. Des. 8 (1994) 323–340.
[10] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[11] N.J. Liverton, J.W. Butcher, C.F. Claiborne, D.A. Claremon, B.E. Libby, K.T. Nguyen, S.M. Pitzenberger, H.G. Selnick, G.R. Smith, A. Tebben, J.P. Vacca, S.L. Varga, L. Agarwal, K. Dancheck, A.J. Forsyth, D.S. Fletcher, B. Frantz, W.A. Hanlon, C.F. Harper, S.J. Hofsess, M. Kostura, J. Lin, S. Luell, E.A. O'Neill, C.J. Orevillo, M. Pang, J. Parsons, A. Rolando, Y. Sahly, D.M. Visco, S.J. O'Keefe, Design and synthesis of potent, selective, and orally bioavailable tetrasubstituted imidazole inhibitors of P38 mitogen-activated protein kinase, J. Med. Chem. 42 (1999) 2180–2190.
[12] Y.C. Martin, 3D QSAR: current state, scope, and limitations, Perspect. Drug Discovery Design 12–14 (1998) 3–23.
[13] D.D. Robinson, P.J. Winn, P.D. Lyne, W.G. Richards, Self-organizing molecular field analysis: a tool for structure–activity studies, J. Med. Chem. 42 (1999) 573–583.