Journal of Computer-Aided Molecular Design, 11 (1997) 601–612.
KLUWER/ESCOM

J-CAMD 420

Alignment of flexible molecules at their receptor site using 3D descriptors and Hi-PCA

Anders Berglund[a,*], Maria Cristina De Rosa[a,**] and Svante Wold[b]

[a]*Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford OX1 3QU, U.K.*
[b]*Research Group for Chemometrics, Department of Organic Chemistry, Umeå University, S-901 87 Umeå, Sweden*

Received 18 April 1997
Accepted 20 August 1997

Keywords: Molecular alignment; Flexible alignment; Grid

Summary

Three categories of molecular flexibility are defined. A novel method of aligning partly flexible molecules with each other is described. The binding mode of one of these molecules to its receptor site was already well known from previous crystallographic studies, and this known binding mode was used to predict the binding mode of the other molecules at their receptor. The predictions were checked by comparison with previous observations, and were correct. Two novel methods were combined in this research. It was necessary to take account of the conformational changes which occur when each ligand molecule binds to the protein, and a new release of programme Grid was used for this. It was also necessary to analyse the Grid results in order to distinguish the role of each chemical group at the receptor site. This was done by applying hierarchical principal component analysis (Hi-PCA) methods to the descriptors obtained from Grid.

Introduction

A common problem in 3D QSAR (three-dimensional quantitative structure–activity relationships) is to predict the alignment of different drug molecules with their receptor site on a biological macromolecule. This problem is particularly difficult when the drug molecules are flexible. The ideal solution to this problem would be to co-crystallise each compound with the protein and observe its structure when bound to its receptor. However, this is only possible after the molecules have been synthesised, and the co-crystallisation method cannot be used to predict the alignment or the binding of a molecule before it has been prepared.

Many procedures have been suggested for aligning a set of molecules so that mutually consistent descriptors can be generated for a 3D QSAR model [1–5]. We propose that each molecule should be subdivided into two distinct parts: a rigid core and a flexible region or regions. In the present work, the porphyrin ring system of heme constitutes the core, and the propionate side chains are flexible. In general, however, one can distinguish at least three different classes of molecular flexibility:

(1) All rigid (AR) molecules such as some steroids in which there is only a core.

(2) Rigid and flexible (RAF) molecules such as the hemes.

(3) All flexible (AF) molecules such as straight-chain paraffins.

The alignment problem is different for each of these classes. If the molecules are AR, and they also have a common core like the steroids, there may be no significant alignment problem at all [6–8]. If the molecules are RAF, and they have a common core, the alignment is achieved by rotation and translation of the core. If the molecules are AF, they may be aligned by NMR or X-ray observations of the bound molecules at their receptor site, and some promising theoretical studies have also been made [9], but we are not aware of any satisfactory method of general applicability.

*To whom correspondence should be addressed at: Research Group for Chemometrics, Department of Organic Chemistry, Umeå University, S-901 87 Umeå, Sweden.

**Present address: Istituto di Chimica e Chimica Clinica, Universita’ Cattolica del S. Cuore, Largo F. Vito 1, I-00168 Rome, Italy.

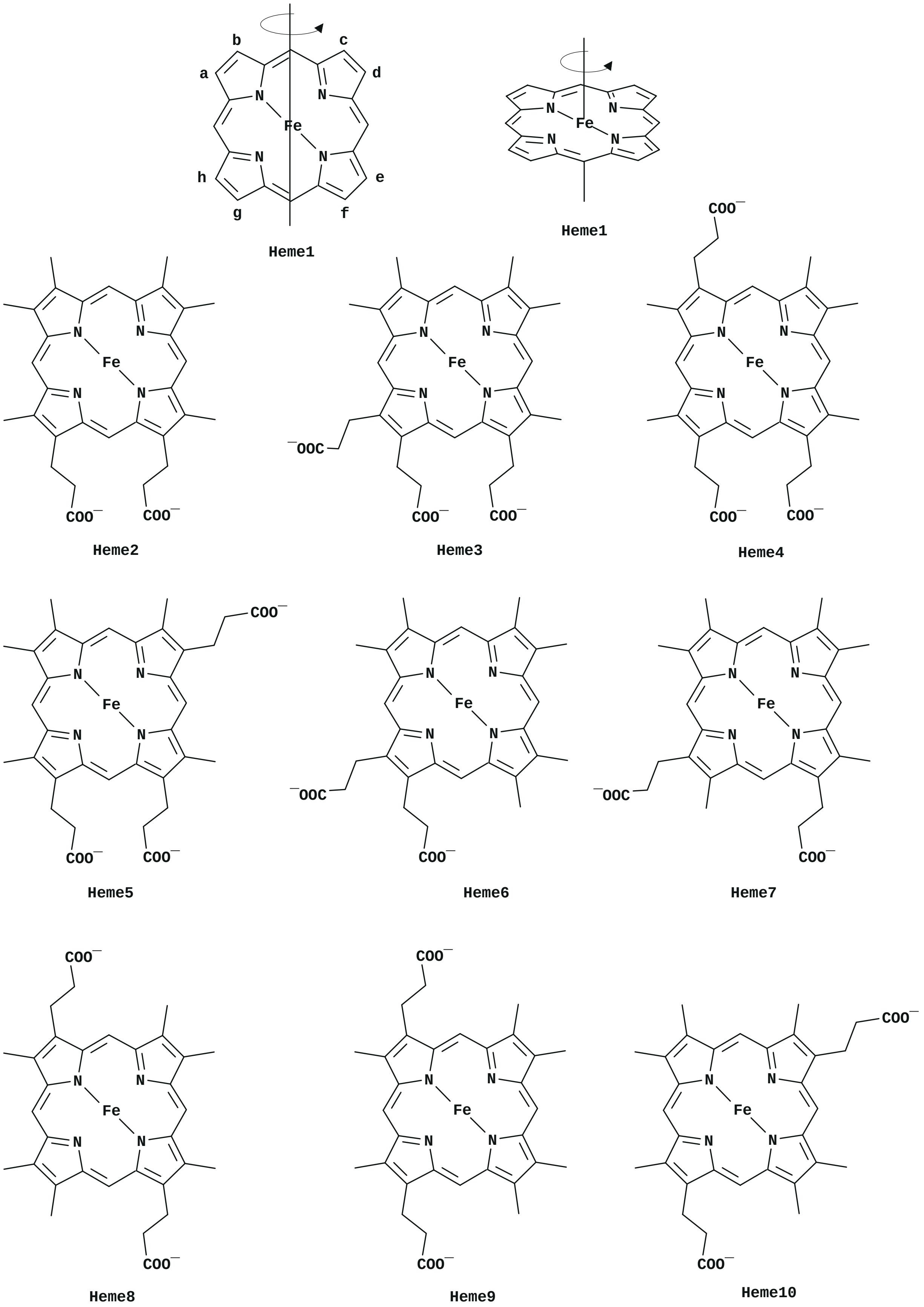

Fig. 1. The nine different modified hemes. The same names have been used as in the paper from Hauksson et al. [24]. The ‘parallel’ and the ‘perpendicular’ axes of heme1 are shown. See main text for further information.

In fact, the alignment of the core is a necessary requirement for RAF alignment, but it is not always sufficient. Flexible side chains such as those of heme can assume a large number of conformations and the correct alignment must allow for this. One approach has been to find the conformation of lowest energy for each molecule, but close contacts in the binding cleft will frequently distort the side chains from these idealised positions, and the descriptors calculated from energy-minimised structures may therefore be misleading. Such methods have worked well in particular cases [10–13], but they often make the implicit assumption that flexible regions have the same conformation in vacuum as in the binding pocket. An alternative approach has been to find all different possible conformations. Jones et al. [14] have applied genetic algorithms for selecting plausible conformations. Cluster analysis has also been used [15,16] to find plausible conformations. Another way to find the best conformation is to align the molecules in the regression step in 3D QSAR [17], although Kroemer and Hecht [18] have shown that spurious models may be found when this method is used.

In the present work, a four-stage approach has been applied to a group of partly flexible molecules. The first step was to find all plausible arrangements of the common core for each heme molecule. The second step was to generate quantitative descriptors in order to describe each of these arrangements. The programme Grid [19,20] was used for this, and the descriptors were then used as input for an hierarchical principal component analysis (Hi-PCA) [21,22] (step 3). One of the molecules was chosen as a template, and the best alignment was found by calculating the distance in Hi-PCA space between that template and each of the other molecules (step 4). In this last step, it was assumed that all the molecules do in fact bind at the same site, and the distance was measured in Hi-PCA space.

There are significant differences between this procedure and another popular method [23]. Our approach was not restricted to a steric/electrostatic description of the molecules, but simulated chemical probes were used with terms for polarisability and directed hydrogen bonding, and the flexibility of the molecules was also taken into account. The relevant information was extracted from the crude data by Hi-PCA, and the resulting latent variables were then used for calculating the distances in Hi-PCA space. The use of Hi-PCA made it possible to identify the important chemical groups in the receptor, and identify various trends in the alignment. It also yielded results which were less sensitive to noise and irrelevant information in the data than using the crude data for calculating the distances.

Data

To choose an appropriate data set, different criteria have been used. Since this is a novel method, the selected data must not complicate the reasoning and understanding of the steps. A bigger and more challenging data set would require more special arguments and assumptions which would complicate the interpretation of the results. The right binding modes must of course also be available for verifying the proposed binding mode. The selected data set fulfils these criteria without being too trivial.

A group of nine different heme analogues (Fig. 1) were considered as ligands of myoglobin. Each of these molecules has a relatively rigid porphyrin core to which either two or three propionate side chains are attached. The number and position of these side chains vary between the different molecules.

The NMR observations of Hauksson et al. [24] were used to check the correctness of the predicted binding mode of each compound to myoglobin, and we have used the nomenclature of Hauksson et al. for the different molecules and binding modes. The influence of altering the distance between Grid points has also been studied. The effects of small changes in the porphyrin structure have been investigated, in order to test if the conclusions depend critically on the assumption that the molecules do have a perfectly rigid common core. This is why two different alignments are presented even though the second is more correct than the first.

Methods

Strategy

In order to deal with the nine hemes studied by Hauksson et al., we have used the following procedure. For each heme, all plausible binding orientations at its receptor were first generated, as detailed in the Results and Discussion section. Due to the high symmetry of heme, a relatively small number of orientations were required for each molecule, and these were all superimposed on each other.

Each molecule was then described by Grid [19,20] using several different probes (Table 1). These were selected in order to mimic the known chemical nature of the binding pocket in myoglobin. The selection ensured that electrostatic, steric and chemical properties, such as hydrogen bonding were all considered. The relevant information from the Grid maps was then extracted with Hi-PCA to give a few principal components describing the differences between all the molecules in all their orientations. These principal component scores were used for calculating the distance between the template and the rest of the molecules.

Multivariate characterisation

Version 15 of Grid [25] has been used in order to deal with the flexible side chains of heme. This version works like previous versions by computing the interaction energy of a probe at every Grid point on an orthogonal matrix

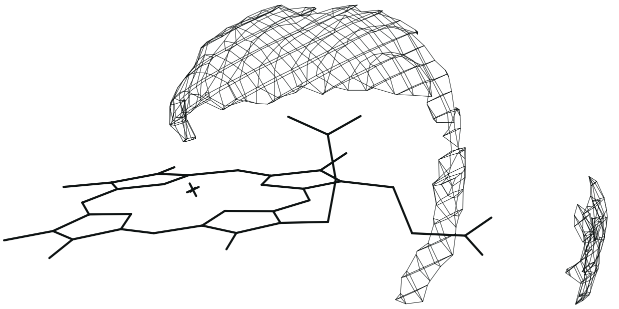

Fig. 2. Grid contours with heme2 as target and N3+ as probe. The side chains are not treated as flexible. The energy contour level is −5.8 kcal/mol.

of points around the molecule of interest. Probes are chemical groups such as methyl, aliphatic hydroxyl or NH3+ amine, and the interaction energy at each Grid point is calculated between the chosen probe and every atom of the molecule. The positions of all the core atoms are fixed, but in v. 15 of Grid flexible side-chain atoms are allowed to move before the energy at a Grid point is computed. If a flexible heme atom is attracted by the probe, it moves closer until it reaches an optimum distance from the probe at its Grid point, or until it cannot move any further because it is constrained by the chain of bonds linking it to the core of the molecule. An appropriate allowance is then made for entropic effects. If, on the other hand, the atom is repelled, it moves away as far as possible from the probe until it is restrained by the chain of bonds.

This approach gives a large volume of good attractive interactions between the molecule and the probe, since the flexible side-chain atoms always move in order to find the most favourable positions for interaction with the probe at each particular Grid point. Figure 2 shows a Grid map for heme with NH3+ as the probe when the new flexible option in v. 15 was *not* used. In this case, the Grid contours show where the probe would make good hydrogen bonds with the carboxyl groups of the propionate side chains in their fixed positions as specified by the input file for heme. Two distinct interaction regions are defined whose shape and position depend critically upon the particular heme conformation chosen for investigation.

Figure 3 shows the findings with the same probe and with the flexible option used. This Grid map was generated at the same energy level as Fig. 2, but the contoured region is now much bigger since the side-chain atoms tend to follow the probe, and move to the best positions for interaction. The contours are also symmetrical because they are not dependent on the initial side-chain positions as defined by the input file, and the computed energies therefore give a more realistic representation of the interactions that may take place in the binding pocket.

The selection of the appropriate Grid spacing is important, and the Grid spacing was varied from 1 Å up to 4 Å in the present work. These different Grid spacings have been compared, and the most suitable Grid spacing has been estimated.

Ten different probes were selected (Table 1), which represent chemical groups that are present in the myoglobin pocket. This will ensure that the molecules are described in a manner that mimics the known chemical nature of the pocket in the best way.

Matrix generation and pretreatments

Each Grid map was unfolded to form a vector. This was done for each probe and for each heme molecule. All the Grid maps for a molecule were concatenated, resulting in a single vector for each molecule, and these vectors were used to build a matrix. Variables with a standard deviation less than 0.05 kcal/mol were deleted. This value was estimated from the known approximations in Grid, and by comparing the solutions from different standard variation cut-offs. The matrix was then columnwise centred before the main calculations began.

Statistical methods

Describing a complex system often generates a mass of data coming from different sources with different properties. It is then usually helpful to subdivide the data into blocks according to the source of the data. The blocks may differ from each other in many ways, some consisting of just a few variables while others might consist of many thousands of variables.

Techniques like principal component analysis (PCA) consolidate the initial data into a form which can be more easily used and understood [26]. The original multidimensional space is reduced down to a few dimensions called principal components which still contain the main

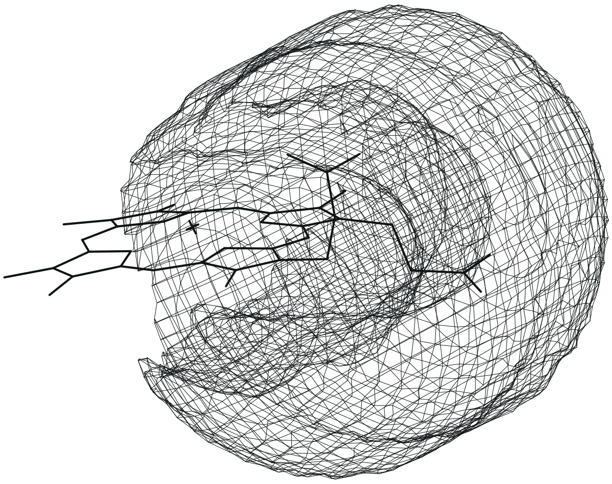

Fig. 3. Grid contours with heme2 as target and N3+ as probe. The side chains are treated as flexible. The energy contour level is −5.8 kcal/mol.

TABLE 1
LIST OF THE 10 SELECTED PROBES

No.	Code	Description
1	C3	Methyl group
2	OH	Phenolic hydroxyl group
3	O1	Aliphatic hydroxyl group
4	N3+	sp^3 cationic NH_3 group
5	O::	Carboxy oxygen atom
6	N2=	sp^2 cationic NH_2 group
7	N1=	sp^2 cationic NH group
8	N:=	sp^2 nitrogen with lone pair
9	N1	Amide NH group
10	DRY	Hydrophobic probe

variation. Each component can be displayed and analysed separately, and its role can often be described in simple terms. However, traditional PCA is not always satisfactory when the variables are subdivided into blocks of differing size, weight and variance.

Hierarchical-PCA (Hi-PCA) provides a good method for handling variables in blocks, and hierarchical models have already been used successfully for analysing process data [27]. In Hi-PCA, each block of input data is modelled separately by a projection model. This accounts for the variation within each individual block. The block scores are then used as 'super variables' on a higher level of the model where a PC-like model estimates the correlated structure of the blocks, i.e. of the super variables. In this way, Hi-PCA operates on two levels, as illustrated in Fig. 4. On the lower level, the relationship between variables in the same block is modelled by block models. On the higher level, the relationship between the blocks is modelled. This provides two levels of interpretation. The lower block level shows the detail and the upper level provides an overview. On both levels these plots, e.g. loading plots or score plots, provide tools for diagnosing the data set.

Hi-PCA is a method suitable for 3D QSAR studies because there is a natural blocking of the data, particularly if more than one probe is used. Hi-PCA will also not penalise probes which happen to have a smaller variance as much as ordinary PCA does. At one level, it is possible to assess the importance of each Grid point for each probe, while the relative performance of the individual probes can be determined at a higher level.

The theory of Hi-PCA is briefly described here (details are in Refs. 21 and 22). Hi-PCA is based on the NIPALS algorithm [28] as is PCA. The steps in Hi-PCA can briefly be described as follows (see Fig. 4). For each block, one sub score-vector (**r**) is calculated and moved up to the super level forming the super X-matrix (**R**). For this new matrix, a new super score-vector (**t**) is calculated with a corresponding super loading-vector (**w**). The super score-vector is used for calculating a sub loading-vector (**p**). This is then used for calculating a new sub score-vector that again is moved up to the super level, and the process is repeated until it converges.

The block loadings in Hi-PCA correspond to the Grid points in the Grid box. These can be displayed, for each probe, in 3D space, generating a map of loadings. These will show important regions around the molecules for that probe. Of course, this can also be done with classical PCA, but with Hi-PCA there is an extra super level, and the super loadings show how each probe is related to each other. This option has not been available before, and improves the understanding and interpretation of the models.

There are also other differences. In Hi-PCA, the super scores are normalised instead of normalising the loadings in classical PCA. The super scores are normalised because the super scores appear both in the sub and the super level, and the super loadings (**w**) are not orthogonal as they are in ordinary PCA.

Estimating the number of significant components

A significant component is one which explains the different positions adopted by the side chains of heme, and the number of significant components for Hi-PCA can be estimated in different ways. In this paper, the explained variance, the length of the super loading vector, and the chemical meaning of the components have been used. The variance in the data set can be divided into three different parts. One part relates to the variation of the side-chain position; this is the major variance and the most interesting one. There is further variance due to the fact that the molecules are not perfectly symmetrical, and there will also be some variance due to noise.

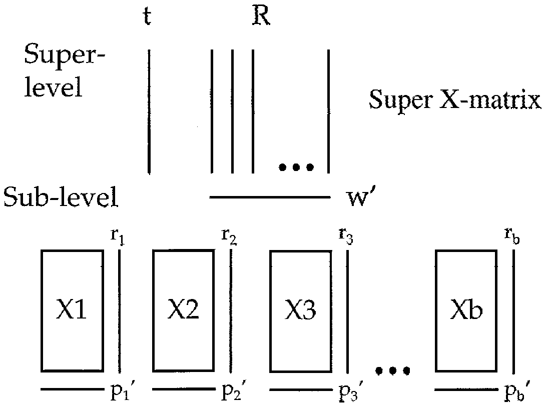

Fig. 4. A schematic of how Hi-PCA is built up of two layers. Each block, on the sub level, corresponds to the Grid maps from one probe, $\mathbf{X}_b$, where b is the block number. For each of these X-blocks, a loading vector ($\mathbf{p}_b$) and a score vector ($\mathbf{r}_b$) are calculated. The score vectors from the sub level are moved up to the super level and form the super X-matrix, **R**. A super loading vector (**w**) and a super score vector (**t**) are calculated from the super X-matrix. The super score vector, **t**, is then projected down on each separate X-block, on the sub level, to give a new sub loading vector. A new sub score-vector is then calculated for each X-block and the whole procedure is repeated until convergence.

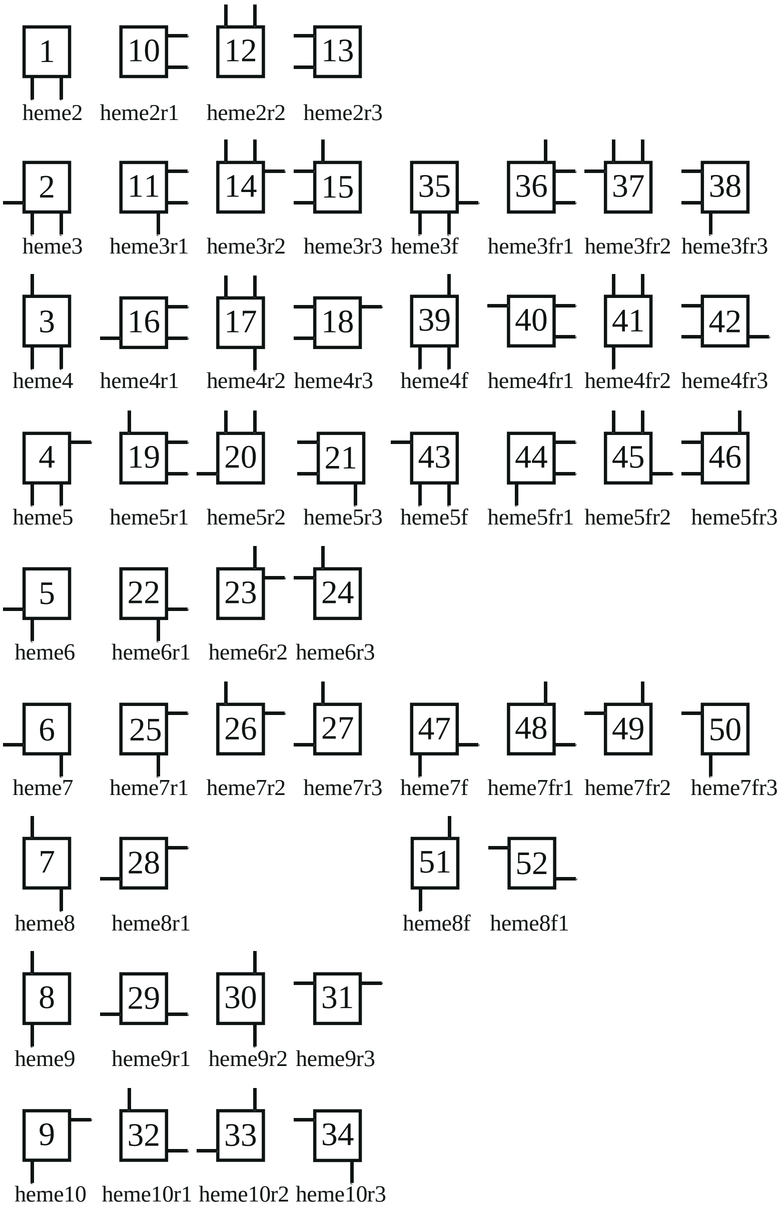

Fig. 5. Graphical representation of the conformers for the nine heme molecules. The square represents the heme skeleton and the lines represent the positions of the propionates. The names are made up of three parts. For example, for molecule heme3fr2, the first part (heme3) is the molecule name and the second part (f) indicates if the molecule has been rotated 180° around the parallel axis. The last part (r2) indicates the number of 90° rotations around the perpendicular axis that have been made.

The first two variances may well be statistically significant since they explain systematic variance between the molecules, while the last will not be statistically significant. To distinguish between the two statistically significant variances, a chemical interpretation must be used.

Calculating the differences between the compounds

The differences between the template (heme2) and the rest of the molecules are calculated using the super score values. The distance between the template and the rest of the molecules for each molecular orientation is calculated in the hyperspace of the super score-vectors.

Results and Discussion

Structure generation and alignment

The 'wild-type' heme structure was taken from myoglobin [29], and heme2 was chosen to be the reference structure since it is most like wild-type heme. The conformation of the propionate chains was not modified, and no modification was made to the porphyrin ring, but the vinyl side chains of wild-type heme were truncated to methyl groups to give the same molecule as Hauksson et al. used. To generate all the other hemes 3–10, the propionate chains were moved into appropriate positions.

This was done by copying one of the propionate chains from the original molecule, and then reattaching it in the appropriate position for each new heme. Since porphyrin is highly symmetrical, the number of plausible conformations for each molecule is limited. It is assumed that the porphyrin ring always has the same position in the myoglobin pocket, and we have therefore generated the four, or eight, conformations necessary for each heme molecule, giving a total of 52 possibilities in all; these are shown graphically in Fig. 5.

Two different approaches were used in order to generate the 52 structures. In the first method (first alignment), each molecule was rotated by 90° increments around the symmetry axis perpendicular to the ring plane which passes through the Fe atom in the centre of the porphyrin ring (Fig. 1). The molecules were also rotated 180° around the parallel symmetry axis, but this was only necessary for heme3, heme4, heme5, heme7 and heme8. Since the porphyrin ring is not perfectly planar, a perfect realignment was not possible after this 180° rotation, and the parallel rotation axis was arbitrarily chosen to lie in the best least-square plane of the four nitrogen atoms. As a result, the nitrogens were almost perfectly superimposed, but the iron and carbon atoms were not. However, the iron makes a large interaction with His^{64} of the apoprotein and it was therefore moved so it had the same position for all molecules. This first alignment is shown for heme3 and heme3f in Fig. 6, which shows that the carbon atoms are not perfectly superimposed on each other when this alignment procedure is used.

A second method of alignment (second alignment) was therefore tried, and this was done in the same way as the first method except that the molecules were not rotated around the parallel axis. Instead the molecules were remade by moving the propionate side chains into the correct position without altering the ring. This generated the same conformers as in the first alignment, with the difference that there were no problems with the curvature of the porphyrin ring, and all molecules therefore have the same curvature in this second alignment.

Grid spacing

Choosing the appropriate resolution is a question of speed, the quantity of data, and the necessary precision for the model. The computer time was perfectly acceptable for this job with 10 probes and 52 targets and a Grid spacing of 1 Å. Ideally, even higher resolutions might be better, but this might provide so much data that the Grid calculations and the analysis of the results would then become too time-consuming. On the other hand, if the resolution was too low some information might be lost since there might be too few points. An energy minimum might be lost, or small differences between molecules might not be identified. There is an optimum between precision and speed, giving enough information with reasonable speed. This optimum resolution depends on the data set, on what the data are to be used for, and also on available computer facilities.

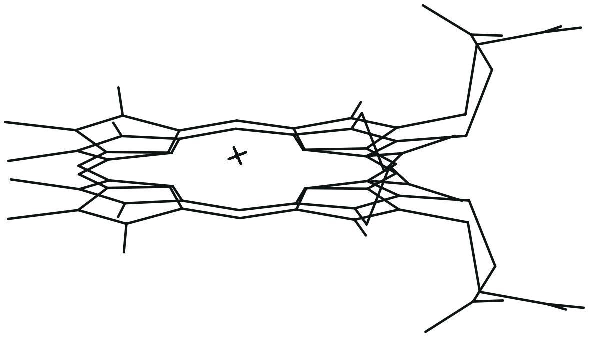

Fig. 6. The first alignment of heme3 and heme3f.

First alignment

Hi-PCA calculations

Five Hi-PCA models were calculated for the different Grid spacings indicated in Table 2. As expected, the results showed that there were multiple eigenvalues. This occurs when the size of the eigenvalues is pairwise similar. This is expected since the molecules have a high degree of symmetry, and the only appreciable variance is related to the position of the propionate chains. The effect this has is that the first two components are not individually well resolved, but seen as a pair they are perfectly defined. It is difficult for the algorithm to find the direction with one largest variance, since there are two directions in which the variance is almost equal.

This can be exemplified by comparing an ellipsoid and a square. In the ellipsoid, there is no problem finding the largest distance between two opposite points, and there is only one solution. The solution for the square is more

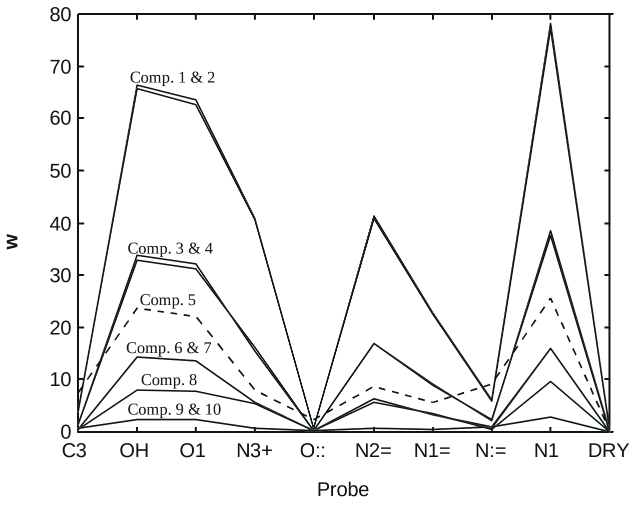

Fig. 7. Super loadings, **w**, for 1 Å Grid spacing for the first alignment. Each line corresponds to a component.

TABLE 2

STATISTICAL RESULTS FROM THE Hi-PCA MODELS FOR THE DIFFERENT GRID RESOLUTIONS FOR THE FIRST ALIGNMENT

A[a]	1 Å model		1.5 Å model		2 Å model		3 Å model		4 Å model	
	SSX[b]	||w||[c]	SSX[b]	||w||[c]	SSX[b]	||w||[c]	SSX[b]	||w||[c]	SSX[b]	||w||[c]
1	0.262	136.0	0.263	133.3	0.259	131.2	0.260	139.5	0.282	130.0
2	0.260	134.5	0.260	131.8	0.259	130.4	0.254	133.6	0.233	105.0
3	0.111	65.3	0.112	64.6	0.110	62.8	0.120	73.1	0.125	65.7
4	0.111	64.0	0.111	62.6	0.109	62.4	0.108	65.4	0.100	50.7
5	0.075	44.9	0.075	43.6	0.078	46.4	0.068	44.5	0.067	38.4
6	0.043	26.9	0.043	26.5	0.044	27.2	0.049	32.3	0.051	28.6
7	0.043	26.9	0.042	26.0	0.042	25.6	0.046	30.3	0.044	23.5
8	0.035	17.1	0.035	16.7	0.034	16.2	0.036	17.1	0.031	14.3

Variables with a standard deviation less than 0.05 were deleted.

[a] Component number.

[b] Explained variance for each component.

[c] Length of the super loading vector, **w**.

complicated, because the two diagonals are both equally large. The fact that porphyrin is square in shape makes this analogy particularly appropriate.

This pairwise effect was stronger in the models with a finer resolution. In those with a lower resolution, the Grid points were not evenly spread around the molecule, and so the different sides of a symmetrical molecule were somewhat different in the analysis, according to the position of the Grid points. This effect can be used to estimate the most appropriate Grid spacing. As seen in Table 2, the first two pairs of components with finer resolution are similar in explained variance and in the length of **w**, ||**w**||. The results show that 1 Å is an adequate spacing for a good model, and from now on all the plots are generated using the 1 Å results.

Super loadings

The super loading plot (Fig. 7) shows the relative importance of each probe for each component. The three most important probes are OH, O1 and N1. All these are non-charged and are able to donate a hydrogen bond which can interact directly with the propionate oxygens. The positively charged probes N3+, N2= and N1= are not as important as the former three, but they are still reasonably important. However, probes which cannot donate hydrogen bonds (C3, O:: and N:=) have small **w** values. A more detailed discussion about different Grid probes and the actual interactions formed in the complexes will be given elsewhere (M.C. de Rosa, A. Berglund and P.J. Goodford, in preparation).

The least selective probe is the DRY probe (hydrophobic) which interacts with lipophilic surfaces of the molecule. This is because the DRY probe gives completely different Grid maps compared to the other probes. It finds the hydrophobic regions of heme, and these are mostly on the common core where they do not differentiate between the locations of the polar side chains. The energy minimum for the DRY probe is also rather flat, and is not clearly located in a well-defined position. Furthermore, the hydrophobic interactions are relatively weak, and their variance is correspondingly small. For all these reasons, therefore, the DRY probe has little influence on the Hi-PCA model.

At first sight, it is perhaps surprising that the N1 probe ranks higher than N3+, because the latter makes much more powerful interactions with carboxy groups. It can donate three hydrogen bonds as opposed to one for N1, and can also attract the carboxy oxygens electrostatically. However, these electrostatic interactions only have a modest distance dependence and they are isotropic, so they do not distinguish very well between the different arrangements of the propionate side chains. The importance of N1 can be explained by the fact that its interaction with the target is more specific. Hydrogen bonding is a very specific interaction both in direction and dis-

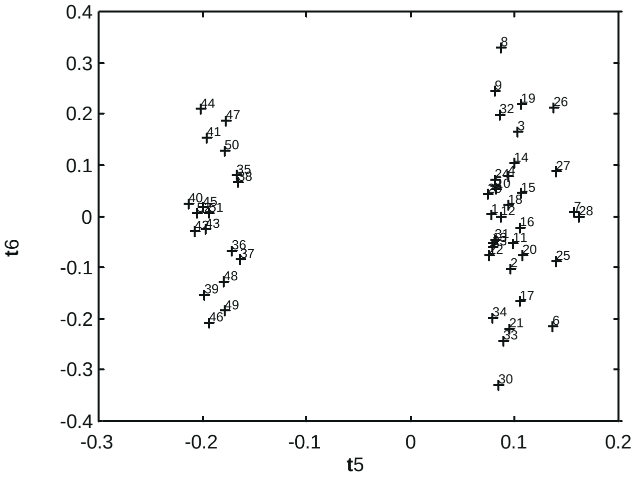

Fig. 8. Super score plot **t**5 versus **t**6 for the first alignment. The plot shows two different classes of hemes. To the right are the molecules which have not been rotated around the parallel axis and to the left are those which have been rotated. This apparent difference is an artefact caused by the first alignment method (see text).

TABLE 3
STATISTICAL RESULTS FROM THE Hi-PCA MODELS FOR THE DIFFERENT GRID RESOLUTIONS FOR THE SECOND ALIGNMENT

A[a]	1 Å model		1.5 Å model		2 Å model		3 Å model		4 Å model	
	SSX[b]	‖**w**‖[c]	SSX[b]	‖**w**‖[c]	SSX[b]	‖**w**‖[c]	SSX[b]	‖**w**‖[c]	SSX[b]	‖**w**‖[c]
1	0.281	141.7	0.283	143.0	0.281	136.5	0.277	139.4	0.304	136.1
2	0.277	139.5	0.277	139.2	0.276	134.3	0.272	134.8	0.243	105.4
3	0.117	65.0	0.118	66.9	0.117	64.1	0.127	73.5	0.124	62.1
4	0.127	66.5	0.114	64.5	0.113	62.2	0.113	64.6	0.101	49.4
5	0.045	27.2	0.045	27.6	0.046	27.6	0.053	33.4	0.054	29.3
6	0.045	27.6	0.045	27.4	0.043	25.3	0.049	30.4	0.046	24.1
7	0.037	17.8	0.032	18.1	0.035	17.0	0.033	21.0	0.028	15.6
8	0.028	16.6	0.033	16.5	0.029	15.8	0.037	16.2	0.033	14.3

Variables with a standard deviation less than 0.05 were deleted.
[a] Component number.
[b] Explained variance for each component.
[c] Length of the super loading vector, **w**.

tance to the target. The electrostatic interaction for N3+ is more blurred since those isotropic interactions have a long range, and this smears out the overall interaction with the target.

Another interesting aspect with the first alignment method is that the first four components appear pairwise (Fig. 7), while the fifth component does not. This can also be seen in Table 2. The C3, O:: and N:= probes have an optimum in this fifth component (Fig. 7), while for the rest of the probes the optimum is in component one. This shows that the fifth component, the broken line in Fig. 7, is qualitatively different compared to the first four, in that it describes steric interactions with the target. Because of this difference, the fifth component super score was investigated further.

Super scores

In Fig. 8 the super scores for principal components five and six are plotted against each other. Each point corresponds to one molecule, and the fifth component clearly divides the molecules into two separate groups. The molecules to the left are those which have been rotated around the parallel axis, and as can be seen in Fig. 6, the carbon skeleton no longer superimposes perfectly after this rotation when the first alignment method is used. Component five therefore seems to detect this inadequacy of the first alignment method, and this interpretation is confirmed when the findings from the second alignment method are similarly analysed (see below).

The super score plots are therefore important for finding structural differences between molecules that are induced by a faulty alignment. Reducing the multi-dimensional Grid space to a few principal components helps to enhance the understanding of the problem. So, for this example, it is clear that the first four principal components are related to the side-chain position while the fifth is induced by the first alignment method. The question then arises as to whether the first four components are enough or if the sixth and seventh components are also needed to make good predictions. In order to answer this question, new models were calculated on the second alignment.

Second alignment

New models were calculated using data from the second alignment method, and the statistics for the different Grid spacings are given in Table 3. The results from this data set are similar to the ones from the previous model, but principal components five and six are now pairwise similar, and there is no unpaired component corresponding to component five in the previous analysis. The relative importance of the different probes is the same in both models, with uncharged hydrogen-bonding probes showing the most significant effect; this was expected since the main difference between the molecules is still in the positions of the side chains.

Estimating the number of principal components

The next requirement is to estimate the number of principal components that are chemically significant. This can be done in different ways. One method is to use the fact that heme2 is symmetrical around the parallel axis, as seen in Fig. 2, which implies that the difference between, for example, heme3 and heme3f should be small, since heme2 does not distinguish between positions e and h. This is true for all flipped compounds and some of the rotated molecules. If it was not true, some of the components would explain variances due to other structural differences. They might come from the fact that the molecules were not minimised before the generations of the different alignments, but this should not be a serious problem as long as the main variation comes from the side-chain positions. The first four principal components were used to calculate the differences, since these do not separate, for example, heme3 and heme3f. These four

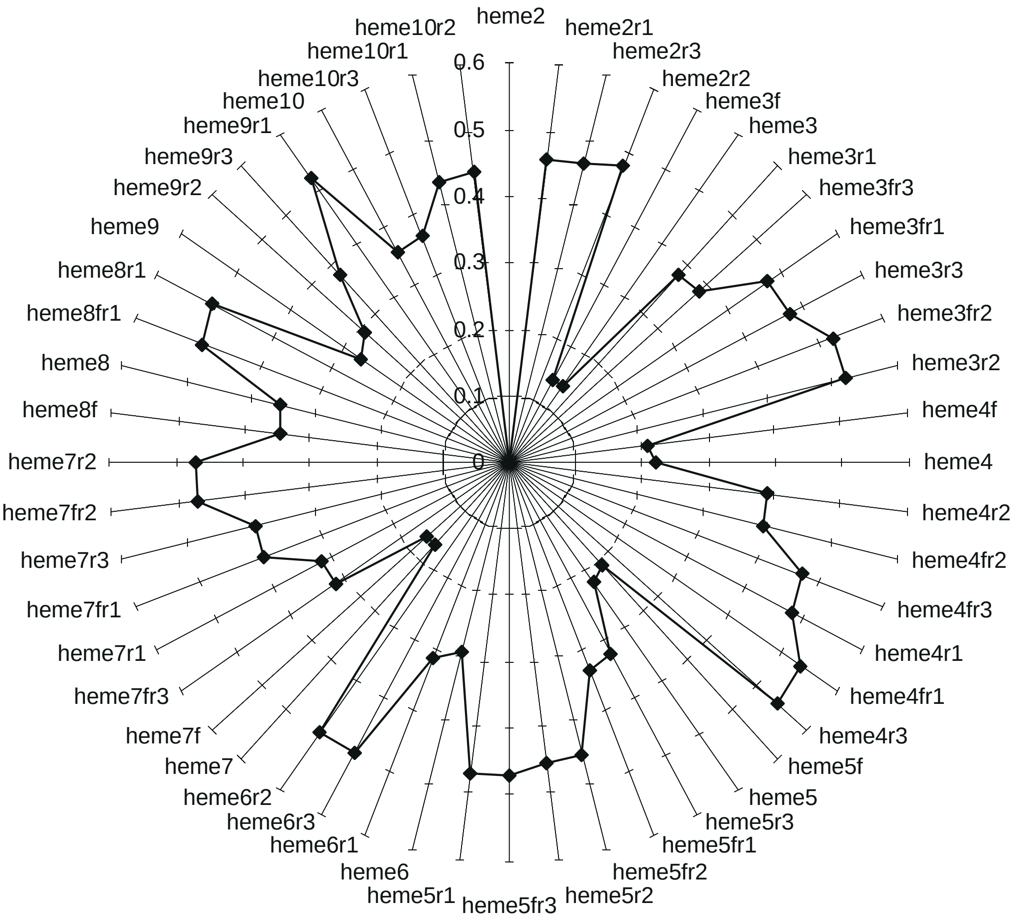

Fig. 9. Distance plot between heme2 and the rest of the 51 molecules. The distances are calculated on the first four components from the 1 Å data for the second alignment. The distance for heme2 itself is zero since it was used as the template.

components explain 79% of the variance in X, and relate almost entirely to differences in side-chain positions.

Calculating the distances

The first four score vectors were used to calculate the distance, in the super score space, between heme2 and each of the other molecules. Figure 9, based on the second alignment, shows these distances. For example, there are eight points in Fig. 9 relating to heme3 in each of the eight possible alignments shown for this molecule in Fig. 5. However, two of these eight points (heme3 and heme3f) are markedly closer to the centre of Fig. 9 than the remaining six, and this shows that alignments 3 and 3f best match the alignment of the reference structure heme2 itself.

Heme3 and heme3f are pairwise related to each other as shown in Fig. 5, and this demonstrates that four components do not completely model the other structural differences. Heme7 and heme7f show a similar mutual relationship, and once again the rotation around the parallel axis makes them both equally similar to heme2, which is a consequence of the properties of heme2 itself, as discussed earlier. There is a slightly different effect with heme9 and heme10, which will be discussed elsewhere (M.C. de Rosa, A. Berglund and P.J. Goodford, in preparation). The limitation due to using non-minimised heme structures can be estimated by comparing compounds (e.g. heme6 and heme6r1) which should be at exactly the same distance from heme2. The actual difference between two compounds that should be identical can give an approximation of how precise the model is. This difference arises from the fact that the crystal structure is not perfectly symmetrical, but the associated error does not affect the interpretation of the results since these differences are always much smaller than other sources of variance. In fact, the presence of these differences actually enhances the overall interpretation of the findings.

Comparing the alignments

In order to compare the two different alignment methods, the distances from heme2 for the first alignment were also determined. These distances were calculated using the first four principal components, since Fig. 8 shows clearly that the fifth component with the first alignment was not related to side-chain position. The distances for the first and second alignments were then plotted

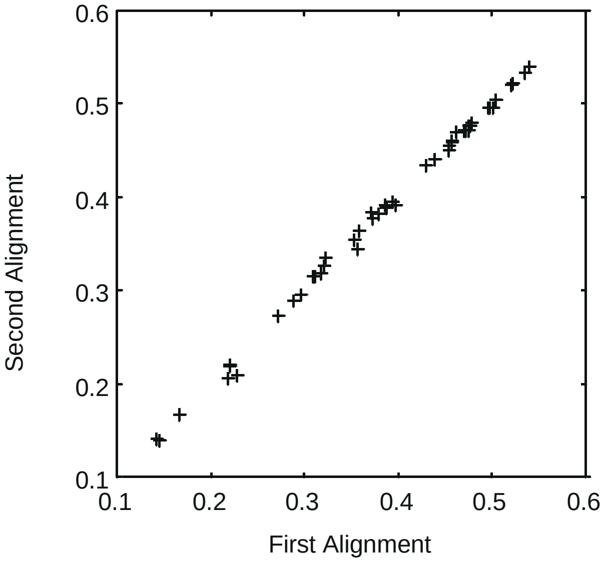

Fig. 10. The distances for the first and second alignment models plotted against each other. This shows that the first alignment artefact identified in Fig. 8 does not appreciably influence the findings based on that alignment procedure (see text).

against each other (Fig. 10). This plot shows that the results with the two methods are similar, which indicates that the model with the first alignment also gives acceptable results. This implies that even such a big artefact in the data as that caused by rotation about the parallel axis in the first alignment method does not necessarily give problems in the final results. The key requirement is to choose the most suitable number of principal components for the final model, and exclude inappropriate components.

We believe that this may be an important general finding. It is possible that too much emphasis has been placed in previous research on the requirement for an accurate rigid alignment of the target molecules. However, modest errors in the alignment of the core may only amount to a few tenths of an Angstrom, while the failure to take account of side-chain movements can introduce an error of several Angstrom in the assumed position of the side-chain atoms. It may therefore be much more important to treat flexible atoms adequately than to make an ideal alignment of the rigid cores.

Experimental verification

The predictions have been verified by comparison with the NMR data of Hauksson et al. [24], and this shows that we can correctly predict the most likely binding conformations for all of the heme molecules, although some redundancy is introduced due to the symmetry of heme2. This symmetry precludes a unique solution for some of the heme molecules, because there may be two conformations which are equally similar to heme2. However, this is not a general problem inherent in the method, and the redundancy could have been avoided in the present work if we had chosen a different template molecule initially.

Conclusions

We have defined three categories of molecular flexibility, and described a novel method for aligning a set of molecules which have a common core and flexible side chains. This was done using v. 15 of Grid which takes account of molecular flexibility, and analysing the Grid data with Hi-PCA which is an extension of ordinary PCA. This combination of Grid and Hi-PCA makes no assumptions about the conformation of the flexible side chains, and is less sensitive to errors in the alignment of the molecules.

The performance of the method was demonstrated on a set of nine heme molecules, one of which was chosen as a template. The aligned conformations of the rest were predicted, and the findings were verified with Hauksson NMR data on each of the molecules when bound to its receptor site in myoglobin. In each case, the observed mode of binding was correctly identified, although some redundant predictions could not be avoided due to our inappropriate selection of the initial template molecule.

Acknowledgements

We would like to thank Dr. P.J. Goodford for valuable discussions, remarks and comments. S.W. would also like to thank the Swedish Science Research Council (NFR) for financial support.

References

1 Klebe, G., Mietzner, T. and Weber, F., J. Comput.-Aided Mol. Design, 8 (1994) 751.
2 Dunn III, W.J., Hopfinger, A.J., Catana, C. and Duraiswami, C., J. Med. Chem., 39 (1996) 4825.
3 Klebe, G. and Abraham, U., J. Med. Chem., 36 (1993) 70.
4 Waller, C.L., Oprea, T.I., Giolitti, A. and Marshall, G.R., J. Med. Chem., 36 (1993) 4152.
5 Waller, C.L. and Marshall, G.R., J. Med. Chem., 36 (1993) 2390.
6 Thomas, B.F., Compton, D.R., Martin, B.R. and Semus, S.F., Mol. Pharmacol., 40 (1991) 656.
7 Loughney, D.A. and Schwender, C.F., J. Comput.-Aided Mol. Design, 6 (1992) 569.
8 Cramer III, R.D., Patterson, D.E. and Bunce, D.E., J. Am. Chem. Soc., 110 (1988) 5959.
9 Gamper, A.M., Winger, R.H., Liedl, K.R., Sotriffer, C.A., Varga, J.M., Kroemer, R.T. and Rode, B.M., J. Med. Chem., 39 (1996) 3882.
10 McMartin, C. and Bohacek, R.S., J. Comput.-Aided Mol. Design, 9 (1995) 237.
11 DePriest, S.A., Mayer, D., Naylor, C.B. and Marshall, G.R., J. Am. Chem. Soc., 115 (1993) 5372.
12 Masak, B.B., Merchant, A. and Matthew, J.B., J. Med. Chem., 36 (1993) 1230.
13 Kato, Y., Inoue, A., Yamada, M., Tomioka, N. and Itai, A., J. Comput.-Aided Mol. Design, 6 (1992) 475.

14 Jones, G., Willett, P. and Glen, R.C., J. Comput.-Aided Mol. Design, 9 (1995) 532.

15 Perkins, T.D.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 7 (1993) 155.

16 Perkins, T.D.J., Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 479.

17 Nicklaus, M.C., Milne, G.W.A. and Burke Jr., T.R., J. Comput.-Aided Mol. Design, 6 (1992) 487.

18 Kroemer, R.T. and Hecht, P., J. Comput.-Aided Mol. Design, 9 (1995) 396.

19 Goodford, P.J., J. Med. Chem., 28 (1985) 849.

20 Goodford, P.J., J. Chemometrics, 10 (1996) 107.

21 Wold, S., Hellberg, S., Lundstedt, T., Sjöström, M. and Wold, H., Symposium on PLS Model Building: Theory and Applications, Frankfurt am Main, Germany, September 23–24, 1987.

22 Wold, S., Kettaneh, N. and Tjessem, K., J. Chemometrics, 10 (1996) 463.

23 Cramer, R.D., Clark, R.D., Patterson, D.E. and Ferguson, A.M., J. Med. Chem., 39 (1996) 3060.

24 Hauksson, J.B., La Mar, G.N., Pandey, R.K., Rezzano, I.N. and Smith, K.M., J. Am. Chem. Soc., 112 (1990) 8315.

25 GRID User Manual, v. 15, Molecular Discovery Ltd., Oxford, U.K., 1997.

26 Jackson, J.E., A User's Guide to Principal Components, Wiley, New York, NY, U.S.A., 1991.

27 McGregor, J.F., Jaeckle, C., Kiparissides, C. and Koutoudi, M., A.I.Ch.E. J., 40 (1994) 826.

28 Wold, H., In Krishnaiah, P.R (Ed.) Multivariate Analysis, Academic Press, New York, NY, U.S.A., 1966.

29 Takano, T., J. Mol. Biol., 110 (1977) 537.