J Comput Aided Mol Des (2006) 20:567–587
DOI 10.1007/s10822-006-9082-y

ORIGINAL PAPER

GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D

Nicola J. Richmond · Charlene A. Abrams · Philippa R. N. Wolohan · Edmond Abrahamian · Peter Willett · Robert D. Clark

Received: 4 June 2006 / Accepted: 21 September 2006 / Published online: 19 October 2006

Summary Alignment of multiple ligands based on shared pharmacophoric and pharmacosteric features is a long-recognized challenge in drug discovery and development. This is particularly true when the spatial overlap between structures is incomplete, in which case no good template molecule is likely to exist. Pair-wise rigid ligand alignment based on linear assignment (the LAMDA algorithm) has the potential to address this problem (Richmond et al. in J Mol Graph Model 23:199–209, 2004). Here we present the version of LAMDA embodied in the GALAHAD program, which carries out multi-way alignments by iterative construction of hypermolecules that retain the aggregate as well as the individual attributes of the ligands. We have also generalized the cost function from being purely atom-based to being one that operates on ionic, hydrogen bonding, hydrophobic and steric features. Finally, we have added the ability to generate useful partial-match 3D search queries from the hypermolecules obtained. By running frozen conformations through the GALAHAD program, one can utilize the extended version of LAMDA to generate pharmacophores and pharmacosteres that agree well with crystal structure alignments for a range of literature datasets, with minor adjustments of the default parameters generating even better models. Allowing for inclusion of partial match constraints in the queries yields pharmacophores that are consistently a superset of full-match pharmacophores identified in previous analyses, with the additional features representing points of potentially beneficial interaction with the target.

Keywords GALAHAD · Hypermolecule · Hyperstructure · Molecular alignment · Pharmacophore · Pharmacostere · Shape similarity

Abbreviations

ATP	Adenosine triphosphate
CDK-2	Cyclin dependent kinase 2
DHFR	Dihydrofolate reductase
GALAHAD	A Genetic Algorithm with Linear Assignment for the Hypermolecular Alignment of Datasets
GASP	Genetic Algorithm Superposition Program
GPCR	G-Protein coupled receptor
HIV-1	Human immunodeficiency virus 1
LAMDA	Linear Assignment for Molecular 50 Dataset Alignment
LAP	Linear assignment problem
MCS	Maximum common subgraph
RMSD	Root mean square deviation
RT	Reverse transcriptase

N. J. Richmond · P. Willett
Krebs Institute for Biomolecular Research, Department of Information Studies, University of Sheffield, 211 Portobello St., Sheffield S7 4DP, UK

C. A. Abrams · PhilippaR. N. Wolohan · E. Abrahamian · R. D. Clark (✉)
Informatics Research Center, Tripos Inc., 1699 South Hanley Road, St Louis, MO 63144, USA
e-mail: bclark@tripos.com

Present Address:
N. J. Richmond
Cheminformatics, GlaxoSmithKline, Gunnels Wood Road, Stevenage SG1 2NY, UK

Introduction

Methods for correlating molecular structure with biological activity have long played a central role in computer-aided molecular design. Early approaches based on physicochemical properties and 2D (topological) structural features are increasingly being superseded by 3D methods that take geometric or molecular field information into account [1]. Most of these 3D methods are dependent, to a greater or lesser degree, on ligand *conformation* (e.g., EVA [2] or GRIND [3]). Others are also dependent on how the ligands are aligned with respect to one another in Cartesian space—i.e., on ligand *configuration*. CoMFA [4] and CoMSIA [5] are examples of the latter class of methods, as are *in silico* docking and pharmacophore elucidation. Here, we focus on how, given a suitable conformation for each, one can align the ligands in a dataset so as to highlight commonalities in the spatial distribution of ligand-protein interactions (the *pharmacophore*) and in molecular shape (the *pharmacostere*). A complementary methodology [6] for generating a set of pharmacophorically and pharmacosterically concordant conformations suitable for such an alignment will be described in detail elsewhere [7].

Many automated alignment methods have been described in the literature (as reviewed by Lemmen and Lengauer [8]); here, we discuss the use of a 3D *hypermolecule*, a 3D analog of a 2D chemical *hyperstructure*, for rigid-body alignment. A 2D chemical hyperstructure is generated by sequentially overlapping each molecular graph onto a hyperstructure, adding new nodes to the hyperstructure to accommodate any atoms that do not overlap. The overlapping is generally carried out so as to minimize the increase in size of the hyperstructure at each stage, which is equivalent to maximizing the overlap between the molecular graph and the hyperstructure. When atom and bond data for the input molecules are retained as features of the relevant hyperstructure nodes and edges, subsequent analysis of the hyperstructure can provide useful insight into aggregate properties of the dataset without losing information about the properties of the constituent molecules. The chemical hyperstructure representation proposed by Vladutz and Gould [9] was originally devised primarily to improve the efficiency of substructure searching [10]. More recently hyperstructures have been applied to the analysis of bioactivity data, with the aim of identifying substructural features that are positively or negatively associated with activity [11]. The molecular field topology analysis (MFTA) technique developed by Palyulin et al. [12] is another example of this approach.

The 2D *hyperstructure* is logically related to the better known maximum common subgraph (MCS) [13] which is the largest subgraph of every chemical graph in a dataset of two or more molecules. The MCS can be regarded as the logical intersection (i.e., the Boolean AND) of the chemical graphs and is thus a substructure of every molecule in the dataset. Conversely, a 2D hyperstructure can be regarded as the logical union (i.e., Boolean OR) of the chemical graphs. Hence, the chemical graph of every molecule in the dataset is a subgraph of the hyperstructure.

A *hypermolecule* is a 3D representation of a dataset that seeks to maximize the degree of molecular overlap between structures while preserving the geometry and molecular connectivity of each molecule in the dataset. It is defined by a set of *hyperatoms* along with their interconnectivity, as well as associated *hyperfeatures*. Each hyperatom and hyperfeature is characterized by a vector encoding the properties of the atoms or features that it represents. Such a 3D hypermolecule identifies and encodes the most important substructural and conformational commonalities between sets of molecules, and can therefore provide an alignment rule for molecules not used in its construction. One application of a 3D hypermolecule is for the alignment of structures as a way to derive structure-activity relationships (SARs).

The LAMDA (linear assignment for molecular dataset alignment) algorithm, originally developed for atom-based alignment of pairs of 3D molecules [14], has now been extended to the molecular alignment of large datasets based on shared pharmacophoric and pharmacosteric features. These alignments can be used to define partial-match 3D search queries. Here we provide a detailed description of how the extended method works, and illustrate its operation from within the GALAHAD program [15] by applying it to ''frozen'' ligands from several literature datasets. In particular, the algorithm was able to reproduce target pharmacophores for the test sets compiled by Patel et al. for evaluating the performance of DISCO, GASP and Catalyst [16].

Methods

The LAMDA algorithm [14] was inspired by a computer vision algorithm developed by Belongie et al. for matching 2D shapes, where each shape is represented by points sampled from the internal and external boundaries [17]. The shape matching algorithm consists of

three principal stages: the identification of a one-to-one correspondence between points representing shape *A* and points representing shape *B*; the determination of a morphing transformation that superimposes corresponding pairs of points; and the calculation of a similarity measure based on the sum of the matching errors for corresponding points and the magnitude of the morphing transformation. Our 3D alignment algorithm, as described in detail by Richmond et al. [14], is necessarily somewhat different. A one-to-one correspondence is first identified between pairs of atoms, one from each of the two molecules. Then, these correspondences are used to calculate a Procrustes transformation that superimposes the two molecules so as to maximize the overlay of the corresponding atoms. This algorithm provides both an effective and a highly efficient way of generating molecular alignments, with individual overlays typically requiring ca. 0.02 CPU seconds on an 800 MHz PC.

In the work described here, we extend the LAMDA algorithm from purely atom-based alignment of pairs of molecules to the alignment of datasets of two or more molecules based on common pharmacophoric and pharmacosteric features and the generation of a hypermolecule. The hypermolecule not only aids in alignment of the dataset, but also supports subsequent generation of partial-match 3D search queries.

The core structure of the algorithm is very similar. However, we now seek molecular alignments that maximize the overlay of similar *features* in terms of feature interaction strengths and steric environment in which each feature finds itself. So we first identify a one-to-one correspondence between pairs of features, one from each of the two (hyper)molecules, then use the set of feature-feature correspondences identified to calculate a Procrustes transformation that superimposes the two (hyper)molecules so as to maximize the overlay of the corresponding features. As is the case with atom-based alignment, this set may include geometrically inappropriate feature-to-feature correspondences that could unduly compromise the quality of the molecular overlay if included in the calculation of the Procrustes transformation. Hence several filter and refinement steps are used to identify and eliminate such deleterious correspondences. One filter discards any correspondences where either feature has a local mirror image within its molecule, and another discards any feature-feature correspondences that do not respect distance constraints.

Order of hypermolecule construction

A hypermolecule of a dataset is generated by aligning successive pairs of molecules or hypermolecules and combining features of the same type that lie in close proximity whilst retaining the individuality of the atoms themselves. Hence, the quality of the hypermolecule produced depends on the quality of each successive overlay, which in turn depends on the order of pairwise alignment. So, to ensure that each alignment step has the best chance of success, we align the most similar structures first. The hypermolecule construction order is defined by a dendrogram resulting from a hierarchical clustering of the dataset. Construction then proceeds by traversing the dendrogram from bottom to top. For the work described here, the similarity between individual molecules *A* and *B* is given by the similarity between the corresponding pharmacophore multiplet bitmaps [18] and the similarity between two hypermolecules ***A*** and ***B*** is the minimum pairwise similarity found between any of the molecules A_i making up ***A*** and B_i making up ***B***. This corresponds to agglomerative clustering using complete linkage [19, 20].

Alternatively, one could align the most active ligands first, then align the next most active to the nypermolecule produced, and so forth. Here, we hew to the definition of a pharmacophore as the spatial disposition of a (more or less) minimal feature set that is shared by *all* active compounds, and so do not distinguish among actives based on potency. To do otherwise risks missing partial match features that do not happen to be shared by the most active ligand.

We illustrate the hypermolecule construction process using the dendrogram shown in Fig. 1, where the basal nodes (leaves) *A* through *J* correspond to the individual molecules in the dataset and the higher level nodes correspond to intermediate hypermolecules. More specifically, the hypermolecule corresponding to node ***K*** is generated from molecules *A* and *B*, then hypermolecule [21] ***L*** is generated from the next most similar pair of molecules, *C* and *D*. Hypermolecules ***M*** and ***O*** are similarly generated from the molecule pairs {*E*, *F*} and {*H*, *I*} respectively. Hypermolecules ***P***, ***Q*** and ***S***, in contrast, are produced by aligning hypermolecules to each other. Finally, hypermolecules ***N*** and ***R*** are each generated by aligning a single molecule *G* and *J*, to hypermolecules ***M*** and ***Q***, respectively.

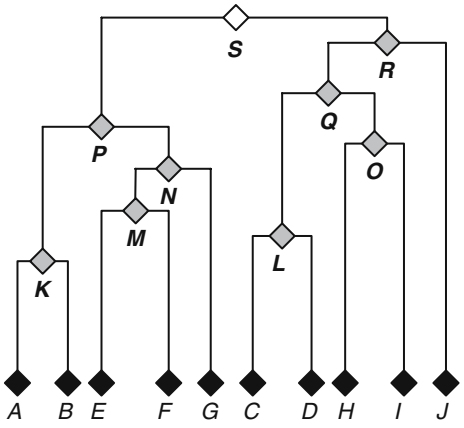

Fig. 1 Dendrogram for constructing a hypermolecule from a set of individual structures. The basal nodes ***A*** through ***J*** (*black symbols*) correspond to the individual molecular structures in the dataset, nodes ***K*** through ***R*** (*gray symbols*) correspond to intermediate hypermolecules, and node ***S*** (*open symbol*) corresponds to the hypermolecule that represents the full dataset

Aligning individual molecules

Identifying an optimal mapping between the features in one molecule and those in another is an example of a linear assignment problem (LAP), which is one of the most important classes of problem encountered in combinatorial optimization. Any such problem can be recast informally as the task of assigning n jobs to n computers so that each job is carried out as efficiently as possible. Every computer must be assigned a job and no job can be assigned more than once.

In the context of molecular alignment, the features in molecule A can be identified with the computers and the features in molecule B can be identified with the jobs. A set of feature-to-feature correspondences can then be computed by solving the LAP with cost matrix whose ijth entry is the cost of matching feature a_i in A with feature b_j in B. This function is minimized if features a_i and b_j are similar in terms of feature interaction strength and geometric environment. An optimal solution to the LAP is then a set of feature-to-feature correspondences $\{(a_i, b_i)\}$ that minimizes the total feature matching cost c given by

$$c = \sum_i \left[\alpha(|\gamma(a_i) - \gamma(b_i)|) + \frac{1}{\beta_i} \sum_j \sum_k \frac{(\lambda_{jk}(a_i) - \lambda_{jk}(b_i))^2}{\lambda_{jk}(a_i) + \lambda_{jk}(b_i)} \right] \quad (1)$$

$$\beta_i = \sum_j \sum_k (\lambda_{jk}(a_i) + \lambda_{jk}(b_i)) \quad (2)$$

where $\gamma(a_i)$ is the interaction strength of feature a_i and $\gamma(b_i)$ is the interaction strength of the corresponding feature b_i; only correspondences between features of the same type are considered. The function $\gamma(a_i)$ depends on the particular substructures of a_i, each of which has a characteristic strength of interaction. The γ functions we use are based on those developed for the GASP program [22, 23]. They are tabulated as *gasp weights*, ranging from 0 to 1, and are given in an external file that defines the various feature types, provided as Supplemental Material. For example, carboxylates are strong hydrogen bond acceptors, so they are assigned a gasp weight of 1.0, whereas ether oxygens are typically much weaker acceptor atoms so they are assigned a gasp weight of 0.35.

The double summation in Eq. 1 gives the χ^2 dissimilarity between the neighborhoods $\lambda(a_i)$ and $\lambda(b_i)$ of the corresponding features (a_i, b_i). These neighborhoods, one for each feature type, are calculated from the 20 bin radial distribution function whose kth bin count $\lambda_{jk}(a_i)$ is the number of features of type j that are between k–1 and k Å from feature a_i. The scaling factor β_i ensures that the total neighborhood mismatch cost is less than or equal to 1 regardless of the total number of features found in the neighborhood of a_i and b_i.

GALAHAD's default configuration files in SYBYL 7.2 cover six feature types: hydrogen bond donor and acceptor atoms (**D** and **A**, respectively); positive nitrogen (**P**); negative and hydrophobic centers (**N** and **H**); and steric features (**S**). Neighborhood calculations for hydrophobic centers take the radial distribution functions of donor and acceptor atoms, positive nitrogen and negative centers as well as other hydrophobic centers and steric features into account. Neighborhoods for donor atoms include the distributions of acceptor atoms, hydrophobic centers and steric features, whereas neighborhoods for acceptor atoms include donor atoms, hydrophobic centers and steric features. Neighborhood dissimilarities for positive nitrogen and negative centers, on the other hand, only take the distribution of steric features into account.

Because correspondences between features differing in type are not of interest, separate linear assignment calculations can be run for each feature type. Only donor:donor, acceptor:acceptor, positive:positive, negative:negative and hydrophobe:hydrophobe correspondences were used at this stage for the work described here; potential correspondences between steric features were not considered. Note, however, that the radial distribution function for steric features is ''seen'' by each of the other feature types. Among other things, this serves to distinguish internal features

in a particular conformation from more peripheral ones. Steric feature correspondences were also used in the subsequent refinement steps described below.

Correspondence filters

Symmetrical substituents are a powerful tool in drug discovery and lead optimization, but the degeneracy they introduce can complicate the step of determining feature-to-feature correspondences. Problematic ''mirror features'' are identified by comparing each molecule to itself, with features in *intra*molecular correspondences that have low matching costs being set aside before making *inter*molecular comparisons. Such correspondences can be revisited if too few good correspondences survive subsequent filtering steps, but this is usually unnecessary.

Figure 2 illustrates the factors affecting the equivalence costs of several intramolecular correspondences for a relatively simple molecule. Here, the cost entailed in matching the hydrophobic centers is calculated as 0.58, of which 0.15 arises from differences in gasp weights (1.0 vs 0.9, with $\alpha = 0.15$) and 0.43 comes from the differences in the feature profiles. The cost of matching the distal phenol (**A1**) to either carboxylate oxygen is even higher at 0.63, which is solely due to feature profile mismatches. The intramolecular correspondence cost for **A2** and **A3** is much smaller, at

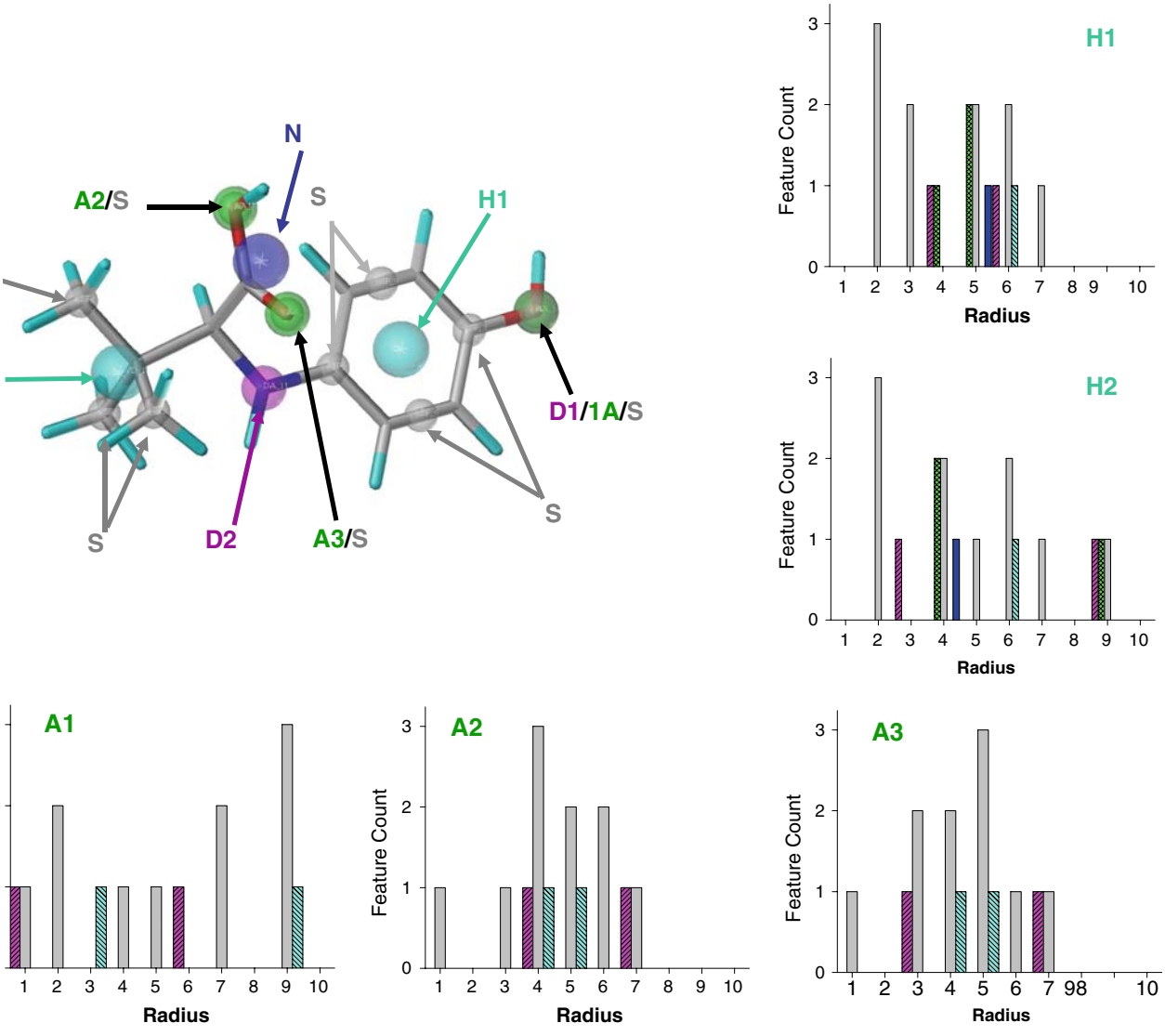

Fig. 2 Radial distribution functions for a simple molecule. *Cyan spheres* indicate hydrophobic centers, *purple* indicates donor atoms, *green* indicates acceptor atoms and *blue* indicates negative centers. *Gray spheres* represent steric features. Abbreviations are as follows: *H* hydrophobic center; *D* donor atom; *A* acceptor atom; *N* negative center; and *S* steric feature. Radial distribution profiles for donor atoms and the negative center are not shown

0.110, most of which arises from asymmetry with respect to **D2**.

In this example, the intramolecular cost was calculated using the total 20 bin radial distribution neighborhoods, but since symmetry is generally a local problem, only the first five 1 Å layers need to be taken into consideration for most *intra*molecular comparisons. For *inter*molecular comparisons, the default number of bins for the neighborhoods is 20 Å. Either default value can be modified by editing the external configuration file.

Once an optimal set of intermolecular correspondences has been identified, those correspondences with large costs are set aside to leave reduced feature sets A' and B'. A *geometric filter*, which can be thought of as an application of the principles upon which the distance geometry approach to pharmacophore elucidation [24] is based, is applied to the reduced sets to remove correspondences that are geometrically incompatible [14]. If the reduced feature sets A'' and B'' include too many correspondences (ten for the analyses described here), they are subjected to a final *geodesic filtering* step to identify those features that have the greatest geometric leverage. The first features picked are those that lie farthest from the Cartesian centroids of A'' and of B''. The list of features to be used is then built up to the specified limit by iterative addition of those features that have the largest minimum spatial separation from those already on the list. Note that this step was not part of the original LAMDA algorithm.

If fewer than three correspondences remain in the reduced feature sets A'' and B'' after the filtering steps, the alignment is deemed to have failed and the molecule with the largest number of features is returned. Otherwise, the remaining correspondences define a unique least-squares transformation that superimposes molecule A onto molecule B such that corresponding features in A'' and B'' are overlaid. The initial overlay so obtained is refined by cycling through the entire process twice more [14]. Only steric features are considered for these refinement steps, and Euclidean distance serves as the cost function for the linear assignment step. Geometric filtering at this stage is based on an externally specified distance cutoff—here, 1.0 and 1.2 Å for the first and second refinement cycles, respectively.

A hypermolecule is then created from each pair of successfully overlaid molecules. Features of the same type that lie within some threshold distance of each other (the default being 0.6 Å) are consolidated into a single *hyperfeature* of that type that is associated with an array made up of the gasp weights of its constituent features. Atoms retain their individuality and connectivity, so each molecule becomes a substructure in the hypermolecule produced.

Aligning hypermolecules

The process described thus far is the alignment of molecules that correspond to the leaf nodes on the construction order dendrogram. The hypermolecules resulting from these alignments correspond to the first-level nodes in the dendrogram. Aligning two hypermolecules (or a hypermolecule with a single molecule) proceeds exactly as described for molecular alignment, except that the cost function c is inversely weighted to favor correspondences between ''large'' hyperfeatures (Eq. 3).

$$c = \sum_i \omega_i \left[\alpha_0 \left| \left\langle \gamma(\mathbf{a}_i) \right\rangle - \left\langle \gamma(\mathbf{b}_i) \right\rangle \right| + \alpha_1 \min_{j,k} \left| \gamma(a_{ij}) - \gamma(b_{ik}) \right| + \frac{1}{\beta_i} \sum_j \sum_k \frac{(\lambda_{jk}(\mathbf{a}_i) - \lambda_{jk}(\mathbf{b}_i))^2}{\lambda_{jk}(\mathbf{a}_i) + \lambda_{jk}(\mathbf{b}_i)} \right] \tag{3}$$

$$\beta_i = \sum_j \sum_k (\lambda_{jk}(\mathbf{a}_i) + \lambda_{jk}(\mathbf{b}_i)) \tag{4}$$

$$\omega_i = \frac{\|\mathbf{A}\|}{\|\mathbf{a}_i\|} + \frac{\|\mathbf{B}\|}{\|\mathbf{b}_i\|} \tag{5}$$

where a_{ij} and b_{ij} are the component features that make up hyperfeatures $\mathbf{a}_i$ and $\mathbf{b}_i$, respectively; $\langle\gamma(\mathbf{a}_i)\rangle$ and $\langle\gamma(\mathbf{b}_i)\rangle$ are the average interaction strengths over the a_{ij} and b_{ij}; hypermolecule sizes $\|\boldsymbol{A}\|$ and $\|\boldsymbol{B}\|$ are the number of molecules used in the generation of $\boldsymbol{A}$ and $\boldsymbol{B}$; and $\|\mathbf{a}_i\|$ and $\|\mathbf{b}_i\|$ are the sizes of the hyperfeatures $\mathbf{a}_i$ and $\mathbf{b}_i$, respectively. The radial distribution bin counts λ_{jk} are weighted by the sizes of each contributing hyperfeature. Note that in the case of aligning individual molecules (i.e., hypermolecules and hyperfeatures of size 1), the cost function in Eq. 3 reduces to that in Eq. 1. The weighting factor ω_i serves to favor matches between features that are large relative to the size of the hypermolecules that contain them.

Query generation

Once alignment is complete, individual features are regenerated for each component 3D substructure in the hypermolecule and clustered by associating each feature with the nearest hyperfeature of the same type, provided that they are close enough together. ''Close enough'' is defined as a Euclidean distance less than 1.2

times the specified initial tolerance threshold, the default threshold being 1.0 Å. For each cluster whose population exceeds the minimum number of molecules t_{min} required to ''hit'', a representative feature is created and placed in the center of the box defined by the extreme x, y and z coordinates of the features in that cluster. The tolerance for each query feature is then set to the maximum distance between this centroid and any single feature in the cluster.

The value of t_{min} is specified separately for each analysis of n ligands, with a value of $\sqrt{n}$ rounded up to the nearest whole number generally serving as a good starting point. To avoid generating overly complex queries, larger values may be needed for complex ligands or large datasets. The DHFR example discussed below is one such case.

Setting t_{min} below the number of ligands in the training set allows the generation of queries even when there are ligands that share too few features to be incorporated into the hypermolecule. It also allows for the inclusion of features found in some ligands but not in all. Each such feature represents a potential ''extra'' interaction with the target protein that is not strictly required for activity. Nonetheless, ligands that bind well generally take advantage of at least one such interaction. Such features are collected into a *partial match constraint*; when the query is subsequently used to carry out a 3D search, a target molecule will ''hit'' so long as it can match some minimal number of the constituent features.

Full-match queries comprised of six or more features are generally over-specified and will only hit molecules very similar in structure to those from which the query was derived. Conversely, when such queries include fewer than four features, they are rarely discriminating enough to be useful in a 3D flexible search as they produce too many non-specific ''hits''. Unfortunately, many published full-match pharmacophore models are comprised of only three features, and closely spaced ones at that [16].

One way to create flexible 3D search queries that are specific and discriminating enough to be useful is to include some ''fuzziness'' in the form of partial match constraints. However, having a single such constraint that includes all features is usually suboptimal. Typically, it is better to have two constraints—a ''tight'' partial match where most or all features are required to match and a ''loose'' one where relatively few features are required. The former may correspond to a generic pharmacophore (e.g., the donor/acceptor/hydrophobe triad characteristic of kinases in general), whereas the latter features are specific to the particular target.

A detailed description of the algorithm used to determine which features should be included in the model query, and how they are partitioned between partial match constraints, is provided in the Appendix. One key objective of the algorithm is to keep features that ''hit'' the same number of model ligand conformations in the same partial match constraint. A second aim is to get an equivalent of three to five full-match features into the ''tight'' constraint.

Throughout the processing described above, acceptor atoms that fall within 2 Å of a negative center are suppressed unless at least one ligand lacks the negative center. Redundant donor atoms lying too close to a positive nitrogen are also suppressed.

Results

The LAMDA algorithm was applied through the GALAHAD interface released with SYBYL 7.2 [25] by freezing all molecules, setting the population size to 2, and setting the number of generations to 1; these represented the minimal requirements of the interface, and were internally overridden in the extraordinary case of *all* ligands being frozen. The program was run under the Red Hat Enterprise Linux WS 3.0 operating system on a 3.2 GHz Intel® Xeon™ or 3.4 GHz Pentium 4 processor. Otherwise default settings were used except as noted. Mirror filtering was turned on at an equivalence threshold cost of 0. The averaging threshold—which determines how close two features have to be to be merged during hypermolecule construction—was set to 0.6 Å unless otherwise indicated, and default refinement thresholds of 1.0 and 1.2 Å were used. The initial query feature tolerance was set to 1.0 Å unless otherwise noted.

All but one of the datasets used were drawn from Patel et al. [16] The GASP analyses reported in that work were carried out on groups of two to four molecules selected from each dataset, with ''rigid'' indicating that the template molecule was frozen and other ligands were allowed to flex freely. Unless otherwise indicated, the results described here were obtained from the full datasets with each molecule held rigid in the conformation found in the corresponding crystal structure.

Minor errors in Patel et al. for the ligand structures from **2dhf**, **1dlr**, **1fin**, and **1d6w** have been corrected for the analyses presented here.

Dihydrofolate reductase (DHFR)

The six ligands in the DHFR dataset fall into two related, structurally homologous classes but pose a

challenging alignment problem nonetheless. This is because two of the ligands—1 (folate) and 2 (5-deazafolate)—are lactams that bind with the pterin ring flipped with respect to the interaction patterns of the four aminopterin analogs. In addition, the steric overlap of the distal dimethoxyphenyl rings in 5 and 6 with the central *p*-aminobenzoyl of the other ligands and with each other is poor. Finally, many of the heteroatoms in the pterin rings are amphoteric—able to serve as donor or acceptor atoms—depending on the protonation and tautomerization state of the ring.

GALAHAD uses an external macro definition file to assign features and allows donor and acceptor features to overlap. These macro definitions take the acidity, basicity and tautomerization of commonly encountered groups into account as well [18]. The protonated N^1 position in **3**, for example, is recognized as an acceptor atom as well as a donor atom because the hydrogen that it bears can undergo a tautomeric shift to N^3 or N^8. Strong acids and strong bases are similarly accommodated, so that carboxylic acids are treated as deprotonated regardless of how they are entered. Hence the carboxylates in **1**, **2**, **3** and **4** are depicted as anions in Fig. 3, in keeping with their exposure to solvent and interaction with Arg70 and Arg32 in their complexes with DHFR. They would be treated that way, however, regardless of whether they were entered in the deprotonated states shown or as carboxylic acids per se [26].

Our method does an excellent job of reproducing the alignment seen when the respective *X*-ray crystal structures are overlaid based on the α carbons of the proteins using the Bioploymer module in SYBYL (Fig. 4). Here, GALAHAD was run with the number of ligand ''hits'' required per feature (t_{min}) set to 4, so the distal carboxylates of **1**, **2** and **3** are aligned nicely but do not produce a negative center in the model query.

As discussed in Patel et al. and in a subsequent paper from the same group discussing results obtained with multi-objective version of GASP [27], it is difficult to quantitatively compare alignments and 3D queries in a meaningful way. That said, visual inspection of the overlays shows that the alignment shown here is as good or better than the models obtained from Catalyst, DISCO or GASP for this dataset [16]. Indeed, the GALAHAD model (Fig. 4a, c) is decidedly more crisp than the overlay obtained from aligned crystal structures (Fig. 4b, d). The slight fuzziness in the latter probably reflects uncertainties in atomic coordinates and incidental variations in overall protein structure due to crystal packing effects rather than true differences in binding mode.

The query shown in Fig. 4a, c is a superset of the donor/acceptor/double hydrophobe consensus target identified by Patel et al. [16] It includes the acceptor-donor-donor-acceptor tetrad highlighted in Fig. 3, along with the hydrophobic center (**H1**) that is common to all six ligands. It includes eight other features as well - a second pterin hydrophobe (**H2**); two amphoteric feature pairs (**A3**/**D3** [28] and **A4**/**D4**); and a distal hydrophobe (**H3**). The features are grouped into partial match constraints—a ''tight'' constraint that is satisfied when at least four of its seven features are matched, and a second, ''looser'' constraint that is satisfied when at least two of its constituent features are matched.

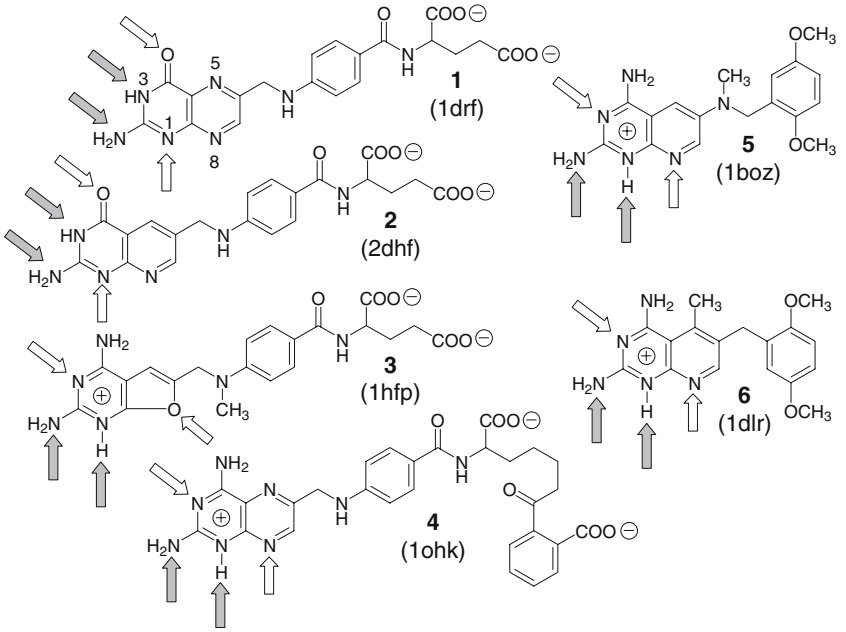

Fig. 3 Structures of the ligands in the DHFR dataset. *Gray arrows* indicate key hydrogen bond donor atoms and *white arrows* indicate acceptor atoms that make up the pattern of interactions shared by the pterin rings and their analogs. The PDB accession code for each ligand complex is given in *parentheses*

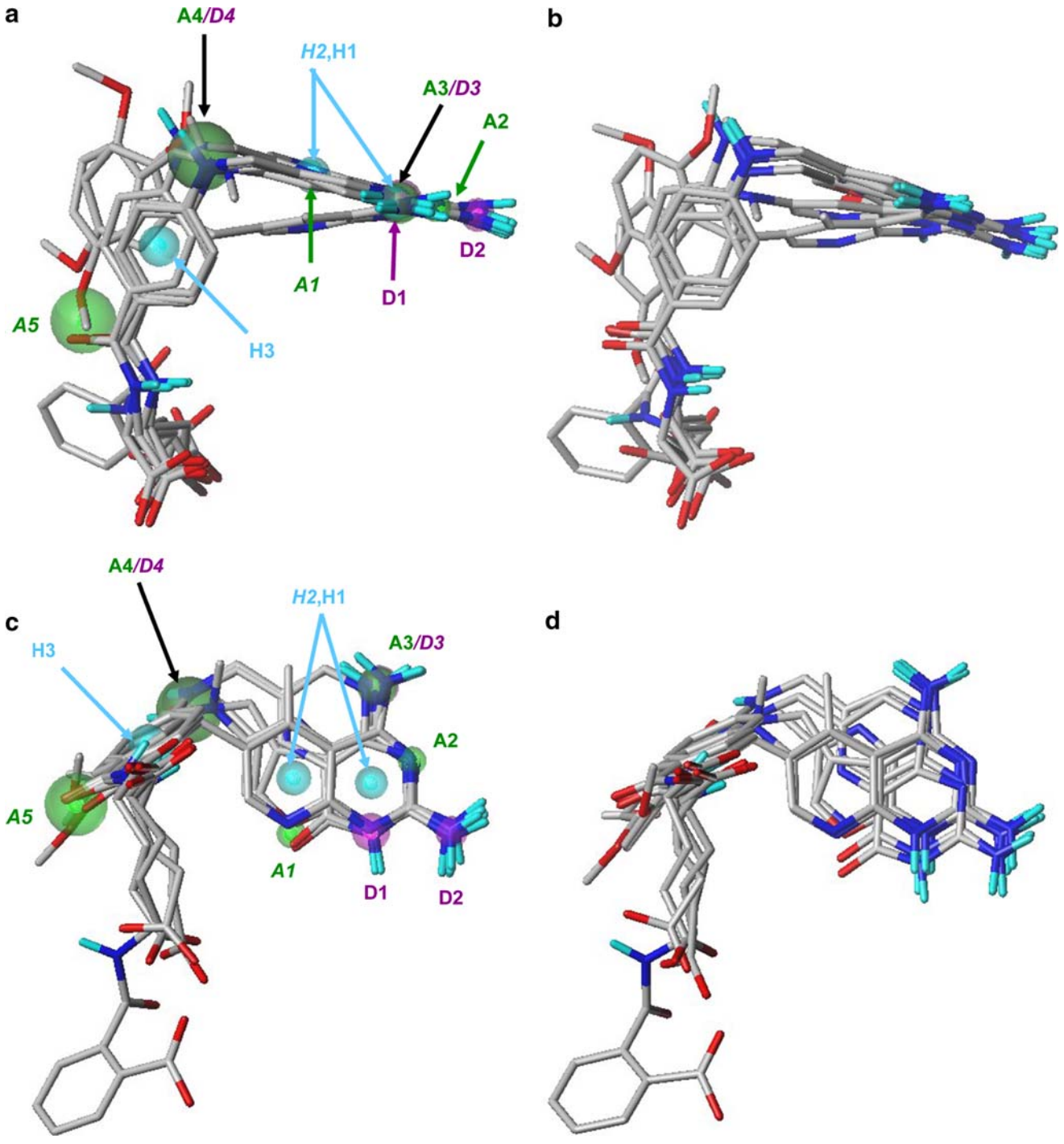

Fig. 4 Roughly orthographic views (**a** and **b** vs **c** and **d**) of overlaid structures from the DHFR dataset. (**a**, **c**) Overlay obtained using GALAHAD to rigidly align the conformations of the ligands found in the crystal structures. Feature labels are *D* for hydrogen bond donor atoms, *A* for acceptor atoms, and *H* for hydrophobic centers. The associated tolerances are colored *purple*, *green* and *cyan*, respectively. A *slash* (/) separates overlapping feature pairs, and *italics* indicates the features in the secondary partial match constraint. Four of seven (4/7) features were required to satisfy the primary partial match constraint, whereas two of the five (2/5) features labeled in *italics* were required to satisfy the secondary partial match constraint. (**b, d**) Overlay obtained by least-squares fitting of the protein α carbons in the corresponding complexes with DHFR

Note that one of the acceptor atoms associated with the pterin rings—**A1**—falls into the secondary partial match, because the deazapterin ring of 6 is sharply tilted with respect to the others in the model (Fig. 4a) as well as in alignment derived from the crystal structures (Fig. 4b). It would be missed in a full-match

query derived from an accurate alignment of the ligands. Note, too, that features associated with the distal benzamide and carboxylate substructures in **1–4** contribute to the alignment despite the fact that they do not appear in the query. This reflects the fact that the query is derived from the alignment, not vice versa.

Thrombin

The thrombin dataset is comprised of the seven inhibitors shown in Fig. 5. Two views of the rigid-body alignment produced are shown in Fig. 6a, c, each paired with the corresponding view of the overlay from the respective crystal structures (Fig. 6b, d). The union volume of the overlaid ligands closely matches that of the active site as reflected in the ligand crystal structures.

Figure 6a, c also show the query produced when t_{min} (the minimum number of ligands a feature must ''hit'' to be included in the query) is set to 4. It consists of a positive nitrogen (**P**), three hydrophobic centers (**H1**, **H2** and **H3**), two donor atoms (**D1** and **D2**), and an acceptor atom **A**. **H2**, **H3** and **A** are strictly required, whereas the four distal features constitute a two out of four (2/4) partial match constraint. **D1** falls just outside the tolerance for the positive nitrogen center. It would have been suppressed as redundant had the centroid of the associated donor atoms in the model fallen slightly closer to **P** than it actually does, and would likely be edited out of the query prior to running any 3D search.

This is a superset of the full-match target pharmacophore identified by Patel et al. [16]: a positive nitrogen and three hydrophobic centers. Actually, **13** lacks a proximal hydrophobic center corresponding to **H1**. Even were a hydrophobic center to be placed on the methylene chain of the arginyl group, it would ''miss'' **H1** in the crystal structure overlay (Fig. 6d). Similarly, the central hydrophobe in **13** ''misses'' **H2**, albeit by less. Hence the ''true'' full match query for these seven ligands consists of only two features—**P** and **H3**—unless unreasonably large tolerances are allowed for **H2**.

The partial match query generated by GALAHAD includes acceptor atom **A**, a feature that is found in most of the ligands but is absent from **11**. It also includes donor atom **D2**, which corresponds to an interaction with thrombin that is absent for **10**, **11** and **13**. Ligand **8** has a donor atom at this position in the crystal structure, but it points in a different direction and so is not scored as such by Patel et al. That both appear in the query reflects the program's ability to identify such partial match features. Note, however, that ligands missing those features are aligned well nonetheless, since each ''sees'' the shape and feature distribution in every other ligand, not just the consensus features in the query.

Ligand **13** is highlighted in green in Fig. 6. Of the seven compounds in this dataset, it alone is seriously misaligned in the model. There are several reasons for this. It lacks any feature corresponding to **D2**. As noted above, it has no hydrophobic center adjacent to its

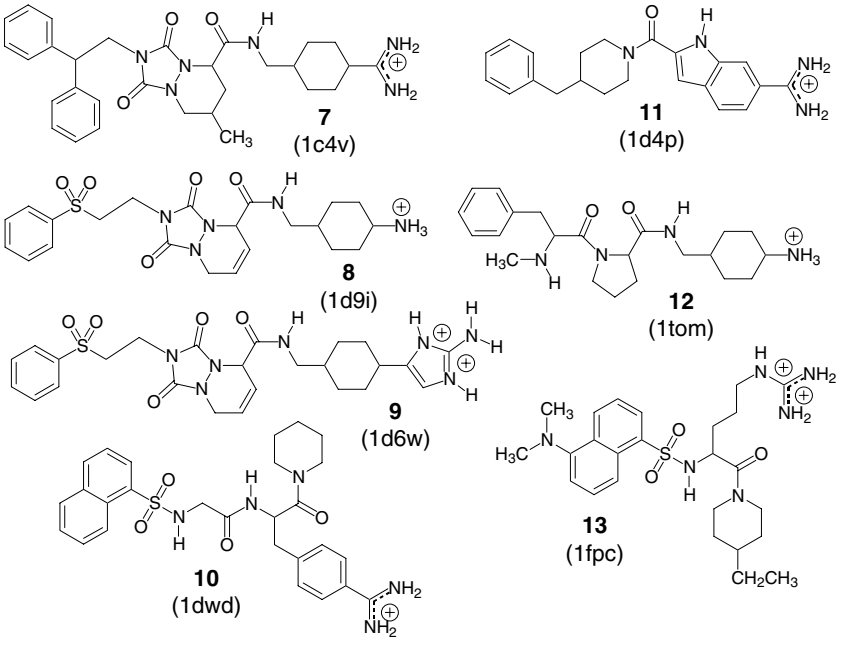

Fig. 5 Structures of ligands in the thrombin dataset from Patel et al

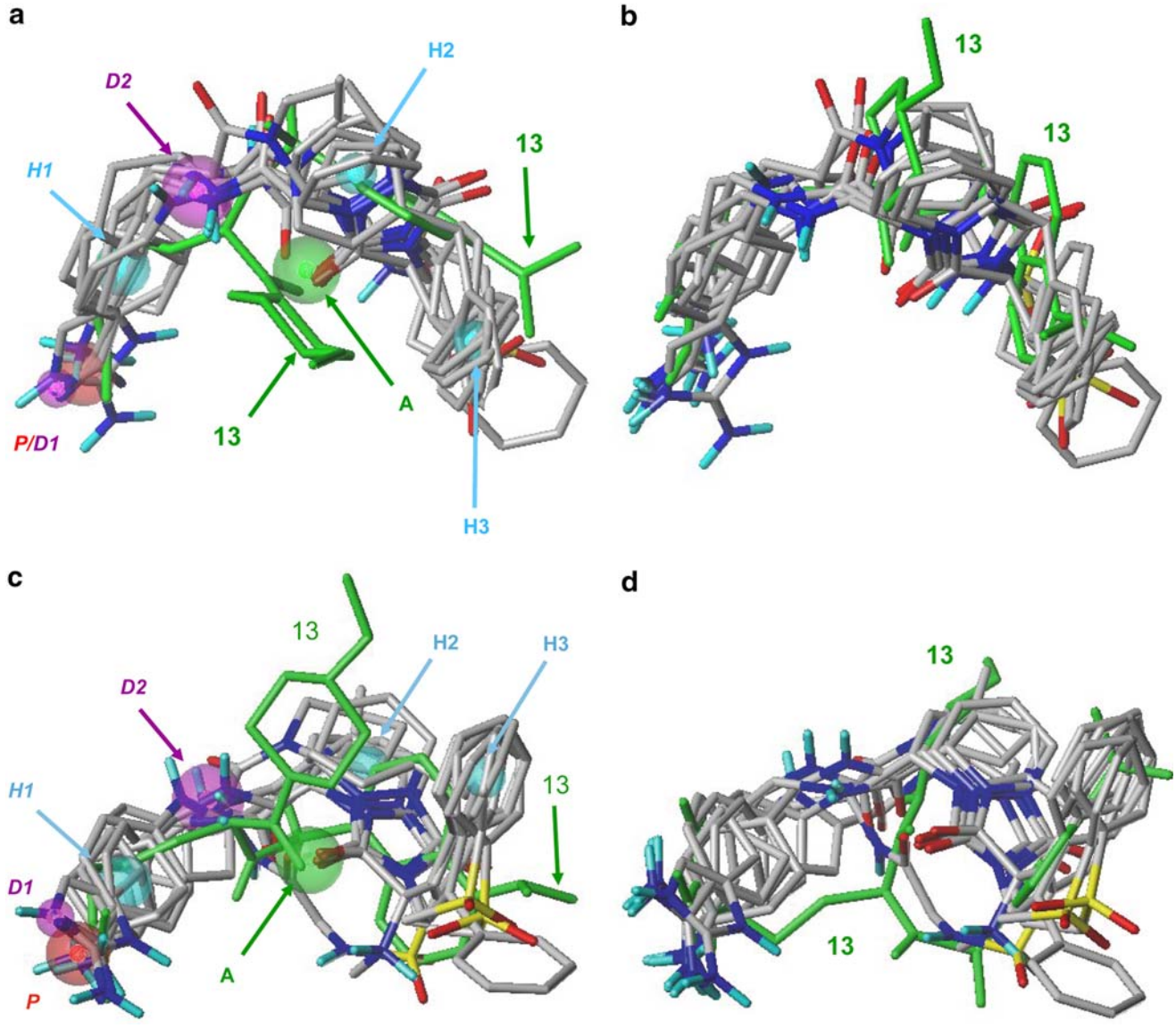

Fig. 6 Roughly orthographic views (**a** and **b** vs **c** and **d**) of overlaid structures from the thrombin dataset. (**a, c**) Overlay obtained using GALAHAD to rigidly align the crystal structure conformations. The positive nitrogen feature *P* is shown in *red*; other feature labels and color coding are as in Fig. 4. Features labeled in regular typeface were each required to ''hit'' the query, whereas the *italicized* features constitute a 2/4 partial match constraint. (**b**, **d**) Overlay obtained by least-squares fitting of the protein α carbons in the corresponding complexes with thrombin

positive nitrogen; its 4-ethylpiperidine ring cannot reach the consensus hydrophobic center **H2**; and neither hydrophobic center from its naphthyl group can overlap with **H3** (Fig. 6b, d). Because only two good feature correspondences—to **P** and **A**—are left to work with, the final alignment is perforce dominated by steric considerations.

A key reason for this nominal ''failure'' is best appreciated by considering the view of the crystal structures shown in Fig. 6d. This perspective points up the fact that the arginyl side chain in **13** falls well outside the union volume of the other ligands. Indeed, only **10** has much pharmacosteric similarity at all to **13** in the crystal structure. Given the poor pharmacophoric and pharmacosteric correspondence between **13** and the other ligands, this alternative alignment is not an altogether unreasonable compromise to that found in the crystal structure, in that it maximizes overall steric overlap in the absence of many good pharmacophoric correspondences. The model as a whole, however, does a good job of reproducing the consensus shape and feature disposition seen in the aligned crystal structures.

Cyclin-dependent kinase 2 (CDK-2)

Patel et al. included **1fvv**, **1aq1** (staurosporine; **15**), **1fin** (ATP; **16**), **1di8, 1e1v** and **1e1x** in their CDK-2 dataset; the corresponding ligand structures are shown in Fig. 7. The target pharmacophore includes the donor/acceptor dyad exemplified by the lactam *syn* amides in **14** and **15** and a central hydrophobic feature. In three of the ligands (**14**, **18** and **19**) this dyad is elaborated into a symmetrical triad of hydrogen bonding features [29], which makes ''hitting'' the pharmacophore easier but complicates getting an alignment that matches the crystal structures. This is especially so given the fact that one ligand—**17**—lacks any donor atom at all in the appropriate area, yet bears a distal phenolic atom which is a strong hydrogen bond donor and acceptor (Fig. 7). The distal sulfonamide donor and acceptors in **14** also represent a serious distraction, as do the oxygen atoms in the polyphosphate group of ATP (**16**), which may or may not be protonated at neutral pH.

Just as Patel et al. noted for Catalyst and GASP, GALAHAD overlays the polyphosphate group of ATP (**16**) with the donor/acceptor pair from the other ligands. ATP was therefore omitted from the subsequent analyses presented here.

The model obtained when the other five ligands were rigidly aligned with t_{min} (the minimum number of ligands required to ''hit'' each feature) set to 3 is shown in Fig. 8a. Staurosporine (**15**) was rejected by the program because fewer than three features could be identified that were shared with the other ligands (see below). The query produced consists of two hydrophobic centers, a donor atom and three acceptor atoms. The donor atom overlapped one of the acceptor atoms, with all six features included in a single 4/6 partial match constraint. All six ligands were ''hit'' by this query in a static 3D search, but the separation between **D** and **A3** is larger than in the target pharmacophore.

The model depicted in Fig. 8a was obtained using the standard feature definition file provided with UNITY 7.2. As noted above, the default macro definitions are set up to accommodate uncertainties due to tautomerization and pK_as that may fall near physiological pH. The inherent tradeoff of specificity against simplicity and speed imposes practical limits on the precision with which such accommodations can be made, however, so the definitions are kept simple and conservative. In particular, anilinic nitrogens are identified as potentially serving as either hydrogen bond donors or acceptors regardless of the substitution pattern borne by the aromatic ring to which they are attached.

The improved model shown in Fig. 8b was obtained when the default feature definitions were modified to prevent the exocyclic nitrogen in conjugated aminopyridines derivatives from serving as acceptors. The associated query now includes the ''classic'' **D**/**A2**/**H2** triad as well as a distal acceptor atom (**A1**) and hydrophobic center (**H1**). The latter two features result from the pharmacophorically symmetrical **18** and **19** being flipped so as to again overlap their cyclohexyl rings.

Both GALAHAD alignments are considerably more compact than is the alignment based on crystal structures (Fig. 8c). This is accomplished by rotating ligands **18** and **19** relative to **14** in the one case (Fig. 8a) and by rotating them about their pharmacophoric

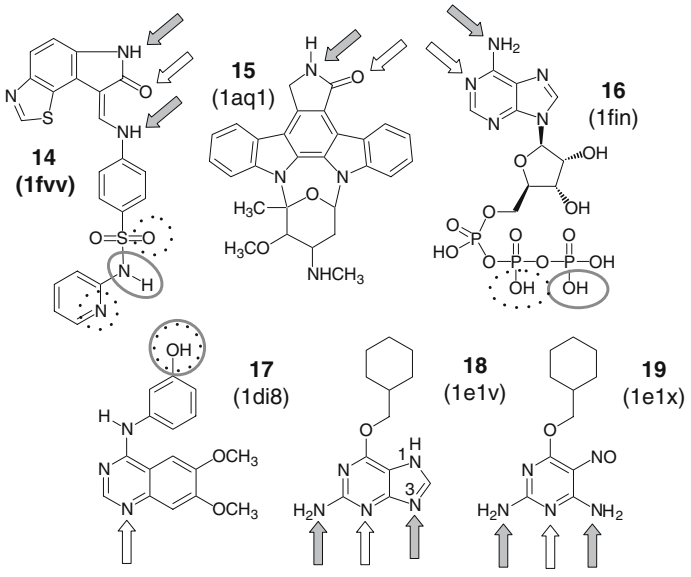

Fig. 7 Structures of ligands from complexes in the CDK-2 dataset compiled by Patel et al. [16]. *Arrows* identify pharmacophoric hydrogen bond donors and acceptors, whereas *circles* and *ellipses* identify similar feature clusters not related to the pharmacophore. *Dotted lines* encircle ''red herring'' acceptor atoms, whereas *solid gray lines* encircle the analogous donor atoms. Otherwise the color coding is as for Fig. 3

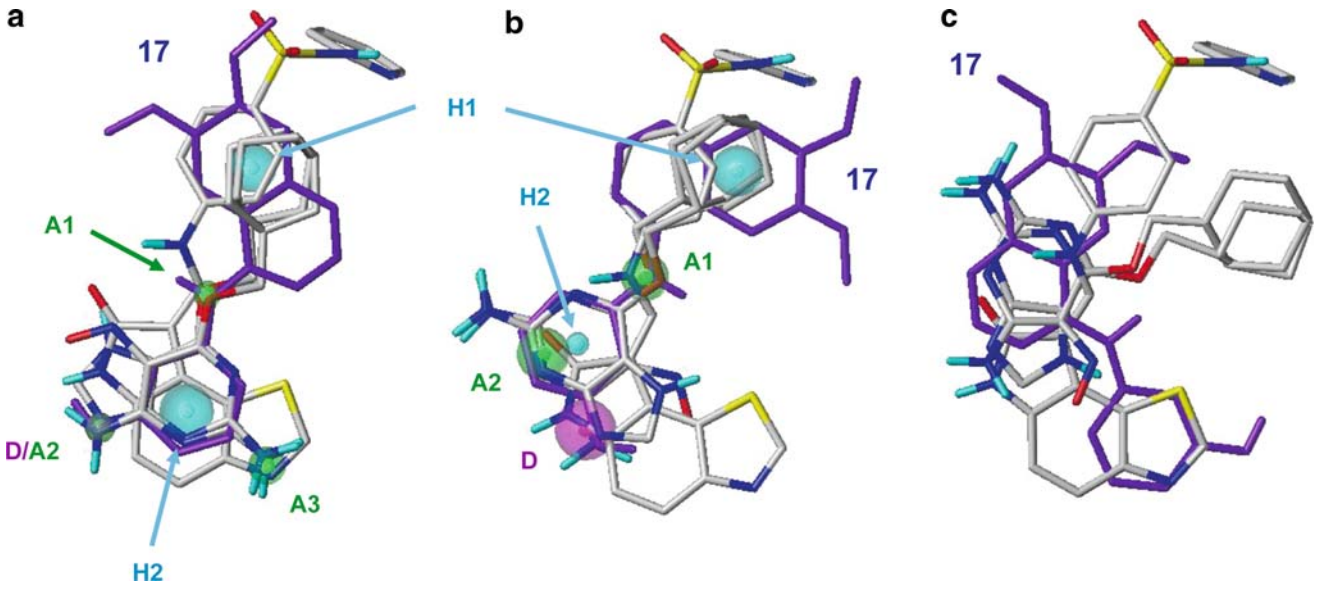

Fig. 8 Alignments of CDK-2 ligands from Patel et al, with **16** omitted from the analysis. Ligand **17** is highlighted in purple. The feature color scheme and naming conventions are as indicated for Fig. 4. (**a**) Rigid-body alignment obtained using default feature definitions. All features fall into a single 4/6 partial match constraint. The program excluded **15** from the final model. (**b**) Rigid-body alignment obtained after modifying feature definition to exclude the exocyclic nitrogen in conjugated amino pyridines from serving as an acceptor atom. All features were covered by a 4/5 partial match constraint. (**c**) Ligands **14**, **17**, **18** and **19** overlaid based on the α carbon coordinates from the respective complexes with CDK-2

symmetry axes in the other (Fig. 8b). This serves to overlap their cyclohexyl rings with the phenyl group in **14** in both cases, which increases the pharmacophoric concordance as well as the steric overlap.

Again, in both cases, ligand **17** (highlighted in purple in Fig. 8) is rotated so that it overlaps its phenolic OH group (a strong donor) with the amino substituents of the pyrimidine ligands. As noted above, this ligand lacks the donor atom normally associated with kinase pharmacophores, so this failure to reproduce the alignment from the crystal structure is not unreasonable.

Langer's CDK-2 data set

Analyses were also run on a set of CDK-2 inhibitors compiled by Thierry Langer for the Fifth European Workshop on Drug Design [30]. Most of these exhibit the full **D**/**A**/**D** triad, with several incorporating ''red herring'' features as well. These ligands, the structures of which are shown in Fig. 9, represent a pharmacophorically and pharmacosterically much tighter ensemble than do those chosen by Patel et al. GALAHAD consequently does a much better job of aligning them, as can be seen from the models presented in Fig. 10.

Figure 10a shows the model generated with default parameters. Here, **20–24** are well-aligned with each other, as are **25** and **26**. The latter pair, which are highlighted in yellow in Fig. 10, are aligned well with each other but are rotated with respect to their X-ray configuration (Fig. 10c). This rotation serves to overlap the bridging amide carbonyls in these ligands with **A2** at the expense of the overlap with **H1**.

Increasing the averaging threshold at which proximal features coalesce from the default of 0.6–0.75 Å yields a much better model (Fig. 10b), as judged on its own merit as well as by comparison with the crystal structures. This effect is traceable to the creation of a hydrophobic hyperfeature from the phenyl and naphthyl rings in the intermediate hypermolecule constructed from **25** and **26** when the granularity is relaxed in this way (Fig. 10b); these features are just slightly too far apart to merge under the more stringent 0.6 Å averaging threshold (Fig. 10a). That drives the subsequent overlap of these rings with **H1**, flipping the two aminopyrazole ligands over so as to overlay the amide nitrogens with **D1**, just as they are in the alignment based on the crystal structures. This loosening of the averaging threshold was suggested by the fuzziness of the original model, as well as the failure to pick up the distal acceptor atoms from the sulfonyl groups in **20–23**. This adjustment did indeed lead to a pair of ''new'' acceptors in the associated query (**A4a** and **A4b** in Fig. 10b); the observed tightening up of the overall alignment is a side benefit.

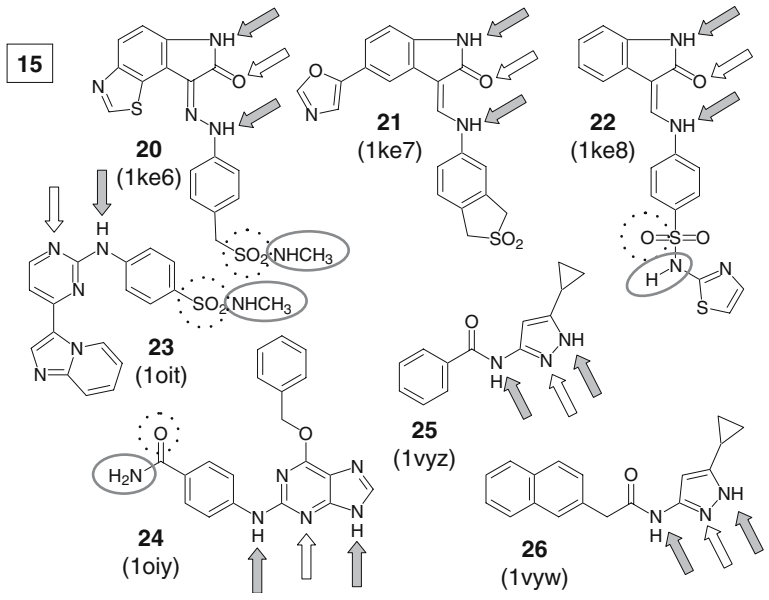

Fig. 9 Ligands from complexes in the CDK-2 dataset compiled by Langer et al. [30]

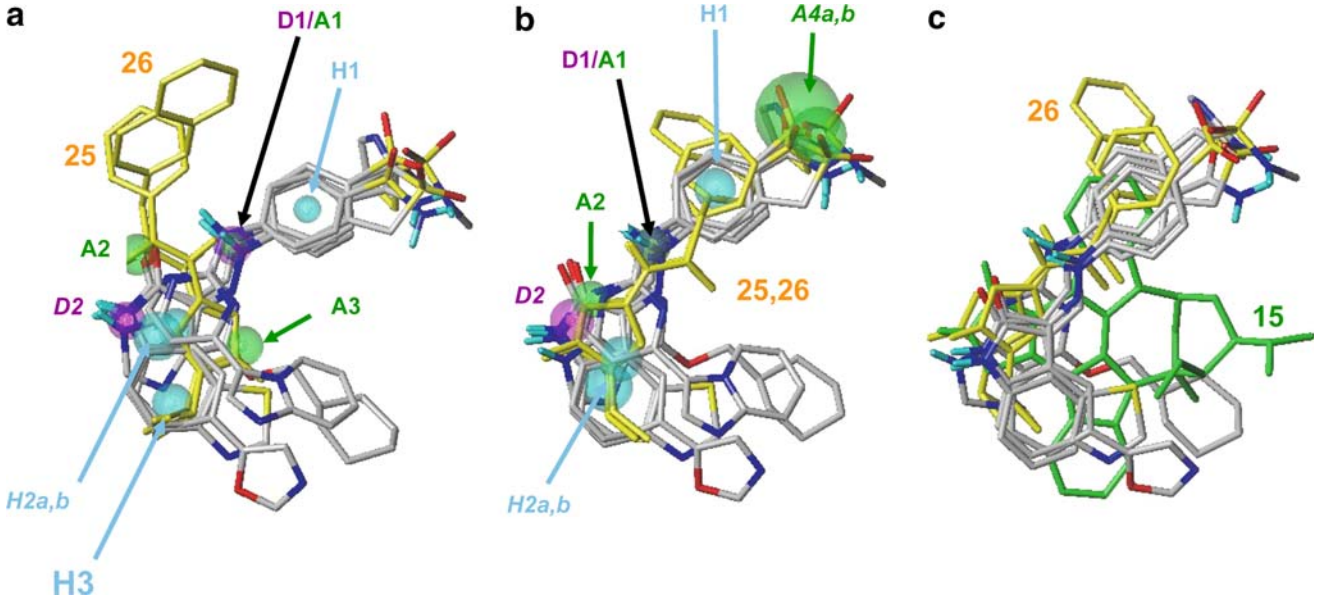

Fig. 10 Alignments of CDK-2 ligands from Langer et al. [30]. Ligands **25** and **26** are highlighted in *yellow*. The feature color scheme and naming conventions are as indicated for Fig. 4. (**a**) Rigid-body alignment obtained from GALAHAD using default feature definitions. All features fall into a single 4/6 partial match constraint. The program excluded **15** from the final model. (**b**) Rigid-body alignment obtained after modifying feature definition to exclude the exocyclic nitrogen in conjugated amino pyridines from serving as a hydrogen bond acceptor. All features were covered by a 4/5 partial match constraint. (**c**) Ligands overlaid based on the α carbon coordinates from the respective complexes with CDK-2

Thermolysin

Structures for ligands from the thermolysin dataset compiled by Patel et al. are shown in Fig. 11. This dataset represents several challenges, not least of which is that the wide range of ligand sizes precludes good steric overlap in the binding site. Nonetheless, GALAHAD successfully identified the target pharmacophore, including the two proximal acceptor atoms that interact directly with the Zn^{++} ion in the binding site (**A1a** and **A1b**), the secondary acceptor **A2**, and the conserved hydrophobic center **H3**. It also highlights a donor atom (**D1**) found in most but not all ligands. One ligand (**27**, colored yellow in Fig. 12) is flipped so that its hydroxamic acid group is overlaid on the isosteric methyl ester of **29**. This alignment allows

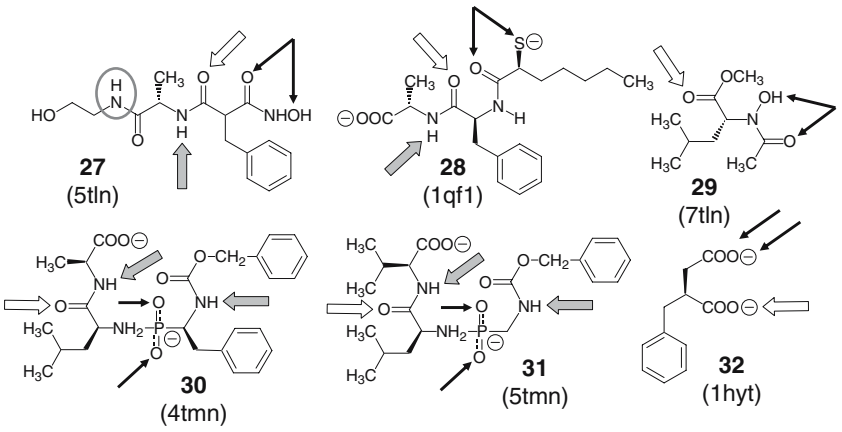

Fig. 11 Ligands from complexes in the thermolysin dataset. *Black arrows* indicate bidentate metal binding sites; other labeling is as indicated for Fig. 6

the hydroxyethylamide nitrogen of **27** to ''hit'' **D1**, an extra interaction that is absent in the corresponding crystal structure. Ligand **28** is also flipped, roughly interchanging the position of the thiolate and amide carbonyl ''teeth'' of the metal interaction site and overlaying its alkyl hydrophobe with **H3**, rather than the benzyl overlay seen in the crystal structures (Fig. 12b).

A second complication for this dataset is that the metal chelating groups that lie at its heart are not explicitly defined as a separate feature class in UNITY. The unsubstituted hydroxamic acid moiety in **27** and the carboxlate group in **32** would undoubtedly both have been anticipated as chelation sites in any such feature definitions, and the former would be assigned a higher gasp weight (i.e., chelation strength) than the latter. The *N*-alkylated hydroxamic acid in **29** and the phosphonamidates in **30** and **31** are much less obvious a priori candidates, however. In any event, adding such a new class of feature after the fact would

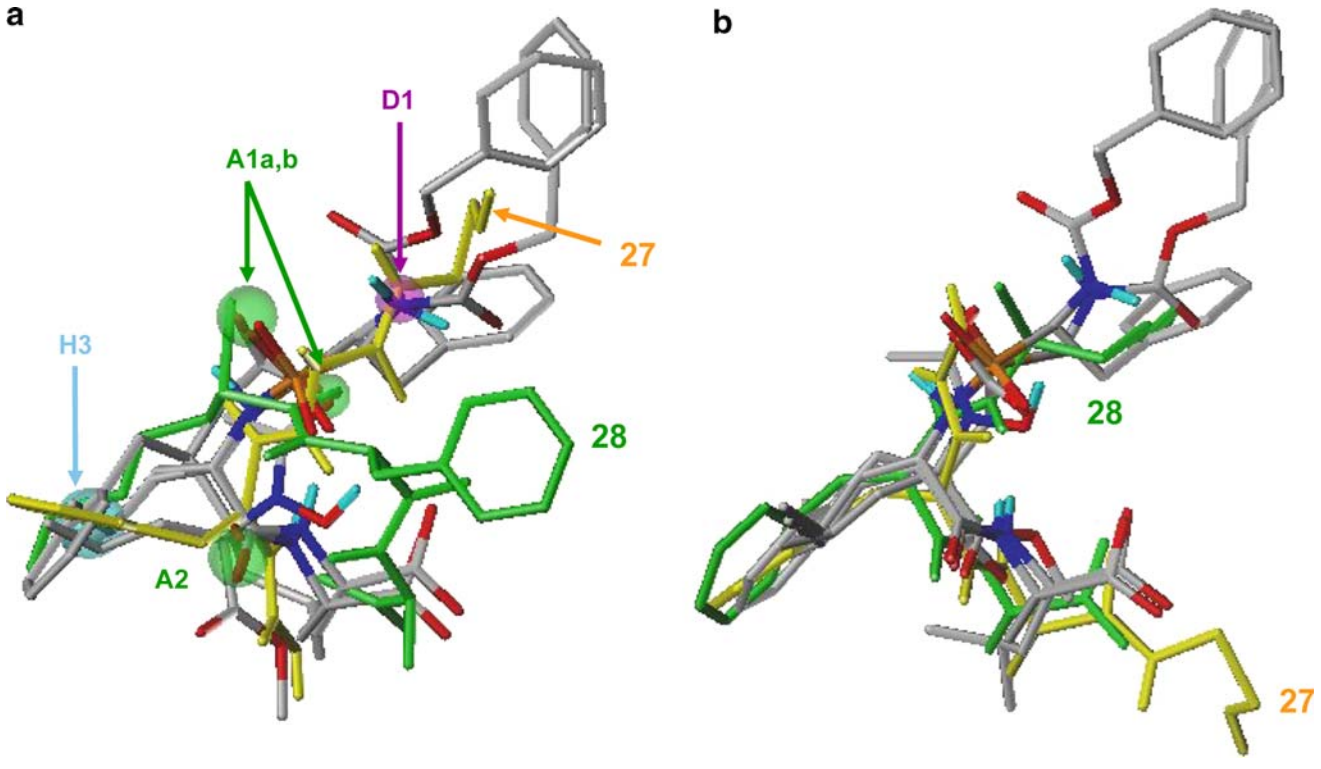

Fig. 12 Alignment of thermolysin ligands. Ligands **27** and **28** are highlighted in *yellow* and in *green*, respectively. The feature color scheme and naming conventions are as indicated for Fig. 4. (**a**) Rigid-body alignment obtained using default feature definitions. All features fall under a single 4/5 partial-match constraint. (**b**) Ligands overlaid based on the α carbon coordinates from the respective complexes with thermolysin

obviate the value of the analysis as an exercise in validation. Using the default feature definitions forced GALAHAD to rely on the presence of proximal acceptors and similarities in their environment to identify appropriate correspondences. This proved remarkably successful (Fig. 12a), especially considering the fact that the chelation strength of these groups does not track their propensity for accepting hydrogen bonds.

HIV reverse transcriptase

Figure 13 shows the structures of the HIV-1 reverse transcriptase (RT) inhibitors used by Patel et al. to evaluate the performance of DISCO, GASP and Catalyst. They noted that the only feature shared by all ten ligands—the ''pharmacophore''—is a rather ill-defined hydrophobic center. A broader target pharmacophore consists instead of a donor, a nearby acceptor and a hydrophobic center. Just as for CDK-2, the geometry of the hydrogen bonding interactions is that characteristic of a *syn* secondary amide or a lactam. Contrary to the literature indication [16], we find that neither the indole carboxamide in **38** nor the carbamate in **41** align with the donor and acceptor features in the pharmacophore evident from the crystal structure overlay. Rather, as Patel et al. noted for the primary amide in **42**, these substructures are located well outside the common interactions with the protein. Nor do these four ligands represent an alternative pharmacophore, since their pharmacophoric features do not overlay with each other.

Finding a pharmacophore shared by only six of the ten ligands in the dataset represents a rather daunting challenge. Moreover, the thiourea in **35** is a symmetrically degenerate donor-acceptor-donor variant of the key hydrogen bonding interactions, and the imide substructure in **36** presents an acceptor-donor-acceptor variant. Taken together, these complications make this a very difficult dataset from which to extract a coherent alignment with much pharmacophoric or pharmacosteric similarity to the overlay based on the corresponding crystal structures.

Applying GALAHAD to the crystal structure conformations yields the model shown in Fig. 14a, c. Two of the ligands that do not fit the pharmacophore—**38** and **41**—are excluded from the model altogether. The other two ligands that lack the target pharmacophore—**37** and **42**—are included in the model but do not ''hit'' the query deduced from them; in the interests of clarity, these have been omitted from Fig. 14.

The model reflects the characteristic bowl shape of the crystal structure overlay, as is evident from the perspective shown in Fig. 14c, d. Only ligand **36** (highlighted in yellow in Fig. 14) departs seriously from the target pharmacostere. Basically, it could have been aligned based on **H1** and **H3** or based on the donor and acceptor features. The former pair prevailed because they corresponded to ''bigger'' hyperfeatures that represent a broader consensus across the training set.

Several relevant donor and acceptor features in the model shown in Fig. 14a, c fall just outside the corresponding query feature tolerances, which suggested

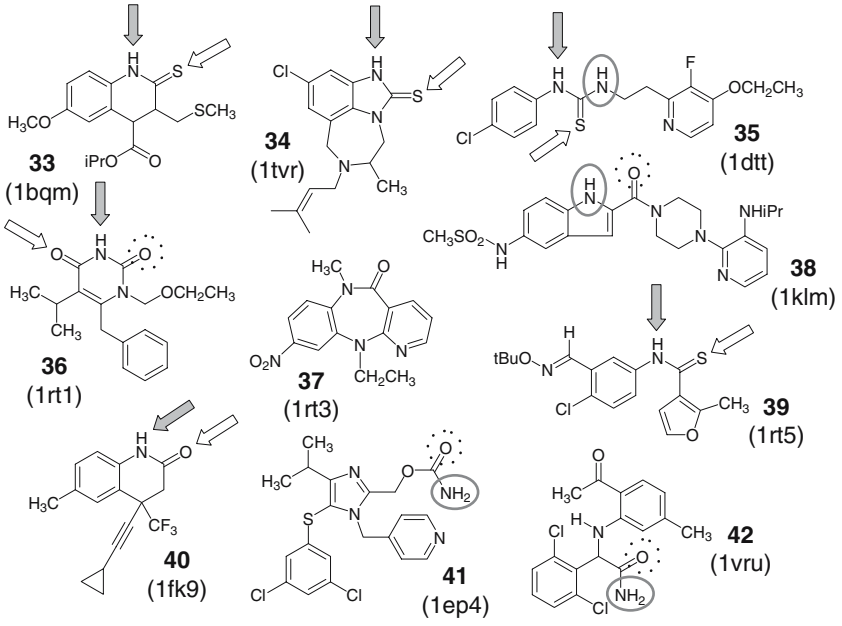

Fig. 13 Ligands from complexes in the reverse transcriptase dataset. Features are labeled as indicated for Fig. 6

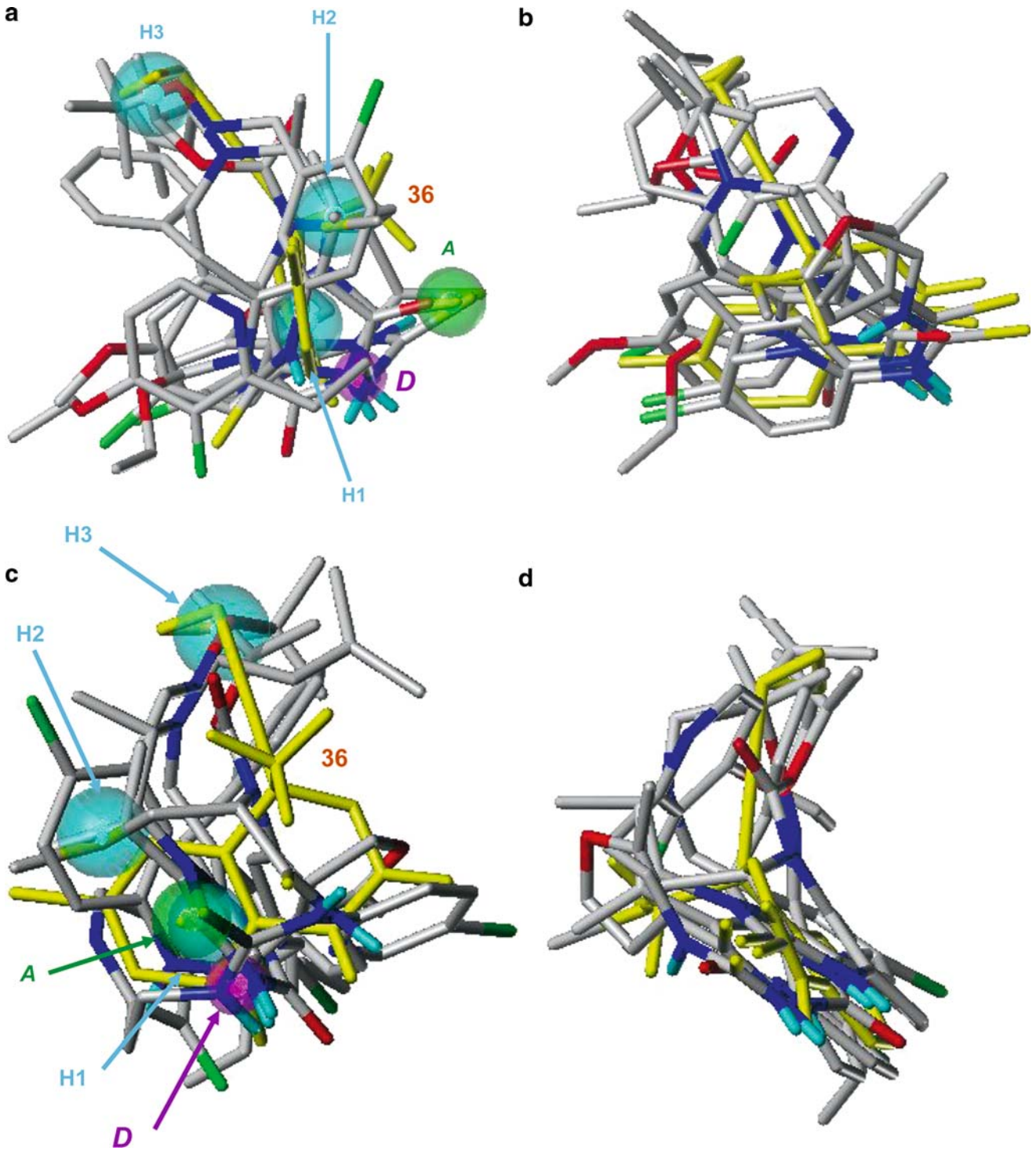

Fig. 14 Two roughly orthogonal perspectives on alignments of reverse transcriptase ligands. Ligand **36** is highlighted in *yellow*. Ligands **37** and **42**, which do not match either the target or the model query, have been omitted for clarity. The feature color scheme and naming conventions are as indicated for Fig. 4. (**a, c**) Rigid-body alignment obtained using default settings. Features *A* and *D* comprise a 1/2 partial match constraint; the three hydrophobic centers are required. t_{min} was set to 3. (**b**, **d**) Ligands overlaid based on the α carbon coordinates from the respective protein complexes

that the model could be improved by loosening the alignment constraints. In fact, rerunning the analysis with an increased averaging threshold (raised to 0.75 Å from the default value of 0.6 Å) and increased initial tolerance (raised to 1.2 Å from 1.0 Å) gave the much crisper alignment shown in Fig. 15. **H2** is retained in

this model, albeit with a slightly increased tolerance, whereas **H1** is shifted and the associated tolerance is tightened up. Ligand **36** is positioned correctly with respect to the other ligands in this model. The consensus hydrophobic center at **H3** disappears once the other ligand amides are more tightly aligned, which causes ligand **39** to be misaligned. It can be aligned based on the two remaining hydrophobic centers or based on the donor and acceptor pair, but those two alternatives conflict. The program opted to align the former rather than the latter, with a result that is at variance with the crystal structure overlay shown in Fig. 15b.

Discussion

LAMDA was originally developed for aligning pairs of molecules based on correspondences between atom position and atomic properties [14]. Such an atom-based technique is well suited to alignment problems involving more or less congeneric structural series, once relevant conformations have been defined. Not surprisingly, the extension to multi-way alignment described here works well in such cases (data not shown). It is less useful for structurally diverse datasets of the sort typically subjected to pharmacophore analysis, particularly when the ligands involved are flexible.

The GALAHAD program, in contrast, was primarily designed to carry out flexible alignment of ligands that exhibit similar interaction patterns and shapes when bound to a target protein—i.e., that share *pharmacophoric* and *pharmacosteric* elements. Unlike other pharmacophore elucidation methods, it operates in two distinct steps [6]. The first is a genetic algorithm that serves to generate sets of concordant conformations for the ligands of interest [7]. Subsequent application of a rigid-body alignment algorithm is then required to derive a common Cartesian frame of reference. The LAMDA methodology proved an excellent way to accomplish the latter task, once it had been extended to operate on pharmacophoric and steric features.

The work described here covers the construction of hypermolecules using the interaction strengths of the various pharmacophoric features as well as the distribution of other features around each. Overlapping features from the individual molecules are merged into hyperfeatures at each step, but the atoms defining those features do not merge. As a result, the final hypermolecule is composed of disjoint substructures—one for each ligand—along with a pharmacophore query constructed from the associated hyperfeatures.

One way to validate such a methodology is to assess its ability to reproduce crystal structure alignments.

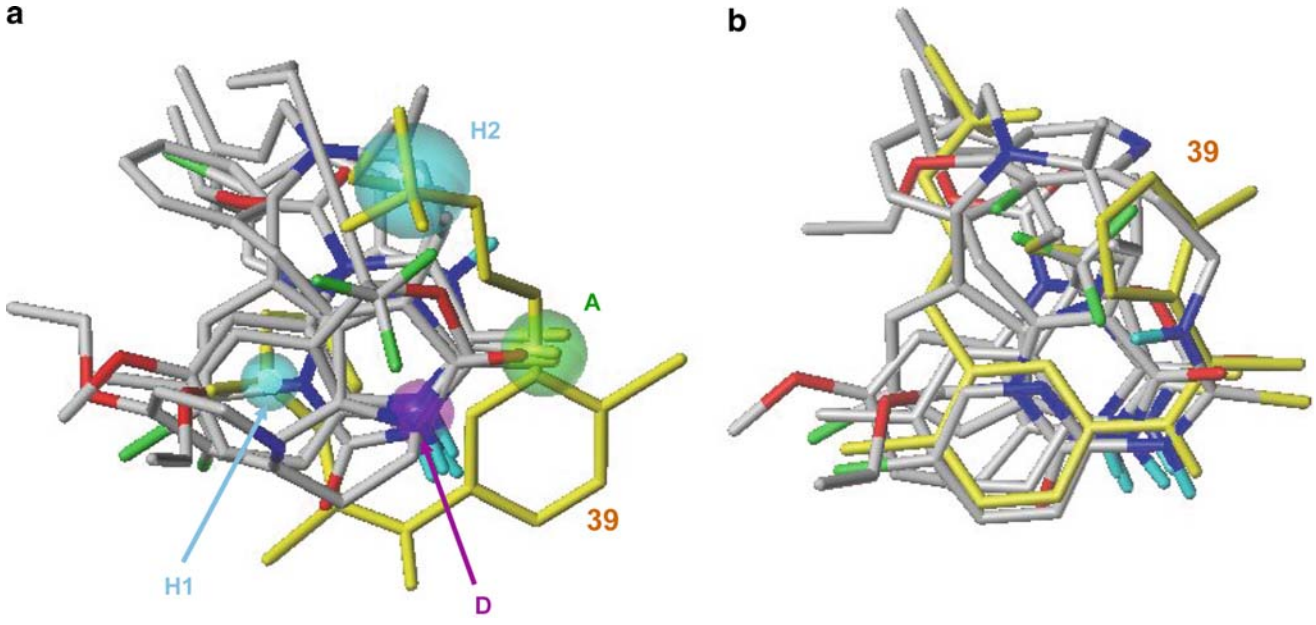

Fig. 15 Alignments of reverse transcriptase ligands. Ligand **39** is highlighted in *yellow*. The feature color scheme and naming conventions are as indicated for Fig. 4. (**a**) Rigid-body alignment obtained from GALAHAD with an averaging threshold of 0.75 Å and the initial tolerance set to 1.2 Å. The corresponding default values for these parameters are 0.6 and 1.0 Å, respectively. t_{min} was set to 3. All four features belong to a single 3/4 partial-match constraint. (**b**) Ligands overlaid based on the α carbon coordinates from the respective protein complexes

Rather than create our own test set, we turned to one from the literature: the one Patel et al. [16] compiled for evaluating the performance of three existing pharmacophore elucidation programs by comparing the models produced to the corresponding crystal structures. We have opted to validate rigid alignment using GALAHAD separately from the more general, flexible fitting applications, which will be described elsewhere [7]. This was done in part because it is difficult to know if good conformations are being produced unless one has confidence that ligands already in good conformations will be aligned well.

In fact, internal testing of the methodology during its development had focused on flexible fitting of GPCR ligands, where crystal structures are not available. Hence carrying out rigid alignments on these relatively naïve literature test sets provided a useful test of that part of GALAHAD which utilizes the LAMDA algorithm. In each case, the queries produced included the desired target pharmacophores, and the ligand overlays reflected the overall pharmacostere seen in the crystal structure overlay. In several cases these were improved significantly by modifying the default parameters for the released program somewhat. This is not altogether surprising, given that the default values used were originally designed for aligning flexible ligands with more consistent bond lengths, bond angles and internal torsions. Indeed, it is somewhat surprising that adjustments as small as 0.2 Å were enough to compensate for potential errors due to limited X-ray structure resolutions.

Further test of the methodology lies in the ability to align conformations generated for new candidate ligands to known bound conformations, and in the ability of the queries obtained to discriminate between known actives and inactives [31]. These and related database searching applications will be explored elsewhere [7, 32].

The adoption of an external test set led to some unexpected complications. In particular, the program seeks to maximize the pharmacophoric and steric overlap among rigid ligands. This goal presumes the existence of a shared interaction pattern and a shared shape—i.e., that a pharmacostere exists as well as a pharmacophore. Some ligand sets fail to meet this criterion because the corresponding target has a very open binding site, exhibits varying amounts of bound water, or both. For the most troublesome cases among the datasets compiled by Patel et al. [16], however, it is because of substantive differences in binding site configuration.

This is illustrated in Fig. 16, which compares the crystal structure overlay for four of the DHFR binding sites with the overlay for four of the reverse transcriptase inhibitors. The backbone and sidechains in the RT binding site are clearly more disordered than in DHFR. Indeed, getting an alignment at all was difficult in the former case; the pairwise root mean square deviations for the backbones ranged from 0.7 up to 3.7 Å across the RT dataset, with the largest deviation being between complexes of two ligands (**33** and **39**) that both exhibit the target pharmacophore. This compares with a heavy-atom RMSD of 3.2 Å for the model shown in Fig. 15a, an RMSD that drops to 2.5 Å when the one obviously mis-aligned ligand is excluded. The analogous overlays for CDK-2 are similar to that for RT, reflecting the large conformational changes many kinases undergo during activation. Such induced-fit variations in protein structure present well known difficulties for docking programs [33], but their interference with ligand-based alignment programs has not been widely appreciated to date.

It is reassuring in this context that the least well-reproduced crystal structure alignments are those in which the α-carbon alignment itself is least robust, i.e., those in which the individual ligands ''see'' the most structurally dissimilar binding sites. Comparing RMSDs for purely ligand-based alignments is meaningless in such circumstances and more subjective criteria become relevant [16, 33].

Validation sets created based on the availability of crystal structures are particularly likely to encounter this problem, in part because flexible binding sites can accommodate a wider structural range of ligands than can more rigid targets. This can in turn lead to increased synthetic activity around the target protein and a wider range of crystal structures becoming available. Conversely, targets for which many complexes are available because of their established, high therapeutic potential are likely to have their range of binding site flexibility more widely explored.

The analyses described here indicate that being able to specify the role of potentially amphoteric features in the training set is likely to improve the ability of GALAHAD to reproduce CDK-2 crystal structure alignments. Being able to specify the location of metal binding sites in thermolysin ligands would likely be useful, as would providing differential weighting to favor mapping of hydrogen bonding features over hydrophobic ones which would probably serve to reorient the misaligned single ligands in the RT models shown in Figs. 14 and 15. Given that several of the test sets present only relatively non-selective pharmacophores and weak pharmacosteres, it seems clear that GALAHAD is likely to perform at least as well on

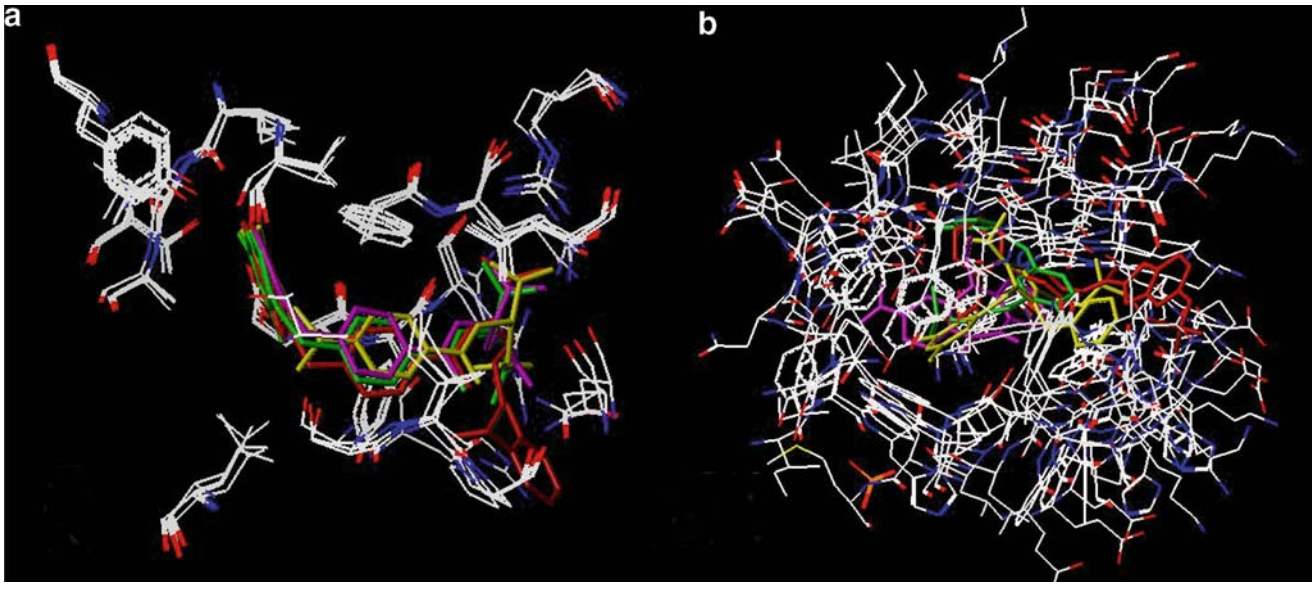

Fig. 16 Overlays of enzyme binding sites based on the coordinates of their α carbon backbones. (**a**) Overlay for **1drf**, **2dhf**, **1hfp** and **1ohk**. Ligands **1–4** are highlighted in *green*, *purple*, *yellow* and *red*, respectively. (**b**) Overlay for **1rt3, 1klm**, **1rt5** and **1ep4**. Ligands **37**, **38, 39** and **41** are highlighted in *purple*, *red*, *green* and *yellow*, respectively

targets of current medicinal chemical interest as on those considered here.

Conclusion

The linear assignment method for dataset alignment (LAMDA) has been extended to support multi-way rigid-body alignments of large datasets. Furthermore, it has been extended from being a strictly atom-based method to being able to operate on ionic, hydrogen bonding, hydrophobic, and steric features. The ability to generate partial-match search queries from the hypermolecules produced has been added as part of its incorporation into the GALAHAD program. Working from frozen conformations, the method was able to generate pharmacophores and pharmacosteres in good agreement with crystal structure alignments for a range of literature datasets. The inclusion of partial match constraints in the queries produced gave pharmacophores that were consistently a superset of full-match pharmacophores identified in previous analyses, with the additional features representing points of potentially beneficial interaction with the target protein.

Acknowledgments Tripos Inc. funded the work described here. We would like to thank Martin Bohl of Tripos GmbH and Jerk Vallgårda, Evert Homan, Anna-Lena Gustavsson, Peter Brandt, and Maria Wirstam of Biovitrum for their encouragement and support during the development of GALAHAD. We would also like to thank anonymous reviewers for their remarkably careful reading of the manuscript and insightful comments.

Appendix: Assigning features to partial match sets

Suppose that clusters of query features have been sorted in decreasing order of how many features they contain—i.e., of how many of the n ligands in a hypermolecule they "hit." Suppose further that one wishes to distribute the features in such a way that features hitting the same number of ligands fall under the same partial match constraint. Then the feature centroids q_i representing the clusters $i = 1,2,...$ can be allocated among partial match constraints in the query by applying the following method:

1. Let $k(q_i)$ be the number features in the cluster represented by q_i—i.e., the number of ligands that hit query feature q_i;
2. Drop any q_i for which $k(q_i) < t_{min}$, where t_{min} is the minimum number of ligands that each query feature must "hit."
3. Initialize the partial match sets $\mathbf{q}_2$ and $\mathbf{q}_1$as empty sets;
4. If the number of features $|q_i| < 5$, set $\mathbf{q}_1 = \{q_i\}$ and go to Step 15;
5. Set $t = \max(k(q_i))$;
6. If $k(q_i) = t$, add q_i to $\mathbf{q}_1$;
7. Set $t = t - 1$;
8. If $t < \min(t_{min}, 0.75n)$, go to step 15;
9. If the cardinality $|\mathbf{q}_1| < 3$ and some features q_i have not been assigned, go to step 6:
10. If $k(q_i) = t$, add q_i to $\mathbf{q}_2$;

11. Set $t = t - 1$;
12. If $t < t_{min}$, go to step 14;
13. If $|\mathbf{q}_1| + |\mathbf{q}_2| < 8$, go to step 10;
14. If $|\mathbf{q}_2| = 1$ and $|\mathbf{q}_1| = 3$, set $\mathbf{q}_1 = \mathbf{q}_1 \cup \mathbf{q}_2$ and set $\mathbf{q}_2$ equal to the empty set;
15. If $|\mathbf{q}_1| \leq 3$, mark all q_i in $\mathbf{q}_1$ as required matches and go to step 18;
16. if $|\mathbf{q}_1| = 4$:
    a. If $k(q_i) = n$ for any q_i in $\mathbf{q}_1$, mark all q_i in $\mathbf{q}_1$ as required matches and go to step 18; else
    b. Set the minimum partial match for $\mathbf{q}_1$ (min_1) to 3 and go to step 18;
17. If $|\mathbf{q}_1| > 4$, set $min_1 = 4$;
18. Set min_2 as follows:
    a. $min_2 = 0$ if $|\mathbf{q}_2| = 1$;
    b. $min_2 = 1$ if $|\mathbf{q}_2| = 2$;
    c. $min_2 = 2$ if $|\mathbf{q}_2| = 3$;
    d. $min_2 = 5 - min_1$ or to 2, whichever is greater, if $|\mathbf{q}_2| > 3$.

References

1. Kubinyi H, Folkers G, Martin YC (eds) (1998) 3D QSAR in drug design. Kluwer/ESCOM, Leiden
2. Ferguson AM, Heritage T, Jonathon P, Pack SE, Phillips L, Rogan J, Snaith PJ (1997) J Comput Aided Mol Des 11:143
3. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S (2000) J Med Chem 43:3233
4. Cramer RD III, Patterson DE, Bunce JD (1988) J Am Chem Soc 110:5959
5. Klebe G, Abraham U, Mietzner T (1994) J Med Chem 37:4130
6. Clark RD, Abrahamian E, Strizhev A, Wolohan PRN, Abrams C (2005) 230th ACS National Meeting, Washington, COMP 137
7. Clark RD, Abrahamian E, Abrams C, Brandt P, Gustavsson A-L, Homan E, Metwally E, Richmond NJ, Strizhev A, Wirstam M, Wolohan P (manuscript in preparation)
8. Lemmen C, Lengauer T (2000) J Comput Aided Mol Des 14:215
9. Vladutz G, Gould SR (1988) In: Warr WA (ed) Chemical structures. The international language of chemistry. Springer, Berlin Heidelberg New York, pp 371–384
10. Brown RD, Downs GM, Jones G, Willett P (1994) J Chem Inf Comput Sci 34:47
11. Brown N, Willett P, Wilton DJ, Lewis RA (2003) J Chem Inf Comput Sci 43:288
12. Palyulin VA, Radchenko EV, Zefirov NS (2000) J Chem Inf Comput Sci 40:659
13. Raymond JW, Willett P (2002) J Comput Aided Mol Des 16:521
14. Richmond NJ, Willett P, Clark RD (2004) J Mol Graph Model 23:199
15. GALAHAD™ is distributed by Tripos, Inc., 1699 S. Hanley Rd., St. Louis MO 63144 USA, http://www.tripos.com
16. Patel Y, Gillet V, Bravi G, Leach AR (2002) J Comput Aided Mol Des 16:653
17. Belongie S, Malik J, Puzicha J (2002) IEEE Trans Pattern Anal Mach Intell 24:509
18. Abrahamian E, Fox PC, Nærum L, Christensen IT, Thøgersen H, Clark RD (2003) J Chem Inf Comput Sci 43:458
19. Murtagh F (1983) Comput J 26:354
20. Barnard JM, Downs GM (1992) J Chem Inf Comput Sci 32:644
21. Individual molecules have conformations but the accompanying position in space—the configuration—is only meaningful in relation to another molecule, e.g., in complex with a protein or in a hypermolecular alignment
22. Jones G, Willett P, Glen RC (1995) J Comput Aided Mol Des 9:532
23. GASP™ is distributed by Tripos, Inc., 1699 S. Hanley Rd., St. Louis MO 63144 USA, http://www.tripos.com
24. Ghose AK, Crippen GM (1982) J Med Chem 25:892
25. SYBYL® is distributed by Tripos, Inc., 1699 S. Hanley Rd., St. Louis MO 63144 USA, http://www.tripos.com
26. It should be noted in passing the N^1-H tautomers given on the Research Collaboratory for Structural Bioinformatics (RCSB) web site (http://www.pdbbeta.rcsb.org/pdb/Welcome.do) for folate (**1**) and 5-deazafolate (**2**) would yield a grossly incorrect alignment if taken literally and used in place of the N^3-H tautomers shown in Fig. 3. The exact choice of tautomer provided to the program may affect the hydrogen bonding strength (gasp weight) assigned to each feature, but will generally have little or no qualitative effect on the models produced
27. Cottrell SJ, Gillet VJ, Taylor R, Wilton DJ (2004) J Comput Aided Mol Des 18:665
28. Under the macro definitions used by default here and in SYBYL 7.2, anilinic nitrogens in general are recognized as both acceptor and donor atoms. This reflects that fact that the pK_a of anilines can fall above or below physiological pH values, depending on exactly how they are substituted. Modifying the macro definitions to account for the fact that the exo nitrogens in aminopterins are not basic has no affect on the results shown in Fig. 3 aside from the disappearance of **A3** from the model query
29. Note that ^3N in **18** is "seen" as both a donor and an acceptor by GALAHAD, because of the potential for tautomerization with ^1NH
30. Langer, T (2005) Fifth European Workshop in Drug Design, Siena, Italy, 29 May–5 June 2005, http://www.unisi.it/EWDD, http://www.inteligand.com/demos/cdk2-workshop-siena.pdf
31. Mason JS, Good AC, Martin EJ (2001) Curr Pharm Des 7:567
32. Shepphird JK, Clark RD (2006) J Comput Aided Mol Design (in press)
33. Cole JC, Murray CW, Nissink JW, Taylor RD, Taylor R (2005) Proteins 60:325