How to Measure the Similarity Between Protein Ligand-Binding Sites?

Esther Kellenberger*, Claire Schalon and Didier Rognan

Bioinformatics of the Drug, UMR 7175 CNRS-ULP (Université Louis Pasteur- Strasbourg I), F-67400 Illkirch, France

Abstract: Quantification of local similarity between protein 3D structures is a promising tool in computer-aided drug design and prediction of biological function. Over the last ten years, several computational methods were proposed, mostly based on geometrical comparisons. This review summarizes the recent literature and gives an overview of available programs.

A particular interest is given to the underlying methodologies. Our analysis points out strengths and weaknesses of the various approaches. If all described methods work relatively well when two binding sites obviously resemble each other, scoring potential solutions remains a difficult issue, especially if the similarity is low. The other challenging question is the protein flexibility, which is indeed difficult to evaluate from a static representation. Last, most of recently developed techniques are fast and can be applied to large amounts of data.

Examples were carefully chosen to illustrate the wide applicability domain of the most popular methods: detection of common structural motifs, identification of secondary targets for a drug-like compound, comparison of binding sites across a functional family, comparison of homology models, database screening.

Keywords: Local similarity, three-dimensional comparison, protein structure, binding site, molecular shape, drug design, function assignment.

INTRODUCTION

The three-dimensional (3D) structure of biological macromolecules (proteins, nucleic acids) is of utmost importance to decipher the molecular machinery of living cells. In drug discovery, rational structure-based approaches are often used to design low molecular-weight compounds (from hereon called ligands) aimed at activating or inhibiting the function of therapeutically relevant proteins. Until the beginning of the nineties, the number of biological macromolecules of known 3D structures had been rather constant and limited. This number is now constantly increasing thanks to spectacular progresses in molecular and structural biology and also thanks to worldwide structural genomics initiatives.

Because adverse reactions of many drugs (side effects, toxicity) are basically achieved by permissive binding to several off-target proteins, anticipating similarities between protein ligand-binding sites is nowadays a strategic advantage in profiling novel drug candidates. If important innovations in algorithmic have taken place over the late nineties for measuring the 3D similarity between low molecular-weight compounds [1] and aligning molecules accordingly [2], other approaches were first conceived for the local 3D comparison of proteins because of their large size and their limited chemical diversity. Amino acid sequence information and secondary structure elements (α-helices, β-strands and turns) unambiguously recognize identical folds [3, 4]. Approaches based on amino acid sequence-alignment (e.g. pvSAOR [5]) were described to find local similarity between proteins, but they imply the identical arrangement in sequence of a common 3D pattern. Structural alignment of well defined 3D templates (e.g. PINTS [6]) have been applied to the study of conserved residues in protein structures [7]. Recently, new methods have focused on simplified protein representations [2] and benefit from the developments made for 3D comparison of ligands, molecular docking and more generally shape matching.

The current review reports progress in 3D methods for locally comparing proteins in order to derive biologically relevant similarity between ligand-binding sites. A first section describes the comparison methods and reports available programs. The second section discusses the problems specific to structural comparison of protein binding sites. The last section classifies the applications described in the literature.

METHODS & PROGRAMS

Structure-based methods for local comparison of unrelated proteins typically use a simplified representation of cavity residues, by partitioning of the surface into small patches which are then treated as geometric patterns or numerical fingerprints.

Methods Based on Geometric Pattern Comparisons

All described methods follow the same three-step flowchart (Fig. **1**). First, the structures of the two proteins are parsed into meaningful 3D coordinates in order to reduce the complexity of the pairwise comparison. Typically, only key residues are considered and described by a limited number of points, which are labeled according to pharmacophoric, geometric and/or chemical properties of their neighborhood. Second, the two resulting patterns are structurally aligned using the transformation that produces the maximum number of equivalent points. Last, a scoring function quantifies the similarity based on aligned features.

*Address correspondence to this author at the Bioinformatics of the Drug, UMR 7175 CNRS-ULP (Université Louis Pasteur- Strasbourg I), 74 route du Rhin, B.P. 24, F-67400 Illkirch, France; Tel: +33-3-90244224; Fax: +33-3-90244310; E-mail: esther.kellenberger@pharma.u-strasbg.fr

1573-4099/08 $55.00+.00

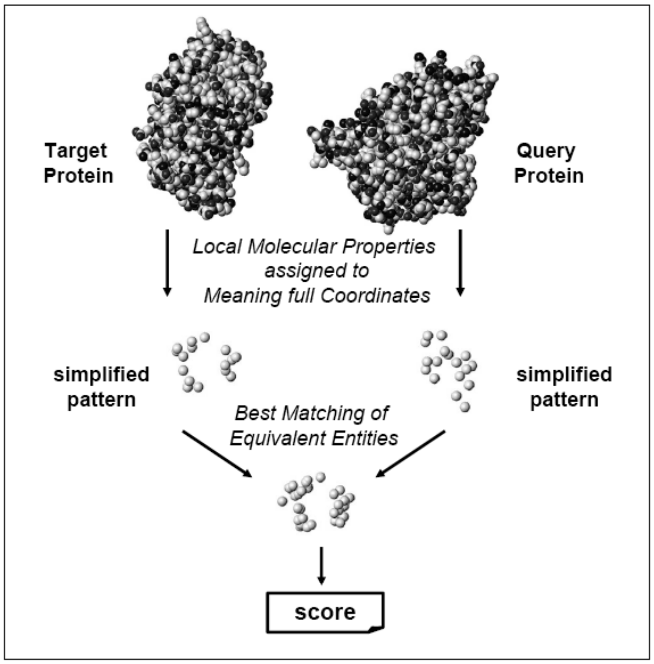
Target Protein
Query Protein
Local Molecular Properties assigned to Meaning full Coordinates
simplified pattern
simplified pattern
Best Matching of Equivalent Entities
score

Fig. (1). General scheme for the comparison of protein binding sites.

Simplified Representation of Protein Cavities

The basic assumption in the search for similarity between active sites is that only residues close to the surface need to be considered. These amino acids may directly or indirectly (*e.g. via* a water molecule) be involved in ligand recognition. Depending of the aim of comparison, all surface residues [8] or only cavity-flanking residues are taken into account. The cavity is either deduced from the analysis of the protein surface shape (ProSurfer [9], SiteEngine [8], CavBase [10]), or defined from the distance of cavity-lining residues (generally from 5 to 7 Å) to a co-crystallized ligand (eF-site [11], SuMo [12], CPASS [13], SitesBase [14]). Whereas the former method only selects solvent accessible surface residues, the latter method may include a few buried residues.

Cavity residues are then transformed into either an irregular or a regular 3D arrangement of points (Fig. **2**). The irregular pattern results from the description of each residue by one (CPASS, SURFACE [15]) or all (ProSurfer, SitesBase) of its atoms or by pseudo-atoms describing a group of atoms (CavBase, SuMo and SiteEngine). In addition to its 3D coordinates, each pattern point is annotated according to physicochemical or pharmacophoric properties of the corresponding atom(s). Molecular information on the binding site may also be encoded into a lattice-based representation of the molecular surface. Points originate from triangulation of the Connolly surface or from the selection of surface atoms-contacting points in a regular-spaced Cartesian grid. Descriptors are assigned to each point according to the property of the closest surface atom or pseudo-atom and/or to surface properties [11] such as electrostatic potential, hydrophobicity and curvature.

Geometric Search for the Best Structural Alignment

Whatever the representation, similarity between the cavities of two proteins is always inferred from the best struc-

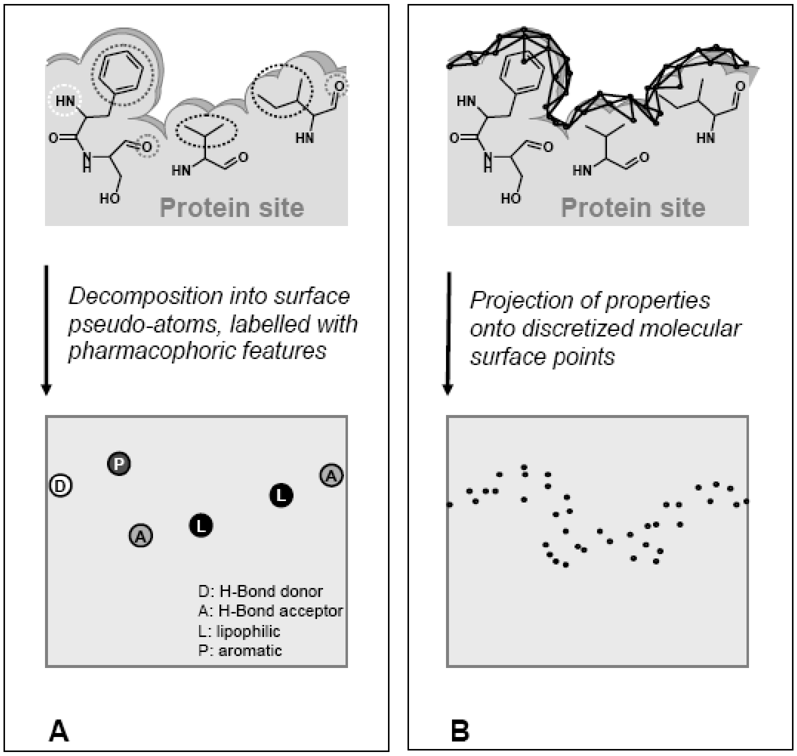
HN
O
N
H
O
HO
HN
O
O
HN
Protein site
Decomposition into surface pseudo-atoms, labelled with pharmacophoric features
D
P
A
L
L
A
D: H-Bond donor
A: H-Bond acceptor
L: lipophilic
P: aromatic
A
Protein site
Projection of properties onto discretized molecular surface points
B

Fig. (2). Description of the protein binding site by pattern of points. (**A**) Example of irregular pattern of points. The protein binding site is represented by pseudo-atoms, also called pseudo-centers, *i.e.* points labelled by the pharmacophoric features of corresponding surface-exposed functional groups. The coordinates of a pseudo-atom are usually close the mass centre of the corresponding functional group. (**B**) Example of regular pattern of points. The protein surface, usually following Connolly's definition [55], is decomposed into triangle mesh vertices. Each vertex is assigned a label according to protein local properties.

tural alignment of corresponding patterns. The search for the best structural alignment of two geometrical patterns consists in seeking the rigid body transformation that optimizes the number of matched elements.

The simplest method is the exhaustive **iterative search for the best translation/rotation** of the query pattern keeping the target pattern fixed; a coarse search identifies the most similar regions which are then more precisely explored (CPASS). This method requires a scoring function to steer the exploration. Interestingly, no connectivities between patterns of points are defined. The method hence does not promote local over global alignment of patterns. It however has the disadvantage of being relatively slow.

Pattern complexity can be reduced by grouping close elements into triplets, then testing all combinations of triplet pairs (ProSurfer, SuMo, SitesBase). The **geometric matching** of triplets is performed "on-the-fly" (e.g. by least-square fitting) and yields several transformations to superimpose query and target patterns. During a subsequent stage, transformations are ranked according to the number of well aligned elements (ProSurfer, SitesBase) or all identified local similarities are simply extended to adjacent congruent triplets (SuMo). An efficient method to compare two sets of triplets is **geometric hashing** (SiteEngine, SitesBase). Geometric hashing simply uses the spatial relationship between the points into the triplet; hence, two triangles are compared based on simple measurements (e.g. length of triangle edges in SiteEngine), regardless of the orientation (Fig. **3A**). The hashing procedure generates a triplet of numbers {x,y,z}, namely the hash key, that uniquely describes a given triangle. A hash table is created for the set of points of the target protein site, and each triplet is stored in a hash bin associated with its key (triplets similar in shape are stored in identical or close bins). To recognize the structural motifs of the target in the query structure, the hash keys generated for all possible query triplets are searched in the target hash table. Similar triplets are aligned, so that each pair of matched triangles defines a transformation. The most frequent transformation corresponds to the largest alignment, and the best structural similarity in terms of numbers of well aligned triplets. The hash tables can easily be stored in a database for later comparisons. In SiteEngine, a triplet of non co-linear points is only considered if the corresponding edge lengths are within a predefined range, and the hash key includes an additional physicochemical index, which encodes the properties of the triangle vertices. If geometric hashing is computationally efficient, appropriate scoring rules are necessary to avoid over-emphasing local geometries.

Another frequently used method to detect similarity between geometric patterns is **clique detection** [16] (CavBase, eF-site (Fig. **3B**)). Query and target patterns are represented as graphs: pattern elements are nodes (pseudocenters in Cavbase, vertices in eF-site) connected by edges if the associated distance is within a defined range (< 12Å in CavBase). A product graph is then obtained by pairing nodes with identical labels from query and target graphs, then connecting pairs if the edges connecting the two elements in the query and the target graphs share the same value. Labels are physicochemical descriptors (CavBase) or electrostatic potentials and curvatures (eF-site). Edge value is the distance between the two elements of the corresponding node. Typically, distance values have to be similar rather than strictly identical, with a fixed tolerance in eF-site (1.5 Å) and a distance dependant tolerance in CavBase [17]. The maximum

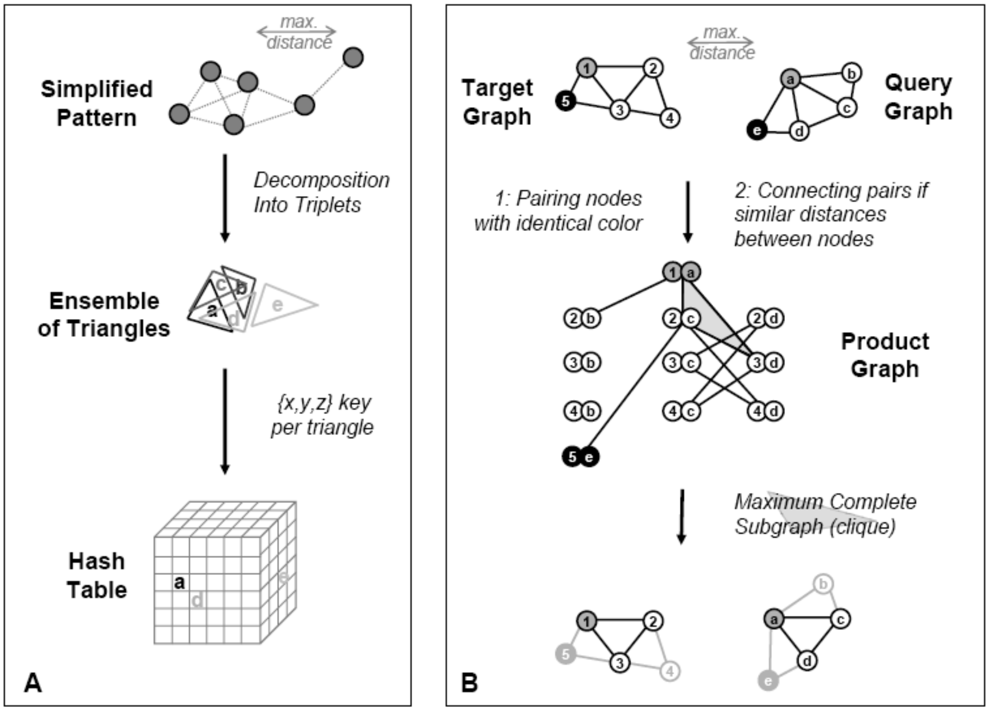

Fig. (3). Search for the best structural alignment of patterns of points. (**A**) Geometric hashing. Because of the inter-point distance threshold used to generate the triplets, no triangles include the far-right point. For the sake of clarity, no labels were associated to points. **(B)** Clique detection. In the graph definition, nodes are connected if closer than a distance threshold. Point's properties are represented by different achromatic colors.

common subgraph isomorphism, i.e. the largest subset of adjacent pairs of target nodes that can be mapped to adjacent pairs of query nodes, corresponds bijectively to the largest clique of the product graph, i.e. the maximal subgraph of the product graph in which every node is connected to every other node and is not contained in any larger subgraph with this property. The Bron-Kerbosh algorithm [18] used in both CavBase and eF-site enables all substructures common to query and target patterns to be identified.

Scoring the Similarity

Scoring the similarity between two cavities is based on the number of aligned features of corresponding patterns. Elementary scoring functions consist in simple counts or derived percentages. Since percentage depends on the ordering of query and target, a Tanimoto-like index is often computed (number of aligned features divided by total number of features in both patterns). A weighted count of matched points can emphasize geometry (coefficient: RMSD) or/and given properties (coefficient depending on descriptors). The most sophisticated scoring schemes involve hierarchical stages with an increased precision of site spatial overlap description. Hence, in CavBase and SiteEngine, the selection of the best alignment of query and target sites involves examination of both the patterns of points and the corresponding molecular surfaces.

Comparison of Cavity Fingerprints

Recently, Schalon *et al.* proposed an alternate strategy to compare binding sites. The program, called SiteAlign [19], transforms user-defined cavity residues into fixed-length integer, i.e. fingerprint (Fig. **4**). Numeric representation of the protein impacts the search for similarity, which does not consist in direct overlay of geometric patterns.

A 1-Å radius sphere, whose surface is discretized into 80 triangles, is placed at the cavity center. A vector is derived from the Cβ carbon of each residue to the sphere center, and 3 topological and 5 pharmacophoric descriptors of the residue are assigned to the sphere triangle hit by the projection. The resulting pattern has a fixed size of 80 elements organized into an invariant 3D spherical arrangement. It is finally converted into a fingerprint of 640 integers (80 triangles × 8 descriptors). The pattern obtained by the projection of the cavity onto the discretized sphere is not unique for a given site. The relative position of triangles hit by the projection, as well as the topological values assigned to residues (typically the distance between the Cβ atom and the sphere center), depend on the position of the sphere into the site. When comparing two sites, the target site yields the reference fingerprint whereas a series of fingerprints are systematically generated for the query site by iterative rotation and translation of the sphere within its cavity. Similarity between sites is evaluated by the average of normalized differences for each descriptor of each triangle between compared fingerprints. Two scores were calculated respectively by considering all triangles with at least one non-null descriptor value in either the target fingerprint or the query fingerprint, and by considering only triangles with non-null descriptor values in

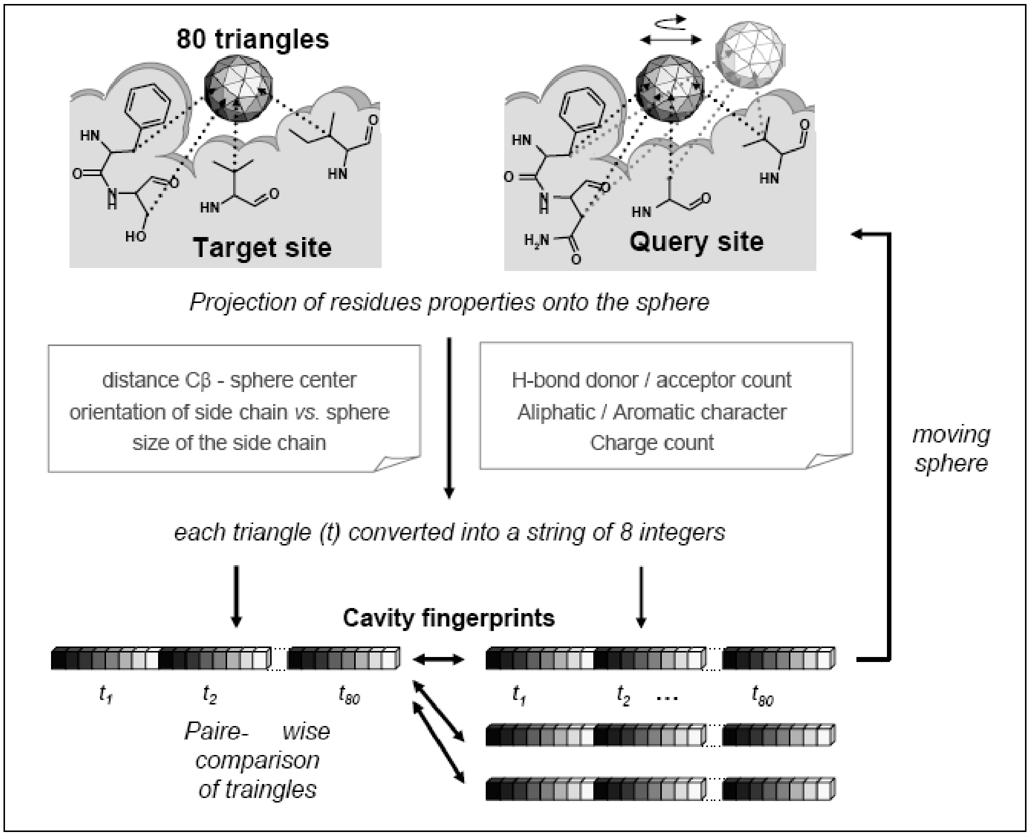
80 triangles
HN
O
N
H
HO
O
HN
O
HN
O
Target site
HN
O
N
H
H₂N
O
O
HN
O
HN
O
Query site
Projection of residues properties onto the sphere
distance Cβ - sphere center
orientation of side chain vs. sphere
size of the side chain
H-bond donor / acceptor count
Aliphatic / Aromatic character
Charge count
moving sphere
each triangle (t) converted into a string of 8 integers
Cavity fingerprints
t1
t2
t80
t1
t2 ...
t80
Paire- wise comparison of traingles

Fig. (4). Fingerprint-based comparison of protein binding sites. The geometrical and pharmacophorical properties of each residue of the site are projected onto a discretized sphere to generate a numerical fingerprint. A series of fingerprints are generated for the query site by rotation/translation of the sphere.

both the target fingerprint and the query fingerprint (a least 7 common triangles are required for a pairwise comparison). The first score is a global measure of similarity and the second one is more suited to detect local similarity among unrelated binding sites.

Available Programs and Databases

Table **1** gives an overview of methods that were published during the last years. Clique detection and geometric matching have been successfully used to find the largest local similarity between sites represented by patterns of

Table 1. List of Programs and Databases, Sorted by Publication Date

Name and Reference	Protein Representation			Search Method	Scoring
	Site Identification	Geometric Pattern	Point Label		
eF-Site [11]	Residues around a selected ligand, surface atoms only	Triangle mesh	Electrostatic potential, surface curvature	Clique detection, applied to small surface patches (edge < 12Å, 1.5 Å tolerance in product graph)	Descriptors-weighted count of matching points
CavBase [10]	Residues around a selected ligand, surface atoms only	Pseudo-atoms	Pharmacophoric features	Clique detection (edge < 12Å, distance-dependent tolerance in product graph)	Relative count of aligned surface-grid points (0.5 Å spacing) (distance ≤ 1 Å)
CSC [26]	Selection of residues	Representative side chain atoms	Element type	Detection of all four-node cliques, then combination of overlapping common structural cliques (CSC)	Count of aligned atoms
SuMo [12]	Residues around a selected ligand	Pseudo-atoms	Pharmacophoric features	Geometric matching of triplets (max. distance < 8Å), stepwise connexion of congruent triplets	Count of matching points, of matching residues
SitesBase [14]	Residues around a selected ligand	All atoms	Element type	Geometric matching of triplets, vote for best transformation	Count of matching points (RMSD < 1Å), statistical measure
SiteEngine [8]	Surface residues	Pseudo-atoms	Pharmacophoric features	Geomatric hashing, hierarchical scheme to select best transfomation (best overlay of patch centers, then surface and last pseudo-atoms)	Weighted counts of aligned patch centers (distance ≤ 3Å), surfaces grid points (distance ≤ 1-2-3 Å) and pseudo-atoms (distance ≤ 3-1 Å)
SURFACE/ Query3D [15]	Residues of surface clefts	All atoms	Residue type	Identification of similar pairs of residues (evolution matrix scores and RMSD), stepwise connexion of congruent similar pairs	Count of matched residues (RMSD ≤ 0.8Å)
CPASS [13]	Residues around selected ligand	Cα atoms	Residue type	Iterative search for best rotation/translation	RMSD-weighted sum of evolution matrix scores for pairs of matched residues
Zhang & Grigorov [30]	Residues around a selected ligand	Geometric center of side chain	Residue physico-chemical class	Clique detection	Count of matching residues / total number of residues
Ramensky *et al.* [45]	Atoms around a selected ligand	All atoms	Force-field chemical atom type	Clique detection	Percent of matched atoms
Park & Kim [31]	Residues around a selected ligand	Cα atoms	Residue type	Clique detection	Blosum62-weighted count of matched residues – penalty per unmatched residues
ProSurfer [9]	Surface atoms of protein cavities	Selection of atoms	Distance–weighted pharmacophoric features of neighborhood	Alignment of atom triads	Descriptor-weighted count of matched atoms (RMSD ≤ 2.5Å)
SiteAlign [19]	Residues around a selected ligand	Sphere discretized into 80 triangles	3 topological and 5 pharmacophoric features per residue projected onto sphere	Comparison of target and query numerical fingerprints ; iterative rotation/translation of sphere into query site	Average of normalized differences between fingerprint values

points in eF-site, CavBase, SitesBase and SiteEngine programs. In these methods, similarity is evaluated at the atomic level and is mainly determined by geometrical criteria.

The CPASS program consists in an iterative search for the best global structural alignment of two sets of Cα atoms. Similarity is evaluated at the residue level and is determined by the optimization of a score based on both the evolution substitution matrix BLOSUM62 and the RMSD between Cα positions.

In approaches taken in SuMo, CSC, SURFACE and ProSurfer, the similarity search starts with the initial selection of anchor points within sites represented by a pattern of points. These methods first look for highly similar elements between the query and a target structures, then only select significant matches and extend the similarity region. In SuMo, local similarities are identified by comparing triplets of pseudoatoms, and extended depending on the geometry of vicinal selected triplets. In CSC, local similarities are identified by comparing quadruplets of side chain atoms, and extended by combining quadrulets which contain identical atom pairs. Extension does not consider spatial overlap of enlarged common patterns. In the SURFACE database, each amino acid of a surface cleft is described by its residue type and represented by its atoms. Upon comparison of cavities, similar pairs of amino acids are identified based on Dayhoff evolution matrix scores and RMSD. The similarity region is extended by connexion of coherent similar pairs. The similarity is finally evaluated by the count of matched residues. Hirota and coworkers [9] proposed to describe protein cavities by surface atoms labelled with a numerical fingerprint that encodes the atomic neighborhood. For each atom, the fingerprint contains the distance-weighted count of close atoms classified by pharmacophoric type. Typically, six different types are considered, giving in a fingerprint length of 6. During the comparison of two sites, the first step is the selection of the most similar points between query and target sites (max 30 per site), by computing Euclidian distance or Tanimoto coefficient between atomic fingerprints. All possible triads of selected target and query points are then superimposed and the resulting transformations are ranked according to the fingerprint similarity of aligned atoms.

Last, SiteAlign uses an invariant geometrical object to represent the binding site, thereby switching the 3D problem towards a comparison of numerical fingerprints.

Most of the cited programs have been implemented into web servers that allow browsing a database of predefined cavities for similarity to any PDB cavity (Table **2**).

WHY IT IS DIFFICULT TO DETECT LOCAL SIMILARITY BETWEEN PROTEIN STRUCTURES

How to Accurately Represent Protein Ligand-Binding Sites?

From a computational perspective, ligand-binding site comparisons imply a simplification of the protein structure. From a biological perspective, the description must contain information about shape and physicochemical properties of sites. Intuitively, one may expect that the highest the level of information, the more precise the similarity prediction.

By using patterns of points, three parameters control the similarity search focus: i) the spatial resolution of the site representation, ii) the distance tolerance used to assign

Table 2. Available programs and Linked Databases

Name	Availability	Comparison	
		Target	Query
eF-site [41]	http://ef-site.hgc.jp/eF-seek/	user defined site	eF-site database: ~ 12,000 sites[a]
CavBase	module in Relibase+ (CCDC, Cambridge, UK)	Single entry of CavBase database	CavBase database: > 80,000 sites[a]
SuMo [49]	http://sumo-pbil.ibcp.fr/	user–defined site or full protein or Single entry of SUMO database	Subset or full SUMO database: > 45,000 sites[a]
SitesBase [50]	http://www.modelling.leeds.ac.uk/sb/	Single entry of SitesBase database	SitesBase database: ~30,000 sites[a]
SiteEngine [51]	http://bioinfo3d.cs.tau.ac.il/SiteEngine/	user–defined full protein or single PDB protein	User-defined site or single PDB site (required co-crystallized ligand)
CPASS	http://bionmr-c1.unl.edu/	user–defined site (required co-crystallized ligand)	CPASS database: ~21,000 sites[a]
SURFACE [52, 53]	http://cbm.bio.uniroma2.it/surface/	Single representative entry of SURFACE database	SURFACE database: 2,425 sites[b]
ProSurfer	http://d-search.atso-net.jp/top	Single representative entry of ProSurfer database	ProSurfer database: ~ 48,000 sites[a]
SiteAlign	http://bioinfo-pharma.u-strasbg.fr/template/jd/pages/download/download.php	user-defined site or sc-PDB site	user-defined site(s) or sc-PDB site(s)[c]

Ligand binding sites were extracted from the PDB[a], the non-redudant PDB[b]or the sc-PDB dataset [c] [54].

matching points and iii) the complexity of point descriptors. Resolution and tolerance both account for the comparison of shape, resolution being the determinant factor. Hence, small structural differences significantly alter grid-based definition of sites while little affecting the Cα atom-based description (Fig. **5**). In intermediate pseudoatoms-based representation (CavBase, SiteEngine), increasing the tolerance value decreases sensitivity to atomic coordinates. But, geometric comparison of small patterns can yield unreasonable matches, typically by underestimating side chain orientation in residue-based comparisons (i.e. matching a residue whose side chain points into the cavity with a residue whose side chain is buried into the protein core) or by ignoring the surface curvature when comparing pseudo-atoms (i.e. matching points from a concave area to that from a convex area (Fig. **6**). In CavBase and SiteEngine, the relevance of the computed similarity is therefore evaluated at the scoring stage. Descriptors of atoms, pseudo-atoms or residues are also crucial in the comparison, since they depict the physico-chemical properties of the sites of interest. Assuming biochemical function is due to ligand recognition mediated by non-bonded interactions, amino acid propensity to establish non-covalent interactions is well described by pharmacophoric features. For example, CavBase provides each residue with a detailed description of all H-bond donors, acceptors and donor-acceptors, including their directionality property, as well as all aliphatic and aromatic centers (each residue is converted in 3-8 key points). Residue types implicitly encode the physico-chemical properties of amino acids (CPASS, SURFACE), but evolution matrix scores (BLOSUM, PAM) used for comparison cannot be easily interpreted in terms of non covalent binding capacity, and emphasize conservation and rare residues (e.g. the frequency of tryptophan, isoleucine and leucine in protein is about 1%, 6% and 9%, respectively. The BLOSUM62 scores of tryptophan conservation, leucine conservation and leucine/isoleucine substitution are 11, 4 and 2, respectively. Consequently upon site comparison, the tryptophan/tryptophan 3D alignment contributes more to the final similarity score than leucine/leucine 3D alignment, which itself has more weight than leucine/isoleucine 3D alignment).

In SiteAlign, protein ligand-sites are described by residue-based fingerprints; the relative spatial arrangement of residues is both encoded by the ordering of strings and by values of the topological descriptors for each projected residue. Whereas the former drives the coarse similarity search, the latter is taken into consideration at the same level than the pharmacophoric features. SiteAlign is thus a properly balanced method which has been shown to tolerate large variation in atomic coordinates (up to 3Å RMSD during a protein unfolding simulation monitored by molecular dynamics).

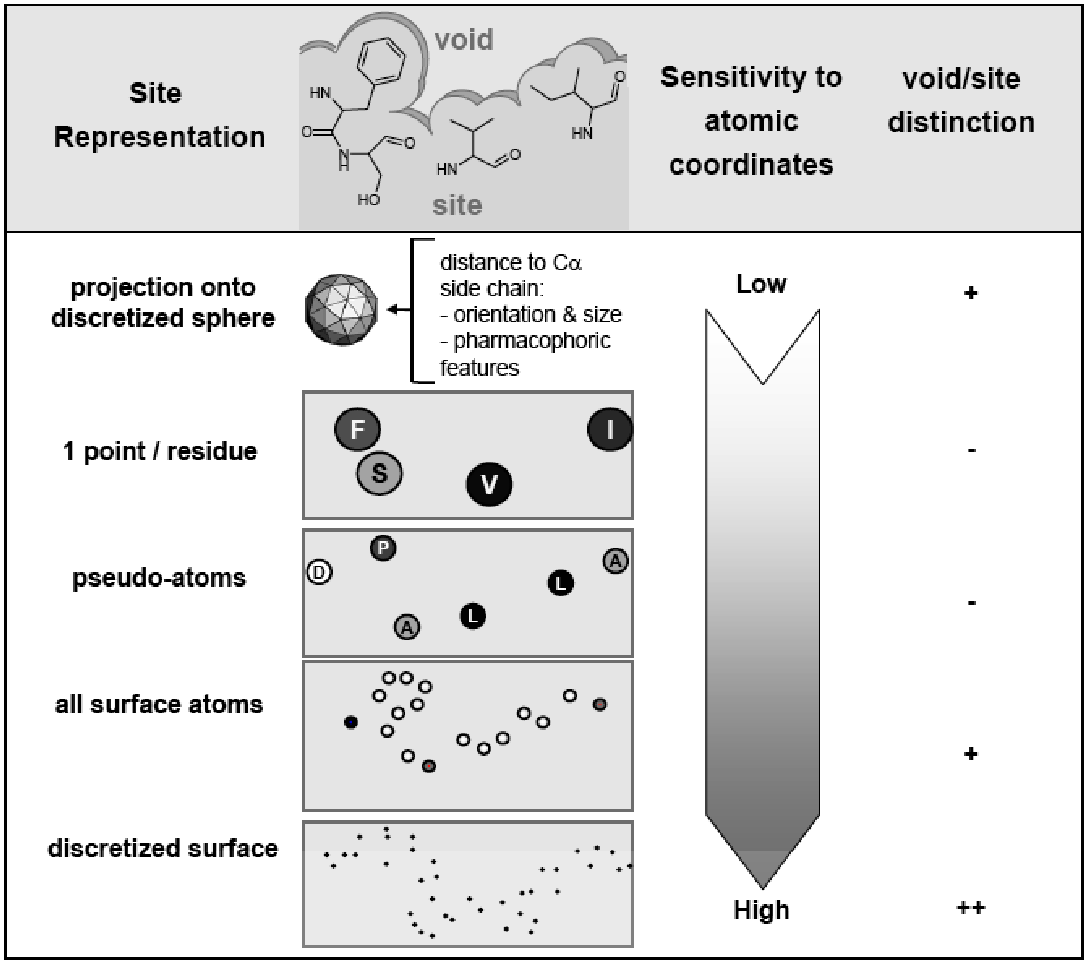

Fig. (5). Influence of site description on the focus of the similarity search.

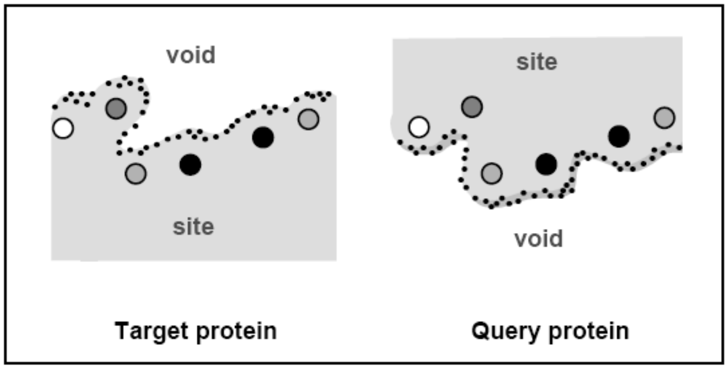

Fig. (6). Example of unrealistic match between two binding sites. The overlay of pseudo-atoms (black, white and grey disks) yields to alignment of a ligand-binding site to void, whereas the comparison of surface points (black dots) unambiguously distinguishes the two protein binding sites.

Proteins are in Constant Motion

The "working" proteins experience conformational changes, ranging from side chain fluctuations to more global motions, like the rearrangement of compact hydrophobic units upon change in the conformation of connecting loops (induced fit and conformational selection theory [20]). In addition, binding sites are often characterized by regions of both high and very low mobility [21, 22].

In all described site comparison methods, the input protein structures are treated as rigid objects. Although it is not feasible to model large scale motions, it is possible to account for the possible rotameric states of site residues by considering a fuzzy representation of the protein site (like in SiteAlign). Moreover, a hazy outline of the site resulting from the residues selection around a co-crystallized ligand may be better than a sharp image of the solvent accessible surface, since the accessibility of key residues may drop due to small reorientation of the side chains of vicinal residues.

The Scoring Issue: What is a Relevant Similarity?

In the comparison of two sites by methods based on clique detection and geometric matching/hashing, similarity scores are used to rank possible solutions by decreasing number of matching atoms or feature points. In both cases, proper ranking is achieved by counting matched points. It has been frequently noted that a minimal number of matched points is necessary to detect reasonable similarity.

As previously noticed, different representations may be used for the search and for the scoring, in order to discard wrong solutions. In CavBase and SiteEngine, ranking of the solutions found by comparing pseudo-atoms patterns typically involves the analysis of the degree of spatial overlap between aligned surfaces.

In iterative search for the best structural alignment of two sites (CPASS, SiteAlign) and in identification of highly similar regions (SuMo, SURFACE, ProSurfer), the score directly determines the explored solutions; distances between aligned pairs of amino acids guides similarity search (distances are generally evaluated by positional RMSD or by the distribution of non-null values in a cavity fingerprint), and descriptors distinguishes purely geometrical matches.

All described scoring schemes assign good values to positive matches, but do not penalize mismatches, which are dismissed or hardly contribute to the final score, like in SiteAlign and CPASS. Hence, they allow detection of similarities between parts of sites or between sites of different sizes, but also of fragmented similarities.

Most programs provide normalized scores, which are suitable for ranking similarities obtained for the comparison of more than two different sites. SiteAlign scores have been designed to range from 0 to 1, CavBase R_3 score is the number of matched pseudo-atoms divided by total number of pseudo-atoms, eF-site coverage index was likewise defined. In SuMo the number of matches is divided by the volume of target site. In SiteEngine and CPASS, the target/query score is divided by the query/query score. Global analysis of the score distributions obtained for extensive site comparisons were carried out using eF-site, SURFACE, ProSurfer and SiteAlign to derive Z-scores (computed by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation). In SitesBase, the statistical significance has been further addressed by removing evolutionary related proteins and generating a background extreme value distribution of scores to calculate the probability of obtaining a given score by chance (P-value) [23] and more recently by computing Poisson Index [24].

Whatever the method, the main difficulty is to estimate the biological significance of the proposed match, which itself depends on the investigation focus. Ultimate purpose of site comparisons may vary from 3D motif detection to ligand binding prediction or structural analysis of a catalytic site within a given functional protein family. As a consequence, scoring evaluation may require task-dependent adjustments. Along the same line, applicability of a normalized score will depend on the diversity of the database population. As an example, SiteAlign score thresholds were calibrated using a homogeneous dataset of druggable binding sites and are unlikely to work for non-druggable large crevices.

SCOPE AND APPLICATIONS

Local structure comparison methods all aim at finding similarity between proteins that are not necessarily close in sequence or fold space, with the final purpose of functional annotation or drug design. Suitable test cases were described for the validation of all described programs. Different approaches were presented and suggested the applicability of

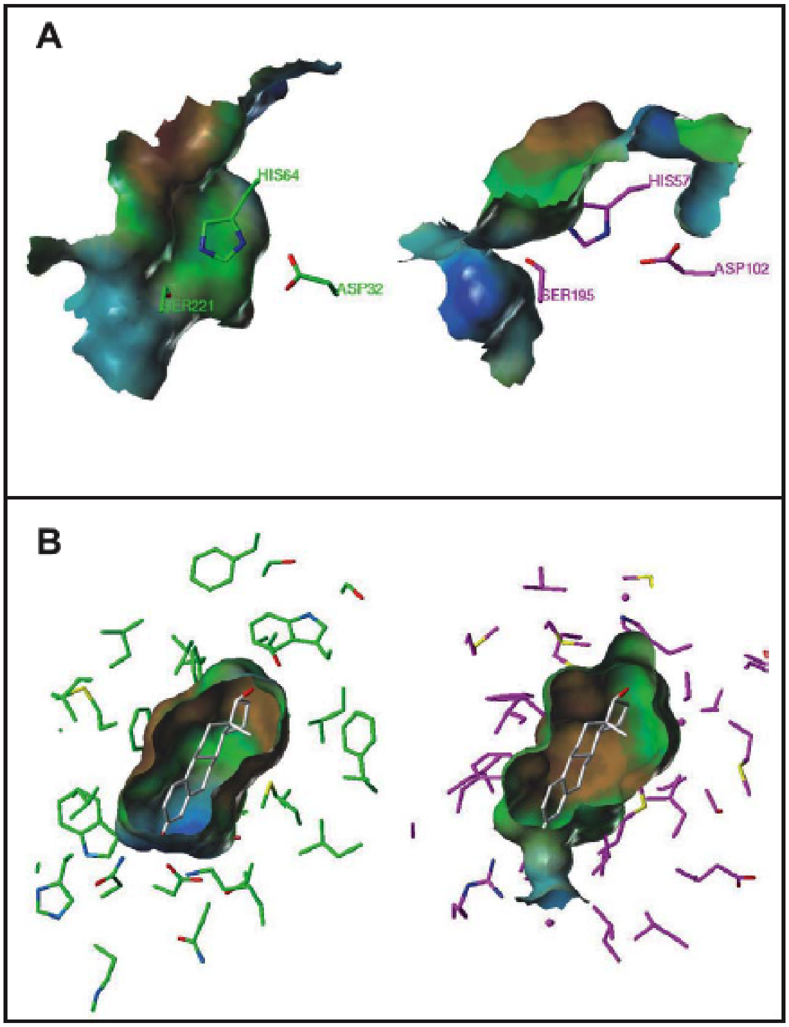

Fig. (7). Two levels of sites similarity illustrated by 1:1 comparison of protein structures. (**A**) Detection of three-dimensional motif. The catalytic triads of subtilisin (PDB: 1scb, left) and γ-chymotrypsin (PDB: 1afq, right) are shown in the same orientation (rigid body fit of side chain atoms performed using Sybyl8.0 [56]). Corresponding sites are represented by the Connolly surface computed for all residues closer than 4.5 Å of any atom of the H-D-S triad and colored according to lipophilic potential (brown: hydrophobic → blue: hydrophilic). Interestingly, the match of catalytic residues does not result in the superimposition of the surface clefts involved in the binding of peptide substrate. (**B**) Identification of proteins able to accommodate the same ligand. The active conformations of estradiol in complex with sex hormone-binding globulin (PDB:1lhu, left) and with estrogen α receptor are shown in the same orientation (rigid body fit of estradiol atoms performed using Sybyl8.0). All protein residues closer than 6.5Å of any ligand atom are represented by their side chain, demonstrating the low conservation between the two protein site sequences. The Connolly surface colored according to lipophilic potential (brown: hydrophobic → blue: hydrophilic) however clearly indicated that both pockets have similar shape and physico-chemical properties. For sake of clarity, depth clipping was applied to show the ligand inside the pocket. Note that due to the high distribution of apolar residue in both sites, SiteEngine comparison yielded a different, but relevant, 3D alignment.

each method: (1) the one to one comparison of selected proteins, (2) database screening for similarity to a query site, (3) classification of binding sites (proteins with different functions or members of a given functional family) or/and (4) comparison of conformers of a selected protein (apo or ligand-bound protein [25], snapshots of a molecular dynamics simulation [19]). Because of the profusion of published applications, illustrative examples were picked in the literature and discussed below.

Recognition of 3D motifs by the SuMo algorithm has been successfully applied to the Ser-His-Asp catalytic triad of serine proteases [12]. Jambon and coworkers compared subtilisin and γ-chymotrypsin. Despite different overall folds, low sequence identity and distinct order of the catalytic residues in the sequence, the two enzymes were shown to share a small common surface patch, which is composed of the catalytic triad and a glycine residue and is located at the centre of dissimilar clefts (Fig. **7A**). Milik and coworkers also compared enzymes that catalyse the same reaction but have different folds [26]. The common structural features extracted from two aminotransferases whose substrates are enantiomers (L- and D- aminoacids), revealed 3D organisations which are mirror images of each others. In contrast to most graph-based methods, the CSC algorithm does not consider atom overlap after 3D alignment by rigid transformation, thus finds pertinent similarity between sites despite conflict in the spatial organisation of corresponding patterns.

Overall comparison of protein surface using SiteEngine revealed similarity between ligand binding pockets [8]. Two databases of complete and annotated protein structures were screened for estradiol-binding sites using as query the sex-hormone-binding globulin (1luh PDB entry, estradiol binding site defined around co-crystallized estradiol). In the test dataset of 126 proteins classified into 12 functional families according to different ligand-binding abilities, 8 of the 11 known estradiol-binding proteins were retrieved among the top ranked 15 entries. In the non-redundant ASTRAL dataset (4,375 proteins structures that have less than 40% sequence identity) [27], the three proteins known to bind estradiol were ranked first, second and 39th. Remarkably, SiteEngine always correctly defined binding sites within the full protein surfaces and well detected similarities between estradiol-binding sites in different proteins, namely sex-hormon binding proteins, estrogen receptor α and 3-Hydroxyacyl-CoA dehydrogenase (Fig. **7B**). This observation suggests that local comparison methods may be suitable to predict the binding profile of permissive ligands or to suggest secondary targets for drugs. Hirota and coworkers performed high-throughput comparisons of surface cavities using ProSurfer in order to predict protein-drug interactions [9]. The all-against-all comparison of 48,347 cavities detected using PASS [28] in 9,708 proteins of the non-redundant PDB subset ended up with 540 predictions for 105 drugs. The putative interaction of ibuprofen, whose primary target is phospholipase A2, with porcine pancreatic elastase was supported by changes in [^{1}H]-NMR spectra of the drug upon protein addition.

Large scale studies of similarity between protein-ligand binding sites were exploited to build networks. All-against-all comparison of 211 protein-ligand complexes of PLD database [29] were carried out by Zhang and Grigorov at site, protein fold and protein sequence levels [30] Network of proteins were constructed based on site similarity values. Due to the limited dataset, *scale-fre*e (nodes and edges distributions are independent of the size) and *small-world* (most nodes are not neighbour of one another, but can be reached from every other by a small number of hops) properties could not be unambiguously evaluated. However, the authors showed that many nodes have few links and identified a few highly connected hubs which referred to immunoglobulins and serine proteases and may provide valuable information about target in structure-based drug design. Park and Kim performed a general analysis of the PDB to construct binding similarity network of ligands [31]. About 15,000 sites (4-49 residues) for 4,208 different ligands were systematically compared; two ligands were connected if they have at least one pair of binding sites within a defined similarity score threshold. Again, although it was not possible to clearly demonstrate its *scale free* property, the resulting network had many nodes with few edges and only few highly connected hubs (e.g. 5 of the 10 top hub ligands are adenosine derivatives, which are frequently encountered in the PDB). The network was further discussed with respect to functional annotations that were collected for each ligand in Enzyme and KEGG databases.

The systematic comparison of sites across a family of targets benefits the knowledge about ligand-protein interaction. Jackson and coworkers widely explored the large and very heterogeneous family of nucleotide-binding proteins (ligands are ATP, GTP, NAD, FAD, FMN principally) [32]. Many nucleotide-bound proteins are present in the PDB, but the ligand-protein relationship is not straightforward: some proteins can bind multiple nucleotides and the same nucleotide can adopt different active conformations related to different recognition modes depending on the target [33, 34]. All-against-all comparison of over 5,000 nucleotides binding sites using SitesBase yielded a classification that matches the SCOP classification by fold. Detailed analysis suggested cases of convergent site evolution of proteins with distinct folds and demonstrated that the comparison method finds the common three-dimensional substructures between sites of different proteins and distinguishes different conformers of the same protein. In a recent study of 258 protein kinase binding sites, Klebe and coworkers demonstrated that the local similarity evaluated using CavBase separates the kinase functional sub-families [35]. In addition, the method was able to capture specific features that are not resolved in sequence space, e.g. the different activated states of a kinase.

Comparing binding sites across a whole family of genes requires the knowledge of all corresponding three-dimensional structures. If the experimental structures are usually solved by biophysical techniques for only few members of the family, these can serve as template for homology modeling to complete the family of structures. SiteAlign was used to compare the canonical antagonist-binding site of human G protein-coupled receptors (GPCRs) in high-throughput homology models [36]. The clustering based on site similarity of the complete family of non-olfactive GPCRs (369 members) mainly fits the functional classes defined by global sequence analysis [37]. The similarity-based tree of GPCRs is nevertheless more precise than the tree constructed using the sequence identity between sites [38] For example, the Duffy antigen chemokine receptor (DARC) is considered as singleton based on sequence yet joins the chemokine receptor's branch in the structure-based classification (Fig. **8**).

The recognition of function can be assisted by searching databases of functionally annotated binding sites with a binding site of interest [39, 40]. This application was shown to contribute to structural genomics projects. Nakamura and coworkers successfully inferred the function of hypothetical proteins from 3D structure using eF-site [25, 41]. For example, the entire surface of the crystal structure of *Thermos thermophilus* TT1542 protein was compared to 22,747 binding sites extracted from ligand-bound proteins in the PDB. The search yielded several putative ligands, most of them sharing similarity with disaccharides. Together with information retrieved from sequence and fold analysis, the protein was predicted to be a disaccharide hydrolase [42]. This assumption was later supported by the structural determination of the homologous protein MshB. CPASS was developed by Powers and coworkers to achieve the same goal, i.e. to assign function to un-annotated proteins with known 3D structure [43]. The strategy is based on the discovery of true binders for genomic targets by NMR screening of a trim functional chemical library. NMR data combined with docking simulations allows the localisation of the binding site, which is then used to query the collection of PDB-derived binding sites for local similarities. This methodology was applied to *Staphylococcus aureus* hypothetical protein SAV1430. No related proteins of known function could be found by sequence homology nor fold com-

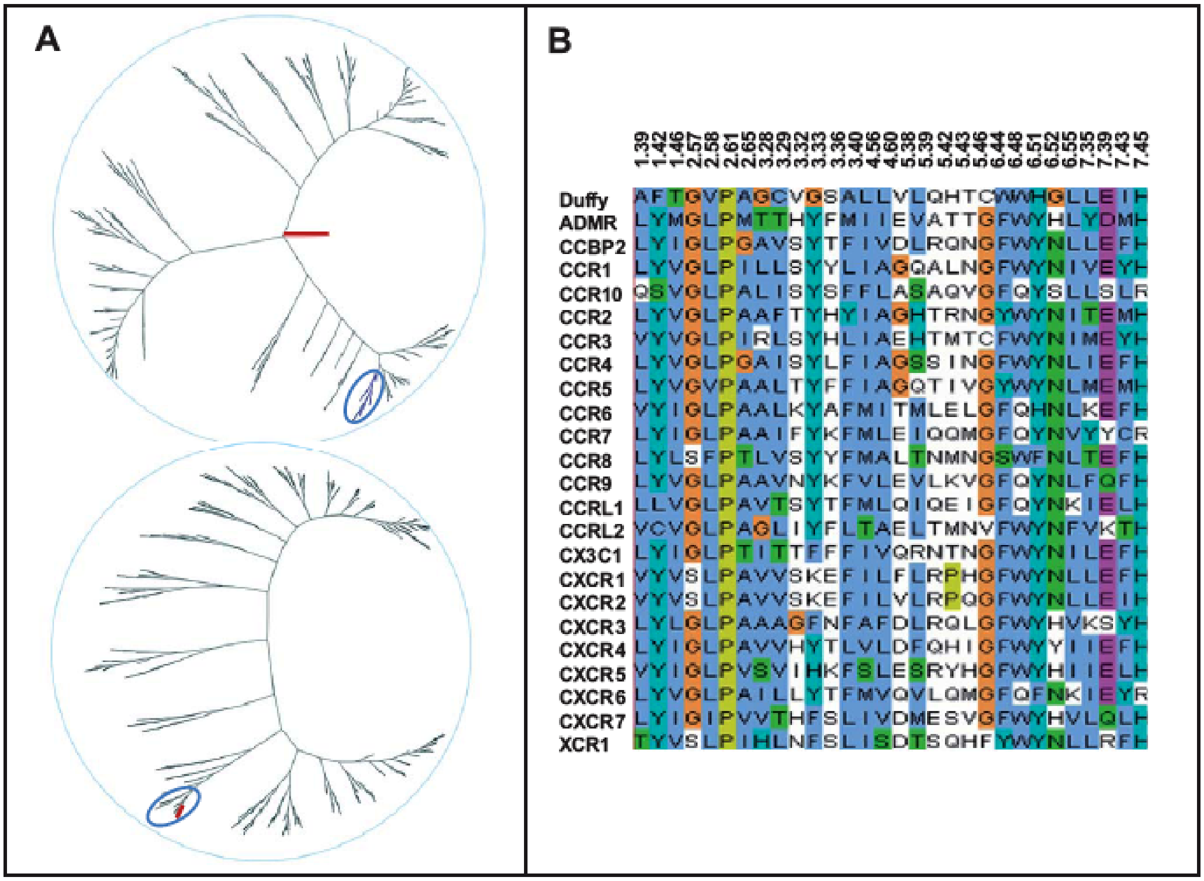
A
B
1.39 1.42 1.46 2.57 2.58 2.61 2.65 3.28 3.29 3.32 3.33 3.36 3.40 4.56 4.60 5.38 5.39 5.42 5.43 5.46 6.44 6.48 6.51 6.52 6.55 7.35 7.39 7.43 7.45
Duffy AFTGVPAGCVGSALLVLQHTCWWHGLLEIH
ADMR LYMGLPMTTHYFMIIEVATTGFWYHLYDMH
CCBP2 LYIGLPGAVSYTFIVDLRQNGFWYNLLEFH
CCR1 LYVGLPILLSYYLIAGQALNGFWYNIVEYH
CCR10 QSVGLPALISYSFFLASAQVGFQYSLLSLR
CCR2 LYVGLPAAFTYHYIAGHTRNGYWYNITEMH
CCR3 VYVGLPIRLSYHLIAEHTMTCFWYNIMEYH
CCR4 LYVGLPGAISYLFIAGSSINGFWYNLIEFH
CCR5 LYVGVPAALTYFFIAGQTIVGYWYNLMEMH
CCR6 VYIGLPAALKYAFMITMLELGFQHNLKEFH
CCR7 LYIGLPAAIFYKFMLEIQQMGFQYNVYYCR
CCR8 LYLSFPTLVSYYFMALTNMNGSWFNLTEFH
CCR9 LYVGLPAAVNYKFVLEVLKVGFQYNLFQFH
CCRL1 LLVGLPAVTSYTFMLQIQEIGFQYNKIELH
CCRL2 VCVGLPAGLIYFLTAELTMNVFWYNFVKTH
CX3C1 LYIGLPTITTFFFIVQRNTNGFWYNILEFH
CXCR1 VYVSLPAVVSKEFILFLRPHGFWYNLLEFH
CXCR2 VYVSLPAVVSKEFILVLRPQGFWYNLLEIH
CXCR3 LYLGLPAAAGFNFAFDLRQLGFWYHVKSYH
CXCR4 LYIGLPAVVHYTLVLDFQHIGFWYYIIEFH
CXCR5 VYIGLPVSVIHKFSLESRYHGFWYHIIELH
CXCR6 LYVGLPAILLYTFMVQVLQMGFQFNKIEYR
CXCR7 LYIGIPVVTHFSLIVDMESVGFWYHVLQLH
XCR1 TYVSLPIHLNFSLISDTSQHFYWYNLLRFH

Fig. (8). Comparison of the transmembrane cavities of 369 GPCR homology models. Analysis was performed on 30 cavity-lining residues known to be critical for antagonist binding. (**A**) Classification based on sequence identity (top) and structural similarity (bottom). The chemokine cluster is highlighted in blue and the DARC receptor is indicated by a red line. (**B**) Sequence alignment of members of the chemokine cluster. Conservation were highlighted using Kalign2.0 [57] and residues are numbered according to Ballesteros [58].

parison. NMR experimental screen identified phosphotyrosine and chemical analogs as ligands, all interacting with a consensus shallow protein cleft. CPASS *in silico* screen then found SAV1430 site similar to the cell signalling Src SH2 protein, more especially its domain accommodating a phospho-tyrosine containing peptide. Authors concluded that SAV1430 is involved in protein-protein interaction. Extensive prediction of function has been carried out in Helmer-Citterich's group [15]. The query of SURFACE database using 257 genomics targets as baits suggested a function for 127 protein chains. About half of the predictions referred to ligand binding ability, the rest being associated to protein signature, typically PROSITE two-dimensional sequence pattern.

In drug design, the detection of local similarities can generate ideas. A practical exercise of inhibitor design was applied by Klebe's group to the protease M^{pro} of the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) [17]. The three-dimensional structure of M^{pro} site for a covalently-bound peptide inhibitor (PDB 1uk4) was decomposed into subpockets. The subpockets S1 and S2 were individually used as CavBase queries to search a collection of 9,446 binding sites extracted from co-crystallized ligand-protein complexes in the PDB. The information about bound-ligand fragments found in the most similar subpockets suggested polar and aliphatic/aromatic chemotypes for S1 and S2 ligands respectively, in line with the substrate specificity of the enzyme. Results guided the design of a library of reversible peptide aldehydes. The library was synthetised, tested and confirmed to contain micromolar inhibitors [44].

Exploiting information about the chemotypes of fragments bound to similar subcavities was shown to amend performance of a scoring function in retrospective docking studies [45]. A collection of 6,129 ligand-bound PDB protein sites was searched with the query site, i.e. the target for the docking experiment. For each matched site, the ligand fragment corresponding to the similarity region was placed into the query site by rigid body motion using the transformation obtained by site comparison. This resulted in a "cloud" of binding patterns, which described the query site. The cloud was used in the docking procedure for initial ligand anchoring and for score calculation, in combination with a mixed knowledge-based and empirical function. The cloud-based approach outperformed standard docking/scoring programs [46].

CONCLUSIONS

We wished to provide here an overview of recent literature about structural alignment of protein sites. Most of the published methods resemble each other in the simplified representation of sites and/or in the algorithm that provides the 3D alignment. They nevertheless all exhibit distinctive characteristics and were used for a wide range of practical applications.

The choice of the program, parameters, score and filtering procedures should be driven by the scientific purpose. In our opinion, it is worth considering: (1) the sensitivity to atomic coordinates (low for the comparison of unrelated proteins, high for the distinction of protein conformers), (2) the kind of expected similarity (detection of local motifs is favoured by the comparison of graphs/triangles, whereas more global comparisons are obtained from an overall superimposition of sites); 3) the cpu time required for a single comparison, which imposes the level of throughput (from 1:1 comparison of two proteins to all-against-all comparison of the full PDB). Computing time increases with the complexity of the protein site description, and geometric hashing is quicker than clique detection, which is itself much faster than an exhaustive exploration by rotation/translation. According to the literature, the time needed for comparing two sites on standard computing architectures ranges from about 50 ms (SitesBase, ProSurfer and SURFACE) to less than 5 minutes (eF-Site).

Many applicative studies involved the comparison of well defined ligand-binding sites, whose size and depth are linked to the shape of a selected ligand in co-crystallized or NMR complexes. Interestingly, local 3D comparison methods can also been applied to protein protein-binding sites, in order to predict the druggability of a cavity formed at the protein interface [47] or to compare surface clefts involved in protein-protein binding [48].

REFERENCES

[1] Good, A.C.; Richards, W.G. *Persp. Drug Discov. Des.,* **1998**, *9,* 321-338.

[2] Lemmen, C.; Lengauer, T. *J. Comput. Aided Mol. Des.,* **2000**, *14,* 215-232.

[3] Carugo, O. *Cur. Prot. Pept. Sci.,* **2007**, *8,* 219-241.

[4] Kolodny, R.; Petrey, D.; Honig, B. *Curr. Op. Struct. Biol.,* **2006**, *16,* 393-398.

[5] Binkowski, T.A.; Adamian, L.; Liang, J. *J. Mol. Biol.,* **2003**, *332,* 505-526.

[6] Stark, A.; Russell, R.B. *Nucl. Acids Res.,* **2003**, *31,* 3341-3344.

[7] Via, A.; Ferre, F.; Brannetti, B.; Helmer-Citterich, M. *Cell Mol. Life Sci.,* **2000**, *57,* 1970-1977.

[8] Shulman-Peleg, A.; Nussinov, R.; Wolfson, H.J. *J. Mol. Biol.,* **2004**, *339,* 607-633.

[9] Minai, R.; Matsuo, Y.; Onuki, H.; Hirota, H. *Proteins: Structure, Function, and Bioinformatics,* 2008.

[10] Schmitt, S.; Kuhn, D.; Klebe, G. *J. Mol. Biol.,* **2002**, *323,* 387-406.

[11] Kinoshita, K.; Furui, J.; Nakamura, H. *J. Struct. Funct. Genomics,* **2002**, *2,* 9-22.

[12] Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. *Proteins: Structure, Function, and Bioinformatics,* **2003**, *52,* 137-145.

[13] Powers, R.; Copeland, J.C.; Germer, K.; Mercier, K.A.; Ramanathan, V.; Revesz, P.; *Proteins: Structure, Function, and Bioinformatics,* **2006**, *65,* 124-135.

[14] Brakoulias, A.; Jackson, R.M. *Proteins: Structure, Function, and Bioinformatics,* **2004**, *56,* 250-260.

[15] Ferre, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M. *BMC Bioinformatics,* **2005**, *6,* 194.

[16] Gardiner, E.; Artymiuk, P.; Willett, P. *J. Mol. Graph. Model.,* **1997**, *15,* 245-253.

[17] Kuhn, D.; Weskamp, N.; Schmitt, S.; Hullermeier, E.; Klebe, G. *J. Mol. Biol.,* **2006**, *359,* 1023-1044.

[18] Bron, C.; Kerbosh, J. *Commun. ACM,* **1973**, *16,* 575–577.

[19] Schalon, C.; Surgand, J.S.; Kellenberger, E.; Rognan, D. *Proteins: Structure, Function, and Bioinformatics,* **2008**, *71,* 1755-1778.

[20] Bosshard, H.R. *News Physiol. Sci.,* **2001**, *16,* 171-173.

[21] Carlson, H.A. *Curr. Pharm. Des.,* **2002**, *8,* 1571-1578.

[22] Teague, S.J. *Nat. Rev. Drug Discov.,* **2003**, *2,* 527-541.

[23] Gold, N.D.; Jackson, R.M. *J. Chem. Inf. Model.,* **2006**, *46,* 736-742.

[24] Davies, J.R.; Jackson, R.M.; Mardia, K.V.; Taylor, C.C. *Bioinformatics,* **2007**, *23,* 3001-3008.

[25] Kinoshita, K.; Nakamura, H. *Protein Sci.,* **2005**, *14,* 711-718.

[26] Milik, M.; Szalma, S.; Olszewski, K.A. *Protein Eng.,* **2003**, *16,* 543-552.

[27] Chandonia, J.-M.; Hon, G.; Walker, N.S.; Lo Conte, L.; Koehl, P.; Levitt, M.; Brenner, S.E. *Nucl. Acids Res.,* **2004**, *32,* D189-192.

[28] Brady, G.P.; Stouten, P.F.W. *J. Comput. Aided Mol. Des.,* **2000**, *14,* 383-401.

[29] Puvanendrampillai, D.; Mitchell, J.B.O. *Bioinformatics,* **2003**, *19,* 1856-1857.

[30] Zhang, Z.; Grigorov, M.G. *Proteins: Structure, Function, and Bioinformatics,* **2006**, *62,* 470-478.

[31] Park, K.; Kim, D. *Proteins: Structure, Function, and Bioinformatics,* **2007**, *71,* 960-971.

[32] Gold, N.D.; Jackson, R.M. *J. Mol. Biol.,* **2006**, *355,* 1112-1124.

[33] Kahraman, A.; Morris, R.J.; Laskowski, R.A.; Thornton, J.M. *J. Mol. Biol.,* **2007**, *368,* 283-301.

[34] Stockwell, G.R.; Thornton, J.M. *J. Mol. Biol.,* **2006**, *356,* 928-944.

[35] Kuhn, D.; Weskamp, N.; Hullermeier, E.; Klebe, G. *ChemMedChem,* **2007**, *2,* 1432-1447.

[36] Bissantz, C.; Logean, A.; Rognan, D. *J. Chem. Inf. Comput. Sci.,* **2004**, *44,* 1162-1176.

[37] Fredriksson, R.; Schioth, H.B. *Mol. Pharmacol.,* **2005**, *67,* 1414-1425.

[38] Surgand, J.-S.; Rodrigo, J.; Kellenberger, E.; Rognan, D. *Proteins: Structure, Function, and Bioinformatics,* **2006**, *62,* 509-538.

[39] Hambly, K.; Danzer, J.; Muskal, S.; Debe, D. *Mol. Divers.,* **2006**, *10,* 273-281.

[40] Watson, J.D.; Laskowski, R.A.; Thornton, J.M. *Curr Op. Struct. Biol.,* **2005**, *15,* 275-284.

[41] Kinoshita, K.; Nakamura, H. *Protein Sci.,* **2003**, *12,* 1589-1595.

[42] Handa, N.; Terada, T.; Kamewari, Y.; Hamana, H.; Tame, J.R.; Park, S.Y.; Kinoshita, K.; Ota, M.; Nakamura, H.; Kuramitsu, S.; Shirouzu, M.; Yokoyama, S. *Protein Sci.,* **2003**, *12,* 1621-1632.

[43] Powers, R.; Mercier, K.; Copeland, J. *Drug Discov. Today,* **2008**, *13,* 172-179.

[44] Al-Gharabli, S.; Shah, S.; Weik, S.; Schmidt, M.; Mesters, J.; Kuhn, D.; Klebe, G.; Hilgenfeld, R.; Rademann, J. *ChemBioChem,* **2006**, *7,* 1048-1055.

[45] Ramensky, V.; Sobol, A.; Zaitseva, N.; Rubinov, A.; Zosimov, V. *Proteins: Structure, Function, and Bioinformatics,* **2007**, *69,* 349-357.

[46] Kellenberger, E.; Muller, P.; Rodrigo, J.; Rognan, D. *Proteins: Structure, Function and Genetics.,* **2004**, *57,* 225-242.

[47] Block, P.; Weskamp, N.; Wolf, A.; Klebe, G. *Proteins: Structure, Function, and Bioinformatics,* **2007**, *68,* 170-186.

[48] Mintz, S.; Shulman-Peleg, A.; Wolfson, H.J.; Nussinov, R. *Proteins: Structure, Function, and Bioinformatics,* **2005**, *61,* 6 - 20.

[49] Jambon, M.; Andrieu, O.; Combet, C.; Deleage, G.; Delfaud, F.; Geourjon, C. *Bioinformatics,* **2005**, *21,* 3929-3930.

[50] Gold, N.D.; Jackson, R.M. *Nucl. Acids Res.,* **2006**, *34,* D231-234.

[51] Shulman-Peleg, A.; Nussinov, R.; Wolfson, H.J. *Nucl. Acids Res.,* **2005**, *33,* W337-341.

[52] Ausiello, G.; Via, A.; Helmer-Citterich, M. *BMC Bioinformatics,* **2005**, *6 Suppl 4,* S5.

[53] Ferre, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M. *Nucl. Acids Res.,* **2004**, *32,* D240-244.

[54] Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D. *J. Chem. Inf. Model.,* **2006**, *46,* 717-727.

[55] Connolly, M.L. *J. Appl. Crystallogr.,* **1983**, *16,* 548–558.

[56] Sybyl, version 7.3 ed, TRIPOS: St. Louis, MO.

[57] Lassmann, T.; Sonnhammer, E.L.L. *Nucl. Acids Res.,* **2006**, *34,* W596-W599.

[58] Ballesteros, J.; Weinstein, H. *Methods Neurosci.,* **1995**, *25,* 366-428.

Received: April 23, 2008 **Revised: May 9, 2007** **Accepted: May 19, 2008**