WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

A novel approach to local similarity of protein binding sites substantially improves computational drug design results

Vasily Ramensky,[1,2*] Alexandr Sobol,[1] Natalia Zaitseva,[1,3] Anatoly Rubinov,[1,4] and Victor Zosimov[1,3]

[1] Algodign LLC, Bolshaya Sadovaya 8, Moscow 123379, Russia
[2] Engelhardt Institute of Molecular Biology of Russian Academy of Sciences, Vavilova 32, Moscow 1199911, Russia
[3] Applied Acoustics Research Institute, Dubna-1, Moscow 141981, Russia
[4] The Institute for Information Transmission Problems of Russian Academy of Sciences, Moscow 127994, Russia

ABSTRACT

We present a novel notion of binding site local similarity based on the analysis of complete protein environments of ligand fragments. Comparison of a query protein binding site (target) against the 3D structure of another protein (analog) in complex with a ligand enables ligand fragments from the analog complex to be transferred to positions in the target site, so that the complete protein environments of the fragment and its image are similar. The revealed environments are similarity regions and the fragments transferred to the target site are considered as binding patterns. The set of such binding patterns derived from a database of analog complexes forms a cloud-like structure (fragment cloud), which is a powerful tool for computational drug design. It has been shown on independent test sets that the combined use of a traditional energy-based score together with the cloud-based score responsible for the quality of embedding of a ligand into the fragment cloud improves the self-docking and screening results dramatically. The usage of a fragment cloud as a source of positioned molecular fragments fitting the binding protein environment has been validated by reproduction of experimental ligand optimization results.

Proteins 2007; 69:349–357.

Key words: virtual ligand screening; similarity of protein binding sites; docking; lead optimization.

INTRODUCTION

Molecular recognition and binding performed by proteins are the background of all biochemical processes in a living cell. In particular, the usual mechanism of drug function is effective binding and inhibition of activity of a target protein. Direct modeling of molecular interactions in protein-inhibitor complexes is the basis of modern computational drug design but is an extremely complicated and far from solved problem.[1–5] Methods that use simple atom–atom scores for estimation of molecular interaction energy often fail to provide adequate accuracy required by drug design,[6,7] whereas methods based on more adequate physical models including force fields and molecular dynamics are restricted by their computational complexity.[5] Fortunately, modeling of binding can be supported by the wealth of known structural data on protein–ligand complexes available, for example, from the Protein Data Bank,[8] taking advantage of the similarity between the protein features responsible for binding. The fact that protein function is related to specific binding regions rather than overall fold or amino acid sequence[9,10] makes the analysis of local similarity of binding sites a basic tool for functional annotation and classification of novel proteins, for development of targeted protein inhibitors in drug design, and for analysis of potential side-effect of developed drugs.[11,12]

In the current paradigm, site similarity is recognized by the existence of chemically and spatially analogous regions from binding sites. The existing programs aim to find maximal regions of this kind.[13–20] However, the assessment of results produced by existing tools raises the fundamental question about the biological significance of observed matches. The reliable interpretation of detected similarity is possible when the region of similarity comprises the complete binding site or a complete environment of a ligand part (e.g., nucleotide base) in one protein. But this region may have a bizarre configuration, for example, consist of small unconnected parts. As a consequence, the interaction of the whole molecule, its fragments, or even single

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Abbreviations: HSV, human simplex virus; PDB, protein data bank; PTP1B, protein tyrosine phosphatase-1B; RMSD, root mean square deviation; SCOP, structural classification of proteins; TK, thymidine kinase.

*Correspondence to: Vasily Ramensky, Engelhardt Institute of Molecular Biology of Russian Academy of Sciences, Vavilova 32, Moscow 1199911, Russia. E-mail: ramensky@imb.ac.ru
Received 30 October 2006; Revised 16 February 2007; Accepted 27 February 2007
Published online 10 July 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21487

atoms with the protein is not determined by their interaction with the region of similarity. An established similarity therefore "does not necessary imply similarity in the binding partners and in the biological functions."[12] We present here a novel notion of binding site local similarity based on analysis of complete protein environments of ligand fragments, and evaluate its applications to computational drug design problems.

In computational drug design, protein–ligand interaction patterns are used within the framework of the so-called knowledge-based approach based on analysis of relative positions of different types of ligand atoms and protein atoms, active groups, or amino acid residues. An observed density of relative positions can be converted into a one-dimensional atom–atom potential[21] or spatial atom-residue potential.[22] As another option, the analysis of relative positions reveals template points where a ligand can form hydrogen bonds or hydrophobic interactions with a particular active group[23] or residue[22] of a target protein. Docking of a ligand or its fragments is thus reduced to embedding into the cloud of template points. The main weakness of the knowledge-based method is its independent accounting of interactions between various ligand and protein elements with neglected cooperative effects. Suppose for example that a hypothetical type of ligand atom is frequently bound by either of two different protein active groups, but never by both of them simultaneously. If both groups are found close to the possible position of such an atom in a particular binding site, this position will be erroneously favorable with knowledge-based score or template points. The key to correct use of analogy-derived information is therefore analysis of complete environments of ligand atom or fragments, which is the main idea of the method proposed here.

When the structures of other ligands co-crystallized with the target protein are already known, one may avoid the problem of site comparison, as implemented in the docking program SDOCKER.[24] It minimizes a combined score which is a sum of a traditional energy-based score and a similarity-based score reflecting the distances between positions of ligand atoms and the nearest atoms of co-crystallized ligands. This improves docking results, but requires a series of crystallographic data for a particular protein. The method presented here is based on a similar approach to ligand positioning, but instead of co-crystallized ligands it uses the ligand fragments from other complexes placed in the target site with the help of site similarity found.

MATERIALS AND METHODS

Novel approach to local site similarity—fragment clouds

Our notion of local site similarity is based on application of the natural idea that similar protein environments will bind the same ligand fragment in similar positions. More precisely, local site similarity is defined as follows. One of two compared sites of known X-ray crystallographic structure (the analog) is specified together with a bound ligand; the structure of this complex also needs to be known, whereas the existence of the ligand in the other site (target) is not required. Local similarity of the two binding sites is observed if there exists a connected fragment (not known beforehand) of the ligand bound by the analog that can be transformed by a rigid motion into the target site so that the protein environments of the fragment in the analog and its image in the target are chemically and spatially similar, as further defined below. When these conditions are satisfied, we expect that the similarity region of the target protein has affinity for the ligand fragment in the image position. We consider this positioned fragment as a binding pattern for the target protein site. As a negative example, suppose that a ligand contacts two sides of a cleft in the analog protein and that there exists a region on the surface of the target site that is similar to one side of the cleft. The definition above does not consider this as similarity. Depending on the degree of similarity between the environments of a ligand atom in the analog protein and its image in the target, we assign a so-called *reliability values* R to the atom of the image fragment, such that $0 < R \leq 1$, see definition below.

A comparison of a target site against a database of protein–ligand complexes creates a "cloud" of binding patterns in the target site (Fig. 1). When a site of a known complex is chosen as a target, one may see that its native ligand is usually completely covered by cloud fragments that are chemically and spatially similar to the target ligand parts. This important feature suggests that the clouds can be used in the following applications.

Ligand scoring

Cloud fragments represent binding patterns, but do not carry information about their contribution to binding energy. Hence, it is necessary to combine a traditional energy-based score with the cloud score representing the quality of embedding of a ligand into the cloud. The latter is maximal if a ligand is completely covered by atoms of cloud fragments that are chemically and spatially similar to the respective covered ligand atoms and have high reliability values. To this end, one can use a two-term score as in Ref. 24. Alternatively, for a simple overall scoring function, one can use another method of combining scores: if a binding energy score of the protein–ligand complex is a sum of energy scores of individual ligand atoms (e.g. atom–atom scores), the energy score of each atom can be weighted by its cloud score. The cloud score of a ligand atom is maximal if there exists a nearby cloud atom with the same chemical type and high reliability value. The mixed score function introduced this way (see more details in the corresponding section) accounts for atom positions rather than complete frag-

 DOI 10.1002/prot

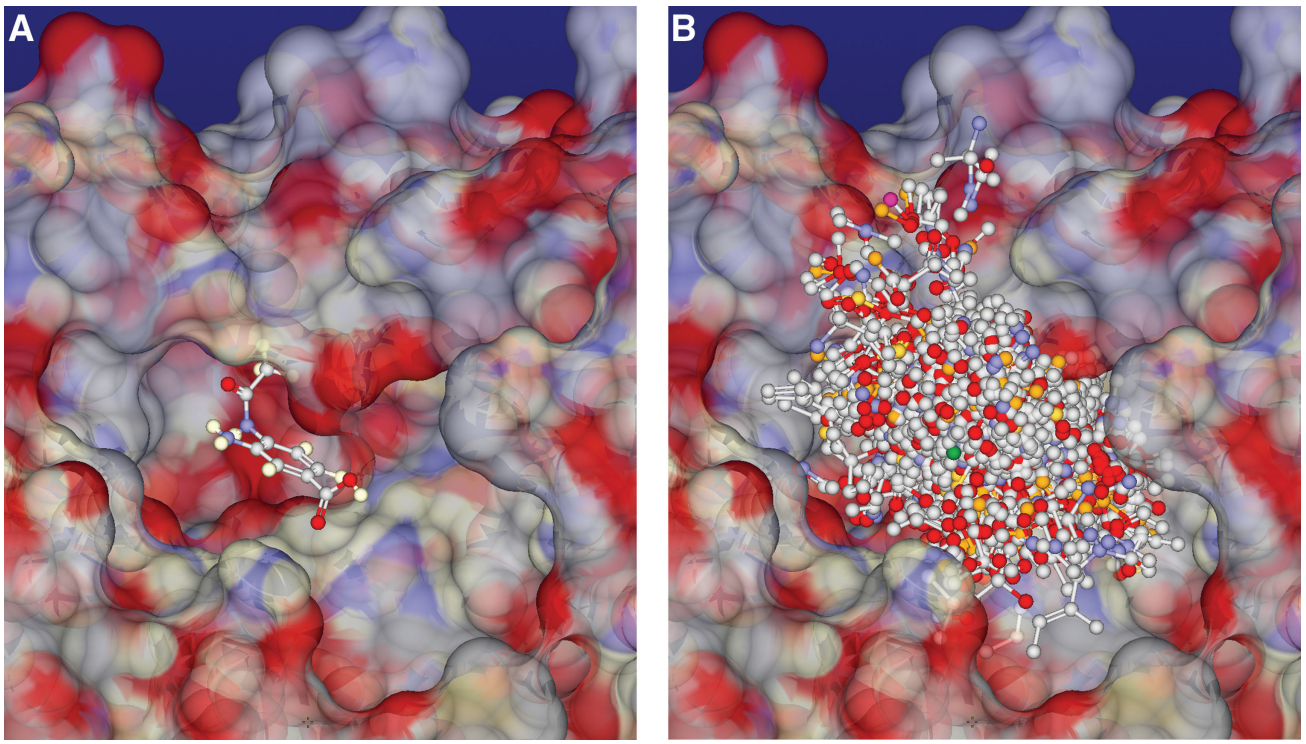

Figure 1

*The fragment cloud for the influenza A subtype N2 neuraminidase. The protein (also known as sialidase; PDB entry 1IVE[25]) is represented by its surface. (**A**) The aromatic inhibitor BANA108 (4-(acetylamino)-3-aminobenzoic acid) in the binding site. (**B**) The fragment cloud built for the binding site. Atoms are colored according to CPK convention. The cloud contains 4838 atoms in 3416 fragments.*

ments but has two essential advantages. First, it can be represented in the grid form, what is important for rapid computation. Second, unlike the combined score mentioned earlier,[24] a high score of the whole ligand is attained only if atoms with high contribution to the binding energy (e.g., donors or acceptors of hydrogen bonds) also have high cloud scores, that is, are properly positioned from the cloud point of view. This mixed energy/cloud score is an improvement of the traditional energy-based score and can replace it in all applicable computational drug design tasks, for example, in ligand screening and *de novo* design, as supported below for the docking problem.

A more complicated score (directed mixed score) accounts not only for atom positions, but also for directions from an atom to its covalently linked neighbors. The directed mixed score can not be represented in the 3D-grid form, so it is not computationally efficient and is currently used only for rescoring of ligand positions generated with mixed score.

Ligand block positioning

Different computational methods of ligand docking include a stage of searching for a diverse set of positions of known ligand rigid blocks.[3,5] The next stage consists of a search for the optimal set of properly positioned blocks forming the whole ligand (fragment-based approach) or for the optimal ligand conformation with one positioned block chosen as the fixed anchor. A fragment cloud can help to perform the first stage. Let the *star* of a molecular atom be the atom set including the atom itself and all its covalently linked neighbors. If a star from a given ligand block is chemically equivalent to a star from a cloud fragment and includes at least three atoms, we can find a rigid motion that relates the stars and determines the position of the block. If the mixed score of the block in this position is satisfactory, the position is accepted.

Search for positioned molecular fragments

In de novo ligand design, techniques similar to the fragment-based approach used for ligand docking are applied, but with the molecular fragments not known in advance.[3] Ligand construction here is reduced to a search for a diverse set of positioned molecular fragments and linking of these fragments to each other. Similarly, in the case of so-called ligand optimization, one needs to find an appropriate molecular fragment to replace a given

chemical group of a positioned ligand to improve ligand characteristics.[5,26,27] A fragment cloud built for the binding site can serve as a source of molecular fragments. Every cloud fragment is extended to comprise a complete rigid fragment from its source ligand together with all potential covalent bonds that serve as linkage handles. An example of a complete fragment is a ring structure or atoms connected by a double bond. For example, if a cloud fragment contains three atoms from an aromatic carbon ring, it will be extended to a full ring with six hydrogen atoms, each of which is considered as a potential covalent linkage site. The extended fragments from the cloud can be filtered leaving only the fragments with good mixed score. Unlike optimization methods which use a precompiled database of pairs of interchangeable fragments (bioisosteric replacements),[28] the set of replacing fragments depends on the binding protein environment of a replaced ligand fragment and not on the fragment itself.

Site comparison and fragment cloud construction

Protein atoms are assigned 43 chemical types similar to those from the Merck Molecular Force Field classification[29] (Supplementary Table I online). For ligand atoms, 10 chemical types described in Ref. 21 are used (Supplementary Table II). Hydrogen atoms are used only for correct type assignment.

First, consider the protein 5 Å-environment $A = \{a_1, a_2, \ldots a_N\}$ of one ligand atom X in the analog protein, that is, all atoms from the binding site that are in the 5 Å-neighborhood of X. Suppose that the complete target binding site T consists of N' atoms: $T = \{t_1, t_2, \ldots t_{N'}\}$ and there exists a subset $T_0 \subseteq T$ of size n ($N' \geq n \geq 4$) such that n atoms from T_0 are similar to n atoms $A_0 = \{a_{i1}, a_{i2}, \ldots a_{in}\} \subseteq A$ in their chemical types and spatial arrangement. The search for A_0 and T_0 is performed using a standard clique detection technique in the graph whose nodes represent pairs (a_i, t_j) of chemically equivalent atoms and edges reflect similarity of corresponding pairwise distances.[12] If the search is successful, the optimal rigid motion superimposing matched protein atoms[30] is applied both to the initial ligand atom X and its complete environment A [Fig. 2(a)]. The atoms are thus transferred to the target binding site. Then we extend the matching between A_0 and T_0 by such atom pairs (a_i, t_i) that a_i and t_i have the same chemical atom type in the coarser 10-type typification mentioned above, and the distance between t_i and the image a_i' of atom a_i is below a threshold.

Next, a *reliability value* R, with $0 \leq R \leq 1$, is assigned to the image X′ of X in the target site and reflects the similarity between the environments of X and its image X′. If the environments are highly similar ($R \approx 1$) we expect that the position of X′ is the place where an atom with chemical type identical to X can be bound by the target, since the environment of X′ contains only atoms required for binding with no "alien" atoms. However, as illustrated in Figure 2(a), the analog site may contain extra binding atoms (shown on the lower side) that decrease the reliability value. In a simple form, the reliability R can be defined as the sum of the number of matched atoms divided by the total number of analog and target atoms in the 5 Å-environments of X and X′, respectively [Fig. 2(b)]: $R = 2n/(N + N')$, using the notation presented above. In fact, we use a somewhat more complicated definition that accounts for the quality of spatial superposition of matched atoms and their distance from X′.

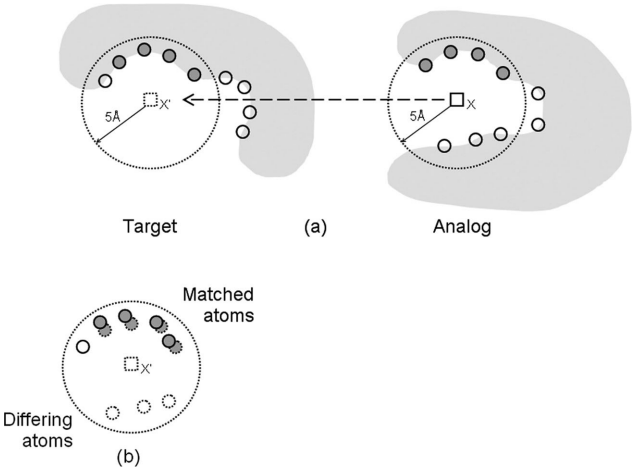

Figure 2

*Transfer of ligand atom X from the analog to the target binding site and calculation of corresponding reliability value. Grey areas denote protein binding regions; circles: protein atoms; squares: ligand atom X and its image X′. (**a**) Superposition of matched atoms (four grey circles) determines the rigid motion that is applied both to the ligand atom X and its complete protein 5 Å-environment A in the analog. (**b**) The reliability* R *assigned to X′ reflects the similarity between the environments of X and its image X′ in the target site. Atom images are denoted by dashed circles; matched atoms are grey, differing atoms are white. In the simplest form,* R *can be defined as the sum of matched atoms divided by the total number of target and transferred analog protein atoms in the 5-Å-environment of X′:* $R = (4 + 4)/(5 + 7) = 0.66$.

Finally, if the calculated reliability of atom X′ is above the threshold $R_{\min} = 0.7$, we try to extend the transferred part of the ligand. The determined rigid motion is applied to all atoms of the ligand and their local environments, and the calculation of reliability is done for every ligand atom image. The extended fragment includes covalently linked atoms with reliability above the threshold. After the extension, the initial rigid transformation is adjusted to provide the best superposition for all matched protein atoms. This results in a ligand fragment transferred to the target binding site.

Database of ligand-site complexes

The database of ligand-site complexes is used for construction of fragment clouds. Complexes are extracted

DOI 10.1002/prot

from PDB structures[8] and contain bound protein ligands and their binding sites. The selected PDB entries are X-ray structures with assigned structural classification from SCOP,[31] resolution no worse than 3 Å, and ASTRAL SPACI score[32] no less than 0.2. Any allowed ligand should have at least six heavy atoms of chemical types (C, O, N, P, S, F, Cl, Br, I) and can be a short (no longer than 10-mer) polymeric chain of amino acid residues, nucleotide bases or other moieties from the PDB Dictionary of heterogroups. A binding site here is defined as a set of protein atoms with distances no greater than 5 Å from ligand atoms.

The current database of ligand-site complexes contains 11,605 complexes from 8650 PDB entries. The database is made non-redundant by keeping only one ligand-site complex from a set of complexes with the same ligand and SCOP family identifier. This database non-redundant at the SCOP family level ("nr_family") contains 6129 complexes with 4043 different ligands from 5271 PDB entries representing 1208 SCOP families from 463 SCOP folds.

Mixed score

The mixed score S is the sum over all ligand atoms: $S = \sum_i E_i C_i$. Here, the energy-based score value of the i-th ligand atom E_i is weighted with a *cloud score* C_i ($0 \leq C_i \leq 1$) that serves to evaluate ligand atom positioning. The cloud score for a ligand atom a_i equals 1 if there exists a spatially close cloud atom b_j with the same chemical type as a_i and $R(a_i) = 1$. More precisely, C_i is defined as the maximum of the product $R(b_j) \times M(a_i, b_j) \times W(d(a_i, b_j))$ over all spatially close cloud atoms b_j with reliability $R(b_j) \geq R_{\min}$. M denotes the chemical type similarity matrix; $W(d)$ weights interatomic distance $d(a_i, b_j)$ and equals 1 for $0 \leq d \leq D_0$, $W_{\min}$ for $d \geq D_1$, and linearly decreases in the interval $D_0 \leq d \leq D_1$. The parameters D_0 and D_1 are usually set to 0.5 and 1.0 Å, respectively. For ligand atoms with positive energy score $E_i > 0$, the cloud score C_i is set to 1 to penalize atom clashes. Penalty terms are also introduced for ligand atoms located out of the cloud. The chemical type similarity matrix M is diagonal except for CA–C2 terms equal to 0.9 and C2–C3, CA–C3, NI–C* terms equal to 0.7.

RESULTS

Ligand docking with fragment clouds

The cloud-based docking method has been implemented in the in-house software AlgoComb, which uses fragment clouds both for initial ligand anchoring and subsequent calculation of the mixed score. The AlgoComb docking procedure starts with decomposition of a ligand into rigid blocks, for example, rings or atoms linked by double bonds. The procedure then proceeds in two stages: positioning of ligand rigid blocks, or anchoring, followed by search for the optimal combination of torsion angles connecting the blocks, see Ref. 33. For anchoring the generated positions are clustered, and the representative with the best mixed score is kept for every cluster. Every such representative serves as an anchor position. The subsequent search for optimal conformation of the ligand is performed by stepwise changes of torsion angles at bonds linking the blocks. The search aims at minimization of the mixed score function, whose energy component is itself obtained from a combination of knowledge-based and empirical approaches, as previously described in Ref. 34.

AlgoComb performance with the mixed score was tested on the self-docking test with 100 protein–ligand complexes[35] that have been used before for the extensive comparative evaluation of eight widely used and readily available docking tools.[7] Descriptions of the complexes are given in the Supplementary Table III online. The programs compared in Ref. 7 are DOCK,[36] FLEXX,[37] FRED,[38] GLIDE,[39] GOLD,[40] SLIDE,[23] SURFLEX,[41] and QXP.[42] Within the course of the evaluation, the ligands are docked into binding sites defined as 12 Å neighborhoods of the native ligands. The performance of a docking program is measured by the percentage of cases in which the RMSD of the best scored position of the ligand does not exceed 2 Å from the X-ray determined native position. FLEXX, GLIDE, SURFLEX, and GOLD show similar results and succeed in docking 50–55% of all cases, whereas the success rate of DOCK, FRED, SLIDE, and QXP is below 40%. AlgoComb has shown comparable 52% performance when run with the energy-based scoring function without cloud usage (Table I).

Table I
Docking Success Rate of Eight Programs Reviewed in Kellenberger[7] and AlgoComb Evaluated on a Set of 100 Protein–Ligand Complexes From Paul[34]

Method	Success rate (%)
AlgoComb with clouds	
Clouds not used	52
Complete clouds with the query complex removed	73
Reduced clouds, Set1: removed species-level and closer layers	69
Reduced clouds, Set 2: removed protein-level and closer layers	68
Reduced clouds, Set 3: removed family-level and closer layers	60
Clouds after the layered filtering	78
Clouds after the layered filtering, results rescored with directed mixed score	82
Removed protein-level and closer layers, atom types are randomly permuted	40
Other docking tools	
FLEXX, GLIDE, SURFLEX, GOLD	50–55
DOCK, FRED, SLIDE, QXP	<40

Success rate measures the number of cases in which the spatial position of the best scored ligand deviates not greater than 2 Å from the native position as measured by RMSD.

When fragment clouds are used as described above, AlgoComb succeeds in 73% of all cases (Table I). This significant improvement is explained by the fact that the cloud-based approach effectively utilizes the structural information from PDB complexes. The whole database of ligand–protein complexes except for the query complex itself has been used here for cloud construction. For 87 entries from this test set the database includes complexes of the same target protein co-crystallized with other ligands. The results of the proposed approach improve with the presence of such complexes or complexes with structural neighbors of the target protein in the database. To evaluate the contribution of structural neighbors of a protein to the docking accuracy and model the situation of docking into a structural "orphan", we prepared three sets of reduced fragment clouds with varying degree of filtration based on SCOP classification of protein structural similarity.[31]

In the first set, the clouds do not include fragments from exactly the same protein co-crystallized with other ligands. This reduces the success rate to 69% (Table I, Set 1) and corresponds to the case when no inhibitors for a drug design target are yet discovered and co-crystallized, but protein–ligand complexes of orthologous proteins from other species are available. When the latter are removed, too, the success rate is 68% (Table I, Set 2). Finally, at the strongest (i.e., most exclusive) family-level filtration, the clouds do not include fragments from analogs that have the same SCOP structural family identifier as the query protein for which the cloud is constructed (Table I, Set 3). The AlgoComb docking success rate for such clouds is 60% (Table I, Set 3).

The experiments with the reduced clouds show that the cloud fragments derived from proteins structurally similar to the target one are very important for docking accuracy, although they form a relatively small cloud fraction (see Supplementary Table IV online). We define a cloud layer as a collection of fragments from analog proteins structurally similar to the target protein at a certain level of SCOP hierarchy (e.g., "family", "fold", "other"). To increase the impact of these layers, we perform a procedure of layered cloud filtering, where all levels except for the "domain" one (the query complex itself) are used and the reliabilities of atoms in the i-th layer are changed by ΔR_i, where ΔR_i equals -0.2 for the "other" layer (proteins from other structural classes than the target), -0.1 for the same SCOP class-layer, 0 for fold and superfamily layers, and 0.3 for family, protein and species layers, respectively. The same reliability threshold $R_{\min} = 0.7$ is used, so more atoms from the structurally similar layers contribute to the mixed score and their weight increases. On the other hand, in the structurally remote layers only atoms with high reliability remain. That is, either "good" atoms from structural neighbors or only the "best" from the remote proteins are accepted. Supplementary Table IV online describes the layer sizes before and after the layered filtering procedure. The fraction of atoms from structurally close analogs (species, protein, and family layers) increases from 4.3 to 14.4% after filtration. At the same time, even in the filtered clouds the structurally remote proteins ("other" layer) provide more than one half of all atoms with high reliability. This observation supports the fact the clouds can effectively utilize the binding patterns from various proteins. The AlgoComb success rate with these filtered clouds is 78%. Rescoring with the directed mixed score of the position set generated by AlgoComb with the filtered clouds improves the success rate to 82%.

To estimate the non-randomness of fragment clouds and their contribution to the docking process, we randomly permuted the atom types in the cloud and performed the calculations for the filtered clouds from Set 2. As a result, the docking performance was decreased from 68 to 40%, clearly worse than if no clouds are used. This test provides strong evidence of high specificity and non-randomness of cloud composition.

Virtual screening with fragment clouds

The performance of fragment cloud-assisted virtual ligand screening was tested on the HSV-1 thymidine kinase (TK, PDB ID 1kim[43]) used as a target for the database of 990 drug-like molecules and 10 known TK inhibitors.[44] Thymidine kinase is known to be a relevant yet difficult drug design target.[44,45] The success rate in virtual screening experiments is determined by a number of real inhibitors in a certain top fraction of the energy scores assigned to all tested compounds. Out of the eight docking tools tested in Ref. 7 the best performance was shown by the SURFLEX software[41] with eight true hits among the 50 top-scored compounds. The screening procedure by AlgoComb consisted of docking with the filtered fragment cloud and subsequent rescoring of the generated positions with directed mixed score, as described above for the self-docking test. The obtained ranks of the 10 true TK inhibitors (see Supplementary Table V online) are in the range 4–24, thus showing the clear improvement over the best results yet attained. The best ranks (4, 5, and 6) are observed for chemically similar mct, dT, and idu, respectively, two of which are known to display a submicromolar binding constant (dT and idu) whereas all others bind to TK with micromolar binding constants.[44]

Ligand optimization with fragment clouds

The binding affinity of an already-known ligand can be substantially increased by replacement of certain chemical groups. This procedure is called ligand, or lead, optimization.[5,26,27] More precisely, the problem of optimization

 DOI 10.1002/prot

Table II
Results of the Test of Cloud-Based Optimization Algorithm on Five Examples of Known Optimization of Inhibitors of Three Proteins

Protein	Initial ligand (PDB id)	Initial fragment (F1)	Optimizing fragment (F2)	Number of suggested fragments	Rank of F2	Number of SCOP families
PTP1B	2bge_T2D	Hydrogen	Phenyl	43	3	80
PTP1B	2bge_T2D	Hydrogen	Methoxy	26	13	46
PTP1B	2bge_T2D	Hydrogen	Methyl	26	17	46
Deaminase	1ndw_FR2	Phenyl	1-Naphthyl	43	2	106
Trypsin	1c5p_BAM	Phenyl	2-Benzothiophenyl	185	21	233

The columns contain the protein name, PDB identifier of initial (non-optimized) protein-ligand complex, names of the initial and optimizing fragments F1 and F2, total number of replacing fragments suggested by our method, mixed score-based rank of F2 among all fragments, and number of SCOP families that provide the replacing fragments.

may be defined as follows: given a known ligand, its position in the binding site (known from X-ray crystallography or predicted by docking), and a ligand functional group (fragment) to be replaced, find a small number (~10–100) of replacement side groups aimed at improving (a) binding affinity, or (b) synthesizability, (c) solubility, (d) side effects, and so forth. A fragment cloud built for the binding site may serve as a source of molecular fragments able to replace the optimized ligand part, keeping in mind the feature to be optimized. Extended molecular fragments are constructed as described above. All their potential bonds are then checked to find those whose spatial position and orientation is close to the bond connecting the optimized part to the rest of the ligand. The fragments are filtered by linkage quality and sorted by mixed score, leaving only the fragments that improve the mixed score of the bound ligand.

The cloud-based optimization algorithm was tested on five known, experimentally verified examples of optimization of inhibitors of human protein tyrosine phosphatase-1B (PTP1B),[46] bovine adenosine deaminase,[47] and bovine trypsin.[48] For the purpose of ligand optimization, the complete ligand-site database containing 11,605 complexes has been used. For PTP1B, there are three independent examples of optimization by replacement of two different hydrogen atoms. In all cases, the structure and position of the initial (nonoptimized) ligand bound by the protein is known from the PDB. The question is, whether the computational optimization is able to reproduce the results of *in vitro* experiments, that is, suggest for replacement of the initial ligand fragment F1 the optimizing fragment F2 described in the original publication with satisfactory linkage and binding quality among the pool of other replacing fragments. In each case, the desired fragment F2 was in fact found, as shown in Table II. We checked the number of different replacing fragments suggested by the optimization algorithm, rank of F2 among the other fragments, and diversity of SCOP families of proteins that are the sources of replacing fragments. The rank of a fragment is based on its mixed score calculated as described in the previous section. More details including the chemical structures of the initial ligand and F2, binding affinities, and so forth, are given in Supplementary Table VI online.

As seen from Table II, the number of suggested replacing fragments varies from 26 to 185, with the maximum attained for the bovine trypsin inhibitor. The number of different SCOP families providing the fragments roughly correlates with the total number of fragments. The observed structural diversity of proteins that provide the replacing fragments emphasizes the fact that the site similarity detected by the method is highly local and does not depend on sequence or overall structural similarity of corresponding proteins. The rank values of F2 vary in the range 2–21, with the other high-scoring replacing fragments being candidates for *in vitro* validation of binding improvement—some of these may turn out to be better than F2 when tested experimentally.

DISCUSSION

We have developed a novel approach to analysis of interaction of proteins and small molecules utilizing the wealth of structural data on protein–ligand complexes. The approach operates on complete environments of ligand fragments in the complexes, in contrast to existing knowledge-based techniques that account for interactions of a ligand atom with different environment regions of the protein independently. This feature presumably accounts for the advantage of the cloud method over previous knowledge-based approaches, and shows that the knowledge-based functions have not extracted all the available information from the available structural data. We have shown that the combined use of the cloud score function with a particular energy score[34] essentially improves scoring selectivity. However, it is worth noting that the cloud score is independent of the energy score and can be combined with any traditional energy scoring function and then implemented in various drug design methods.

The performance level of 82% by AlgoComb for the independent Rognan's self-docking test set[35] substantially exceeds the plateau of 50–55% reached by a number of other docking programs.[7] In the virtual screening experiment

with 10 known inhibitors of HSV-1 thymidine kinase and 990 random drug-like molecules,[44] the method has also substantially improved the best ranks of the real binders.[7] The successful usage of cloud fragments for ligand optimization has been shown on known inhibitors of human protein tyrosine phosphatase-1B, bovine trypsin and bovine adenosine deaminase.

The experiments with reduced clouds show that the benefits of the suggested approach depend on the number of complexes with different ligands and structural relatives of the target protein in the complexes database. Note that the results described above have been obtained for the proteins that belong to the diverse families thoroughly represented in the PDB. As a result, approximately one half of cloud fragments after the layered filtering procedure originate from the structurally similar proteins (see Supplementary Table IV online). So we can expect similar results only for widely presented targets. However, the advance of the structural genomics projects[49] and the limited number of protein folds[50] means that in the future there will be fewer and fewer "orphan" target proteins.

The important advantage of the cloud-based approach to optimization and *de novo* ligand design over traditional methods is in the usage of the molecular fragments matching the binding protein environment instead of precompiled fragment databases.

The notion of local site similarity presented here along with the method for its detection can be applied to other areas, such as functional annotation and classification of novel proteins. Fragment clouds are also a powerful tool for analysis of site selectivity, protein–protein interactions, and side effects of developed drugs.

ACKNOWLEDGMENTS

The authors would like to thank S. Nikitin and M. Subbotin for valuable discussions and C. Queen for careful review of the manuscript. V.R. has performed part of the calculations at the Engelhardt Institute of Molecular Biology on the hardware funded by a Cellular and Molecular Biology grant from the Russian Academy of Sciences.

REFERENCES

1. Abagyan R, Totrov M. High-throughput docking for lead generation. Curr Opin Chem Biol 2001;5:375–382.
2. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
3. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. J Comput Aided Mol Des 2002;16:151–166.
4. Brooijmans N, Kuntz ID. Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 2003;32:335–373.
5. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 2004;3:935–949.
6. Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 2003;46:2287–2303.
7. Kellenberger E, Rodrigo J, Muller P, Rognan D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 2004;57:225–242.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
9. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM. Protein folds and functions. Structure 1998;6:875–884.
10. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: approaches and limitations. Nat Struct Biol 2000;7:991–994.
11. Jones S, Thornton JM. Searching for functional sites in protein structures. Curr Opin Chem Biol 2004;8:3–7.
12. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339:607–633.
13. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 1994;243:327–344.
14. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
15. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 2002;323:387–406.
16. Barker JA, Thornton JM. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 2003;19:1644–1649.
17. Stark A, Sunyaev S, Russell RB. A model for statistical significance of local similarities in structure. J Mol Biol 2003;326:1307–1316.
18. Shulman-Peleg A, Nussinov R, Wolfson HJ. SiteEngines: recognition and comparison of binding sites and protein–protein interfaces. Nucleic Acids Res 2005;33:W337–W441.
19. Jambon M, Andrieu O, Combet C, Deleage G, Delfaud F, Geourjon C. The SuMo server: 3D search for protein functional sites. Bioinformatics 2005;21:3929–3930.
20. Gold ND, Jackson RM. Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. J Mol Biol 2006;355:1112–1124.
21. Ozrin VD, Subbotin MV, Nikitin SM. PLASS: protein–ligand affinity statistical score–a knowledge-based force-field model of interaction derived from the PDB. J Comput Aided Mol Des 2004;18:261–270.
22. Moreno E, Leon K. Geometric and chemical patterns of interaction in protein–ligand complexes and their application in docking. Proteins 2002;47:1–13.
23. Zavodszky MI, Sanschagrin PC, Korde RS, Kuhn LA. Distilling the essential features of a protein surface for improving protein–ligand docking, scoring, and virtual screening. J Comput Aided Mol Des 2002;16:883–902.
24. Wu G, Vieth M. SDOCKER: a method utilizing existing X-ray structures to improve docking accuracy. J Med Chem 2004;47:3142–3148.
25. Jedrzejas MJ, Singh S, Brouillette WJ, Laver WG, Air GM, Luo M. Structures of aromatic inhibitors of influenza virus neuraminidase. Biochemistry 1995;34:3144–3151.
26. Bohm HJ, Banner DW, Weber L. Combinatorial docking and combinatorial chemistry: design of potent non-peptide thrombin inhibitors. J Comput Aided Mol Des 1999;13:51–56.
27. Olesen PH. The use of bioisosteric groups in lead optimization. Curr Opin Drug Discov Dev 2001;4:471–478.

 DOI 10.1002/prot

28. Wagener M, Lommerse JP. The quest for bioisosteric replacements. J Chem Inf Model 2006;46:677–685.
29. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization and performance of MMFF94. J Comp Chem 1996;17:490–519.
30. Kearsley SK. Structural comparisons using restrained inhomogeneous transformations. Acta Cryst 1989;A45:628–635.
31. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;7:536–540.
32. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28: 254–256.
33. Nikitin S, Zaitseva N, Demina O, Solovieva V, Mazin E, Mikhalev S, Smolov M, Rubinov A, Vlasov P, Lepikhin D, Khachko D, Fokin V, Queen C, Zosimov V. A very large diversity space of synthetically accessible compounds for use with drug design programs. J Comput Aided Mol Des 2005;19:47–63.
34. Muryshev AE, Tarasov DN, Butygin AV, Butygina OY, Aleksandrov AB, Nikitin SM. A novel scoring function for molecular docking. J Comput Aided Mol Des 2003;17:597–605.
35. Paul N, Rognan D. ConsDock: A new program for the consensus analysis of protein–ligand interactions. Proteins 2002;47:521–533.
36. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 2001;15:411–428.
37. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261:470–489.
38. Vigers GP, Rizzi JP. Multiple active site corrections for docking and virtual screening. J Med Chem 2004;47:80–89.
39. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 2004;47:1739–1749.
40. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein–ligand docking using GOLD. Proteins 2003;52: 609–623.
41. Jain AN. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. J Med Chem 2003;46: 499–511.
42. McMartin C, Bohacek RS. QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 1997;11:333–344.
43. Champness JN, Bennett MS, Wien F, Visse R, Summers WC, Herdewijn P, de Clerq E, Ostrowski T, Jarvest RL, Sanderson MR. Exploring the active site of herpes simplex virus type-1 thymidine kinase by X-ray crystallography of complexes with aciclovir and other ligands. Proteins 1998;32:350–361.
44. Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 2000;43:4759–4767.
45. Wurth C, Kessler U, Vogt J, Schulz GE, Folkers G, Scapozza L. The effect of substrate binding on the conformation and structural stability of Herpes simplex virus type 1 thymidine kinase. Protein Sci 2001;10:63–73.
46. Black E, Breed J, Breeze AL, Embrey K, Garcia R, Gero TW, Godfrey L, Kenny PW, Morley AD, Minshull CA, Pannifer AD, Read J, Rees A, Russell DJ, Toader D, Tucker J. Structure-based design of protein tyrosine phosphatase-1B inhibitors. Bioorg Med Chem Lett 2005;15:2503–2507.
47. Terasaka T, Kinoshita T, Kuno M, Nakanishi I. A highly potent non-nucleoside adenosine deaminase inhibitor: efficient drug discovery by intentional lead hybridization. J Am Chem Soc 2004;126: 34–35.
48. Katz BA. Structural basis for selectivity of a small molecule, S1-binding, submicromolar inhibitor of urokinase-type plasminogen activator. Chem Biol 2000;7:299–312.
49. Brenner SE. A tour of structural genomics. Nat Rev Genet 2001;2: 801–809.
50. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002;420:218–223.