Recognition of Binding Patterns Common to a Set of Protein Structures

Maxim Shatsky[1⋆], Alexandra Shulman-Peleg[1], Ruth Nussinov[2,3] and Haim J. Wolfson[1]

[1] School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[2] Sackler Inst. of Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel
[3] Basic Research Program, SAIC-Frederick, Inc, Lab. of Experimental and Computational Biology, Bldg. 469, Rm. 151, Frederick, MD 21702, USA

Abstract. We present a novel computational method, MultiBind, for recognition of binding patterns common to a set of protein structures. It is the first method which performs a multiple alignment between protein binding sites in the absence of overall sequence, fold or binding partner similarity. MultiBind recognizes common spatial arrangements of physico-chemical properties in the binding sites. These should be important for recognition of function, prediction of binding and drug design. We discuss the theoretical aspects of the computational problem of multiple structure alignment. This problem involves solving a 3D k-partite matching problem, which we show to be NP-Hard. The MultiBind method, applies an efficient Geometric Hashing technique to detect a potential set of multiple alignments of the given binding sites. To overcome the exponential number of possible multiple combinations it applies a very efficient filtering procedure which is heavily based on the selected scoring function. Our method guarantees detection of an approximate solution in terms of pattern proximity as well as cardinality of multiple alignment. We show applications of MultiBind to several biological targets. The method recognizes patterns which are responsible for binding small molecules such as estradiol, ATP/ANP and transition state analogues. The presented computational results agree with the available biological ones.
Availability: `http://bioinfo3d.cs.tau.ac.il/MultiBind/`.

Keywords: multiple structure alignment of binding sites; consensus binding patterns; pattern matching; pattern discovery, recognition of functional sites; k-partite matching;

Introduction

Binding sites with similar physico-chemical and geometrical properties may perform similar functions and bind similar binding partners. Such binding sites may

⋆ To whom correspondence should be addressed, email: maxshats@cs.tau.ac.il.

be created by evolutionarily unrelated proteins that share no overall sequence or fold similarities. Their recognition has become especially acute with the growing number of protein structures determined by the Structural Genomics project. Multiple alignment of binding sites that are known to have similar binding partners allows recognition of the physico-chemical and geometrical patterns that are responsible for the binding. These patterns may help to understand and predict molecular recognition. Moreover, multiple alignment of binding sites allows analysis of the dissimilarities of the binding sites which are important for the specificity of drug leads.

Sequence patterns have been widely used for comparison and annotation of protein binding sites [1]. Several methods search for patterns of residues that are conserved in their 3D positions and in amino acid identities [2–4]. However, there are numerous examples of functionally similar binding sites that are neither sequence order dependent nor share common patterns of amino acids [5–7]. Several methods have been developed for protein multiple structural alignment [8–12]. To overcome the alignment complexity of large protein structures these methods apply a variety of heuristics as well as some assumptions on properties of protein backbone, e.g. sequentiality of some backbone fragments. However, similar binding site patterns may appear in proteins with different overall folds. In addition, such patterns may be relatively small and can be easily missed when applying heuristic approaches used for protein backbone alignment. Methods for recognition of a pharmacophore common to a set small ligands [13, 14] share some methodological aspects with problems of protein backbone or binding site alignment. However, most developed methods for common pharmacophore detection are optimized for its specific problem definition, e.g. assume a tree-like ligand topology. Consequently, the methods for protein backbone alignment or common ligand pharmacophore detection are generally not suitable for recognition of common patterns of protein binding sites.

Several works have analyzed complexes of different proteins with the same ligand. The superimposition between the binding sites has been obtained by alignment of their common ligands [6,15]. This approach has several limitations. First, it can analyze only protein structures with exactly the same partner ligands. Second, the same ligand can bind in alternative modes even to the same protein binding site [6]. Therefore, alignment according to ligands may fail to recognize the pattern.

Computational methods have been developed for direct alignment of protein binding sites. From the algorithmic standpoint, this involves solving a problem of spatial labeled/unlabeled pattern detection. Most of the methods apply clique detection algorithms. Some recent examples are the methods described by Kinoshita et al. [16], Schmitt et al. [17] and Shulman-Peleg et al. [7] However, all of these methods perform a comparison of only two molecules. Pairwise alignments may contain large number of features that are not necessarily required for the binding. Multiple alignments of binding sites with the same function may help to recognize the smallest set of features, a *consensus*, that is essential to achieve the desired biological effect. In creation and screening of databases, such *consen-*

sus binding patterns may facilitate the development of efficient ranking schemes and database architectures [7]. Although it is possible to combine the results of various pairwise comparisons, high scoring pairwise solutions do not necessarily lead to a high scoring solution for a set of molecules [18].

Below we review the progress made in Computational Geometry in studying this problem. We start with a description for the case of two structures and continue to multiple structures. Our emphasis is only on subjects related to molecular structures. The problem of pattern detection answers the following question. Given two point sets A and B, find a subset of A that is similar to some subset of B. The optimization problem is to maximize the cardinality of similar subsets. One common way to define the similarity between two point sets is by the *bottleneck* metric [19, 20]. Similar sub-sets are called ϵ-congruent if the maximal distance between the matched points is less than ϵ. The optimization problem, called the *Largest Common Point Set* (LCP) problem, involves finding a transformation, e.g. Euclidean motion, that maximizes the size of two ϵ-congruent sub-sets. For the *bottleneck* metric in 3D it can be solved in $O(n^{32.5})$ time [21], where $n = max(|A|, |B|)$. Obviously, its time complexity is not practical even for small point sets. Therefore, more efficient methods are required.

Approximation techniques can significantly reduce time complexity at the price of solution accuracy. A simple *alignment technique* [22] constructs a finite set of transformations by *aligning* each triplet of points from the first structure with each ϵ-congruent triplet from the second one. For each transformation we can apply the maximal bipartite matching algorithm to compute a bijective mapping of points that are within ϵ distance from each other. Such *alignment technique* guarantees finding LCP under 8ϵ-congruence of cardinality at least of LCP under ϵ-congruence, if the latter exists. The time complexity is $O(n^{8.5})$. This technique was first developed for the Hausdorff distance [23] and later applied for the *bottleneck* distance [19]. Instead of approximating ϵ-congruence the same technique can be applied to approximate the LCP size [24]. Interestingly, an optimal algorithm for solving LCP for a group of transformations limited to rotations only, can improve the approximation factor of the LCP problem for general Euclidean transformations from 8ϵ to 2ϵ, while preserving the complexity of $O(n^{8.5})$ [24]. However, an implementation of such a technique is more complicated and the constant factors of the time complexity become larger.

Extension of the problem to detect a common point set between a set of K structures (from now on the term *point set* and *structure* will be used alternatively) has many important applications for the analysis of protein and drug molecules. However, even in one dimensional space for the case of exact congruence ($\epsilon = 0$) the problem is NP-Hard and it is hard to approximate within the factor $n^{1-\delta}$, for any $\delta > 0$, where n is the size of the smallest structure [18]. The problem is further complicated by the fact that in practice it is impossible to work with *zero-congruence*. Therefore, we face another combinatorial problem. Namely, given a set of superimposed structures, compute the largest common ϵ-congruent sub-set. We will call this problem *K-partite-3D* matching. While for two structures ($K = 2$) it can be solved by bipartite matching, for $K > 2$ struc-

tures it can be solved by *K-partite* matching. However, this problem is known to be NP-Hard even for $k = 3$ in general and hyper graphs [25, 26]. Here, we show that the *K-partite-3D* matching problem is also NP-Hard.

In this paper we present an efficient, practical, method, MultiBind, for identification of common protein binding patterns by solving the multiple structure alignment problem. The problem we aim to solve is NP-Hard, therefore our goal is to find a trade-off between practical efficiency and theoretical bounds of solution accuracy, while, most importantly, validating the biological correctness of the results. We represent the protein binding site as a set of 3D points that are assigned a set of physico-chemical and geometrical properties important for protein-ligand interactions. The implementation of our method includes three major computational steps. The first one is a generation of 3D transformations that *align* the molecular structures. Here we apply the time efficient Geometric Hashing method [27]. The advantage of this method is that it enables to avoid processing of points that can not be matched under any transformation. In other words, its time complexity is proportional to the number of potentially matched points included in the defined set of transformations. The second step is a search for a combination of 3D transformations that gives the highest scoring common 3D core. For this step we provide an algorithm that guarantees to find the optimal solution by applying an efficient filtering procedure which practically overcomes the exponential number of multiple combinations. The final step is a computation of matching between points under multiple transformations, namely *K-partite-3D* matching. Here, we give a fast approximate solution with factor K. The overall scheme guarantees to approximate the ϵ-congruence as well as the cardinality of multiple alignment. We apply MultiBind to some well studied biological examples such as estradiol, ATP/ANP and transition state analogues binding sites. Our computational results agree with the available biological data.

The Largest Common Point Set Problem

We start from the definition of a pure geometric problem and in the next section extend it to the biology related problem. Objects are represented by point sets in 3D Euclidean space. Object S_1 is $\epsilon - congruent$ to S_2 if there exists an Euclidean transformation T and a bijective mapping $m : S_2 \to S_1$ such that for each point $s \in S_2$, $d(m(s), T(s)) \leq \epsilon$, where $d(.,.)$ is the Euclidean metric. Such similarity measure is called the *bottleneck matching measure*. For simplicity, we will also say that $d(S_1, T(S_2)) \leq \epsilon$ if the two objects are of the same cardinality, $S_1 = \{p_i\}_1^n$ and $S_2 = \{q_i\}_1^n$, and $d(p_i, T(q_i)) \leq \epsilon$, $i = 1, ..., n$, that is the set S_2 is preordered according to some bijective mapping.

Problem 1. Largest Common Point Set (LCP) between 2 Sets. *Given $\epsilon > 0$, and two point sets S_1 and S_2, find a transformation T and equally sized subsets $S_i' \subseteq S_i$ (i=1,2) of maximal cardinality such that $d(S_1', T(S_2')) < \epsilon$.*

Assuming that $|S_1|, |S_2| \leq n$ this problem can be exactly solved in $O(n^{32.5})$ time [21]. Now, we define an approximation version for this problem. Assume

that L is the size of the largest common point set of S_1 and S_2 with an error ϵ. An *(ϵ,β,γ)-approximation* of the LCP problem, $\epsilon \geq 0, \beta \geq 1, \gamma \geq 1$, is to find a common point set of size at least L/γ with an error of at most $\beta\epsilon$. Consider a simple alignment method that works as follows. For each triplet of points from S_1 and for each triplet from S_2 construct a 3D transformation that *aligns* the second triplet with the first one (it is enough to consider only pairs of triangles with maximal triangle side difference $\leq 2\epsilon$). Apply this transformation on S_2 and construct a bipartite graph where the vertices are the points of S_1 and of transformed S_2, and edges are created between the points with distance less than $\beta\epsilon$. Apply a maximal bipartite matching algorithm to compute the largest set of aligned points. The algorithm works in $O(n^3 * n^3 * n^{2.5})$. For this method the approximation ratio depends on the following alignment rule for construction of a 3D transformation based on two triplets of points (p_1, p_2, p_3) and (q_1, q_2, q_3).

Local Reference Frame, $LRF(p_1, p_2, p_3)$: Define a local right hand coordinate system s.t.: $p_1 = (0, 0, 0)$, $(p_2 - p_1)/|(p_2 - p_1)| = (1, 0, 0)$, $(p_2 - p_1) \times (p_3 - p_1)/|(p_2 - p_1) \times (p_3 - p_1)| = (0, 0, 1)$.

Alignment Rule: Define a transformation T' that superimposes $LRF(q_1, q_2, q_3)$ onto $LRF(p_1, p_2, p_3)$, i.e. $p_1 = T'(q_1)$, $(p_2 - p_1)/|p_2 - p_1| = (T'(q_2) - T'(q_1))/|T'(q_2) - T'(q_1)|$ and $sign((p_2 - p_1) \times (p_3 - p_1)) = sign((T'(q_2) - T'(q_1) \times (T'(q_3) - T'(q_1)))$

This rule gives the approximation ratio $\beta \leq 8$ ($\gamma = 1$) [19, 23]. In this work we extend the LCP problem to multiple sets and we call it the mLCP problem. We also define the multiple LCP problem with respect to a *pivot* structure and we call it the pmLCP problem.

Problem 2. **(mLCP).** Largest Common Point Set between K Sets. *Given $\epsilon > 0$, and K point sets S_i, $i = 1, ..., K$, find transformations $\{T_i\}$, $i = 2, ..., K$, and equal sized sets $\{S'_i \subseteq S_i\}$, $i = 1, ..., K$, of maximal cardinality such that $d(T_i(S'_i), T_j(S'_j)) < \epsilon$ ($i \neq j$, $i, j = 1, ..., K$, where T_1 is identity transformation).*

Problem 3. **(pmLCP).** Largest Common Point Set between K Sets with a Pivot Structure. *Given $\epsilon > 0$, a pivot set S_1 and $K - 1$ point sets S_i, $i = 2, ..., K$, find transformations $\{T_i\}$, $i = 2, ..., K$, and equal sized sets $\{S'_i \subseteq S_i\}$, $i = 1, ..., K$, of maximal cardinality such that $d(S'_1, T_i(S'_i)) < \epsilon$, $i = 2, ..., K$.*

Not surprisingly, both problems are NP-Hard, even in one dimensional space for the case of exact congruence, i.e. $\epsilon = 0$ [18]. In addition, for $\epsilon > 0$ we face another combinatorial problem. Consider a reduced mLCP/pmLCP problem where the transformation search is omitted, i.e. the position of the structures is fixed. Then, for two structures the problem is easily solved by a bipartite matching algorithm. However, for K structures it requires to solve a K-dimensional matching in Euclidean 3D space. In general graphs this problem is NP-Hard even for three sets [25, 26]. We show that it is still NP-Hard even for graphs defined on 3D structures, where edges between the nodes from different partitions (structures) are created if and only if the distance between the nodes is less than ϵ.

Definition 4. (K-partite-3D graph). *Given $\epsilon > 0$ and K point sets S_i, $i = 1, ..., K$, a K-partite-3D graph $G(S_1, ..., S_K) = (V, E)$ is defined as $V = \{\cup_{i=1}^{K} S_i\}$ and $E = \{(p_i, p_j) : i \neq j, p_i \in S_i, p_j \in S_j, d(p_i, p_j) \leq \epsilon\}$. A matching of a K-partite-3D graph is a set of disjoint K-tuples $\{(p_{t_1}, ..., p_{t_K}) : p_{t_i} \in S_i, p_{t_j} \in S_j, (p_{t_i}, p_{t_j}) \in E\}$.*

Definition 5. (K-partite-3D-pivot graph). *Given $\epsilon > 0$ and K point sets S_i, $i = 1, ..., K$, of which S_1 is the pivot, a K-partite-3D-pivot graph $G(S_1, ..., S_K) = (V, E)$ is defined as $V = \{\cup_{i=1}^{K} S_i\}$ and $E = \{(p_1, p_j) : j > 1, p_1 \in S_1, p_j \in S_j, d(p_1, p_j) \leq \epsilon\}$. A matching of a K-partite-3D-pivot graph is a set of disjoint K-tuples $\{(p_{t_1}, ..., p_{t_K}) : p_{t_1} \in S_1, p_{t_j} \in S_j, (p_{t_1}, p_{t_j}) \in E\}$.*

Theorem 6. *The maximal cardinality matching problem in K-partite-3D and K-partite-3D-pivot graphs is NP-Hard.*

A Sketch of the Proof. First, we briefly present a reduction from *3-SAT* to *3-partite matching* in general graphs (for more details see [25]), and then extend it for the instances of K-partite-3D and K-partite-3D-pivot graphs.

An instance of the 3SAT problem includes a set of variables $U = \{u_1, u_2, ..., u_n\}$ and a set of clauses $C = \{c_1, c_2, ..., c_m\}$. Each clause contains three literals of variables U. The goal of the reduction is to construct three disjoint sets S_1, S_2 and S_3 of equal cardinality, and a set of edges $M \subseteq S_1 \times S_2 \times S_3$ such that M contains a perfect matching if and only if C is satisfiable.

Three classes of edges are created, T - "truth setting and fan-out", C - "satisfaction testing" and G - "garbage collection". The components of T are constructed for each variable u_i. Denote $u_i[j]$ to be a variable u_i in clause j.

$$T_i^t = \{(\bar{u}_i[j], a_i[j], b_i[j]) : 1 \leq j \leq m\}$$
$$T_i^f = \{(u_i[j], a_i[j+1], b_i[j]) : 1 \leq j \leq m\} \cup \{(u_i[m], a_i[1], b_i[m])\}$$
$$\bar{u}_i[j], u_i[j] \in S_1, \quad a_i[j] \in S_2, \quad b_i[j] \in S_3$$

The component T forces a matching to choose between setting u_i true and setting u_i false. Any perfect matching will have to include either all triplets from T_i^t or all triplets from T_i^f, see Figure 1 (a). Next, for each clause c_j a component C_j aims to select a truth setting for one of its three literals: $C_j = \{(u_i[j], s_2[j], s_3[j]) : u_i \in c_j\} \cup \{(\bar{u}_i[j], s_2[j], s_3[j]) : \bar{u}_i \in c_j\}$, where $s_2[j] \in S_2$ and $s_3[j] \in S_3$.

Thus, only one triplet can be contained in any matching assigning the clause c_j to true setting. Finally, the "garbage collection" component aims to compensate the unequal number of nodes created so far in S_1 and in other two partitions S_2 and S_3: $G = \{(u_i[j], g_2[k], g_3[k]), (\bar{u}_i[j], g_2[k], g_3[k]) : 1 \leq k \leq m(n-1), 1 \leq i \leq n, 1 \leq j \leq m,\ g_2[j] \in S_2,\ g_3[j] \in S_3\}$.

To summarize, the edges are defined as: $T = \cup_{i=1}^{n}(T_i^t \cup T_i^f)$, $C = \cup_{j=1}^{m} C^j$, $M = T \cup C \cup G$. This completes the reduction from *3-SAT* to *3-partite matching*. Next, we adapt the above reduction for K-partite-3D type graphs.

Notice that the constructed graph M does not belong to the K-partite-3D type of graphs. Only the component T can be drawn in 2D to satisfy this property, i.e. only the point triplets from T can be placed within ϵ distance one from

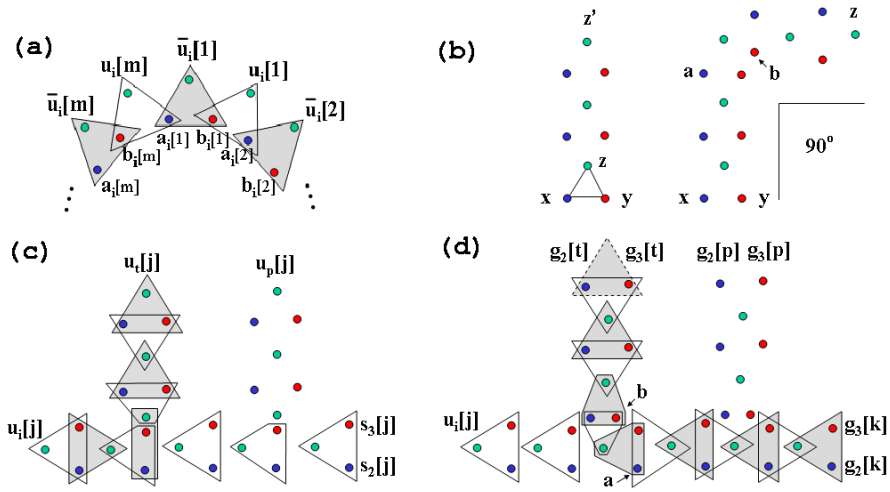
each other (see Figure 1 (a))[4]. The problem is that the nodes of type s_2, s_3 and g_2, g_3 can not be placed in 3D so that their distance from the different nodes of type u_i is less than ϵ. To resolve this problem we introduce *long-distance-edge* gadgets. The basic principle is illustrated in Figure 1 (b). The edge (x, y, z) can be elongated to any distance (z transforming to z') preserving the property for matching. Also, we can *bend* to 90^o any *long-distance-edge*. We can bend an edge in two different ways. The first one, as illustrated in Figure 1 (b), continues the edge in the same 2D plane. The second option, bends the edge so that its two parts are in two perpendicular 2D planes (not illustrated). Care should be taken at the bending part, e.g. nodes a and b from Figure 1 (b) should be placed at a distance larger than ϵ, otherwise it introduces ambiguity for matching. The second gadget aims to *split* edges going from nodes of type s_2, s_3 (g_2, g_3). Assume we have three edges $(u_i[j], s_2[j], s_3[j])$, $(u_t[j], s_2[j], s_3[j])$ and $(u_p[j], s_2[j], s_3[j])$. Figure 1 (c) illustrates how these three edges can be constructed. Triangles illustrate possible matching. This gadget guarantees that any perfect matching will select only one node $u[j]$ with combination of nodes from this gadget. However, there are many *long-distance-edges* coming to nodes of type $u[j]$ and there is a need to join them. The *split* gadget is not suitable for this task, therefore we introduce a *join* gadget (see Figure 1 (d)). The *join* gadget guarantees that any perfect matching will connect a node $u[j]$ to only one pair of type s_2, s_3 (g_2, g_3). To complete the construction we need to show how to place in 3D all the *long-distance-edges* and connections between them. The idea is to place the component T in the plane $(x, y, 0)$ (*zero level*), to place the components C_j on *negative levels* $(x, y, -l_j)$ and the components G_k on *positive levels* (x, y, l_k). The *long-distance-edges* are constructed between the levels like water pipes. The whole construction requires polynomial number of components. Due to lack of space we omit the exact details, which will appear elsewhere. Notice, that by selecting the first partition S_1 as a pivot structure and deleting the edges between S_2 and S_3 the same reduction works as well for the instances of K-partite-3D-pivot graphs.

The MultiBind Algorithm

Input Representation: Physico-Chemical Properties. Selection of the proper representation is crucial for the biochemical significance of the recognized patterns. Given the atomic coordinates of a protein structure, we follow Schmitt et al. [17] and for each amino acid we group atoms with similar physicochemical properties to functional groups. These are localized by 3D points in space, denoted as pseudocenters. Each pseudocenter represents one of the following properties important for protein-ligand interactions: *hydrogen-bond donor, hydrogen-bond acceptor, mixed donor/acceptor, hydrophobic aliphatic and aromatic(pi) contacts*. Since both backbone and side-chain atoms are considered,

[4] In the *3-SAT* to *3-partite matching* reduction, the definition of a hypergraph edge as a triplet of points (a,b,c) is equivalent to three edges (a,b), (a,c) and (b,c) in a regular graph.

Fig. 1.



each amino acid is represented by a set of such pseudocenters. We construct the smooth molecular surface as implemented by Connolly [28] and retain only pseudocenters that represent at least one surface exposed atom. When considering binding sites, we refer only to the surface regions that are within 4Å from the binding partner. In practice, a comparison of the spatial locations of the retained pseudocenters is not sufficient for the accurate prediction of protein-ligand interactions. Thus, we are interested in the maximal number of matching pseudocenters that are most similar in all the physico-chemical and geometrical aspects. For each pair of pseudocenters, p and q, we define a scoring function *PC-Score(p,q)* which measures the similarity of the properties important for the specific type of interaction in which they can participate ($PC\text{-}Score(S_1, S_2)$ is defined as a sum of the matched point scores). The exact calculations and default parameters are detailed in Appendix A. Therefore, practically, we look for a solution for a weighted *pmLCP* problem that we define as[5]:

Problem 7. **(Max-Min Weighted** *pmLCP* **Problem)** *Given $\epsilon > 0$, a scoring function PC-Score, a pivot set S_1 and $K-1$ point sets S_i, $i = 2, ..., K$ find transformations $\{T_i\}$, $i = 2, ..., K$, and equal sized sets $\{S_i' \subseteq S_i\}$, $i = 1, ..., K$, such that $d(S_1', T_i(S_i')) < \epsilon$, $i = 2, ..., K$, and $min_i PC\text{-}Score(S_1', T_i(S_i'))$ is maximal.*

The Pattern Matching Algorithm. There are three major computational steps: (1) generation of 3D transformations and potential points for matching; (2) combinatorial search for a combination of 3D transformations that gives the highest scoring common 3D core (*Traversal stage*); and (3) computation of *K-partite-3D-pivot* matching.

[5] In our implementation we consider only the pmLCP problem since the K-partite-3D matching of the mLCP problem introduces additional complications even for greedy approaches. We'll address this problem somewhere else.

In our approach we follow the efficient strategy of the Geometric Hashing method [27]. The Geometric Hashing method consists of two stages, *preprocessing* and *recognition*. At the *preprocessing* stage each triplet of pseudocenters, (a, b, c), from each molecule except the pivot is considered as a local reference frame $r = LRF(a, b, c)$. The coordinates of the other points are calculated with respect to the local reference frame r. This information is stored in a *Geometric Hash Table*. The key to the hash table is (x^r, y^r, z^r, p), where (x^r, y^r, z^r) are point coordinates with respect to the local reference frame r, and p is the physico-chemical property of the pseudocenter. Only pseudocenters with the same property can be matched [6]. The data stored in the hash table includes the key itself and the identifiers of the molecule and the reference frame.

In the *recognition* stage the same process as in the *preprocessing* stage is repeated for the pivot molecule. However, instead of storing data in the hash table, all entries close to the key within radius ϵ and with the same physico-chemical property are retrieved. For each reference frame r of the pivot structure a voting table is created. It counts the number of matched points for each reference frame stored in the hash table. For simplicity, we explain the method for the pure geometrical case, i.e. for the pmLCP problem. If a reference frame r' from structure i received v votes that means the following. Define a 3D transformation $T_{r,r'}$ that superimposes the triplets of points r' on r according to the *Alignment Rule*. Applying $T_{r,r'}$ on S_i will result in v point pairs from the pivot and i structure that are within ϵ distance. Thus, the size of a maximal matching between S_{pivot} and $T_{r,r'}(S_i)$ is less than v. Therefore, for the next step it is enough to consider only reference frames that have received a number of votes equal or greater than M^*, the size of the largest multiple solution found so far (initially $M^* = 0$). For each survived transformation T we store the list of matched points, $\{(p, q) : p \in S_{pivot}, q \in S_i, |p - T(q)| < \epsilon\}$.

Traversal stage. For each reference frame of the pivot structure we create a combinatorial bucket that contains transformations that received a high number of votes. Namely, a combinatorial bucket for the reference frame r is defined as $CB_r = \{T^2, T^3, ..., T^K\}$, where $T^i = \{T_{i_j}\}$ is a set of transformations for structure i that received $v > M^*$ votes. A multiple alignment is a combination of $K-1$ transformations, $(T^2_{i_2}, T^3_{i_3}, ..., T^K_{i_K})$. The number of all possible combinations equals to $|T^2| * |T^3| * ... * |T^K|$, which is exponential with K. However, we have implemented a branch-and-bound traversal method which in practice is very efficient. First we provide some definitions. Given a transformation vector of the first t structures, $T = (T_{i_2}, ..., T_{i_t})$, create a *t-partite-3D-pivot* graph, $G(T) = G(S_1, T_{i_2}(S_2), ..., T_{i_t}(S_t))$. Define single sides of the graph $G(T)$, $G(T)[j] = \{p_j : p_j \in S_j,\ \exists p_1 \in S_1\ (p_1, p_j) \in G(T)\ and\ \forall k \leq t\ \exists p_k \in S_k\ (p_1, p_k) \in G(T)\}$. Let $M(G(T))$ be a maximal *t-partite-3D-pivot* matching of the graph $G(T)$. Obviously, $M(G(T)) \leq M(G(S_{pivot}, T_{i_j}(S_j))) \leq |G(T)[j]|$.

Given a combinatorial bucket $CB = \{T^2, T^3, ..., T^K\}$ we iteratively traverse it in the following manner. Assume that we have created a vector $T =$

[6] Pseudocenters that can function both as hydrogen bond donors and acceptors are encoded twice, once as donors and once as acceptors.

$(T_{i_2}, T_{i_3}, ..., T_{i_t})$, $T_{i_j} \in T^j$. We try to extend it with a transformation $T_{i_{t+1}} \in T^{t+1}$, $T^* = (T_{i_2}, T_{i_3}, ..., T_{i_t}, T_{i_{t+1}})$. Clearly, $|G(T^*)[j]| \leq |G(T)[j]|$, $j = 2, ..., t$. Therefore, if for some index j holds $|G(T^*)[j]| \leq M^*$, then we can disregard the vector T^* and start to build another combination of transformations. Essentially, we continue with the vector T and try to add another transformation from T^{t+1}, and so on. The number of traversals may be exponential, however in practice the $M(G(T))$ drops very quickly below M^* as the algorithm advances in iterations in the *recognition* stage [7]. Still, the theoretical bound is $O(n^3 n^{3(K-1)})$.

K-partite-3D-pivot Matching. During the traversal stage, once we reach the last bucket we have a uniquely defined K-partite-3D-pivot graph. The next step is to solve the matching problem. As we have shown above this problem is NP-Hard. We apply a greedy method, which iterates over pivot points and selects K-tuples, from non-selected points. This method gives a K approximation to the largest matching since at each greedy selection of K-tuples it may violate at most K-1 nodes that may belong to the optimal matching[8]. In the context of molecular structures for small ϵ (around 3Å) the maximal node degree is bound by a small constant. Therefore the time complexity of the greedy method is $O(Kn)$.

Theorem 8. *MultiBind algorithm is an* $(\epsilon, 8, K)$*-approximation*[9] *for* **Problem 3** *and has time complexity* $O(n^{3K} nK)$.

In practice, when solving the *Max-Min Weighted pmLCP* we introduce the following modifications. First, we define M^* to be the highest physico-chemical score of the multiple solution found so far. Given equally sized sets $(S_1, ..., S_t)$ the physico-chemical score M is defined as in Problem 4 by $M = min_j PC\text{-}Score(S_1, S_j)$, $j = 2...t$. When traversing the combinatorial buckets, instead of looking at the cardinality of the side j, $|G(T^*)[j]|$, we estimate the upper bound of $PC\text{-}Score(S_1, T_j(S_j))$ as $PC\text{-}Score(G(T^*)[j]) = \sum_{q \in T_j(S_j)} max_p PC\text{-}Score(p, q)$. Therefore, we disregard vectors T^* for which $PC\text{-}Score(G(T^*)[j]) \leq M^*$ for some j. We retain a user defined number of high scoring solutions, which are then evaluated by an additional *Overall Surface Scoring* [7] function which compares the corresponding surfaces of the binding sites.

Biological Results

Below we present examples of application of MultiBind for recognition of patterns required for binding of different ligands. In each of the presented examples,

[7] In the second example from the *Results* section, the total number of combinations for all combinatorial buckets is about $1.3 \cdot 10^{11}$, which shows the exponential nature of the problem. The filtering procedure leaves only 246310 combinations of multiple alignments. Most filtering is done already at the third structure ($t = 3$).

[8] The best known approximation algorithm for hyper-graphs gives K/2 ratio [29]

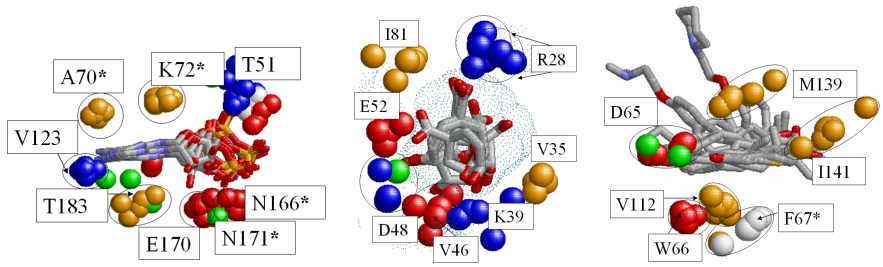
[9] It is possible to reduce the $\beta = 8$ approximation to any accuracy $c\beta$, $c \leq 1$, by applying a discretization technique of the transformational space [30, 31]. However, the payoff is increasing the time complexity factor proportional to $(1/c)^6$.

we describe the details of a single solution that received the highest score. An additional example of application of MultiBind to proteins of trypsin and subtilisin folds is presented by Mintz et al [32]. The running times are measured on a standard PC, Intel(R) Pentium(R) IV 2.60GHz CPU with 2GB RAM. The default distance threshold for the ϵ-congruence is 3.0Å .

ATP/ANP Binding Sites of Protein Kinases. To validate the performance of the method on a well studied example we have selected a set of ATP/ANP binding sites extracted from 5 different protein kinases: cAMP-dependent PK (1cdk), Cyclin-dependent PK, CDK2 (1hck), Glycogen phosphorylase kinase (1phk), c-Src tyrosine kinase (2src), Casein kinase-1, CK1 (1csn). We applied MultiBind to perform a multiple alignment of the corresponding ATP/ANP binding sites. These were recognized to share 14 pseudocenters, 4 of which are created by amino acids with the same identity (see Figure 2(a)). The RMSD between the adenine moieties (which are not a part of the input and are used for verification only) under these transformations is less than 1.4Å . The average binding site size is 76 pseudocenters, and the running time is 58 minutes. It must be noted that since these proteins share similar overall folds, the 3D superposition problem of the binding sites can be solved by multiple backbone alignment methods [11, 12]. However, these methods do not give solution to the *K-partite-3D* matching problem of physico-chemical features (since these are not-ordered on the protein surface). Below we present two examples for which both the superimposition and the matching problems can not be solved by standard protein backbone alignment methods.

Transition State Analogue Binding Sites. We have selected five binding sites complexes with endo-oxabicyclic transition state analogues (TSA/BAR). The binding sites were extracted from proteins of three different folds: (1) Chorismate mutase II (1ecm, 4csm, 3csm); (2) Bacillus chorismate mutase-like (2cht); (3) Immunoglobulin-like beta-sandwich (1fig). Figure 2(b) presents 8 functional groups that were recognized by MultiBind to be shared by all the binding sites. Two of the compared proteins (1ecm and 4csm) were previously aligned by Schmitt et al [17]. Most of the pseudocenters recognized by MultiBind are indeed a subset of those obtained by their pairwise alignment method (except for two donors contributed by 1ecm:Arg28). However, 10 of the functional groups common to a pair of chorismate mutases according to their study, were not recognized to be common to the five structures compared by MultiBind. Alignment of multiple structures with different folds helps to identify the minimal set of features required for the binding of endo-oxabicyclic transition state analogues. The average size of a binding site is 29, and the running time is 8 minutes.

Estradiol Binding Sites. Estradiol molecules are known to bind to protein receptors with different overall sequences and folds. The dataset of this study was comprised of the binding sites of 7 proteins from 4 different folds: (1) Nuclear receptor ligand-binding domain (3ert, 1a52, 1err, 1qwr); (2) NAD(P)-binding Rossmann-fold (1fds); (3) Concanavalin A-like lectins/glucanases (1lhu); (4) P-loop containing nucleoside triphosphate hydrolases (1aqu). Two of these structures were crystallized with Raloxifen (1err) and 4-hydroxytamoxifen (3ert),



(a) (b) (c)

Fig. 2. Multiple alignments done by MultiBind. Matched pseudocenters are represented as balls. Hydrogen bond donors are blue, acceptors - red, donors/acceptors - green, hydrophobic aliphatic - orange and aromatic - white. Matched pseudocenters (backbone or side-chain) from identical amino acids are marked by *. The ligand molecules are presented for verification purpose only and are not a part of the input to MultiBind. **(a)** Multiple alignment of 5 ATP/ANP binding sites, the labeling is according to 1cdk. **(b)** Multiple alignment of five endo-oxabicyclic transition state analogue binding sites. The labeling and the surface (depicted in dots) is according to 1ecm. **(c)** Multiple alignment of eight estradiol binding sites, the labeling is according to 3ert.

which are different from estradiol. In spite of the conformational changes required to accommodate these ligands, MultiBind has recognized 6 functional groups shared by all the binding sites (see Figure 2(c)). One of them is a conserved Phenylalanine (1lhu:Phe67) with an aromatic property shared by all the binding sites. The mean binding site size is 44 pseudocenters and the running time is 15 minutes.

In order to compare the presented results with those obtained by superimposition of ligand molecules, we performed such an alignment for the above mentioned examples (for the complexes with the same binding partners). In the last two cases alignment by ligands failed to recognize any significant pattern (less than 3 pseudocenters), while MultiBind identified patterns of size 8 and 6.

Conclusions

We have presented a novel computational method, MultiBind, for recognition of physico-chemical binding patterns. The method is practically efficient for multiple alignment of protein binding sites and guarantees to detect an approximate solution for the case of pure geometrical problem. We have shown that the matching problem of K-partite-3D/K-partite-3D-pivot graphs is NP-Hard. We have presented an efficient filtering procedure which in our applications practically overcomes the exponential number of multiple combinations.

We have applied MultiBind to several biological targets, such as the binding sites of estradiol, ATP/ANP and transition state analogues. MultiBind is the

first method that performs multiple alignment of binding sites in the absence of overall sequence, fold or binding partner similarity. To the best of our knowledge, the presented results can not be obtained by any other existing computational method. We hope that it will be a useful tool in prediction of molecular recognition and in identification of *consensus binding patterns*. These are important for improvement of architectures of databases of binding sites and development of efficient ranking schemes.

However, from the biological standpoint the method has several limitations. First, there is no explicit treatment of protein flexibility which is introduced only through a set of thresholds to allow variability in locations. Second, due to the hardness of the problem the method is practically limited to point sets of size about 100. Third, scoring functions are known to be one of the major problems in all types of *in silico* predictions. The scoring function of MultiBind suffers from the same limitations [7]. We intend to address these challenges in our future research.

Acknowledgments.

We thank D. Schneidman for contribution of software and O. Dror and M. Landau for their critical reading. The research of M.S. is supported by a PhD fellowship in "Complexity Science" from the Yeshaya Horowitz association. This research has been supported in part by the "Center of Excellence in Geometric Computing and its Applications" funded by the Israel Science Foundation. The research of H.J.W. is partially supported by the Hermann Minkowski-Minerva Center for Geometry at TAU. The research of R.N. has been funded in whole or in part with Federal funds from the NCI, NIH, under contract number NO1-CO-12400. The content of this publication does not necessarily reflect the view or policies of the Dep. of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

Appendix A: Physico-Chemical Scoring

Let p and q be the two matched pseudocenters.

• $dist(p,q)$ - the distance between p and q after the superimposition. Default threshold for the maximal distance is ϵ =3.0Å.

• $chem(p)$ - the physico-chemical property of the point p. There are three types of properties: Hydrogen Bonding (HB), Aliphatic Hydrophobic (ALI) and Aromatic (PII).

• $charge(p)$ - the partial atomic charge of the atom p, which can form hydrogen bonds. $charge(p,q) = |charge(p) - charge(q)|$.

• $shape(p)$ - the average curvature of the surface region created by p. Calculated as an average of the solid angle shape functions [33] with spheres of radius 4,5,6 and 7Å . The sphere centers are located at projection point of p to the surface. $shape(p,q) = |shape(p) - shape(q)|$.

• $n_S(p)$ - normal vector at projection point of p to the surface, $n_S(p,q) = n_S(p) \cdot n_S(q)$.

• $n_{PII}(p)$ - for aromatic pseudocenters denotes the normal to the plane of the aromatic ring. $n_{PII}(p,q) = n_{PII}(p) \cdot n_{PII}(q)$.

• $v_{ALI}(p,q)$ - the overlap of the hydrophobic group spheres of p and q, approximated by the difference between sum of radiuses and the distance between the centers.

Each pair of matched pseudocenters is assigned a score according to the similarity of

the properties important for the specific type of interaction:

$$PC\text{-}Score\ (p,q) = \begin{cases} 0, & dist(p,q) > \epsilon\ or\ chem(p) \neq chem(q) \\ 0, & shape(p,q) > 0.2\ or\ n_S(p,q) > 0.2 \\ dist(p,q)/(1+charge(p,q)) & chem(p) = HB \\ dist(p,q)/(1+shape(p,q)+n_{PII}(p,q)) & chem(p) = PII \\ (dist(p,q)+v_{ALI}(p,q))/(2+20*shape(p,q)) & chem(p) = ALI \end{cases}$$

References

1. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., Bairoch, A.: The PROSITE database, its status in 2002. Nucleic Acids Res. **30** (2002) 235–238
2. Wallace, A.C., Laskowski, R.A., Thornton, J.M.: Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Science **5** (1996) 1001–1013
3. Russell, R.: Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J. Mol. Biol. **279(5)** (1998) 1211–1227
4. Artymiuk, P.J., Poirrette, A.R., Grindley, H.M., Rice, D.W., Willett, P.: A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J. Mol. Biol. **243** (1994) 327–344
5. Moodie, S.L., Mitchell, J.B.O., Thornton, J.M.: Protein recognition of adenylate: An example of a fuzzy recognition template. J. Mol. Biol. **263** (1996) 486–500
6. Denessiouk, K.A., Rantanen, V., Johnson, M.: Adenine Recognition: A motif present in ATP-,CoA-,NAD-,NADP-, and FAD-dependent proteins. PROTEINS: Structure, Function and Genetics **44** (2001) 282–291
7. Shulman-Peleg, A., Nussinov, R., Wolfson, H.J.: Recognition of functional sites in protein structures. J. Mol. Biol. **339(3)** (2004) 607–633 `http://bioinfo3d.cs.tau.ac.il/SiteEngine/`.
8. Russell, R., Barton, G.: Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. PROTEINS: Structure, Function and Genetics **14** (1992) 309–323
9. Taylor, W.R., Flores, T., Orengo, C.: Multiple protein structure alignment. Protein Science **3** (1994) 1858–1870
10. Leibowitz, N., Nussinov, R., Wolfson, H.: MUSTA-a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. J Comput Biol. **8** (2001) 93–121
11. Shatsky, M., Nussinov, R., Wolfson, H.: A method for simultaneous alignment of multiple protein structures. Proteins: Structure, Function, and Genetics **56(1)** (2004) 143–156 `http://bioinfo3d.cs.tau.ac.il/MultiProt/`.
12. Dror, O., Benyamini, H., Nussinov, R., Wolfson, H.J.: MASS: multiple structural alignment by secondary structures. Bioinformatics **19 Suppl. 1** (2003) i95–i104 `http://bioinfo3d.cs.tau.ac.il/MASS`.
13. Lemmen, C., Lengauer, T.: Computational methods for the structural alignment of molecules. J. of Computer-Aided Mol. Design **14** (2000) 215–232
14. Dror, O., Shulman-Peleg, A., Nussinov, R., Wolfson, H.J.: Predicting molecular interactions in silico: I. A guide to pharmacophore identification and its applications for drug design. Curr. Med. Chem. **11** (2004) 71–90

15. Kuttner, Y.Y., Sobolev, V., Raskind, A., Edelman, M.: A consensus-binding structure for adenine at the atomic level permits searching for the ligand site in a wide spectrum of adenine-containing complexes. PROTEINS: Structure, Function and Genetics **52** (2003) 400–411
16. Kinoshita, K., Nakamura, H.: Identification of protein biochemical functions by similarity search using the molecular surface database ef-site. Protein Science **12** (2003) 1589–1595
17. Schmitt, S., Kuhn, D., Klebe, G.: A new method to detect related function among proteins independent of sequence or fold homology. J. Mol. Biol. **323** (2002) 387–406
18. Akutsu, T., Halldorson, M.M.: On the approximation of largest common subtrees and largest common point sets. Theoretical Computer Science **233** (2000) 33–50
19. Akutsu, T.: Protein structure alignment using dynamic programming and iterative improvement. IEICE Trans. Information and Systems **E79-D** (1996) 1629–1636
20. Efrat, A., Itai, A., Katz, M.J.: Geometry helps in bottleneck matching and related problems. Algorithmica **31** (2001) 1–28
21. Ambuhl, C., Chakraborty, S., Gartner, B.: Computing largest common point sets under approximate congruence. In: Proc. of the 8th Ann. European Symp. on Alg., Springer-Verlag (2000) 52–63
22. Huttenlocher, D., Ullman, S.: Recognizing solid objects by alignment with an image. International Journal of Computer Vision **5(2)** (1990) 195–212
23. Goodrich, M.T., Mitchell, J.S.B., Orletsky, M.W.: Practical methods for approximate geometric pattern matching under rigid motions: (preliminary version). In: Proc. of the 10th Ann. Symp. on Comp. Geom., ACM Press (1994) 103–112
24. Chakraborty, S., Biswas, S.: Approximation algorithms for 3-d commom substructure identification in drug and protein molecules. In: Proc. 6th Int. Workshop on Algorithms and Data Structures, Vancouver, Can., Springer-Verlag (1999) 253–264
25. Garey, M.R., Johnson, D.S.: Computers and Intractability. W. H. Freeman, San Francisco (1979)
26. Hazan, E., Safra, S., Schwartz, O.: On the Complexity of Approximating k-Dimensional Matching. In: Approximation, Randomization, and Combinatorial Optimization. Volume 2764 of LNCS., Springer (2003) 83–97
27. Wolfson, H.J.: Model-Based Object Recognition by Geometric Hashing. In: Proc. of the 1^{st} European Conf. on Comp. Vision (ECCV). LNCS, Springer-Verlang (1990) 526–536
28. Connolly, M.L.: Analytical molecular surface calculation. J. Appl. Cryst. **16** (1983) 548–558
29. Hurkens, C.A.J., Schrijver, A.: On the size of systems of sets every t of which have an sdr, with an application to the worst-case ratio of heuristics for packing problems. SIAM J. Discret. Math. **2** (1989) 68–72
30. Heffernan, P.J., Schirra, S.: Approximate decision algorithms for point set congruence. Comput. Geom. Theory Appl. **4** (1994) 137–156
31. Gavrilov, M., Indyk, P., Motwani, R., Venkatasubramanian, S.: Combinatorial and experimental methods for approximate point pattern matching. Algorithmica **38** (2004) 59–90
32. Mintz, S., Shulman-Peleg, A., Wolfson, H.J., Nussinov, R.: Generation and analysis of a protein-protein interface dataset with similar chemical and spatial patterns of interactions. (submitted) (2004)
33. Connolly, M.L.: Measurement of protein surfaces shape by solid angles. J. Mol. Graph. **4** (1986) 3–6