Journal of
CHEMICAL INFORMATION
AND MODELING

Article

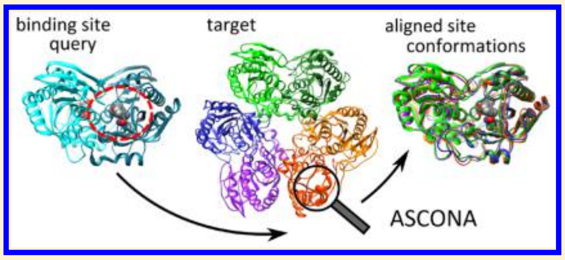
pubs.acs.org/jcim

ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations

Stefan Bietz and Matthias Rarey*

University of Hamburg, Center for Bioinformatics (ZBH), Bundesstrasse 43, 20146 Hamburg, Germany

S *Supporting Information*



ABSTRACT: The usage of conformational ensembles constitutes a widespread technique for the consideration of protein flexibility in computational biology. When experimental structures are applied for this purpose, alignment techniques are usually required in dealing with structural deviations and annotation inconsistencies. Moreover, many application scenarios focus on protein ligand binding sites. Here, we introduce our new alignment algorithm ASCONA that has been specially geared to the problem of aligning multiple conformations of sequentially similar binding sites. Intense efforts have been directed to an accurate detection of highly flexible backbone deviations, multiple binding site matches within a single structure, and a reliable, but at the same time highly efficient, search algorithm. In contrast, most available alignment methods rather target other issues, e.g., the global alignment of distantly related proteins that share structurally conserved regions. For conformational ensembles, this might not only result in an overhead of computation time but could also affect the achieved accuracy, especially for more complicated cases as highly flexible proteins. ASCONA was evaluated on a test set containing 1107 structures of 65 diverse proteins. In all cases, ASCONA was able to correctly align the binding site at an average alignment computation time of 4 ms per target. Furthermore, no false positive matches were observed when searching the same query sites in the structures of other proteins. ASCONA proved to cope with highly deviating backbone structures and to tolerate structural gaps and moderate mutation rates. ASCONA is available free of charge for academic use at http://www.zbh.uni-hamburg.de/ascona.

■ INTRODUCTION

Structural variability of proteins is still one of the most challenging phenomena of structural molecular biology. Although only rudimentarily understood, protein flexibility has a high impact on enzymatic mechanisms and protein function in general. In practice, often different snapshots of proteins are used to investigate their structural variability or to represent their conformational space in computational simulations. For instance, the use of multiple protein conformations (ensemble docking) is capable of improving the performance of rigid-protein docking approaches.[1−5] Other applications of protein ensembles can be found in pharmacophore generation,[6,7] hot spot analysis within flexible binding regions,[8] *de novo* ligand design,[9,10] or ligand-induced side chain flexibility analysis.[11−13] Usually, these scenarios require a structural superimposition or at least a residue-wise alignment of the protein data. Even if this is a trivial task in theory, since the employed structures usually describe exactly the same protein, inconsistent data annotation and structural artifacts like residue gaps or missing peptide chains complicate this problem (cf. Figure 1a). In principle, a valid residue mapping can be achieved by global sequence- or structure-based alignment techniques (see Fariselli et al.,[14] Kalaimathy et al.[15] for reviews). Therefore, a common strategy is the transfer of the active site definition on the basis of a previously generated alignment.[16−20] Nevertheless, this offers various opportunities for optimization. First, most workflows only result in one single alignment for each pair of structures. Especially structures of oligomeric proteins often contain duplicated subunits and therefore also various alignment solutions at equal quality. A consideration of alternative alignments might enrich the variety of structural information. In the case of a symmetrical orientation of equivalent subunits, it can further facilitate the investigation of a protein's internal symmetry (cf. Figure 1b). Second, most alignment procedures, in particular structure-based methods, are designed for the more complicated challenge of aligning different proteins even with low sequence similarity. In turn, this means computational overhead in the case of almost identical structures. Third, in many application scenarios of protein ensembles, it would be sufficient to obtain an alignment of the active site residues only, since the rest of the protein structure is often neglected in downstream calculations. The generation of global alignments therefore consumes an avoidable amount of computation time for the generation of mostly unexploited data. Moreover, the consideration of superfluous data carries a higher risk of producing erroneous results. Accordingly, various methods have been developed that focus on the structural alignment, the prediction, or the comparison of protein binding sites (see Kellenberger et al.,[21] Pérot et al.,[22] and Nisius et al.[23] for reviews). As in the case of global analysis, these procedures

Received: April 15, 2015
Published: June 22, 2015

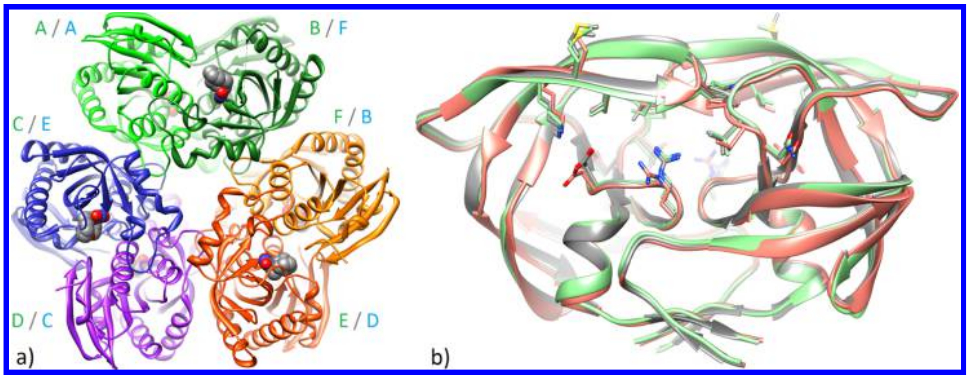


Figure 1. (a) Homohexameric structure of uridine phosphorylase. The capital letters indicate the inconsistent chain annotation in the PDB structures 1U1C (green letters) and 1SJ9 (blue letters). This inconsistency impedes the transfer of binding site information as each site is formed by two neighboring subunits. (b) Alternative alignments of the symmetrical binding site of HIV protease. The reference PDB structure 3VFB is depcited in gray. The intuitive alignment of chain 3VFB A with chain 1MSM A and 3VFB B with 1MSM B (red) results in a worse superposition than the alignment of 3VFB A with 1MSM B and 3VFB B with 1MSM A (green).

generally aim at the comparison of rather distantly related proteins. The detection of different binding site conformations of the same target is only rarely discussed. Schmidtke et al.[24] describe an automated application of fpocket, a shape analyzing pocket detection approach[25] based on α-sphere theory, to pocket tracking on homologous PDB structures and protein ensembles from molecular dynamics. Similarly, Craig et al.[26] use the grid-based PocketAnalyser in combination with Principal Component Analysis for the selection of geometrically diverse pockets from an ensemble of protein structures. Surprisingly, we did not find a method that is specifically optimized for an alignment of active site conformations or closely related binding sites.

Here, we present ASCONA (Active Site CONformation Aligner), a novel approach specially geared to the detection of arbitrarily defined binding sites in multiple protein conformations. ASCONA makes use of a widespread alignment technique that is composed of a fragmentation of the query, followed by a separate search for every fragment and eventually an assembling of single fragment matches for the reconstruction of entire query matches. Variations of this concept have been successfully applied in various structure based alignment scenarios with different similarity measures and optimization procedures used for fragment comparison and assembling.[27−32] We adopt this technique to the alignment of protein binding site conformations by the integration of a fast sequence alignment of short peptide fragments, a reduced but computationally cheap geometric similarity measure, and an efficient assignment procedure for the recombination of single binding site fragments. The combination of these features results in a rapid and highly accurate search engine that is able to deal with large backbone variation and moderate mutation rates.

METHODS

The search strategy implemented in ASCONA consists of three consecutive steps. First, the active site of the query protein is being transformed into a set of connected residue fragments. The second step searches for occurrences of these fragments in the set of target structures using a sequence alignment procedure. Finally, the relative geometric orientation of all identified fragments is analyzed and the target binding sites are reconstructed by a fragment assembly approach. If required, ASCONA uses the resulting active site alignments for a superposition of the target onto its reference structure. The following sections illustrate these steps in detail.

Sequence Fragment Generation. First of all, the query binding site is extracted from the reference protein structure as a set of residues which may result from a binding site prediction algorithm or from a geometric selection around a bound ligand. Considering backbone peptide bonds, this set can be divided into linearly connected residue fragments (see Figure 2a). The length of a fragment is of great importance for the sequence alignment step, as the probability of generating a random incorrect match decreases rapidly with an increasing number of residues in a fragment. Therefore, fragments being shorter than a predefined threshold are symmetrically extended by additional adjacent residues in N- and C-terminal directions. The additional residues are only used in the sequence alignment step and will be discarded afterward.

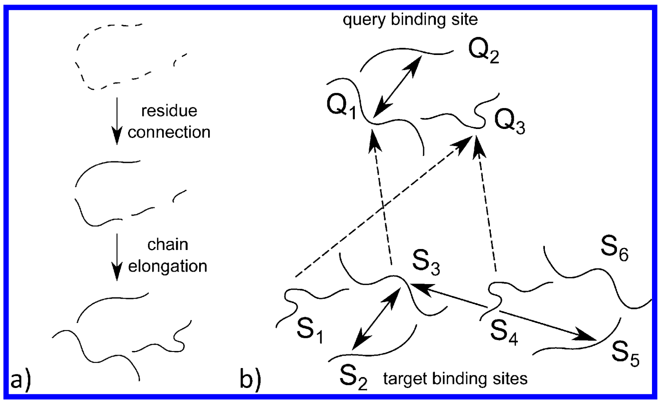


Figure 2. (a) Binding site transformation from single residues to augmented residue fragments. (b) Schematic illustration of interfragment distances (solid arrows) and translation vectors resulting from fragment superposition (dashed arrows).

Sequence Alignment. In the second step, all query binding site fragments are searched in the target protein sequence with a slightly modified version of an approximate string matching algorithm by Ukkonen.[33] The same variation has also been described by Galil and Park[34] as a preliminary basis of their solution to the approximate pattern matching problem and is therefore only briefly summarized here.

Its basic concept is the calculation of the Levenshtein distance[35] between two strings of length m and n (with $m \leq n$)

by a dynamic programming algorithm that fills a distance matrix **D**, the elements of which represent pairwise substring distances. The underlying recurrence ensures that d_{mn} eventually holds the Levenshtein distance of both complete strings. A string matching, or in our case a sequence alignment, can be obtained by storing minimizing edit operations and following a path from the element d_{mn} back to d_{00}. In contrast to earlier algorithms, Ukkonen's procedure calculates the matrix elements on diagonals resulting in a worst case runtime of $O(kn)$ where k denotes the Levenshtein distance between the strings. Galil and Park's variation allows for the detection of all local instead of all global matchings by evaluating $n - m + 2k + 1$ diagonals in $O(mn)$ time (see Figure 3). In addition to this modification, we use an early abort criterion: When a first valid solution has been identified, our variation terminates as soon as all equivalent solutions have been found. Although other solutions to this problem with a lower worst case runtime have been published ($O(kn)$ Ukkonen,[36] $O(kn)$ Galil and Park[34]), this approach still seems to be reasonable since only a low number of expensive diagonal calculations is expected. Furthermore, random matches are rather unlikely due to the relatively large amino acid alphabet and the low k values usually applied.

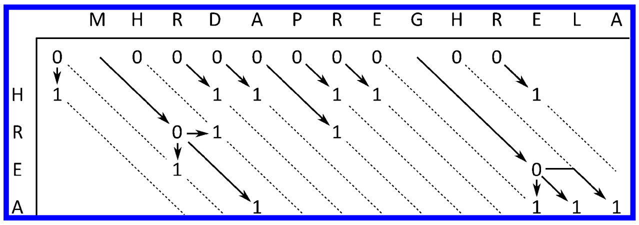


Figure 3. Schematic depiction of distance matrix entries calculated by the variation of the Ukkonen algorithm. The character strings represent a binding site fragment (vertical) and a target protein chain (horizontal). The arrows indicate from which entries in the previous iteration the matrix entries were derived. The entries in the last row indicate the Levenshtein distance as well as the end points of successful string matchings. In contrast to preceding algorithms, less matrix entries need to be calculated to obtain all optimal solutions. For algorithmic details, please see Ukkonen[33] and Galil and Park.[34]

Fragment Accumulation. If there is at least one equivalent of the query binding site present in a target structure, the sequence alignment step results in a set of matches where each match M_x is represented by a fragment pair (Q_x, S_x) with Q_x being a fragment in the query and S_x its matched fragment in the target structure. In the easiest case, every query fragment has exactly one match, and therefore only one combination of the target fragments exists. Provided that no false positive fragment matches occur, this combination directly represents the target binding site. However, if random fragment matches occur, these cannot be identified without considering the relative spatial orientation of the matched fragments. Furthermore, each query fragment might also have multiple matches since proteins often form oligomeric structures. We will refer to a set of homologous fragments which are mapped to the same query fragment as a *fragment class*. In such cases, a procedure for reconstructing the binding site from single target fragments is required. If the whole site belongs to one and the same subunit, this could be in principle solved by a simple matching of chain names or a Depth First Search if the protein is represented by a molecular graph. Nonetheless, this would fail in cases where the binding sites are formed by multiple subunits. An even more complex scenario appears at symmetrical interfaces, e.g., in homodimers, as in these cases certain fragments might occur multiple times in the same binding site. Since all these scenarios can be found in prominent target proteins, a structure-based assignment procedure has been developed.

In general, the choice of an appropriate geometric measure and a suitable assignment strategy is highly dependent on the accuracy of the fragment matching technique. Since the intended application scenario of ASCONA implies a high sequence similarity of the query and the target structures, the sequence alignment parametrization can be set up quite strictly. Thus, we can assume a low rate of random fragment matches and therefore apply a relatively cheap geometry measure. We use a combination of two structural descriptors that measure the probability of two fragments belonging to the same binding site. First, a potential superposition for each fragment S_x onto its reference Q_x from the query binding site is calculated on the basis of the backbone coordinates and represented in the form of a rotation quaternion and a translation vector. Instead of actually applying the superposition, the Euclidean distance of two rotation quaternions δ_{quat} is then used in a first distance measure δ_{rot}:

$$\delta_{\text{quat}}(q_i, q_j) = \sqrt{\sum_{n=1}^{4} (q_i(n) - q_j(n))^2}$$

$$\delta_{\text{rot}}(M_i, M_j) = \min(\delta_{\text{quat}}(q_i, q_j), \delta_{\text{quat}}(q_i, -q_j))$$

where q_x denotes the rotation quaternion of the superposition of S_x onto Q_x and n represent its indices. The minimization considers the fact that q_x and $-q_x$ represent the same spatial rotation. Note that the translation is not considered in this measure, as neither the lengths nor the directions of the translation vectors for the fragments of a distinct binding site necessarily exhibit sufficient similarities. Figure 2b illustrates an exemplary case where the translation vector of fragment S_3 is obviously more similar to that of S_4 than to its actually associated fragment S_1. Therefore, an alternative term (δ_{dist}) is applied describing the normalized difference of interfragment distances within the query and target sites:

$$\delta_{\text{dist}}(M_i, M_j) = \begin{cases} \infty, & \text{if } Q_i = Q_j \\ \left| \dfrac{d_{\text{abs}}(Q_i, Q_j) - d_{\text{abs}}(S_i, S_j)}{d_{\text{abs}}(Q_i, Q_j)} \right|, & \text{otherwise} \end{cases}$$

where d_{abs} denotes the absolute geometric distance of the backbone atom location centroids of two fragments. The distance measure for a pair of fragment matches is eventually given by

$$\delta(M_i, M_j) = \delta_{\text{rot}}(M_i, M_j) + \delta_{\text{dist}}(M_i, M_j)$$

Both terms are expected to be close to zero in the case that S_i and S_j belong to the same binding site, even if its backbone conformation deviates substantially from the query structure. If S_i and S_j belong to different sites, both values are usually much higher, allowing a clear differentiation in most cases. We combine both terms in order to accurately handle more complicated circumstances: If the target structure consists of several identical subunits which are equally oriented, the resulting rotations will hardly be discriminable. However, the δ_{dist} values should still allow for a correct assignment. In the case of symmetrical binding sites in homo-oligomeric protein complexes, every query fragment might have several matches within one target binding site which cannot be clearly distinguished by their relative distance, but generally by deviating rotations. A coincidental failing of both measures leading to an incorrect assignment is, at least for realistic data and sensible parameter settings, extremely unlikely.

The assignment procedure starts with an estimation of the number of matching binding sites in the target structure. In the ideal case, every fragment class has the same size, and it is therefore likely that this size corresponds to the number of binding sites. However, sequence alignment artifacts like missing fragments or additional incorrect matches may result in differently sized fragment classes. In this case, the number of sites can be estimated by analyzing how often the individual fragments are matched in the target structure (fragment frequency), how many residues reach a certain frequency (covered residues), and how the portion of covered residues decreases with increasing fragment frequencies. For the hypothetical example given in Figure 4, 100% of the binding site residues are matched at least twice. In other words, a binding site coverage of 1.0 is reached for fragment frequencies up to a value of two. At a fragment frequency of 3, still 87% of the binding site are covered. Only from a fragment frequency of 4 does the portion of covered residues fall below the applied example threshold of 0.8. This shows that the minimum portion of covered residues that is required to reconstruct a target binding site having a certain minimum site identity limits the highest possible fragment frequency and therefore the expected amount of matching binding sites. In the example, at most three binding sites could fulfill the desired identity. If the resulting estimation does not coincide with the cardinality of any fragment class, it is decreased accordingly.

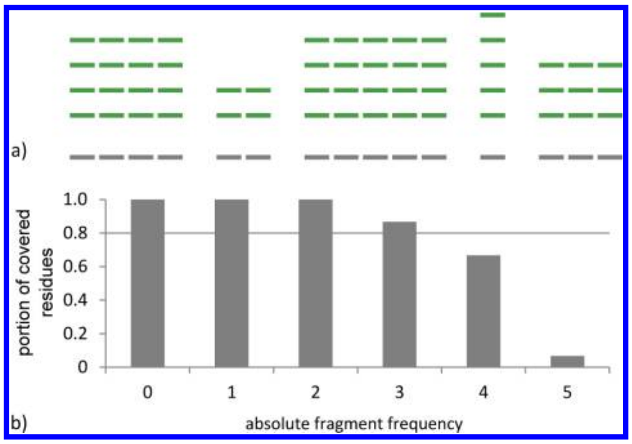


Figure 4. Approximation of the expected number of binding sites in a target structure. (a) Hypothetic example of sequence alignment results. Fragments of the query and the target binding sites are depicted in gray and green, respectively. All fragments are illustrated as dashed lines, indicating their number of residues within the active site. (b) Portion of covered binding site residues as a function of absolute fragment frequency in the target structure. The values correspond to the hypothetical example given in part (a). The line indicates the minimum residue portion that is required to reconstruct a target binding site having a minimum site identity of 0.8.

In the next step, a single fragment class is selected for which the number of fragments is equal to the estimated number of binding sites. The fragments of this class are then considered as seeds for reconstructing the target binding sites. For each remaining fragment class, the fragments are assigned to the seeds as follows: First, for each class member, the δ distance to any of the seed fragments is calculated and a sorted list of possible single fragment assignments is created (see Figure 5a). Note that the length of this list therefore equals the class size times the number of seed fragments. Then, each list is processed by assigning its elements to the seed fragments in ascending distance order. Once a fragment is assigned, all its other elements are removed from the list in order to prevent multiple assignments of the same fragment. This procedure is repeated until the list is empty or every seed has received a fragment of the current class.

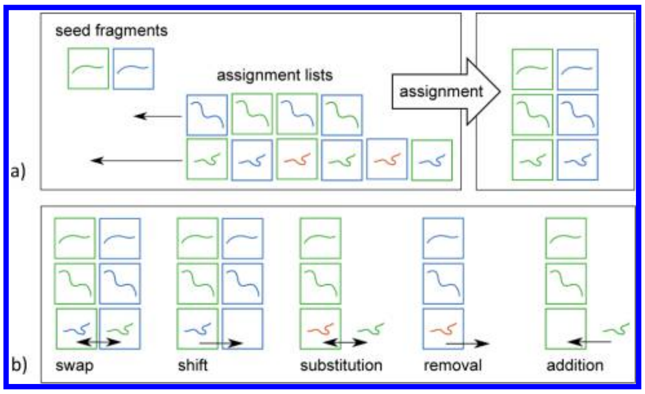


Figure 5. Fragment assembling. (a) Initial assignment procedure. The fragment color encodes a fragment's actual binding site, while its frame color indicate the seed fragment to which it would be assigned. (b) Assignment optimization operations.

Although this basic assignment procedure results in correct solutions in most cases, exceptions to this rule can still occur due to a efficiency-related shortcoming of the initial assignment procedure, which only evaluates the distance values of a fragment to all seeds but neglects the geometrical relations of all other fragment pairs. Therefore, a post-optimization procedure has been developed that aims at the minimization of all mean fragment δ distances within the binding sites. The assignment score $t(A)$ for an assignment A simply sums up the mean fragment δ distances of all fragments S_i in all binding sites B_k and adds a penalty p for each of all w nonassigned fragments:

$$t(A) = \sum_{B_k \in A} t(B_k) + \sum_{i=1}^{w} p$$

$$t(B_k) = \sum_{i=1}^{n_k} t(S_i, B_k \backslash S_i)$$

$$t(S_i, B_k) = \frac{\sum_{S_j \in B_k} d(S_i, S_j)}{n_k}$$

S_i denotes the ith fragment in the binding site B_k; n_k is its number of fragments. Furthermore, we make use of five different change operations (cf. Figure 5b) during the optimization process:

(a) Swapping two fragments S_i and S_j that are part of the same fragment class but assigned to different binding sites $B_{b(i)}$ and $B_{b(j)}$

(b) Shifting a fragment S_i from binding site $B_{b(i)}$ to another site B_k that does not yet contain a fragment from the corresponding fragment class of S_i

(c) Substituting an assigned fragment S_i by a nonassigned competitor S_p from the same fragment class

(d) Removing an assigned fragment S_i from its binding site $B_{b(i)}$

(e) Adding a nonassigned fragment S_p to a binding site b_k

During each iteration, the algorithm evaluates all possible change operations across all fragments and finally applies the one which leads to an maximum decrease of $t(A)$ (see the Math section for calculation details). This procedure is repeated until either no operation yields a further improvement or a maximum number of iterations is reached.

■ RESULTS AND DISCUSSION

The Astex Non-Native set[17] was used to investigate the ability of ASCONA to find correct active site alignments. This data set contains 1112 protein−ligand complex structures on the basis of 65 diverse target proteins taken from the Astex Diverse set.[37] All structures are sequence identical to their reference within 6.0 Å around the Astex reference ligand. This guarantees the existence of at least one valid active site alignment for each of the 1112 structures with their corresponding references from the Astex Diverse set. Furthermore, all structures provided by the Astex Non-Native set are superimposed onto the reference. In order to avoid a bias and to create a more realistic test scenario, we downloaded all original structure files from the PDB. Five structures were no longer available (PDB codes: 2C5P, 2NMW, 1KYE, 2H6W, 2CDD; download 12/9/2014).

Sensitivity. For the first analysis, the remaining 1107 Astex Non-Native structures were aligned to their reference active sites (queries) under variation of the minimal fragment length (l_f) and the maximal fragment Levenshtein distance (k_f). The minimum site identity (i_s) was fixed to 1.0. Figure 6 shows that a success rate of more than 99% can be achieved for most variable combinations. Moreover, several different combinations result in a successful alignment for all 1107 structures. Interestingly, there is a clear correlation between both variables. Obviously, a success rate of 100% requires k_f to be approximately half the size of l_f. A more detailed investigation of those cases for which the alignment procedure fails with deviating settings exhibited that this correlation results from gaps in the protein chain that occur in the direct neighborhood of active site residues. In such cases, half of the fragment alignment needs to consist of mismatches or deletions.

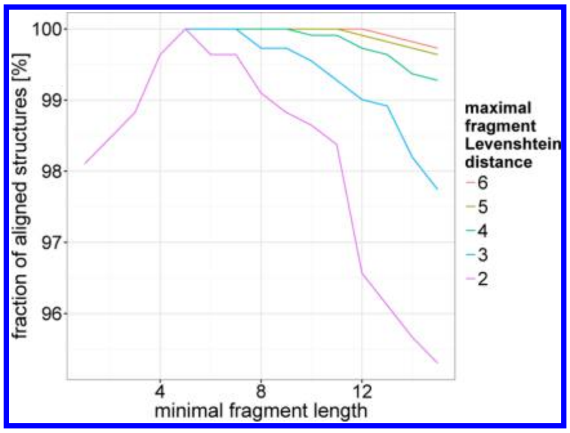


Figure 6. Fraction of successfully aligned Astex Non-Native structures as a function of l_f and k_f. See Figure S1 for additional data series. Fixed parameter: i_s = 1.0.

For 36% of the target structures, more than one match of the query active site from the Astex reference structure were identified. These include both multiple nonintersecting alignments, for instance on different protein chains, and duplicated but permuted matches of symmetrical binding sites. While one could easily avoid the generation of symmetric duplicates by filtering highly intersecting alignments, their analysis could facilitate interesting insights into the mutual interference of symmetrically oriented domains in homo-oligomeric proteins. In total, we found 1562 valid alignments. In this case, it is admittedly not directly possible to define the maximum number of valid alignments for the calculation of success rates, as this would require either absolute knowledge about the protein's quaternary structure and all structural artifacts like residue gaps or at least a highly reliable reference method. Nevertheless, we do not expect additional insights with respect to the sensitivity of our method, as the absolute number of alignments (see Figure S2) show very similar trends compared to the relative number of matched structures (cf. Figure 6 and Figure S1).

Specificity. We also checked all query binding sites against the Astex Non-Native structures of all other proteins. As expected, ASCONA does not detect alignments at sensible parameter settings. Mistakenly detected alignments are only observed at unrealistically high mismatching rates (see Figure S3).

Alignment Accuracy. By definition, our algorithm guarantees that any site identified has the same amino acid sequence as the query if i_s = 1.0. Still, this does not ensure that all residues are correctly matched. We first analyzed a subset of the Astex Non-Native set for which we assume that at least one alignment can be derived, and thus also verified, from the residue annotation in the PDB file. A non-native structure is part of this set if it contains a residue with corresponding annotation for each amino acid from the query active site in terms of amino acid type, sequence id, insertion code, and chain name. Note that our alignment procedure does not make use of the latter three values. The resulting subset, to which we will refer as the VPA set (Verification by PDB Annotation), contains 843 structures, which correspond to 76% of the Astex Non-Native structures and 54% of all identified active sites. By defining an alignment to be erroneous if at least one aligned residue deviates from its reference, errors only occur for parameter settings with unreasonably low fragment length (see Figure S4). For most parameter settings, only two differences between PDB annotation and the calculated alignment were identified. Interestingly, both always involve the same non-native structure (2FPZ). Like its reference (2BM2), it describes a symmetrical homotetrameric complex of human β II tryptase but differs in its chain name annotation. A similar case is depicted in Figure 1a. Visual inspection revealed that the respective calculated alignments are correct, while the PDB annotation indicates an incorrect assignment and cannot be used for verification in this special case. Thus, the alignment procedure proposes no single erroneous alignment for the VPA set when a sensible minimal fragment length is used.

For evaluating the correctness of the remaining alignments (Non-VPA set), we considered various options. In principle, one could apply geometric measures like RMSD values or Distance Matrix Errors (DME) to validate the correctness of the proposed alignment. However, geometric measures presuppose the definition of a threshold for which a trade-off between error detection accuracy and protein flexibility consideration has to be made. We would like to demonstrate this on the basis of a DME-related measure, to which we will refer as the Maximal Mean Distance Deviation (MMDD):

$$\mathrm{MMDD}(C, A) = \max_{c_i \in C}\left(\frac{\sum_{c_j \in C \setminus c_i} |d(c_i, c_j) - d(A(c_i), A(c_j))|}{|C| - 1}\right)$$

where A represents an alignment, C the set of covered residues in the query binding site, c_i the ith residue in C, $A(c_i)$ the residue from the target structure that is aligned to c_i, and d the spatial distance between the α-carbons of two residues. Compared to the DME, which measures the average deviation over all pairwise residue distances, this value seems more suitable in the given scenario, as our algorithm has to deal with single fragment mismatches rather than with a completely false active site assignment.

Figure 7 shows the distribution of MMDD values observed in the VPA set, subdivided by the previously introduced error classification. If l_f = 1 (Figure 7a), most of the 26 resulting erroneous alignments indeed exhibit increased MMDD values. However, there is a considerable amount of correctly aligned sites with likewise MMDD values for which visual inspection indeed revealed correspondingly large backbone shifts. An example is given in Figure 8. Equivalent distributions can be observed for higher l_f values, with the exception of a lower error rate and more correctly aligned structures with high MMDD values (Figure 7b). This complicates the definition of a sensible threshold if the MMDD values should be used for error detection. If the loss of correct alignments is not acceptable, geometry-based error detection seems to be only reasonable when the flexibility of the target under consideration is known to be low. The distribution of MMDD values within the Non-VPA set, which is similar to that of the VPA set (see Figure S7), suggests that this is not the case for all structures.

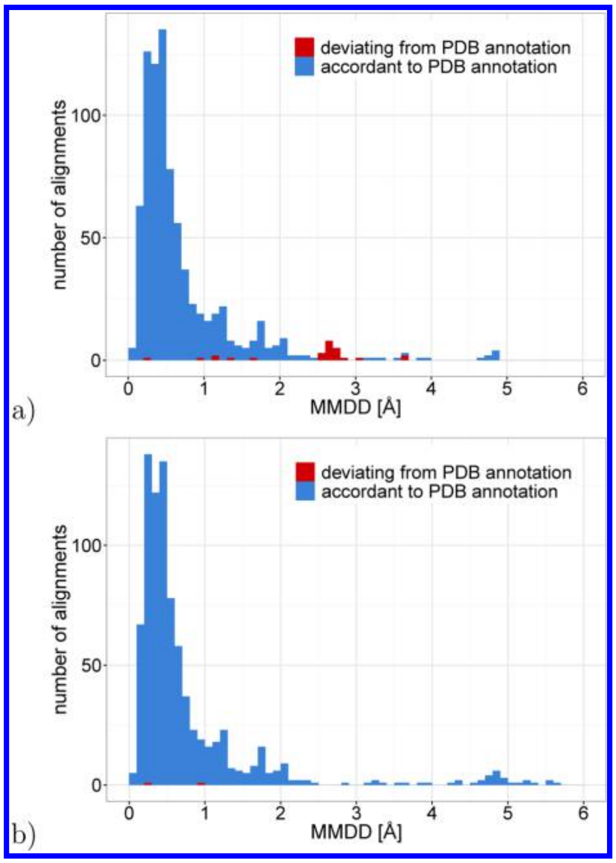


Figure 7. MMDD distribution over all alignments detected in the VPA set. Applied parameters: (a) i_s = 1.0, l_f = 1, k_f = 1; (b) i_s = 1.0, l_f = 7, k_f = 3.

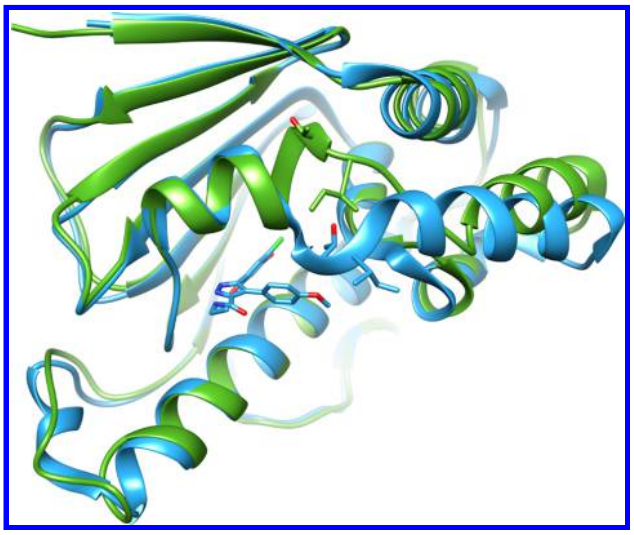


Figure 8. Superimposed heat shock protein 90 structures 2BSM (blue) and 1BYQ (green) exhibiting a MMDD value of 5.3 Å.

Alternatively to the use of a geometric measure, a combination of automatic annotation-based analysis and visual inspection for the classification of the Non-VPA set was used. For that, an alignment was declared to be conclusive if all chain-internal sequence number intervals in the aligned site correspond to their counterpart in the query. The remaining alignments were evaluated by visual inspection. As Figure S7b shows, we did not find any further erroneous alignment in the case of a sensible parametrization, while the error rate resulting from l_f = 1 (Figure S7a) is comparable to that of the VPA set.

Mutation Analysis. ASCONA is also suited to detect sites with lower sequence identity. In order to analyze its capabilities and limitations in this scenario, we selected PDB structures of four diverse proteins for which we expect different mutation rates. Proteins were specified by EC numbers (see Table S1 for details), and sets of sequentially distinct structures were selected by the 100% (global) sequence identity filter as implemented at the PDB Web site.[38] The first set (DHFR) contains 12 structures of the human dihydrofolat reductase. Set two (HIV) comprises 167 structures of HIV-1 protease. For a third set (CA), 129 carbonic anhydrase structures from mammals were selected. Finally, a set of 447 distinct human protein−tyrosine kinases compose the fourth test set (Kinases). For each data set, a query site was selected as follows: For the DHFR and the CA set, we used MONA[39] to select that structure which contains the ligand with the highest molecular weight from those fulfilling Lipinski's Rule of Five.[40] For the HIV set, we solely applied the highest molecular weight criterion as only a minor portion of all ligands matches the Rule of Five. In the case of the kinase set, the lexicographically first structure in complex with ATP was selected. For all data sets, the residues within a radius of 6.0 Å around the ligand define the active site. Further details including a list of all PDB codes can be found in Tables S2 and S3.

The resulting alignments for all four data sets were analyzed in terms of sequence identity and alignment coverage. While the identity represents the relative amount of residue matches within the active site, the coverage is defined as the combined frequency of residue matches and replacements or, in other words, as the inverse of the gap rate. Figure 9 shows the coverage against the sequence identity for all structures contained in the mutation data sets.

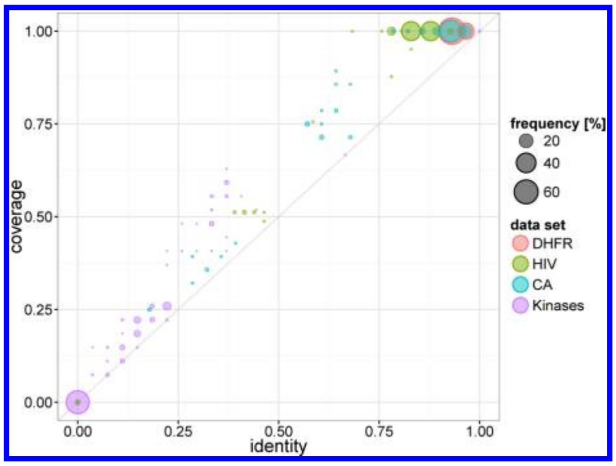


Figure 9. Alignment coverage against identity for all structures from four different mutation data sets. If more than one alignment was detected for a certain structure, only the one with the highest identity was used in the plot. Applied parameters: $i_s = 0.0$, $l_f = 10$, $k_f = 4$.

The results demonstrate that ASCONA is capable of finding alignments at various degrees of mutations within the binding site. For the DHFR set, it can detect an alignment for all structures at a coverage rate of 1.0 and an identity above 0.9. In the case of the CA and HIV set, the mutation rates show a wider range compared to the DHFR set. An ideal coverage can be observed for 81% of the CA and 90% of the HIV structures. Some of the remaining structures can only be aligned at very low coverage and identity levels. Therefore, alignments with a coverage below 1.0 were investigated in more detail. Overall, we found three reasons for these outliers. First, some of the HIV structures only contain one subunit of the homodimer, which usually forms the binding site. Hence, the corresponding alignments are limited to lower sequence identity and coverage values in these cases. Second, some of the selected PDB entries in the HIV set contain wrong EC numbers and actually encode different HIV proteins. In case of the CA set, all outlying structures describe different isoforms compared to the reference, which is a representative of carbonic anhydrases II, the most common isoform in the set. The fact that different isoforms can exhibit substantial sequence deviations is obviously the reason for the lower alignment quality in these cases. Third, we identified in total three alignments for which the lower coverage results from minor shortcomings of ASCONA. In two of these cases, the target structures describe chemically synthesized HIV proteases (PDB codes 4EQ0 and 3NXE) in which the two subunits contain different numbers of mutations and are further connected by a short linker fragment. The combination of these circumstances leads to a scenario where ASCONA can only detect one of two possible matches for the fragment that covers the differently mutated area within the symmetrical binding site. This is a consequence of the early abort criterion as the sequence alignment only detects matches of equal quality within a certain peptide chain. The third case is also related to a synthetically produced HIV protease (PDB code 3NWQ) which contains an α-hydroxyl acid instead of an amino acid. Although ASCONA is able to recognize and align nonproteinogenic amino acids, other types of chemical peptide linkers are not considered and therefore induce alignment gaps. Nevertheless, both scenarios are caused by artificial conditions and are not expected for natural protein structures. A complete list of all structures from the HIV and the CA set with suboptimal alignment coverage including a case classification is given in Table S4.

The kinase set exhibits, compared to the other sets, a significantly lower success rate (the number of fails is indicated by the circle at zero coverage) and also considerably reduced coverage and identity levels. Although the coverage is in most cases distinctly higher than the identity, the utility of those alignments is limited. Even if they might still be useful for application scenarios like site superposition, a detailed residue-wise analysis is no longer possible at such low covering rates. However, we would like to emphasize that the challenge of aligning structures of different proteins, as is the case with the kinase set, is beyond the intended purpose of ASCONA.

Besides the required change of the minimal site identity parameter i_s, the fragment parameters that were applied to generate the alignments for the mutation data sets also deviate from the default setting for aligning sequentially identical conformations (mutation settings: $l_f = 10$, $k_f = 4$; conformation settings: $l_f = 7$, $k_f = 3$). The parameter change is intended to compensate for the increased risk of missing and random fragment matches for mutated structures. Applying the conformation settings leads to a decreased ideal coverage rate for the HIV set (85%, cf. Figure S8 and Table S6). No coverage differences were observed for the DHFR set. For the CA set, deviations only occur for the mentioned outliers. In the case of the kinase set, the number of alignments and their average coverage is increased, but at the cost of a higher mismatching risk. Furthermore, randomly inspected test cases confirmed the expected trends that decreased l_f and increased k_f parameters as well as lower sequence identity involve a higher risk of sporadic mismatches and avoidable gaps.

Computation Times. In order to analyze the runtime behavior, we measured the computation time of the alignment procedure for all experiments. Computation times for file IO, parsing, and the initialization of the underlying protein data structure were separately analyzed since these steps are only performed once for all different parameter settings. All experiments were executed on a single core of an Intel Core i7-4790 CPU with 3.6 GHz and 32 GB of memory.

Figure 10 shows the average alignment runtimes on the Astex Non-Native set. For most cases, the runtime predominantly depends on k_f and just to a very limited extent on l_f. This is in accordance with the expected runtime behavior of the sequence alignment step, which apparently dominates the runtime in these cases. Furthermore, we observed a significantly increased number of postoptimization steps for $l_f = 1$, which explains the higher computation times of the corresponding test runs. In addition, the runtime analysis for the mutation data sets (see Figure S10) reflects the expected dependency on the total number of residues (cf. Table S2). Compared to a successful identification of equivalent binding sites, the screening of all Astex Non-Native structures against the query binding sites of

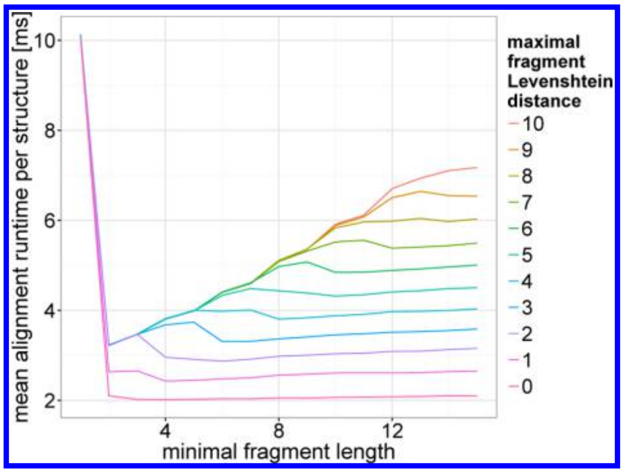


Figure 10. Average alignment runtime for an Astex Non-Native structure as a function of l_f and k_f. Applied parameters: i_s = 1.0.

all other proteins requires considerably higher computation times (see Figure S9). The time differences grow with increasing k_f. Most likely, this is a result of the early abort criterion of the sequence alignment which solely affects the former case. However, reasonable parameter settings still result in very low computation times, e.g. a mean of 8 ms for $l_f = 7$ and $k_f = 3$. The preprocessing steps required an average runtime of 27 ms per structure for the Astex Non-Native set and 28 ms for the mutation data sets.

Application Range. A typical application scenario of ASCONA is the consideration of protein flexibility in molecular docking approaches. In general, ASCONA could make valuable contributions to all phases of a docking workflow. During the preprocessing, its low runtime facilitates an efficient extraction of appropriate alternative active site conformations from large protein structure data sets. If the resulting ensemble should be reduced to a set of representative active site conformations, the necessary structure comparison is commonly based on descriptors like RMSD[41,42] or distance deviations[43] which depend on an accurate residue-wise mapping of the active sites. On the basis of ASCONA, we are currently working on a Web service for the automated compilation of PDB structure ensembles satisfying a user's case specific requirements.

The actual docking process may benefit from ASCONA by applying the alignments for the definition of the active site across all protein structures, either by directly transferring a residue-based site definition or by a sequence alignment-based structure superposition. The active site ensembles can be directly used for ensemble docking or for the construction of a more elaborate flexible protein model. In the postprocessing phase, alignments are required for pose clustering and result visualization as these rely on a proper site superposition. The residue mapping provided by ASCONA could, e.g., also be applied for pose comparison on the basis of molecular interaction fingerprints.

Besides molecular docking, ASCONA can also be used in other ensemble-based applications like protein flexibility assessment, flexible pharmacophore generation, or *de novo* ligand design.

■ CONCLUSIONS

ASCONA is a new algorithm for the detection and alignment of protein−ligand binding sites in structures with high sequence similarity. Various experiments on large data sets highlight its high accuracy and very efficient performance. Moreover, ASCONA is able to correctly align binding sites with high backbone flexibility, a challenging problem for purely structure-based applications. In contrast to many global sequence alignment techniques, it facilitates the detection of multiple binding sites in oligomers and binding site symmetry, e.g., in homodimers. The combination of these features turns ASCONA into a perfectly suited tool for fully automated alignment of protein active site ensembles, filtering wrongly annotated structures or searching new binding site conformations in large selections of protein structures.

■ MATH

During the assignment optimization, a change of $t(A)_x$ upon any operation x can be efficiently calculated on the basis of the following equations:

$$\begin{aligned}\Delta t(A)_{\mathrm{swap}(S_i,S_j)} &= 2(t(S_i, B_{b(j)}\backslash S_j) + t(S_j, B_{b(i)}\backslash S_i)\\ &\qquad - t(S_i, B_{b(i)}\backslash S_i) - t(S_j, B_{b(j)}\backslash S_j))\end{aligned}$$

$$\begin{aligned}\Delta t(A)_{\mathrm{shift}(S_i,B_k)} &= \Delta t(B_{b(i)}) + \Delta t(B_k)\\ &= \left(\frac{(t(B_{b(i)}) - 2t(S_i, B_{b(i)}\backslash S_i))(n_{b(i)} - 1)}{n_{b(i)} - 2} - t(B_{b(i)})\right)\\ &\quad + \left(\frac{t(B_k)(n_k - 1)}{n_k} + 2t(S_i, B_k) - t(B_k)\right)\\ &= \frac{t(B_{b(i)}) - 2t(S_i, B_{b(i)}\backslash S_i)(n_{b(i)} - 1)}{n_{b(i)} - 2}\\ &\quad + 2t(S_i, B_k) - \frac{t(B_k)}{n_k}\end{aligned}$$

$$\Delta t(A)_{\mathrm{substitution}(S_i,S_p)} = 2(t(S_p, B_{b(i)}\backslash S_i) - t(S_i, B_{b(i)}\backslash S_i))$$

$$\Delta t(A)_{\mathrm{removal}(S_i)} = \frac{t(B_{b(i)}) - 2t(S_i, B_{b(i)}\backslash S_i)(n_{b(i)} - 1)}{n_{b(i)} - 2} + p$$

$$\Delta t(A)_{\mathrm{addition}(S_p,B_k)} = 2t(S_p, B_k) - \frac{t(B_k)}{n_k} - p$$

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information contains additional alignment analyses, details on the applied data sets, and a detailed case classification concerning all suboptimal coverage alignment observed in the mutation experiments. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00210. The ASCONA software is available for Linux OS via the Internet at http://www.zbh.uni-hamburg.de/ascona.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financially supported by the BMWI-ZIM Project KF2563701. Molecular graphics were generated with the UCSF Chimera package.[44]

REFERENCES

(1) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular Docking to Ensembles of Protein Structures. *J. Mol. Biol.* **1997**, *266*, 424−440.

(2) Huang, S.-Y.; Zou, X. Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins: Struct., Funct., Genet.* **2007**, *66*, 399−421.

(3) Rao, S.; Sanschagrin, P. C.; Greenwood, J. R.; Repasky, M. P.; Sherman, W.; Farid, R. Improving Database Enrichment Through Ensemble Docking. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 621−627.

(4) Craig, I. R.; Essex, J. W.; Spiegel, K. Ensemble Docking into Multiple Crystallographically Derived Protein Structures: An Evaluation Based on the Statistical Analysis of Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 511−524.

(5) Korb, O.; Olsson, T. S. G.; Bowden, S. J.; Hall, R. J.; Verdonk, M. L.; Liebeschuetz, J. W.; Cole, J. C. Potential and Limitations of Ensemble Docking. *J. Chem. Inf. Model.* **2012**, *52*, 1262−1274.

(6) Carlson, H. A.; Masukawa, K. M.; Rubins, K.; Bushman, F. D.; Jorgensen, W. L.; Lins, R. D.; Briggs, J. M.; McCammon, J. A. Developing a Dynamic Pharmacophore Model for HIV-1 Integrase. *J. Med. Chem.* **2000**, *43*, 2100−2114.

(7) Damm, K.; Carlson, H. Exploring Experimental Sources of Multiple Protein Conformations in Structure-Based Drug Design. *J. Am. Chem. Soc.* **2007**, *129*, 8225−8235.

(8) Landon, M. R.; Amaro, R. E.; Baron, R.; Ngan, C. H.; Ozonoff, D.; Andrew McCammon, J.; Vajda, S. Novel Druggable Hot Spots in Avian Influenza Neuraminidase H5N1 Revealed by Computational Solvent Mapping of a Reduced and Representative Receptor Ensemble. *Chem. Biol. Drug Des.* **2008**, *71*, 106−116.

(9) Todorov, N. P.; Buenemann, C. L.; Alberts, I. L. De Novo Ligand Design to an Ensemble of Protein. *Proteins: Struct., Funct., Genet.* **2006**, *64*, 43−59.

(10) Dean, P. M.; Firth-Clark, S.; Harris, W.; Kirton, S. B.; Todorov, N. P. SkelGen: A General Tool for Structure-Based Denovo Ligand Design. *Expert Opin. Drug Discovery* **2006**, *1*, 179−189.

(11) Gutteridge, A.; Thornton, J. Conformational Change in Substrate Binding, Catalysis and Product Release: An Open and Shut Case? *FEBS Lett.* **2004**, *567*, 67−73.

(12) Zavodszky, M. I.; Kuhn, L. A. Side-Chain Flexibility in Protein-Ligand Binding: The Minimal Rotation Hypothesis. *Protein Sci.* **2005**, *14*, 1104−1114.

(13) Gaudreault, F.; Chartier, M.; Najmanovich, R. Side-Chain Rotamer Changes upon Ligand Binding: Common, Crucial, Correlate with Entropy and Rearrange Hydrogen Bonding. *Bioinformatics* **2012**, *28*, i423−i430.

(14) Fariselli, P.; Rossi, I.; Capriotti, E.; Casadio, R. The WWWH of Remote Homolog Detection: The State of the Art. *Briefings Bioinf.* **2007**, *8*, 78−87.

(15) Kalaimathy, S.; Sowdhamini, R.; Kanagarajadurai, K. Critical Assessment of Structure-Based Sequence Alignment Methods at Distant Relationships. *Briefings Bioinf.* **2011**, *12*, 163−175.

(16) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: Efficient Molecular Docking Considering Protein Structure Variations. *J. Mol. Biol.* **2001**, *308*, 377−395.

(17) Verdonk, M. L.; Mortenson, P. N.; Hall, R. J.; Hartshorn, M. J.; Murray, C. W. Protein-Ligand Docking Against Non-Native Protein Conformers. *J. Chem. Inf. Model.* **2008**, *48*, 2214−2225.

(18) Corbeil, C. R.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model.* **2009**, *49*, 997−1009.

(19) Francis-Lyon, P.; Gu, S.; Hass, J.; Amenta, N.; Koehl, P. Sampling the Conformation of Protein Surface Residues for Flexible Protein Docking. *BMC Bioinf.* **2010**, *11*, 575.

(20) Xu, M.; Lill, M. Utilizing Experimental Data for Reducing Ensemble Size in Flexible-Protein Docking. *J. Chem. Inf. Model.* **2011**, *52*, 187−198.

(21) Kellenberger, E.; Schalon, C.; Rognan, D. How to Measure the Similarity between Protein Ligand-Binding Sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209−220.

(22) Pérot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A.-C.; Villoutreix, B. O. Druggable Pockets and Binding Site Centric Chemical Space: A Paradigm Shift in Drug Discovery. *Drug Discovery Today* **2010**, *15*, 656−667.

(23) Nisius, B.; Sha, F.; Gohlke, H. Structure-Based Computational Analysis of Protein Binding Sites for Function and Druggability Prediction. *J. Biotechnol.* **2012**, *159*, 123−134.

(24) Schmidtke, P.; Le Guilloux, V.; Maupetit, J.; Tufféry, P. Fpocket: Online Tools for Protein Ensemble Pocket Detection and Tracking. *Nucleic Acids Res.* **2010**, *38*, W582−W589.

(25) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinf.* **2009**, *10*, 168.

(26) Craig, I. R.; Pfleger, C.; Gohlke, H.; Essex, J. W.; Spiegel, K. Pocket-Space Maps to Identify Novel Binding-Site Conformations in Proteins. *J. Chem. Inf. Model.* **2011**, *51*, 2666−2679.

(27) Taylor, W. R.; Orengo, C. A. Protein Structure Alignment. *J. Mol. Biol.* **1989**, *208*, 1−22.

(28) Alexandrov, N. N.; Takahashi, K.; Gō, N. Common Spatial Arrangements of Backbone Fragments in Homologous and Non-Homologous Proteins. *J. Mol. Biol.* **1992**, *225*, 5−9.

(29) Holm, L.; Sander, C. Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* **1993**, *233*, 123−138.

(30) Shindyalov, I. N.; Bourne, P. E. Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path. *Protein Eng., Des. Sel.* **1998**, *11*, 739−747.

(31) Holm, L.; Park, J. DaliLite Workbench for Protein Structure Comparison. *Bioinformatics* **2000**, *16*, 566−567.

(32) Schenk, G.; Margraf, T.; Torda, A. E. Protein Sequence and Structure Alignments within One Framework. *Algorithms Mol. Biol.* **2008**, *3*, 4.

(33) Ukkonen, E. Algorithms for Approximate String Matching. *Inform. Control* **1985**, *64*, 100−118.

(34) Galil, Z.; Park, K. An Improved Algorithm for Approximate String Matching. *SIAM J. Comput.* **1990**, *19*, 989−999.

(35) Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Phys.-Dokl.* **1966**, *10*, 707−710.

(36) Ukkonen, E. Finding Approximate Patterns in Strings. *J. Algorithms* **1985**, *6*, 132−137.

(37) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(38) PDB Search Website. http://www.pdb.org/pdb/search/advSearch.do?search=new (accessed: 10/06/2015).

(39) Hilbig, M.; Urbaczek, S.; Groth, I.; Heuser, S.; Rarey, M. MONA-Interactive Manipulation of Molecule Collections. *J. Cheminf.* **2013**, *5*, 38.

(40) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(41) Cheng, L. S.; Amaro, R. E.; Xu, D.; Li, W. W.; Arzberger, P. W.; McCammon, J. A. Ensemble-based Virtual Screening Reveals Potential Novel Antiviral Compounds for Avian Influenza Neuraminidase. *J. Med. Chem.* **2008**, *51*, 3878−3894.

(42) Xu, M.; Lill, M. A. Utilizing Experimental Data for Reducing Ensemble Size in Flexible-Protein Docking. *J. Chem. Inf. Model.* **2011**, *52*, 187−198.

(43) Bolstad, E. S.; Anderson, A. C. In Pursuit of Virtual Lead Optimization: Pruning Ensembles of Receptor Structures for Increased

Efficiency and Accuracy During Docking. *Proteins: Struct., Funct., Genet.* **2009**, *75*, 62−74.

(44) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.