JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

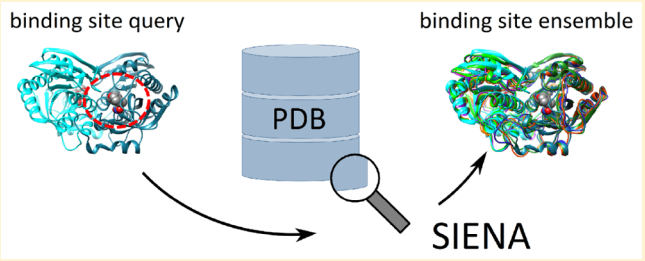
SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles

Stefan Bietz and Matthias Rarey*

Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

Supporting Information

ABSTRACT: Structural flexibility of proteins has an important influence on molecular recognition and enzymatic function. In modeling, structure ensembles are therefore often applied as a valuable source of alternative protein conformations. However, their usage is often complicated by structural artifacts and inconsistent data annotation. Here, we present SIENA, a new computational approach for the automated assembly and preprocessing of protein binding site ensembles. Starting with an arbitrarily defined binding site in a single protein structure, SIENA searches for alternative conformations of the same or sequentially closely related binding sites. The method is based on an indexed database for identifying perfect *k*-mer matches and a recently published algorithm for the alignment of protein binding site conformations. Furthermore, SIENA provides a new algorithm for the interaction-based selection of binding site conformations which aims at covering all known ligand-binding geometries. Various experiments highlight that SIENA is able to generate comprehensive and well selected binding site ensembles improving the compatibility to both known and unconsidered ligand molecules. Starting with the whole PDB as data source, the computation time of the whole ensemble generation takes only a few seconds. SIENA is available via a Web service at www.zbh.uni-hamburg.de/siena.



INTRODUCTION

Experimentally derived protein structures are one of the most important sources of information for a wide range of applications in structural biology. The fact that the amount of publicly available structures is steadily increasing not only allows the investigation of novel proteins but also improves the understanding of already resolved targets. The comparison of alternative structures of the same protein can, e.g., illustrate the impact of protein flexibility. Structures describing a protein bound to different ligands may facilitate a knowledge gain on ligand-induced conformation changes. Furthermore, structures of proteins only differing by point mutations allow insights in the structural foundation of protein stability and mutation-induced activity changes.

Protein structure ensembles are frequently applied for flexibility modeling in protein−ligand docking. Many studies have demonstrated that using properly selected ensembles can improve the docking performance.[1−7] Furthermore, protein ensembles can also be applied for the investigation of protein flexibility,[8−10] pharmacophore generation,[11,12] or de novo ligand design.[13,14]

However, the preparation of protein ensembles is still a rather elaborate process. This is mainly due to inconsistent annotation of protein structures as well as structural artifacts which can disturb the performance of ensemble employing applications. If, e.g., an ensemble that is used for molecular docking contains structures which lack single binding site residues, this might introduce unintended bias into the docking calculation. Therefore, a careful selection and preparation of ensemble structures often requires a substantial amount of manual interaction and the integration of various applications which need to be mutually geared to each other. Mostly, this includes the search and selection of appropriate structures, the alignment, and eventually the superposition of all ensemble members.

While other structure-related search tasks like identification of conserved structures across deviating sequences, binding pocket comparison, or the search for structural motifs are well supported, only few studies and databases deal with simplifying the search process of alternative conformations. One of the few exceptions is the Pocketome project,[15,16] an extensive selection of binding sites ensembles with additionally annotated information. The Pocketome comprises ligand binding sites that are reviewed in the Uniprot and for which at least two alternative structures are available at the Protein Data Bank (PDB).[17] Furthermore, at least one of the PDB structures needs to contain a complexed ligand molecule. All entries are regularly updated and therefore also contain recently published entries. The structures are superimposed and additional information is provided including clash and interaction details of all pairwise combinations of site conformation and ligand molecules as well as geometry- and interaction-correlated similarity analyses of all pairwise site conformations. For this

Received: September 25, 2015
Published: January 13, 2016

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

reason, the Pocketome is a rich resource on protein flexibility information that could be used for variable applications. However, to the best of our knowledge, the Pocketome Web interface has only limited functionality to customize the ensemble generation process and does not support a direct access to the preprocessed structures in PDB format or another commonly distributed format.

The CoDNaS database (Conformational Diversity of Native State)[18] and the preceding PCDB (Protein Conformational Diversity Database)[19] provide statistical summaries of various protein flexibility descriptors like average and maximum RMSD or the number of available conformers. Providing several filter criteria for specifying the targets a user is interested in, both databases supply a general statistical overview on the experimentally observed flexibility of user-selected proteins. Although the CoDNaS database serves a list of all PDB codes applied for a certain target analysis and additionally a pairwise chain comparison and structural superposition, it does not provide a direct access to the structure alignments of all protein conformations that could be used as conformational ensembles for downstream applications.

The Astex Non-native Set[20] constitutes a comprehensive collection of protein ensembles and comprises 1112 PDB structures of 65 diverse proteins. Created for docking performance assessment of non-native protein conformations, the structures are sequentially identical within the binding sites and were superimposed with a focus on structurally conserved binding sites. Protonation states, tautomeric forms, and residue flips are optimized and manually verified for the reference structures taken from the Astex Diverse Set[21] and were furthermore transferred to the other structures of the same protein (non-native conformations). While this is a sensible feature in the context of docking the reference ligands to non-native protein conformations, as described by Verdonk et al.,[20] it constitutes a bias for evaluating other ligand molecules. Moreover, the data set describes a static state of knowledge at the time of publication (2008). Given its elaborate data preprocessing, a completion of the Astex Non-native Set by more recent structures requires substantial additional efforts.

Although the application of conformational ensembles is generally capable of increasing the representation accuracy of computational protein models, the quality of many predictive applications also relies on a careful selection of the ensemble members, as the false positive rate is known to increase with the ensemble size.[2] Therefore, various protocols have been proposed for reducing the number of ensemble structures. A commonly applied strategy is the clustering of structures on RMSD values.[22−26] Other methods apply different geometry-based structure comparison approaches,[27−29] structural descriptors,[30] or docking scores.[4] Rueda et al.[31] and Xu and Lill[32] compare different descriptors for the selection of protein conformations. Both studies propose a selection of protein structures on the basis of small scale virtual screening (VS) results using small training data set of known binder and nonbinding molecules. Korb et al.[7] compare three selection protocols which are based on small scale VS, ligand size, and ligand similarity. On average, the latter performed better than the two former approaches in this study.

Here, we introduce SIENA, a new computational approach for the generation of protein binding site ensembles. Its central purpose is to simplify the process of selecting and superimposing appropriate protein structures under consideration of case-specific requirements. This is realized by the integration of different modules which specifically address the typical obstacles of ensemble generation like the handing of inconsistent data annotation, structurally unresolved and chemically modified residues, or selecting a small set of diverse conformations. For this, SIENA comprises a quick search for alternative conformations of user-defined binding sites, an alignment algorithm specially geared to binding site conformations, a comprehensive set of filters for user-specific tailoring of the ensemble, and a rigid-region detection for structure superposition. SIENA also contains a new interaction-based ensemble reduction procedure which takes up the idea of learning from known ligand binders but circumvents the computationally expensive virtual screening on the training set by a direct transfer of the ligands to alternative structures. All components are fully integrated in a single software application what simplifies its usage and avoids conversion issues. SIENA is available via a Web service and provides the user with a sequence alignment of the binding site as well as superimposed PDB structures which are, apart of the transferred coordinates, equal to the original files from the PDB and thus contain all structural details and further information. With this, SIENA is perfectly suited to create protein binding site ensembles for single target studies, but also for the consistent compilation of large ensemble data sets which can, e.g., be used for benchmark analyses of tools predicting protein flexibility.

METHODS

SIENA has been developed for the instant identification and automatic preparation of alternative binding site conformations including structures exhibiting point mutations. Its automatic ensemble generation process consists of four steps (see Figure 1). First, a database query is used to reduce the set of available protein structures to a subset of structures that exhibit a certain sequence identity with the query binding site and furthermore fulfill additionally specified search criteria such as minimal resolution (candidate selection). In a second step, the topological and geometrical conformity of each candidate with the query binding site is evaluated and nonmatching structures are discarded (structure validation). Third, structure-depended selection criteria are applied to adapt the results to the user's requirements (filtering and ensemble reduction). Finally, a common structural conserved core is identified and used for structure superposition (rigid region superposition).

Candidate Selection. The first and most effective reduction step is based on an SQLite database that stores characteristic properties of each available protein structure. For now, this comprises the following simple scalar values: resolution, EC numbers, electron density availability, and publication year. Furthermore, the database query shall also evaluate the sequence similarity of the query binding site to the stored structures. In the given application scenario, accurate alignment techniques are generally too expensive for an extensive search in large data sets. Therefore, a comparison of short sequence snippets, or *k*-mers, serves as a first heuristic for the selection of similar structures. Similar techniques are the basis of the efficiency of many sequence search tools like FASTA,[33] BLAST,[34] or BLAT.[35] During database generation, each library protein is divided into its connected peptide chains. For each chain, a window of fixed size (four by default) is slid over its sequence and thereby produces a set of overlaying *k*-mers. These are stored in an indexed database table that links them to an identifier of their original structure enabling

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

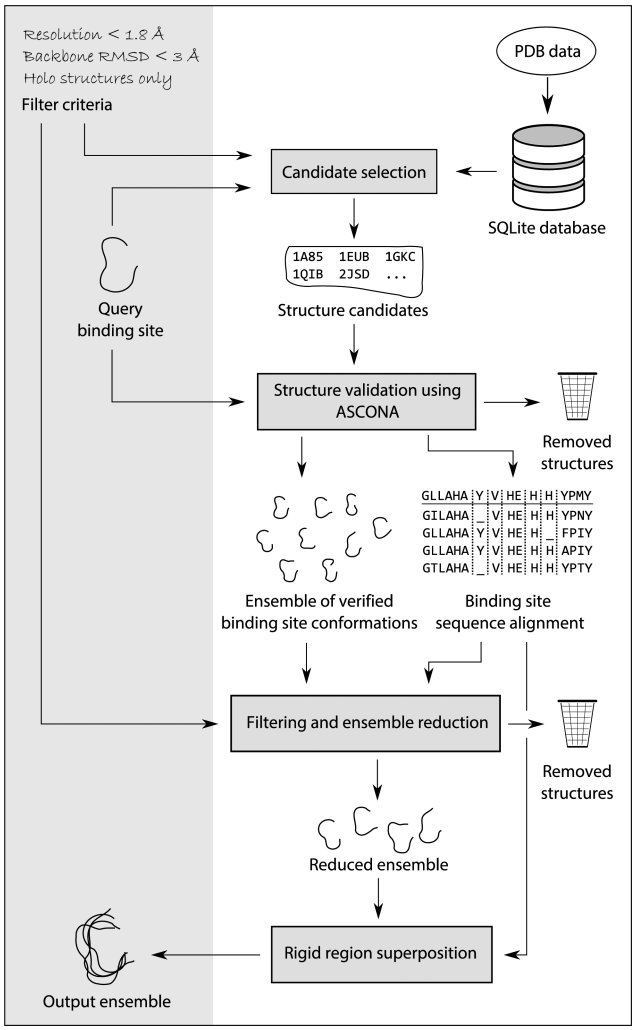


Figure 1. Schematic overview on the ensemble generation strategy implemented in SIENA. The gray shaded area on the left-hand side indicates which elements belong to the user interface. The gray boxes on the right-hand side represent the four major steps of SIENA.

immediate access to all structures that contain a certain sequence pattern.

For query construction, the user defines the residues that form the binding site. This process is supported by an automated selection of all residues around a geometric reference but could be also realized with a site detection algorithm. For each binding site residue, a k-mer describing its surrounding is generated by extending the residue in N- and C-terminal direction until the snippet size used for database generation is reached. Based on this site representation, the identifiers of all structures which contain a user-defined portion of *k*-mers and further fulfill the user's search criteria are extracted from the database. The threshold defining the required portion of *k*-mers can be used to specify a reasonable trade-off between runtime efficiency and the algorithms capability to tolerate structural deviations like missing residues, additional loops, and mutations. The lower the threshold, the more structural deviations are accepted at this stage. However, low thresholds may result in false positives which then need to be filtered out in the following more expensive structure validation step. By default, SIENA uses a threshold of 30%.

Structure Validation. Next, SIENA examines which candidates actually contain at least one equivalent of the query binding site using our novel binding site alignment algorithm ASCONA.[36] Since its functionality is described elsewhere, we only briefly summarize its most important aspects here.

ASCONA represent the binding site by short peptide sequences to which we will refer as binding site fragments. For each candidate, ASCONA searches the binding site fragments of the query with an efficient sequence alignment algorithm. For reconstructing the target binding sites from fragment matches ASCONA employs a structure-based assembly approach. The fragment hits are recombined on the basis of a rather fuzzy geometry measure which involves spatial distance deviations and Euclidean distances of rotation quaternions. This facilitates the consideration of the fragments' relative orientation, while it still allows for dealing with highly flexible binding sites. The result is a set of residuewise sequence alignments that describe the mapping of the query binding site residues onto all detected binding site instances within the target structure. On the one hand, this residuewise mapping allows investigation of whether all residues of the query binding site are also present in a certain target structure. On the other hand, it is also required for an accurate calculation of flexibility descriptors like RMSD values and the structure superposition.

In contrast to most other alignment techniques, ASCONA has been specially developed to meet the requirements of aligning protein binding site conformations. Other algorithms mostly address different problems like the identification of distantly related proteins. Due to this deviating focus, they are often not fully suitable for the alignment of protein ensembles. For instance, purely sequence-based alignment techniques cannot always be used to align binding sites that occur at the interface of different subunits in homo-oligomeric proteins, as this mostly requires also structural information. Purely structure-based approaches often operate on abstract geometric descriptors and can thus not provide a residuewise mapping. Moreover, the analysis of structurally similar regions does not necessarily lead to sequentially correct residue mapping.

Filtering and Ensemble Reduction. Depending on the application scenario, the generated structure ensembles need to satisfy different requirements. Therefore, SIENA allows the user to control the structure selection by applying the filters and ensemble reduction procedures summarized in Table 1.

In addition to simple filters which are based on feature calculation like the number of mutations or the binding site RMSD, SIENA also provides more elaborate clustering-based and interaction-based ensemble reduction procedures.

Interaction-Based Ensemble Reduction (IBER). Starting from a set of binding site conformations that have been extracted during the previous database query and filtering steps (ensemble candidates) and the set of complexed ligands contained therein, the IBER method investigates the compatibility of each of these structures to each ligand in terms of steric overlaps (atom clashes), interaction preservation, and decrease in covered hydrophobic surface area with respect to the ligand's native protein conformation (which is given by the structure the ligand has been extracted from). Based on these descriptors, the approach selects the ensemble structures with a greedy approach.

First, for each ensemble candidate, protonation states, tautomers, and hydrogen orientations of both protein and ligand are optimized with Protoss.[39] Additionally, the ensemble

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

Table 1. Structure Selection Criteria

A. Filters	
minimum site identity	solely maintains structures with a relative frequency of residue matches within the binding site alignment above a certain threshold
minimum site coverage	describes the minimum required relative frequency of residue matches and replacements within the binding site alignment or, in other words, the inverse of the gap rate
maximum number of site mutations	limits the number of accepted replacements within the binding site alignment
identical global sequence	removes all structures which do not contain a complete equivalent of the reference structure in terms of the amino acid sequence described in the PDB ATOM section; in other words, the global alignment of maintained structures may contain additional residues but no mutated or missing residues
no global mutations	removes all structures for which the best global alignment to the reference structure contains replacements (mutated residues); in contrast to the *identical global sequence* filter, this filter does not remove structures with missing residues
diverse sequence	removes structures with identical sequences within the binding site and thus produces a diverse ensemble with respect to sequence
maximum RMSD	defines the maximum accepted RMSD of a structure with respect to the query considering the α-carbons or all heavy atoms within the binding site
holo structures only	keeps only binding sites which are in complex with a ligand molecule; solvent, buffer molecules, and small ions are not regarded as ligands in this context
B. Reduction Procedures	
diverse ligands	selects a set aligned binding sites with different ligands; identical ligands are identified on the basis of normalized protonation states[37] and an internal molecule representation similar to Unique SMILES[38]
diverse conformation	applies a complete linkage clustering for extracting diverse conformations considering the α-carbons or all heavy atoms within the binding site (cf. Clustering-Based Ensemble Reduction)
IBER	selects a set of structures which covers the ligand binding motifs contained in the set of unfiltered structures as well as possible (cf. Interaction-Based Ensemble Reduction)

candidates are superimposed onto the query binding site using a rigid region that is conserved across all structures (see Rigid Region Superposition). This superposition is used to transfer the ligand molecules from their native structures to other ensemble candidates (transferred binding poses). This includes only those ligand molecules that fulfill the following conditions. First, the molecule needs to intersect with the geometric reference (e.g., a reference ligand) that has been used for the query binding site definition. If this is true for more than one molecule, only the one with the most intersecting atoms is selected. Furthermore, the ligand must not be covalently bound to the protein and must contain more than five heavy atoms. A molecule is also not considered if it appears in a precompiled list containing typical cofactors, solvents, buffers, or crystallization agents (see Table S1). For each combination of a selected ligand and an ensemble candidate the descriptors mentioned above are calculated as follows:

Steric Overlaps. Steric overlaps are estimated by simply analyzing whether the spatial distance d of any ligand atom a_l and any protein atom a_p falls below a threshold defined by the sum of their van der Waals radii ($r(a_p)$ and $r(a_l)$) multiplied with a softening factor f. The degree of steric disruption for a certain binding pose is measured by the number of protein residues r in binding site S for which overlaps with any atom of ligand l occur:

$$n_{\text{overlap}} = |\{r \in S \mid \exists\, a_p \in r, \exists\, a_l \in l: d(a_p, a_l) \leq f(r(a_p) + r(a_l))\}| \quad (1)$$

By default, f is set to 0.8 and an ensemble candidate is considered as not compatible to the respective ligand if at least one steric overlap is found. However, in order to obtain a clash-score (σ_c) for the candidate selection process, the number of overlapping residues n_{overlap} is reversely mapped onto the range $]0,1]$:

$$\sigma_c = \frac{1}{1 + n_{\text{overlap}}} \quad (2)$$

Interaction Preservation. Interaction preservation is calculated on the basis of interaction sets and a comparison of these for a ligand's native and transferred binding poses. For calculating the interaction set, all protein residues in the binding site are split into four segments: the backbone's nitrogen, the backbone's oxygen(s), polar side-chain atoms, and aromatic ring systems. Ligands are simply split into single atoms as we do not investigate ligand flexibility here and therefore do not need to take any ambiguities like topological symmetry into account. Aromatic ring systems in the ligand molecule are mapped onto an arbitrary member atom. The interaction set contains all protein−ligand segment pairs, for which an interaction is detected in the corresponding complex on the basis of interaction spheres.[40] To further account for minor changes, the interaction parametrization used to generate the interaction sets for transferred binding poses is setup less strictly than that for the native structure. Eventually, the following term is applied to calculate the relative amount of preserved interactions:

$$\sigma_{ia} = \frac{|R \cap T|}{|R|} \quad (3)$$

where R and T denote the interaction sets of the native reference and the transferred binding pose, respectively. The binding site alignment calculated with ASCONA is applied to obtain a residue mapping for comparison of interaction sets from different protein structures.

Covered Hydrophobic Surface Area. The decrease in covered hydrophobic surface area of the ligand is estimated as a third descriptor. This is done by a procedure developed in the context of the HYDE scoring function which is described elsewhere in detail.[41] In summary, this approach creates a net of surface dots describing the solvent accessible surface at a discrete level. Each surface dot is assigned a portion of the ligand's Connolly surface area, which allows for assessing partial surfaces by adding up the surface area increments of the corresponding surface dots. Analyzing which surface dots overlap with any protein atoms thus leads the covered portion

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

of the molecular surface. In contrast to the HYDE scoring function, this is done only for the hydrophobic part of the ligand. Comparing the covered surface area of a transferred binding pose with its respective native interaction geometry facilitates the detection of an unfavorable opening of the binding site. Therefore, a decrease of the covered hydrophobic surface area upon the ligand transfer to a different ensemble candidate leads to penalization of the respective transferred binding pose which is measured by the following equation:

$$\sigma_s = 1 - \frac{\max(0, s_n - s_t)}{s_n} \quad (4)$$

where s_n and s_t represent the covered hydrophobic surface portions for the native and the transferred binding poses, respectively.

The latter two descriptors are both designed in a way which implies that the features they are describing cannot be improved in comparison to the native state. Although this might be a disadvantage in case of poorly refined native structures, it clearly helps to guide the ensemble optimization to the improvement of poorly supported ligands and to prevent an overestimation of artificial poses.

For each transferred binding pose, i.e., each combination of a ligand l in a non-native protein conformation c, a score is calculated by a Hölder mean variation of the three descriptors:

$$\sigma_{l,c} = \sqrt[p]{\frac{\sigma_c^p + \sigma_{ia}^p + \sigma_s^p}{3}} \quad (5)$$

By default, p is set to −8. Ensemble scores for a certain ligand l and an ensemble E_k are realized by maximizing over the σ-scores of all k ensemble members e_i:

$$\sigma_{l,E_k} = \max_{i=1}^{k}(\sigma_{l,e_i}) \quad (6)$$

Based on this scoring scheme, a greedy algorithm is used for ensemble candidate selection: During each iteration, that candidate is selected as a new ensemble member which best complements the current ensemble. This is realized by choosing the candidate with the highest weighted sum of scores over all transferred ligand poses L:

$$e_k = \arg\max_{c \in C}\left(\sum_{l \in L} w_{l,k} \cdot \max(0, \sigma_{l,c} - \sigma_{l,E_{k-1}})\right) \quad (7)$$

where e_k is the ensemble structure selected in the kth iteration, C the set of the remaining ensemble candidates, E_{k-1} the current ensemble, and $w_{l,k}$ the weighting factor reflecting the priority of ligand l in the kth iteration. The priority values depend on the best score of the respective ligand with any of the already selected ensemble structures. The better it is already scored, the lower its priority. Initially all weights are set to 1.0. After the first iteration, they are defined as follows:

$$w_{l,k} = \frac{1}{1 + \sigma_{l,E_{k-1}}} \quad (8)$$

The algorithm terminates if the ensemble size has reached a user-defined maximum or none of the remaining candidates yields an improvement of the ensemble.

Clustering-Based Ensemble Reduction. Additionally, we implemented a clustering-based selection approach using binding site RMSD values and complete linkage clustering. In order to avoid a combinatorial explosion that would result from considering side-chain symmetry for calculating the optimal superposition of two compared binding sites, a superposition calculated on α-carbons only is used instead as a heuristic solution. Based on this superposition, we then calculate the pairwise RMSD considering all heavy atoms of the residues including side-chain symmetry effects. Cluster representatives were selected by choosing the member with the smallest average RMSD to all other cluster members.

Rigid Region Superposition. The last step of the ensemble preparation is the superposition of all selected conformations onto the given query binding site. This predominantly supports the visual comparison of the output structures and its direct applicability to continuative calculations like ensemble docking or flexibility assessment. Besides that, the following approach is also used for the pairwise structure superposition required for interaction-based filtering as described above.

The usage of a rigid, well conserved region for ensemble superposition is a widespread alternative to considering the whole structure for calculating the transformation. On the one hand, it simplifies the visual recognition of flexible substructures. On the other hand, it is also essential for an accurate evaluation of superposition-dependent ligand comparison like the RMSD calculation of docking poses in non-native protein conformations or a shape-based similarity assessment of ligands from different complex structures. In contrast to a frequently applied iterative refinement approach, we followed the local structural domain approach first described by Kelley et al.[42] for the analysis of NMR structures. Based on a preselection of relatively rigid α-carbon atoms, this approach creates a distance matrix for each structure and afterward calculates the variance of every pairwise α-carbon distance over all structures. The variances are used as a distance measure for an average linkage clustering of the carbon atoms resulting in a partition of the protein into rigid regions. Compared to the iterative approach, the latter technique is better applicable to ensembles consisting of more than two members. We employ this concept to detect the largest rigid region across an ensemble of binding site conformations and use the resulting atom selection as the basis for binding site superposition. By default, a variance threshold of 1.0 Å^2 is used for clustering. Since we are only interested in the largest rigid region and do not need to distinguish between small rigid and highly flexible parts, the preselection can be omitted.

RESULTS AND DISCUSSION

Conformation Mining. In a first experiment we demonstrate that SIENA is capable to produce sensible and complete ensembles. In order to obtain a diverse, unbiased, challenging, and clearly representable test data set, 10 target proteins from the Astex Non-native Set were selected: Five targets featuring the most conformations and the five most flexible proteins measured by RMSD values as listed by Verdonk et al.[20] For each target, the respective reference ligand from the Astex Diverse Set was used to define the binding site by selecting all residues within a radius of 6.0 Å around the ligand. These served as input for several SIENA queries against the whole PDB with varying filter criteria (see Table S2 for details). For comparison, the respective ensembles from the Astex Non-native Set were employed. Two structures (2NMW, 2C5P) have meanwhile been superseded by newer versions (3OXC and 2WEV) and were therefore substituted for this analysis. One structure (2CDD) has been removed from the PDB

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

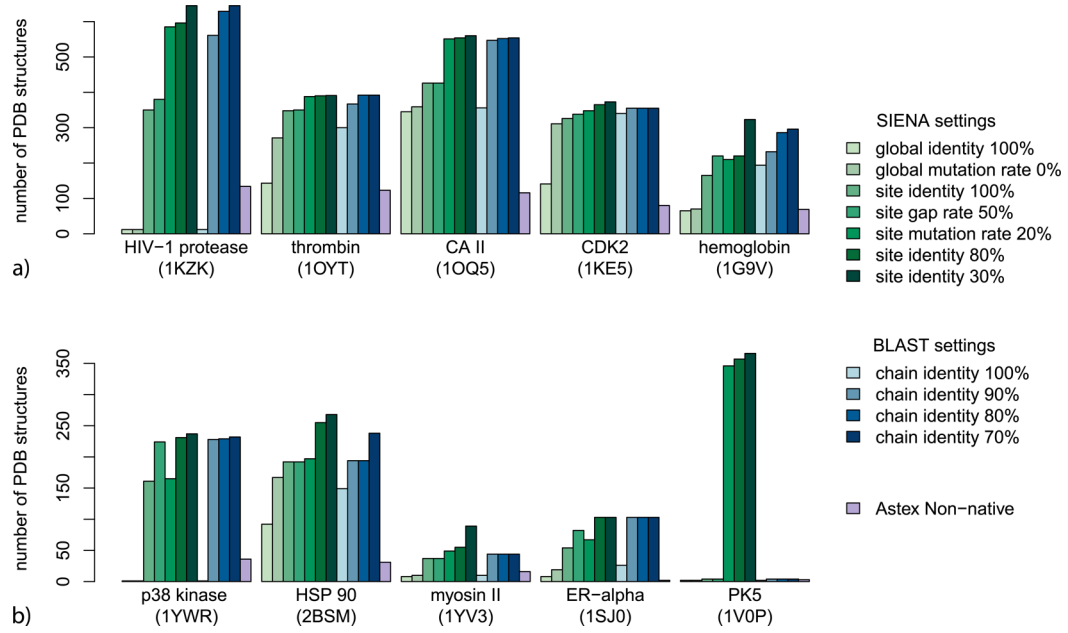


Figure 2. Number of PDB structures that are contained in the Astex Non-native Set and in the ensembles generated by BLAST and SIENA for (a) the five most frequent and (b) the five most flexible Astex Non-native targets.

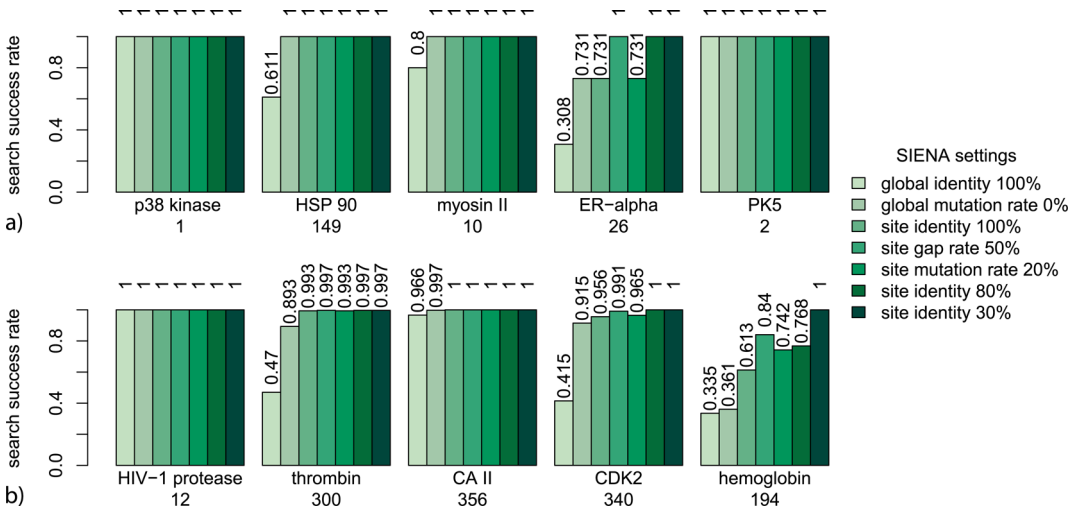


Figure 3. Portion of structures that are identified by SIENA with different settings for (a) the five most frequent and (b) the five most flexible Astex Non-native targets. The reference sets are defined by PDB BLAST searches using a chain identity of 100%. Their sizes are indicated by the number below the protein names.

without substitution and was therefore also eliminated from the reference set. Additional structure ensembles were compiled using the BLAST sequence similarity search as implemented on the PDB Web site[43] using varying sequence identity thresholds. Since the PDB BLAST interface expects a single protein chain sequence as input, we selected chain "A" for all targets besides thrombin, for which chain "H" was chosen. The *Expect* value was set to 100, and the *Low Complexity* filter was switched off.

Figure 2 summarizes the sizes of the resulting ensembles. For all targets, the different ensembles generated by SIENA exhibit substantial differences in size which demonstrates its capability for case specific ensemble selection. Two cases are particularly conspicuous. First, the ensemble size for HIV protease increases from 12 structures for both global search queries to 350 structures for the ensemble generated with a site identity filter of 100%. This is due to six nonbinding site mutations by which the Astex reference structure (1KZK) differs from the wild type. Second, the ensemble sizes detected for Protein Kinase 5 remain on a very low level for the first four depicted SIENA settings which is in accordance to the reference values from BLAST and the Astex Non-native Set. However, applying a lower site identity threshold results in a tremendous increase in ensemble size. In this case, most of the members of the larger ensembles describe the closely related CDK2.

Comparison to the Astex Non-Native Set. The overall comparison with the references from BLAST and the Astex Non-native Set demonstrates that SIENA produces correct results in a fully automated fashion. Although it is not expected that all three approaches produce exactly the same ensembles as they apply different filter criteria, there are certain boundary conditions that should be fulfilled if SIENA works correctly. First, the Astex Non-native ensembles are supposed to form a

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

subset of the SIENA 100% site identity ensembles, as the Astex Non-native Set also contains structures with identical site sequences only. Indeed, a comparison of the ensembles showed that SIENA correctly identifies all structures of the 10 selected targets (cf. Figure S1) using the 100% site identity setting. For most of the 10 targets, the Astex ensemble is considerably smaller than its counterpart produced by SIENA. On the one hand, this might be a result of additional filters applied by Verdonk et al.[20] On the other hand, the data set was compiled in 2008 and since then the PDB has been grown from approximately 54 500 structures in 2008 to more than 112 000 in September 2015.[44]

Comparison to the PDB BLAST Search. Second, the PDB BLAST search using a 100% sequence identity criterion and the SIENA query applying the global identity filter are expected to produce similar ensembles if the protein structure describes a monomer or a homo-oligomeric complex. Otherwise, if the protein is a hetero-oligomer, both approaches do not correspond to each other as the applied BLAST query only makes use of one chain sequence, while SIENA's global identity filter considers the whole protein structure. To the best of our knowledge, the PDB BLAST interface does not offer a direct way to search for structures by simultaneously applying more than one sequence query for matching different chains within one structure. Indeed, the ensemble sizes of both approaches differ substantially for the two hetero-oligomeric proteins, namely thrombin and hemoglobin. 9% of the thrombin and 44% of the hemoglobin structures identified by BLAST carry mutations in the additional chains that were not used for the BLAST query. Therefore, it is reasonable that they are not selected by SIENA using the global identity settings.

Figure 2 also shows that SIENA collects considerably less structures for CDK2 (monomer), HSP 90, and ER-alph (both homodimers) as well as slightly smaller ensembles for CA II and myosin II (both monomers). In order to analyze the reasons for the latter deviations, we calculated the percentage of BLAST-retrieved structures which are not part of the respective SIENA ensemble and further investigated whether these structures can be identified with less restrictive SIENA settings. Figure 3 shows that many nonretrieved structures can be detected if the 100% sequence identity criterion is restricted to the binding site and gaps are accepted for the global alignment (global mutation rate 0%) or only an active site alignment is calculated, allowing for gaps and mutations in the rest of the protein structure (site identity 100%). In some cases, the success rate can be further improved when mutations or gaps are also accepted in the binding site. These deviations are results of different structural artifacts which are not considered in the BLAST search as it obviously operates on the nominal protein sequence.

First, some of the structures lack single residues of the protein which are present in the query structure. This leads to the gaps in the global alignments and, according to which setting is applied, also to elimination during the filtering process. In most of these cases, the missing residues belong to the terminal chain regions and do not directly interfere with the binding site. Thus, allowing for gaps in the global alignment (cf. global mutation rate 0%) results in a successful matching of these structures. If binding site residues are missing in the structure, the site identity threshold needs to be decreased to identify these cases with SIENA (cf. site gap rate 50% or site identity 80%). Second, in the case of hemoglobin, the binding site of the Astex reference ligand is formed at the interface of three subunits. However, some of the structures identified by the BLAST search do only contain one or two subunits inducing an incomplete binding site and are therefore correctly discarded by SIENA. In order to align them properly, ASCONA has to accept a high gap rate in the binding site (cf. site gap rate 50%). For most application scenarios, these structures would be preferably removed from the ensemble. Third, some structures of CDK2 and CA II contain modified amino acids. SIENA recognizes these residues as amino acids but considers them as mutations. The identification of these cases makes especially sense in the context of structure-based applications, since modified residues can affect the protein function. For instance, the posphorylation of TYR 15 in CDK2 causes an inhibition of the enzymatic activity.[45] A detailed overview of all modified residues is given in Table S3.

Similar observation can be made for analogous analyses using lower BLAST identity cutoffs (cf. Figure S2 and S3). Applying a minimal site identity of only 30%, 2859 of 2863 structures collected by BLAST with a chain identity threshold of 70% can also be identified with SIENA. Three of the remaining four structures contain only α-carbon atoms (3HTC, 2HVP, and 1IAN). Therefore, SIENA is not able to properly analyze the proteins' structure in these cases. The fourth structure (2HRP) was identified by BLAST for the HIV-1 protease query but actually describes an antibody bound to a 10 amino acid long fragment of HIV-1 protease.

The results show that SIENA's diverse configuration opportunities allow a precise adaption to different use cases. In particular, eliminating structures with missing, modified, or mutated residues might be essential for generating accurate unbiased conformational ensembles in the context of highly structure-sensitive applications. In such scenarios, using the PDB BLAST search for ensemble generation would require subsequent filtering and verification steps, as it obviously employs the nominal protein sequence instead of the actual structurally present residues. We would like to emphasize that we do not consider this as a general shortcoming of the PDB BLAST search routine, but rather as a demonstration of the different focus of both methods. Clearly, there are application scenarios for which focusing on the nominal protein sequence offers opportunities which cannot be achieved by only considering residues that are actually present in the structure. Nevertheless, it also highlights the utility of SIENA, since a scenario where the structural consistency of alternative conformations is critical for the target application is not sufficiently covered by purely sequence-based search methods.

The reverse experiment shows that the results of the BLAST 100% chain identity query contain all structures that are found by SIENA for a global identity threshold of 100% (cf. Figure S4), with the exception of four structures (1H8D, 1H8I, 3PYK, and 3K97) for which the nominal sequence (as given in the FASTA files from the PDB) do not contain the full reference sequence of the query chain. However, all residues which are present in the respective query structures also occur in these four structures and are therefore identified by SIENA. If SIENA's 100% and 80% site identity settings are used as reference, the BLAST threshold needs to be decreased to 70% and 40% respectively, to obtain ensembles that contain all reference structures (cf. Figure S5 and S6). This may be reasoned in the phenomenon that binding site sequences are often better conserved than the rest of the protein sequence. However, for most targets these ensembles are considerably larger than their references (cf Figure S7). Assuming that

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

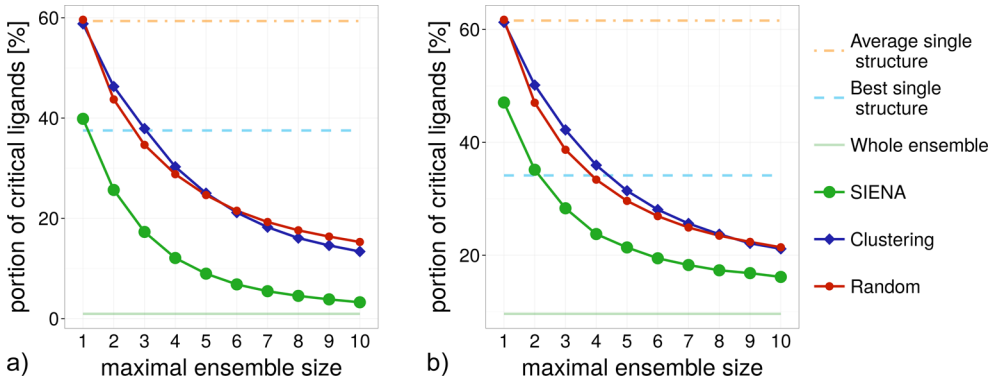


Figure 4. Portion of critical ligands of all NBSE targets as a function of the maximal number of structures in a reduced ensemble (maximal ensemble size). As reference values, the average and best performance of single structures are depicted. Additionally, the optimal ensemble performance is given by the performance of the nonreduced training set ensemble. (a) Results for all training sets. (b) Results for all test sets. Results for higher ensemble size thresholds are depicted in Figure S9.

SIENA finds all structures that fulfill the given site identity criterion, which is indicated by the comparison to the Astex Non-native Set, the difference to the BLAST ensembles constitutes the fraction of superfluous structures that would need to be filtered out during a postprocessing step.

Applicability Limitation. In addition to the comparison between BLAST and SIENA ensembles generated at high sequence similarity thresholds, also SIENA's coverage of BLAST ensembles at lower sequence identity was analyzed to investigate its applicability boundaries. The results are depicted in Figure S8. For BLAST thresholds from 100% to 70%, SIENA preforms highly accurate and only lacks the four structures mentioned above. For lower threshold, SIENA's performance quickly decreases. However, this is in good accordance with its theoretical limitations. On the one hand, the database search relies on the exact matching of a certain percentage of k-mers forming the query binding site (30% by default). Clearly, the probability to reach a sufficient matching rate rapidly decreases with the sequence similarity of the proteins. On the other hand, the ASCONA alignment procedure is also based on matching short sequence fragments. Although in this case an approximate sequence matching algorithm is applied, allowing for mutations and gaps in the individual fragments, a detection of low sequence similarity is not feasible. This is a result of ASCONA's focus on aligning conformations of highly flexible binding sites, which requires the usage of relatively short fragments as well as a low rate of random matches and therefore also a highly accurate fragment matching.

Although SIENA is not suited for searching distantly related proteins, a setting of low site identity threshold as used in the above experiment is still sensible for detecting structures lacking a considerable portion of the binding site as in the case of hemoglobin. This could be for instance useful for protein−protein docking or the investigation of complex formation.

Ensemble Reduction. In order to investigate SIENA's ability to select a reduced set of binding site conformations, we compiled a comprehensive test set of different binding site ensembles on the basis of the sc-PDB,[46] which constitutes a collection of pharmacologically relevant targets. Here 63 ligand molecule files and 1 protein file (1NLM) were adapted in order to circumvent file interpreting errors. Four structures were not available in the PDB anymore and were therefore removed from the set. For each of the remaining structures, the respective reference ligand from the sc-PDB was used to determine the binding site. The resulting binding sites were searched with SIENA in the PDB under consideration of a 100% site identity criterion and a resolution cutoff of 2.5 Å. Since some targets are represented by multiple sc-PDB structures, ensemble duplicates were subsequently removed by the following filtering process. Initially, clusters were formed by linking ensembles which share at least one common PDB structure. Then, a cluster representative was selected by choosing the member with the lexicographically first PDB code in the set of all largest ensembles within one cluster. The set of cluster representatives was then further reduced to those ensembles that contain at least 12 unique ligands (see Table 1 for unique ligand filter definition). In total, this led to a collection of 182 nonintersecting binding site ensembles (NBSE set). A list of all NBSE targets is given in Table S4.

The quality of an ensemble shall be measured by its complementarity to ligands with known binding modes. At this point, we assume that the individual scoring measures applied in the selection process are appropriate for detecting unfavorable binding poses. We focus on analyzing how well the binding site flexibility is represented by reduced ensembles of different size.

Applying SIENA, we generated reduced ensembles for all 182 NBSE ensembles and analyzed the resulting number of critical ligands. A ligand is considered as critical if its relative orientation in a non-native binding site conformation, which is derived from the superposition provided by SIENA, causes atom clashes, the loss of interactions, or a reduction of covered hydrophobic surface area that exceed the threshold of 0.8. Note that in contrast to both other measures, the clash criterion is not defined in relation to the native binding pose of the ligand. Therefore, also the native binding pose can be considered as critical, if a close atom contact occurs that falls below the threshold applied for clash detection. In the case of an ensemble, a ligand is considered as critical, if none of its members is capable of fulfilling all three criteria.

In order to measure the quality of the reduced ensembles, the original NBSE ensembles were split into test and training sets. For each input ensemble, the ligand set was randomly partitioned into four distinct and at most equally sized subsets which were each used once as test set for cross validation. The remaining ligands served as training set for the generation of reduced SIENA ensembles which were build up from the set of holo structures cocrystallized with the training set ligands. A 10 times repeated random ensemble selection was executed as reference for the clustering and interaction-based selection

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

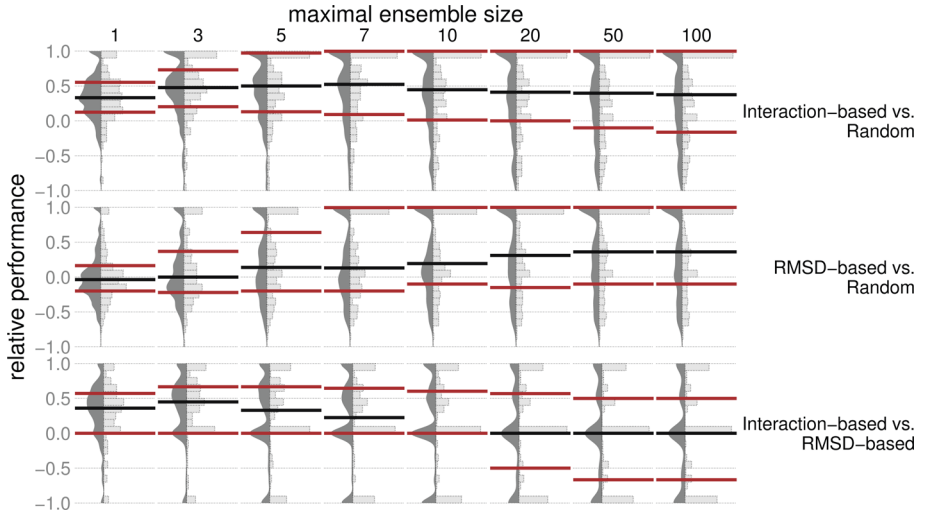


Figure 5. Relative performance of different ensemble reduction approaches as a function of the maximal number of structures in a reduced ensemble (maximal ensemble size). The depicted data represents the result from the test set evaluation. The distribution of the relative performance across all NBSE targets is depicted as an approximated density on the left-hand side and in the form of a histogram on the right-hand side of each subfigure. The black and red bars indicate the median values and the 25% and 75% quantiles, respectively.

approach. For all three approaches, the ensemble size was varied between a minimum of one and a maximum of half of the training set size. Additionally, the number of critical ligands for an average and the best single binding site conformation from the input ensemble were analyzed. (The best single binding site corresponds to the one with the fewest number of critical ligands.) The optimal ensemble performance is given by the number of critical ligands for the whole training set ensemble.

Figure 4 depicts the results for all 182 targets. For the test sets, the portion of critical ligands for an average single conformation (62%) is considerably higher than the performance of the best binding sites (34%). The results also illustrate, that the usage of ensembles generally facilitates a significant improvement over single structures. Compared to the random ensemble selection, the interaction-based approach is clearly better suited for selecting a single structure and further generally improves the performance of smaller ensembles, although its benefit is considerably lower for the test set than for the training set. However, the optimal ensemble performance is also distinctly worse for the test set (10%) than for the training set (1%), which illustrates that the applicability of known protein conformation for unknown ligands is limited for a respective fraction of the considered data. The overall results of the clustering-based ensemble reduction show only minor deviations to random selection. Nevertheless, a more detailed analysis of these measures for each individual target (see Figure S19) revealed that this is rather a result of averaging over all ligands than a general rule. The individual analyses show that there are multiple targets for which the clustering performance is distinctly better or worse than random selection. Similar cases can be found for the interaction-based ensemble reduction.

In order to quantify these observations, the relative performance $\rho_{A,B}(x)$ for a reduction approach A with respect to another approach B at a certain ensemble size x were analyzed for all individual targets on the basis of the following equation:

$$\rho_{A,B}(x) = \begin{cases} 0, & \text{if } \pi_A(x) = 0 \wedge \pi_B(x) = 0 \\ \dfrac{\pi_B(x) - \pi_A(x)}{\max(\pi_A(x), \pi_B(x))}, & \text{otherwise} \end{cases}$$

where π_X denotes the difference of critical ligands between reduction approach X and the optimal ensemble performance.

Figure 5 depicts the distribution of $\rho(x)$ across all 182 individual targets for various ensemble size thresholds. The results illustrate that the ligand-based reduction approach yields an improvement compared to random selection for the majority of targets especially when the ensemble size is limited to smaller numbers. For increasing ensemble size thresholds, the performance of the best quarter of targets improves constantly and reaches its optimum at an ensemble size threshold of six (cf. Figure S10). The $\rho(x)$ median also slightly increases until a threshold of seven but decreases again for higher ensembles size thresholds. The number of targets for which the ligand-based reduction performance is worse than random ($\rho(x) < 0$) also rises with increasing ensemble size threshold. For the RMSD clustering-based approach, there are almost equally sized fractions of targets for which the performance is better and worse than random ensemble selection when low ensemble size threshold are applied. Nevertheless, its performance improves for higher thresholds. The direct comparison of interaction-based and RMSD-based ensemble reduction confirms these trends. The interaction-based approach performs superior for single structures and small ensembles, while the RMSD-based reduction works slightly better at high ensemble size thresholds. An obvious explanation for the decreasing performance of the interaction-based approach at higher ensemble size thresholds is that it does not necessarily exhaust the maximum ensemble size if an increase of ensemble size does not increase the performance on the trainings set. If the test set contains binding geometries which differentiate substantially from those on the test set, this might be a drawback in comparison to the RMSD clustering-based reduction and the random selection, which both produce exhaustive ensembles, as larger ensembles naturally increase the probability to match the native binding geometry. While this is a general limitation of the interaction-based approach, the

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

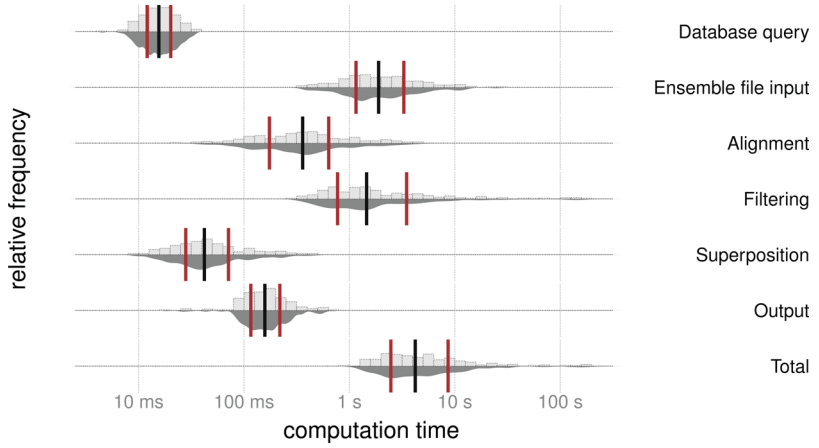


Figure 6. Computation times for all major steps in the SIENA's ensemble generation process for the reduced ensembles from the NBSE set. The distribution of the computation times across all NBSE targets is depicted as an approximated density on the lower half and in form of a histogram on the upper half of each subfigure. The black and red bars indicate the median values and the 25% and 75% quantiles, respectively.

performance of all approaches at higher ensembles sizes is less relevant for most application scenarios since larger ensembles were often observed to increase false positive rates. Therefore, commonly recommended ensemble sizes vary between three and five structures.[31,32]

In general, the results illustrate the high relevance that proper modeling of protein flexibility has on the prediction accuracy of ligand binding.

Computation Times. For all experiments, the computation time of all major steps in the ensemble generation workflow were measured on a single core of an Intel Core i7-2600 with 3.4 GHz and 16 GB of memory.

Figure 6 summarizes the computation times for the generation of the reduced ensembles by the IBER procedure for all NBSE targets using an ensemble size threshold of five. For the majority of targets, the most time-consuming step is the ensemble file input and structure initialization followed by the conformation selection process. The database query, alignment, and superposition steps only require a minor portion of the total computation time. The median of the computation times for all NBSE targets is four seconds. The mean computation time, which is more strongly influenced by considerably higher ensemble reduction runtimes for a few outliers having large ligand sets, is 11 s. If SIENA is solely used for conformation enrichment without ensemble reduction, the mean value reduces to six seconds (see Figure S11), while the median is still four seconds. In this case the file output becomes another expensive step, as more data needs to be written. Computation times for the 10 Astex Non-native targets show similar results (cf. Figures S12−S18).

Limitations. While the usage of ensembles generated from experimental structures has the great advantage of supplying verified protein conformations, it certainly has the limitation that it is only applicable for well studied target proteins. Possible conformations which are not covered by the set of accessible structures can not be considered. Alternatively, flexible protein models can be created on the basis of molecular dynamics[11,24,26] and normal-mode analysis.[47,48] Given the necessity of predictive analyses of protein flexibility, these tools and also their evaluation could clearly benefit from SIENA's ability to produce large, consistently compiled ensembles, which can be applied as benchmark data sets. A further limitation is that the interaction based ensemble reduction is currently not yet suited to consider covalently bound ligands.

CONCLUSION

SIENA has been developed for the automatic enrichment, preprocessing and case-specific selection of flexibility information for protein binding sites. We demonstrated that its initial search process is able to extract exhaustive ensembles of alternative conformations. Using a variety of different filters further facilitates to reduce these ensembles to meet user specific requirements. For instance, this allows to eliminate structures with missing, modified, or mutated residues which is especially relevant for applications which rely on the structural consistency of the applied ensemble. Furthermore, a new interaction-based ensemble reduction algorithm for the generation of small sets of relevant binding site conformations was proven to cover known ligand interaction geometries better than equally sized ensembles generated on the basis of RMSD clustering and in most cases also exhibited an improved ensemble compatibility to unconsidered ligands. It was also shown, that a complete ensemble generation process with SIENA on the basis of a comprehensive database like the PDB takes only a few seconds, which is competitive to established sequence search engines like BLAST. However, in contrast to BLAST, this does not only comprise the detection of similar sequences but also a complete ensemble preprocessing including the verification of structural conformity, conformation selection, and superposition. Moreover, its efficiency and fully integrated workflow enable a quick access to known binding site flexibility of a desired target as well as a consistent compilation of large up-to-date ensemble data sets which, e.g., can be used for benchmarking or large scale flexibility analyses. In summary, SIENA is a well-suited tool for all application scenarios which require an accurate preprocessing of protein binding site conformations.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00588.

Additional data series concerning the mutual ensemble coverage of SIENA, BLAST, and the Astex Non-native Set, further runtime statistics, additional results of the

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

ensemble reduction evaluation, a description of the applied SIENA settings, a detailed case classification of structures with modified residues, a list of those targets which form the NBSE set, and a collection of molecules that are not considered as ligands (PDF)

The ASCONA software is available for Linux OS via the Internet at http://www.zbh.uni-hamburg.de/ascona. The SIENA method described here is available as part of the Proteins*Plus* Web service at the Center for Bioinformatics Hamburg at http://www.zbh.uni-hamburg.de/ProteinsPlus. Further information on SIENA, for example the superimposed structures of all 182 protein binding site ensembles, can be downloaded from http://www.zbh.uni-hamburg.de/siena.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The project is part of the Biokatalyse2021 cluster and is funded by the Federal Ministry of Education and Research (BMBF) under grant 031A1838. We would like to thank Rainer Fährrolfes for the integration of SIENA into the Proteins*Plus* server.

REFERENCES

(1) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft Docking and Multiple Receptor Conformations in Virtual Screening. *J. Med. Chem.* **2004**, *47*, 5076−5084.

(2) Barril, X.; Morley, S. D. Unveiling the Full Potential of Flexible Receptor Docking Using Multiple Crystallographic Structures. *J. Med. Chem.* **2005**, *48*, 4432−4443.

(3) 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.

(4) Bottegoni, G.; Rocchia, W.; Rueda, M.; Abagyan, R.; Cavalli, A. Systematic Exploitation of Multiple Receptor Conformations for Virtual Ligand Screening. *PLoS One* **2011**, *6*, e18845−e18845.

(5) Li, Y.; Kim, D. J.; Ma, W.; Lubet, R. A.; Bode, A. M.; Dong, Z. Discovery of Novel Checkpoint Kinase 1 Inhibitors by Virtual Screening Based on Multiple Crystal Structures. *J. Chem. Inf. Model.* **2011**, *51*, 2904−2914.

(6) Cosconati, S.; Marinelli, L.; Di Leva, F. S.; La Pietra, V.; De Simone, A.; Mancini, F.; Andrisano, V.; Novellino, E.; Goodsell, D. S.; Olson, A. J. Protein Flexibility in Virtual Screening: The BACE-1 Case Study. *J. Chem. Inf. Model.* **2012**, *52*, 2697−2704.

(7) Korb, O.; Olsson, T. S.; 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.

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

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

(10) 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.

(11) 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.

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

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

(14) 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*, 179−89.

(15) An, J.; Totrov, M.; Abagyan, R. Pocketome Via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752−761.

(16) Kufareva, I.; Ilatovskiy, A. V.; Abagyan, R. Pocketome: An Encyclopedia of Small-Molecule Binding Sites in 4D. *Nucleic Acids Res.* **2012**, *40*, D535−D540.

(17) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(18) Monzon, A. M.; Juritz, E.; Fornasari, M. S.; Parisi, G. CoDNaS: A Database of Conformational Diversity in the Native State of Proteins. *Bioinformatics* **2013**, *29*, 2512.

(19) Juritz, E. I.; Alberti, S. F.; Parisi, G. D. PCDB: A Database of Protein Conformational Diversity. *Nucleic Acids Res.* **2011**, *39*, D475−D479.

(20) 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.

(21) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; 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.

(22) Daura, X.; van Gunsteren, W. F.; Mark, A. E. Folding-Unfolding Thermodynamics of a β-Heptapeptide from Equilibrium Simulations. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 269−280.

(23) 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.

(24) Amaro, R. E.; Baron, R.; McCammon, J. A. An Improved Relaxed Complex Scheme for Receptor Flexibility in Computer-Aided Drug Design. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 693−705.

(25) Xu, M.; Lill, M. A. Significant Enhancement of Docking Sensitivity Using Implicit Ligand Sampling. *J. Chem. Inf. Model.* **2011**, *51*, 693−706.

(26) Campbell, A. J.; Lamb, M. L.; Joseph-McCarthy, D. Ensemble-Based Docking Using Biased Molecular Dynamics. *J. Chem. Inf. Model.* **2014**, *54*, 2127−2138.

(27) Amaro, R. E.; Schnaufer, A.; Interthal, H.; Hol, W.; Stuart, K. D.; McCammon, J. A. Discovery of Drug-like Inhibitors of an Essential RNA-Editing Ligase in Trypanosoma Brucei. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17278−17283.

(28) 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.

(29) 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.

(30) Ben Nasr, N.; Guillemain, H.; Lagarde, N.; Zagury, J.-F.; Montes, M. Multiple Structures for Virtual Ligand Screening: Defining Binding Site Properties-Based Criteria to Optimize the Selection of the Query. *J. Chem. Inf. Model.* **2013**, *53*, 293−311.

(31) Rueda, M.; Bottegoni, G.; Abagyan, R. Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *J. Chem. Inf. Model.* **2010**, *50*, 186−193.

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259

(32) Xu, M.; Lill, M. A. Utilizing Experimental Data for Reducing Ensemble Size in Flexible-Protein Docking. *J. Chem. Inf. Model.* **2012**, *52*, 187−198.
(33) Pearson, W. R.; Lipman, D. J. Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 2444−2448.
(34) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(35) Kent, W. J. BLAT-the BLAST-like Alignment Tool. *Genome Res.* **2002**, *12*, 656−664.
(36) Bietz, S.; Rarey, M. ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations. *J. Chem. Inf. Model.* **2015**, *55*, 1747−1756.
(37) Urbaczek, S.; Kolodzik, A.; Rarey, M. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *J. Chem. Inf. Model.* **2014**, *54*, 756−766.
(38) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J.Chem.-Inf.Comput.Sci.* **1989**, *29*, 97−101.
(39) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminf.* **2014**, *6*, 12.
(40) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(41) Schneider, N.; Lange, G.; Hindle, S.; Klein, R.; Rarey, M. A Consistent Description of HYdrogen Bond and DEhydration Energies in Protein-Ligand Complexes: Methods Behind the HYDE Scoring Function. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 15−29.
(42) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An Automated Approach for Defining Core Atoms and Domains in an Ensemble of NMR-Derived Protein Structures. *Protein Eng., Des. Sel.* **1997**, *10*, 737−741.
(43) PDB Advanced Search. http://www.pdb.org/pdb/search/advSearch.do?search/new (accessed Sep 16, 2015).
(44) PDB Growth Statistics. http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content/total&seqid/100 (accessed Sep 16, 2015).
(45) Welburn, J. P.; Tucker, J. A.; Johnson, T.; Lindert, L.; Morgan, M.; Willis, A.; Noble, M. E.; Endicott, J. A. How Tyrosine 15 Phosphorylation Inhibits the Activity of Cyclin-Dependent Kinase 2-Cyclin A. *J. Biol. Chem.* **2007**, *282*, 3173−3181.
(46) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. Sc-PDB: A 3D-Database of Ligandable Binding Sites - 10 Years On. *Nucleic Acids Res.* **2015**, *43*, D399−D404.
(47) Cavasotto, C. N.; Kovacs, J. A.; Abagyan, R. A. Representing Receptor Flexibility in Ligand Docking through Relevant Normal Modes. *J. Am. Chem. Soc.* **2005**, *127*, 9632−9640.
(48) Sperandio, O.; Mouawad, L.; Pinto, E.; Villoutreix, B. O.; Perahia, D.; Miteva, M. A. How to Choose Relevant Multiple Receptor Conformations for Virtual Screening: A Test Case of Cdk2 and Normal Mode Analysis. *Eur. Biophys. J.* **2010**, *39*, 1365−1372.

DOI: 10.1021/acs.jcim.5b00588
J. Chem. Inf. Model. 2016, 56, 248−259