Integrating Sequence and Structure Information for Efficient Retrieval and Alignment of Flexible Protein Binding Sites

Stefan Bietz and Matthias Rarey
Center for Bioinformatics, University of Hamburg, Hamburg, 20146, Germany
rarey@zbh.uni-hamburg.de

The consideration of protein flexibility is long known to be one of the most challenging aspects in computational structural biology. Experimental structures are often used to construct flexible protein models or serve as a reference for flexibility predicting techniques. Therefore, the steadily increasing amount of structural data further improves the fundamental basis of these techniques. However, due to inconsistent annotation or structural deviations in the experimental data, alignment techniques are generally required as an essential prepressing step for the usage of multiple experimental protein structures. In principle, various sequence- and structure-based approach have already been developed which can be applied in these [FRCC07, KSK11]. However, their applicability depends on the particular application scenario. Many structure-based approaches like molecular docking, pharmacophore generation, or the investigation of enzymatic mechanisms focus on protein binding sites and neglect the rest of the protein structure for efficiency reasons. For these applications, it is most relevant that the alignment of the active site is highly reliable. Furthermore, purely sequence-based alignment techniques are not always applicable if the binding site is located at a subunit interface of an oligomeric protein, as in these cases, the assignment of corresponding subunits is not necessarily unambiguous. Structure-based alignment methods mostly target the identification of geometrically conserved motifs which complicates the identification of analogous protein regions exhibiting structural flexibility.

We recently introduced ASCONA [BR15a], an automated approach for the detection and alignment of protein binding site conformations. While most other alignment techniques deal with the generation of structure or sequence alignments of rather distantly related proteins, ASCONA puts the accurate detection of highly deviating binding site conformations into focus. Given an arbitrarily defined binding site of a query structure, ASCONA locates all occurrences of the respective query in a target structure and generates a residue-wise mapping in form of a sequence alignment. It also facilitates the generation of multiple alignments of a certain query in case of an oligomeric target structure.

The underlying algorithm is based on a partition of the query binding site into a set of short peptide fragments which are searched in the target sequence using an efficient approximate string matching algorithm. The fragment hits are recombined on the basis of a two-step fragment assembly approach and a geometry measure that analyses the distance and relative orientation of the fragment hits. Since the typical application scenario assumes a high sequence similarity of query and target structures, the fragment matching step can be set up quite strictly, which results in low rate of random (false positive) fragment matches. In turn, this allows for applying a tolerant geometry measure during the fragment assembly and thus facilitates an accurate detection of binding sites with highly deviating conformations. ASCONA was evaluated on the Astex Non-native dataset [VMH$^+$08] and proved to correctly align all contained binding sites including those with considerable structural deviations. A major advantage of ASCONA is that it only needs to search for the protein region of interest, e.g. a ligand binding site or a protein-protein interface, and thus achieves considerably low computation times. For instance, the alignment of a structure from the Astex Non-native dataset took on average 4 milliseconds.

Besides details on its algorithmic background and the evaluation experiments demonstrating its general functionality, we will present further information on sensible application scenarios. For instance, we developed a server for collecting protein binding site ensembles from the PDB [BR15b]. Starting with a user defined query, the search initially extracts structure candidates from a database that has been specially geared to this purpose. In a second step, ASCONA is used to detect appropriate binding sites within the set of candidates. This step highly benefits from ASCONA's accuracy and efficient runtime behavior. The remaining structures can be further filtered to adapt the set of identified binding

site conformations to the user's requirements. This can, e.g., incorporate the application of RMSD thresholds, mutation rate constraints, or the selection of diverse conformations. These filters also depend on an accurate alignment of the query binding site and the ensemble candidates. Finally, the selected conformations are being superimposed on the basis of a common rigid region.

In summary, ASCONA is a perfectly suited tool for the collection and automatic preprocessing of alternative protein binding site conformations and can support any application that relies on an accurate mapping of the residues in the protein binding site.

References

[BR15a] Stefan Bietz and Matthias Rarey. ASCONA: Rapid Detection and Alignment of Protein Binding Site Conformations. *Journal of Chemical Information and Modeling*, 2015. (accepted).

[BR15b] Stefan Bietz and Matthias Rarey. Efficient Search of Experimentally Derived Structures for the Selective Compilation of Protein Binding Site Ensembles. 2015. (in preparation).

[FRCC07] Piero Fariselli, Ivan Rossi, Emidio Capriotti, and Rita Casadio. The WWWH of remote homolog detection: The state of the art. *Briefings in bioinformatics*, 8(2):78–87, 2007.

[KSK11] Singarevelu Kalaimathy, Ramanathan Sowdhamini, and Karuppiah Kanagarajadurai. Critical assessment of structure-based sequence alignment methods at distant relationships. *Briefings in bioinformatics*, 12(2):163–175, 2011.

[VMH+08] Marcel L Verdonk, Paul N Mortenson, Richard J Hall, Michael J Hartshorn, and Christopher W Murray. Protein- ligand Docking against Non-native Protein Conformers. *Journal of Chemical Information and Modeling*, 48(11):2214–2225, 2008.