Bachelor's Degree in Bioinformatics (UPF-UPC-UB)

Final Grade Project

A method to identify common features among similar protein binding sites.

Jordi Busoms

Scientific director: Xavier Jalencas[1], Jordi Mestres[2]

[1]Chemotargets., [2]Systems Phamacology group.

Abstract

Motivation: Unexpected binding of drugs to proteins beyond the intended protein target is one of the main causes of adverse drug reactions. This drug polypharmacology occurs mainly with protein members of the same family as the protein target, but it can also happen with proteins more distantly related yet exposing similar features in their binding cavity. Computational approaches to profiling compounds over thousands of proteins rely mostly on fast ligand-based methodologies. Structure-based methods are still computationally demanding. The main objective of this project is to investigate more efficient structure-based approaches to anticipating the likely binding of small molecules to thousand of proteins without the need to process them all individually. The underlying hypothesis is that members of the same protein expose similar features to small molecules. If we were able to capture these protein family signatures we would be more efficient in identifying the likely families to which small molecules may bind.

Results: We have obtained a fully working algorithm that is able to get groups of similar binding sites from three-dimensional protein structures and generate representative signatures of the entire group of proteins. Those signatures represent the common features exposed by groups of similar proteins and have thus the potential to be used as a more efficient strategy to virtual target profiling.

Supplementary information: Supplementary material and code available at Github link, https://github.com/jordibusoms/Final-Grade-Project

1 Introduction

Many computational tools have been built in order to detect, characterize, and compare protein binding sites[1]. This interest in binding sites can be related to different biological problems apparently unrelated such as predicting protein function or finding distant homologs. To solve those problems using binding sites is possible because, being there where ligands bind, the region tends to be more conserved than other areas, their characteristics being complementary to those found in the ligand. For the same reason, the approaches used are diverse and each of them attempts to solve a different problem. For example, approaches using genomic sequences have been developed to detect binding sites[2]. However, to predict what kind of ligand a protein might bind, its three-dimensional (3D) structure and chemical properties provide valuable information that might be considered.

The main aim of this work was to design a means to compare protein binding sites that allows for, on one side, extract the signatures of similar proteins and, on the other side, identify potential chemoisosteric sites, that is, protein environments present in either phylogenetically related or distant proteins that nonetheless bind to the same chemical fragment[3]. To find all the proteins a fragment or ligand can bind, even if they are not phylogenetically related, the method to compare binding sites cannot rely on evolutionary conservation. Because of that, we needed our method to use chemical and shape properties instead of sequence, as different sequences may actually produce similar cavities. Moreover, such comparison method should be independent of initial location and orientation of the cavities accommodating the ligand, as if the two proteins sharing similar binding sites are unrelated, aligning their backbones will not align their binding sites. Such type of comparison necessarily requires data on the 3D structure of the proteins. In the last few years, the amount of available data in structure databases, such as the Protein Data Bank, has increased drastically[4], providing a high and growing amount of data to work with.

What are the main advantages that predicting chemoisosterism would be offering? As proteins play a role in many biological and chemical processes of the organism, many drugs produce their therapeutic effects by modifying the function of one or several proteins. In addition, chemoisosterism is a significant source of side-effects for those drugs[5]. A study[6] based on data from 1992 to 2002 showed that side effects caused more than 90% of market withdrawals and were the second main cause of project terminations at the clinical phase. In consequence, predicting protein chemoisosterism through binding site comparison might help to avoid side effects on patients, while saving both time and costs in clinical trials. Moreover, if differences between binding sites can be retrieved, modifications could accordingly be done to the drug to prevent its binding to the protein producing the side effect. In the same direction, all discarded drugs could be retrospectively explored to see if their side effects are due to chemoisosterism and, if so, rationally design such modification.

With this objective in mind, our group decided to use a cavity comparison algorithm on all the 3D structures found in the Protein Data Bank[7] database which, with almost 150,000 entries, is the leading public source for protein structures. However, to perform such an amount of pairwise comparisons (150,000*(150,000-1)/2) taking into account the chemical properties of binding sites and their distribution in space, it would be computationally very expensive with current methods. Binding site signatures were developed in this work as a solution to that problem and aimed at extracting the conserved chemical properties of a protein binding site among several PDB structures. This strategy does not only allow for reducing the number of comparisons but also to limit noise and the size of the binding site representation, making comparisons considerably faster and allowing the comparison of binding sites across the entire structural proteome.

In summary, we considered chemoisosterism as a source of possible side effects for drugs that might be reflected by binding site similarity. Our primary objective was to predict it by comparing all different binding sites available in the PDB. However, given the high amount of data, a protocol had to be built to generate signatures that allowed us to reduce the number of comparisons retrieving the relevant information and ignoring the noise. At the same time, faster binding site comparison methods were explored.

2 Methods

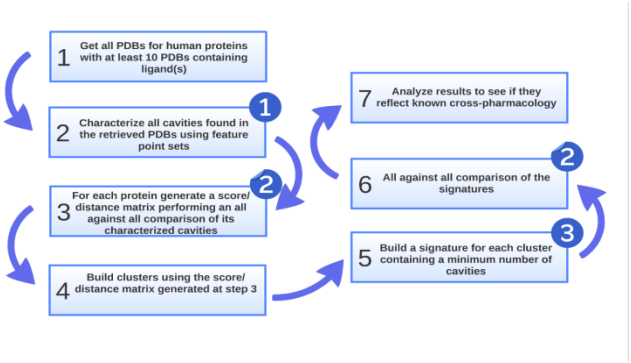

Figure1. General Schema of the protocol used to generate binding site signatures. White labels represent problems that required the development of specific algorithms

As mentioned above, the main objective of this work is to detect cross-pharmacology between proteins from a structural point of view. We choose to use a structural approach because while it is true that similarity on the protein sequences is normally translated to the ability to bind similar molecules, this ability does not necessarily imply sequence similarity. For that reason, the use of sequences to perform this type of searches may loose some cross-pharmacology relations. That also discards some structural approaches based on protein superimpositions as unrelated proteins with different backbones could still expose similar features. Despite this difficulty, different approaches have been used to compare binding pockets[8]. This work is based on the binding site comparison via surface feature points developed by Jalencas and Mestres[3]. This methodology represents binding sites in terms of surface feature points of six different types, namely, hydrophobic, aromatic, hydrogen bond donor, hydrogen bond acceptor, positively charged and negatively charged, which can be compared by a subgraph matching method.

On the downside, the comparison method based on a clique detection algorithm, showed up to be too slow for our objective of detecting cross-pharmacology analyzing all available binding sites in the Protein Data Bank. This method has three main steps. The first one searches for all pairs of feature points of the same type that can be retrieved by taking one from each of the binding sites to be compared. Those pairs are used as graph nodes, and on the second step, they get connected if the distances between their respective integrants are similar for both cavities. Finally, a clique detection algorithm applied and the score is computed by the size of cliques over the square root of the product of the number of feature points on the matching area on both cavities. Cliques are fully connected subgraphs and to find them is an NP-complete problem, which means that there is no efficient algorithm to solve it. Because of that, we devised a new approach described in Figure 1. That protocol consists in building signatures for each different protein cavity and compare them. Signatures are understood as a set of weighted feature points that is representative of a group of similar cavities. By generating signatures, we get some advantages; the first one is that it reduces the number of final comparisons to be performed, as there are fewer signatures than cavities. The second one is that only points that are present in several cavities of the same protein are representative and noise is not usually replicated, so the noise is reduced. In consequence, signatures tend to have fewer feature points than cavities, and that is translated into faster comparisons.

As Figure 1 shows, this new protocol needs to solve three different problems: to characterize protein binding sites, to compare characterized binding sites or signatures, and to build signatures from a group of similar binding sites. Nevertheless, despite only having to solve these three problems, we decided to use two different methodologies to solve one of them. That problem was the comparison of characterized binding sites or signatures, and we decided to do so because, despite the fact that we build signatures to reduce the number of comparisons, building them using not similar cavities makes no sense, so cavities still need to be compared. However, the comparison technique to detect similar enough cavities to build signatures only needs to detect very similar cavities, so it does not need to be as accurate as the cli-

que-based algorithm. In that direction, we tried to get an algorithm that while being accurate enough to detect similar enough cavities, is as fast as possible, exploring several different algorithms. From those explored algorithms (annex1), two of them showed promising results.

Those techniques consisted of a prefiltering step to reduce graph size in the clique-based procedure and a completely new method based on feature points neighbours. Despite being new, neighbours based algorithm shares two characteristics with the clique-based one that may explain why it got promising results while other approaches failed. The first and probably most important one is that both algorithms are able to perform local alignments. The other one is the formula used to compute score once the local matches have been found.

Neighbours based comparison technique

The neighbor's based technique (Figure 2) that was developed to compare cavities was inspired by the "words" methodology used by blast algorithm[10]. That methodology basically tries to find small sequence fragments or "words" which are present on the query sequence and in the target. As cavities are represented by a set of feature points that have no index but a coordinate in space and have type as their only property, they are difficult to compare. Taking that into account, most of the comparison algorithms attempted, neighbours based included, try to map that set of surface feature points into another entity that is easier to compare.

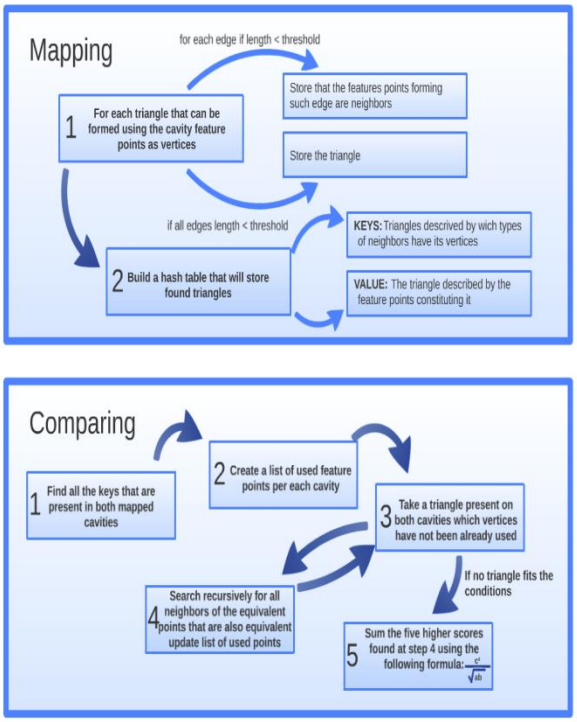

Figure 2. Schema of a method used to compare featured cavities based on neighbours. On step 5 of the comparing method, c represents the common points while a and b represent the total number of points on the matching area on each cavity.

To implement something similar to blast but applied to cavities, the mentioned entity used to represent them must allow local alignments. For that reason, surface feature points can not be characterized by properties that depend on the cavity far ends. Accordingly, the method used to characterize points was an array of six booleans indicating whether the point has or not a feature point of each type nearer than a given threshold. The "words" used were triangles made of near feature points where vertices were sorted accordingly with its representation. However, a problem arises, given that there are cases where one same "word" is found in many different locations of the same binding site. That makes that each "word" location of one cavity could correspond to any "word" location on the cavity being compared. That exponentially raises the number of possibilities to be checked and, subsequently, the computation time. Such a problem was solved by a parameter fixing the maximum number of equal "words" in the same binding sites was set to five. On the same direction, a fourth point was also used to describe "words". That fourth point was the one nearest to the triangle center that does not belong to the triangle itself, and it helps to differentiate words that were previously taken as equivalent. Once the cavities are mapped and the words retrieved, the method to find "word" matches was a hash table. Those matches give an excellent reference to start searching for equivalent points between both cavities. While comparing two cavities, that reference is used and all the neighbours of the triangle used as "word" are explored. If some of the neighbors of one "word" are equivalent to the ones of the matching "word", those are included as matching points. Matching points neighbours are then examined iteratively until the matching areas have no more matching neighbors.

The resulting matching areas are treated as the cliques on the clique-based technique: the total matching points are squared and divided by the square root of the neighbors considered in cavity A times the neighbors considered in cavity B. If many different matching regions found only the one scoring higher is taken into account. Notice that the fact that we use the scoring higher found area is decreasing a lot the effect of the maximum equal "words" parameter on accuracy. That is so because the scoring higher matching area will most likely be the one containing more matching "words" so even if some words are not considered that area is likely to be found.

Prefiltering in comparison method

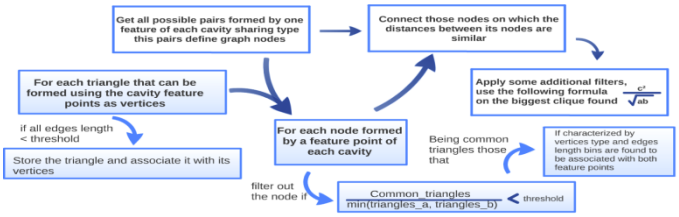

Figure 3. General schema on the clique-based technique showing new prefiltering method detailed and marked with a blue background. On the clique-based procedure (white background) last step, c represents the common points while a and b represent the total number of points on the matching area on each cavity.

Prefiltering in comparison method attempts to quicken the clique-based technique by reducing the size of the graph were cliques must be searched. As clique-detection algorithms are NP-complete, to reduce the graph number of nodes reduces a lot the computation time. Given that the nodes of the graph on the clique-based technique are formed by a pair of feature points

which should be equivalent, the idea is to compare them using some faster method and discard those that are too different. The faster technique was based on triangles and already been tried with good results on similar projects[11]. However, using it alone instead of as a prefilter did not show the desired results for comparing featured cavities on our test set. Such a method was to find all the triangles with edges sizes below a threshold on which the described feature point is involved. The triangles are classified by vertices types and edges size (binned). If the number of common triangles over the maximum possible common triangles (total number of triangles associated with the feature point with fewer triangles) is below a threshold, the node is filtered out.

Binding Site Clustering

For a group of featured binding sites, a score matrix can be obtained using one of the explained comparison methods. Such a score matrix or its equivalent distance matrix can be used as the input for many different clustering techniques. Among them, we firstly discarded those that need to know the number of clusters ab-initio such as k-means. Then we tried DBSCAN[12], but we finally had to drop it because, as it is density based, its final clusters depend a lot on the amount of data available. Being that the case two cavities which not being clustered together can end on the same group if more cavities are included. That made us discard this method because to determine whether or not two cavities should be taken as equivalents, the presence of other cavities should not influence. So, once discarded density based approaches, we tried hierarchical clustering[13]. Once a dendrogram is build, clusters were retrieved using a threshold, and each branch crossing that threshold formed a group.

Building signatures

Either the neighbours based technique or the clique-based one with or without prefilter can obtain a score after comparing two featured cavities. Additionally, those methods can also generate a list of equivalent points between those cavities that can be used to overlap one cavity with the other. To build signatures, we used that overlapping method to overlap all similar binding sites (those clustered together on step 4 in general schema -Figure1-) on the same template. The template was one of those on the cluster, chosen without a specific criterion. We considered building a more elaborate template independent method to overlap featured cavities but given that the overlap is done between cavities of the same protein that need to be similar, using a template would give good enough results. Once the sets of feature points characterizing similar binding sites have been overlapped, signatures can be built setting feature points on the areas in which many cavities have feature points. That can be easily achieved by clustering and getting groups centers if groups can be assumed to occupy a limited uniform area.

To cluster feature points into groups with those characteristics, we used a greedy algorithm trying to make groups with as many individuals as possible. This simple approach ensures distant points never to be clustered together as it could happen on other clustering techniques based in point densities.

The developed algorithm starts considering all points as possible group centers. For each of them computes what feature points would be in that particular group. Points to be included must fit three conditions: to be of the same type as the group center, to be closer to the center than a given threshold and to be the only feature point coming from its cavity in that group. Then the algorithm chooses the biggest group from the possible ones, erases included points from other possible groups and iterates until no more points left. Noise will generally be expressed by singletons or small groups. As one of the objectives of building signatures was to reduce such noise, groups with less than 5 points were not taken into account. The centers were recomputed to be the geometrical mean of all included points in the group, so technically the maximum distance between a point and its group center is twice the distance threshold the algorithm used. The numbers of features per group were used to weigh signature feature points.

3 Results and Discussion

We have been able to develop the protocol to build and compare binding site signatures, and we tested it on a small set of proteins. Before launching that protocol for a bigger dataset, we wanted to ensure it was working correctly, so we designed some tests to evaluate the newly code.

Binding sites comparison methods evaluation

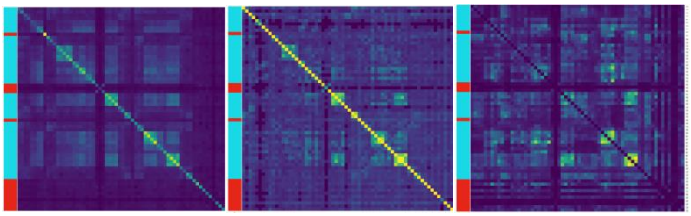

Figure 4. Heatmap representation of the score matrices obtained using Clique-based algorithm, Clique-based algorithm with the prefilter and neighbours based algorithm respectively on the same set. On the color bar, red represents noise/random cavities and blue, similar cavities. Given the small space available labels and scale have been hidden, see full-size version among with other methods results in annex1.

Most of the tried methods to compare cavities were discarded as a visual inspection of the resulting score matrix showed they were performing poorly. The explained ones, prefilter in the clique-based process and neighbours based were by far the ones obtaining most similar matrix to the Clique-based one. This can be seen in Figure 4. To evaluate the new developed methods and using the clique-based comparison method as a reference, adjusted Rand score[14] can be applied to see how equivalent the generated dendrograms are. We computed the score with different numbers of clusters and using those clusters obtained using the clique-based algorithm as reference. We applied that analysis to two sets. The first one was a manually extracted set, including 54 PARP cavities and ten random ones. PARP is a protein superfamily widely studied because it has been found to be related to tumors[15]. The second set was formed with all PDB structures associated with the protein SETD7 (UniProt id Q8WTS6), a transcription initiation factor with two ligand sites.

Looking at the resulting graph (Figure 5), it may be difficult to get a firm conclusion due to the extreme variations in score among the different number of clusters. For that reason, I focus my attention where the number of expected clusters occurs. That would be between 10 and 20 for PARP (10 to differentiate random cavities + noise in the set + possible PARP classifications) and around 3 for SETD7 (2 different cavities + 1 noise coming

from glycerol taken as a ligand). Once focusing on those regions, we can see that despite the visual representations of the score matrices (Figure 4) both methods were obtaining acceptably similar scores, the prefiltering method is achieving a better performance. Having that into account and given that when comparing cavities of a single protein (SETD7), the expected clusters were retrieved identically, we decided to use the prefiltering technique while building the clusters that will generate signatures and the clique-based method without prefilter to compare different signatures.

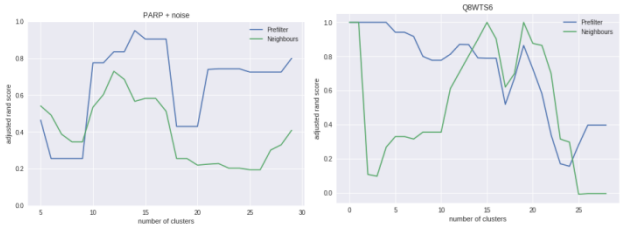

Figure 5. adjusted rand score computed using clusters obtained using the score matrix generated by the clique-based method as a reference. The adjusted Rand score can range from -1 to 1, exact clusters would get 1 and random ones 0.

Moreover, the neighbours based technique tends to be slower for comparisons of similar cavities (annex 2), that should be very frequent when applying this protocol, given that cavities from the same protein are being compared. However, given that it is very fast while comparing different cavities, it could be interesting to use in some other situations. For example, it could be used to perform a search on compatible cavities for a ligand, or in a virtual screening. In that case, most comparisons would be scoring low and consequently the method would be faster. Once hits were found, energy computation could be used to obtain a definitive score.

Validation on the signatures

The first we did to validate signatures was a visual inspection of the signature obtained for the PARP subset. Using a pymol visualization we checked that original cavity feature points were close to the signature feature point they formed and that signature feature points near the ligand rings generally got higher scores than those near the tail or further apart regions, so the area is more conserved among featured binding sites. That is expected as the tail of the ligand varies highly among PARP structures while it is conserved in the other region.

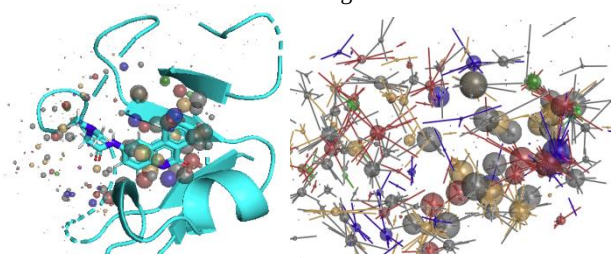

Figure 6. Pymol representation of a PARP signature. Spheres represent signature weighed feature points, with the size showing the score. On the first image, the cavity of the structure used as a template for the overlap, 3gey, and its ligand were included. On the second one, lines were drawn connecting signature feature points with the feature points from the different binding sites that build it.

However, to prove that the signatures were being properly built, we compared it with the cavities used to build it to ensure it is representative of them. On this checking, we used all available PARP(1,2,3,5A and 14) 3d representations in the PDB. We excluded other PARP molecules because of a lack of structures and PARP5B because it got so many that would have slowed down the whole test protocol too much. Then we applied the protocol to obtain signatures, so per each PARP molecule several clusters were obtained and each of those groups generated a signature object. Afterwards, for each signature, we used a manually chosen threshold to classify all used cavities in addition with some randomly chosen ones as belonging or not to that signature. Computing the ratio of times it chosen well over the total number of comparisons, the signature obtaining lower ratio got 0.9466, which confirmed that the signatures retained the information from the binding sites that were used to build them.

Full protocol applied to a diverse set

Finally, after some approximations on the needed time (annex 3), we decided to try our full protocol on a diverse set of proteins. With that objective, we choose 25 proteins from SwissProt having more than 20 and less than 40 associated PDBs and added PARP1. The result was a scoring matrix(annex 4) that showed low scores between different signatures. The largest score was obtained between the signature of Histone-lysine N-methyltransferase SETD2 (Q12962) and Histone-lysine N-methyltransferase SETD7 (Q8WTS6), a result that was expected as they are proteins from the same family. No other significant similarities were found for the remaining signatures.

4 Conclusions

With our current results, we can say that our protocol can build signatures that represent a group of similar binding sites. Those signatures keep the crucial information from those sites while reducing their noise . To group binding sites, the clique-based method was improved with a prefiltering technique that enhances its speed. That also implies that the way used to overlap binding sites, although not being template independent performs correctly, and the way of clustering feature points also does although being greedy. On the other hand, they show that the neighbours based technique is not accurate enough to build the needed groups for the proposed protocol. However, being faster as it is, it could be considered for other tasks, if accuracy were less important or bad results could be filtered out.

Next steps should include the application of the developed protocol to the whole PDB database and to detect cross pharmacology between unrelated proteins and see if it is a known phenomenon or could be related to known side effects. The fast newly developed neighbors based comparison algorithm could be applicable in some other problems where accuracy can be traded by speed such as in virtual screening.

References

1. Henrich, S. *et al.* Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.* **23**, 209–219 (2010).

2. Kel, A. E. *et al.* SITEVIDEO: a computer system for functional site analysis and recognition. Investigation of the human splice sites. *Comput. Appl. Biosci.* **9**, 617–627 (1993).
3. Jalencas, X. & Mestres, J. Chemoisosterism in the proteome. *J. Chem. Inf. Model.* **53**, 279–292 (2013).
4. Bank, R. P. D. RCSB PDB - Content Growth Report. Available at: http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total. (Accessed: 12th June 2019)
5. Reddy, A. S., Srinivas Reddy, A. & Zhang, S. Polypharmacology: drug discovery for the future. *Expert Review of Clinical Pharmacology* **6**, 41–47 (2013).
6. Schuster, D., Laggner, C. & Langer, T. Why Drugs Fail - A Study on Side Effects in New Chemical Entities. *Antitargets* 1–22 (2008). doi:10.1002/9783527621460.ch1
7. Burley, S. K. *et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2019).
8. Ehrt, C., Brinkjost, T. & Koch, O. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLOS Computational Biology* **14**, e1006483 (2018).
9. Butenko, S. & Wilhelm, W. E. Clique-detection models in computational biochemistry and genomics. *European Journal of Operational Research* **173**, 1–17 (2006).
10. Altschul, S. Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215**, 403–410 (1990).
11. Weill, N. & Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein−Ligand Binding Sites. *Journal of Chemical Information and Modeling* **50**, 123–135 (2010).
12. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN Revisited, Revisited. *ACM Transactions on Database Systems* **42**, 1–21 (2017).
13. Rafsanjani, M. K., Varzaneh, Z. A. & Chukanlo, N. E. A Survey Of Hierarchical Clustering Algorithms. *Journal of Mathematics and Computer Science* **05**, 229–240 (2012).
14. Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**, 193–218 (1985).
15. Amé, J.-C., Spenlehauer, C. & de Murcia, G. The PARP superfamily. *BioEssays* **26**, 882–893 (2004).