JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

SiteBinder: An Improved Approach for Comparing Multiple Protein Structural Motifs

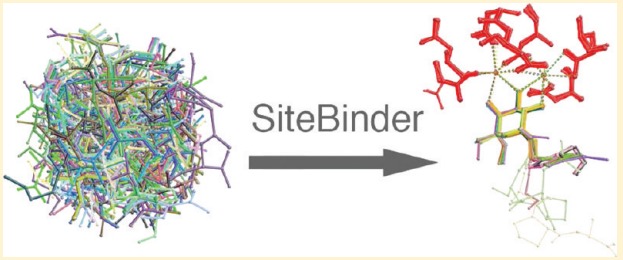
David Sehnal,[†] Radka Svobodová Vařeková,[†,*] Heinrich J. Huber,[‡] Stanislav Geidl,[†] Crina-Maria Ionescu,[†] Michaela Wimmerová,[†] and Jaroslav Koča[†,*]

[†]National Centre for Biomolecular Research, Faculty of Science and CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic

[‡]Centre of Systems Medicine, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, 123 St Stephens Green, Dublin 2, Ireland

Supporting Information

ABSTRACT: There is a paramount need to develop new techniques and tools that will extract as much information as possible from the ever growing repository of protein 3D structures. We report here on the development of a software tool for the multiple superimposition of large sets of protein structural motifs. Our superimposition methodology performs a systematic search for the atom pairing that provides the best fit. During this search, the RMSD values for all chemically relevant pairings are calculated by quaternion algebra. The number of evaluated pairings is markedly decreased by using PDB annotations for atoms. This approach guarantees that the best fit will be found and can be applied even when sequence similarity is low or does not exist at all. We have implemented this methodology in the Web application SiteBinder, which is able to process up to thousands of protein structural motifs in a very short time, and which provides an intuitive and user-friendly interface. Our benchmarking analysis has shown the robustness, efficiency, and versatility of our methodology and its implementation by the successful superimposition of 1000 experimentally determined structures for each of 32 eukaryotic linear motifs. We also demonstrate the applicability of SiteBinder using three case studies. We first compared the structures of 61 PA-IIL sugar binding sites containing nine different sugars, and we found that the sugar binding sites of PA-IIL and its mutants have a conserved structure despite their binding different sugars. We then superimposed over 300 zinc finger central motifs and revealed that the molecular structure in the vicinity of the Zn atom is highly conserved. Finally, we superimposed 12 BH3 domains from pro-apoptotic proteins. Our findings come to support the hypothesis that there is a structural basis for the functional segregation of BH3-only proteins into activators and enablers.



INTRODUCTION

Nowadays, a large amount of information about the 3D structure of proteins is available, and more and more structures are being solved every year because of advances in experimental techniques and their increased availability. This amount of data provides the opportunity to compare large sets of protein structural motifs like binding sites, secondary structure elements, cavities, and tunnels. Such analyses can help identify the main characteristics of important protein motifs. The obtained characteristics can subsequently be used as patterns in drug discovery,[1,2] to understand the relationship between a protein's structure and its function and even predict its function,[3−5] to classify proteins,[6,7] to identify evolutionary relationships between proteins,[8−10] etc. Collecting large sets of protein structural motifs is a fairly simple task. This task can be accomplished by employing available software tools or in-house scripts that retrieve data from structural databases on the basis of primary or secondary protein structure queries. The more sophisticated challenge is to perform the comparison of these large sets of protein structural motifs, as this requires specifically adapted algorithms and software tools. Such a comparison is a particular topic because, on the one hand, these motifs are small compounds, but on the other hand, the motifs are parts of proteins. To our knowledge, no software tool available to date can process hundreds of protein structural motifs at one time and allow for a straightforward comparison within these large sets of structures. Therefore, our goal was to develop and implement a new methodology for comparing large sets of protein structural motifs in an efficient, flexible, and intuitive manner.

The comparison of 3D structures is a complex topic that can be divided into several subtopics. We distinguish between methods that compare compounds with identical (or very similar) 2D structure, as opposed to methods dealing with compounds for which the 2D structure differs significantly. The term "2D structure", as it is introduced in chemoinformatics,[11] refers to the topology of the molecule, meaning the nature and connectivity of the atoms contained in the molecule. We also

Special Issue: 2011 Noordwijkerhout Cheminformatics

Received: September 19, 2011
Published: February 1, 2012

differentiate between the methods on the basis of the type of molecules they process—organic molecules, proteins, or protein motifs.

Organic molecules with different 2D structures can be compared by two principal methods, namely, the rigid body approach and the flexible body approach.[12,13] Rigid body methods[12,14] keep the structure of both molecules fixed and try to find an alignment by maximizing some kind of volume overlap (i.e., van der Waals overlap, electron density overlap, electrostatic potential overlap, etc.). The overlap optimization methods range from simplex optimization, gradient optimization, and Fourier space methods to Monte Carlo optimization. Flexible (or semiflexible) body methods[13,15] change the structure of one molecule during the comparison, thus simulating the process of how the molecule adapts its shape when undergoing a chemical reaction.

The comparison of proteins with different 2D structures can be classified as global or local.[16] The algorithms and available software for both of these approaches were reviewed by Gherhardini et al.[17] Global comparison approaches use various algorithms, such as dynamic programming,[18] double dynamic programming,[19] branch and bound approach,[20,21] subgraph isomorphism,[22,23] or extension of seed matches.[24] Global comparison is used to classify protein structures and to identify evolutionary links between distant homologues. Nevertheless, the function of a protein usually depends more on the identity and location of a few residues comprising the active site than on the overall structure. In order to directly analyze and compare the residues involved in protein function, local (as opposed to global) structural comparison methods have been developed. These methods focus on detecting a similar 3D arrangement of a small set of residues, possibly in the context of completely different protein structures. Local structure comparison approaches are mainly based on algorithms that employ geometric hashing,[25,26] subgraph isomorphism,[27,28] recursive search connected with the branch and bound algorithm,[29,30] and graph-based heuristics.[31] To identify local similarities within two entire protein structures such algorithms can be applied without any a priori assumption or by using a predefined structural template to screen a structure. The structural templates can be user defined.[32] A special case of local structure comparison is searching for a structural motif in a protein by comparing the motif with a relevant part of the protein. These approaches are reviewed in a recent paper.[33]

The development of comparison methods for protein structural motifs with different 2D structures has become an important topic of research within the past few years.[34,35] These comparisons are, among others, necessary for the functional annotation of proteins.[36,37] General purpose software tools able to compare all types of molecules with different 2D structure (i.e., organic molecules, proteins, and protein motifs) are also available (e.g., Bauer et al.[38]).

Superimposition or superposition[16] is the comparison of molecules with identical (or very similar) 2D structures. Superimposition can be applied to study different conformers of one molecule, and these conformers can be obtained from experiment, from molecular dynamics simulations, or from different databases of 3D structures. Likewise, superimposition is often useful to study substructures that were obtained by the analysis and comparison of the 2D structure of molecules or the primary structure of proteins. Superimposition approaches are similar for organic molecules, proteins, and protein motifs.

In brief, superimposition consists of several interdependent stages.[39] First, it is necessary to find the correspondence between the atoms coming from different structures. We will refer to this first step, as well as to its results, as atom pairing or simply pairing. Using an atom pairing is necessary so that the structures can be processed as sequences of points in the 3D space. In the second step, the sets of paired 3D points are fitted together as tightly as possible by a geometrical transformation. We will refer to this step as optimal fitting because its final result gives the coordinates of the superimposed structures. The last phase of the superimposition is to evaluate the quality of the fit. This is done by computing the root-mean-square deviation (RMSD) between the sets of 3D coordinates belonging to the structures that have been superimposed. We further discuss the currently available methodology for performing the steps of atom pairing and optimal fitting.

From the mathematical point of view, pairings are bijections, which are functions where every element from the first set is assigned to exactly one element from the second set. For structures with n atoms, $n!$ such bijections can be constructed, and therefore, $n!$ pairings may exist. It is desirable to find the best pairing, meaning the pairing that will eventually lead to the lowest RMSD between the superimposed structures. Finding the best pairing requires testing all constructed pairings and is therefore very time demanding. Nevertheless, an incorrect atom pairing can lead to a poor superimposition. There are several heuristics and algorithms to solve this problem, such as implicit pairing,[40,41] employing sequence alignment,[42,43] systematic approach,[44] or subgraph matching.[45,46] We briefly describe these below.

Implicit pairing associates atoms with the same index or position (i.e., pairing the i-th atom of the first molecule to the i-th atom of the second molecule). Pairing atoms by this algorithm is extremely fast. An additional advantage is that the subsequent fitting will only be performed once because only one pairing is produced. However, implicit pairing is suitable only when the atoms in both molecules are indexed or ordered identically, as in the case of conformers resulted from molecular dynamics simulations. Many state of the art programs that offer the superimposition of organic molecules (e.g., Chimera,[41] VMD,[47] Gromacs,[40] gOpenMol,[48] Pymol[49]) use implicit pairing.

Employing sequence alignment provides an improvement on the implicit pairing approach. First, the sequence alignment is performed by a selected algorithm (e.g., Needleman and Wunsch alignment,[50] ICM ZEGA alignment,[51] etc.). Afterward, the atoms from the aligned residues are paired using an implicit pairing. This approach is applicable only for the superimposition of proteins or protein sequences with a reasonable degree of sequence similarity. Several drug design packages (e.g., MOE,[42] Discovery Studio,[52] ICM,[43] etc.) implement this approach.

The systematic approach finds all possible pairings and is therefore very robust. However, because the fitting will have to be performed for a large number of pairings, this method is time consuming and therefore useful mainly for small molecules. It can be sped up by backtracking,[53] a procedure that is able to discard possible solutions as soon as they appear unfeasible. Further decrease in computational complexity can be achieved by pairing only atoms that have corresponding chemical element symbols and/or come from comparable chemical neighborhoods.

Subgraph matching, which was originally developed for processing molecules with different 2D structure, can also be used for finding a relevant pairing (reviewed by Raymond et al.[46]).

This approach identifies the largest possible atom sets that can be superimposed.

When an atom pairing has been found, the sequences of paired 3D points can be fitted by performing a geometrical transformation (composition of a translation and a rotation in the 3D space). Finding the transformation that will lead to the optimal fit is a fairly cumbersome task. An iterative solution to this problem was published by McLachlan et al.,[54] while a closed form solution that utilizes rotation matrices was published by Kabsch et al.[55] This rotation matrix approach was later reformulated using quaternion algebra. Many authors over the past 20 years have "rediscovered" the application of quaternions in the superimposition of 3D points (i.e., Horn,[56] Diamond,[57] and Kearsley[58]). However, within the community of computational chemists and biologists, quaternions were introduced by Coutsias et al.[39] and are still a topic of research.[59] All the closed form solutions have linear space and time complexity in the number of atoms. These solutions work by translating both structures to their common origin and then using singular value decomposition in the case of rotation matrices or eigenvectors in the case of quaternions.

Superimposition can be performed for two or more structures at once, depending on the nature of the investigation. Superimposing two molecules or motifs is a very useful task if the purpose is the in-depth structural comparison and characterization of the two compounds under investigation. Many software tools offering the superimposition of two compounds are available,[40,41,47−49] all using implicit pairing. Nevertheless, one often needs to compare the structures of tens or hundreds of compounds at a time in order to find structural trends or peculiarities. In this case, it is necessary to perform a multiple superimposition, which is a fairly more complex procedure than the superimposition of only two structures at a time. The quality of a multiple superimposition procedure can be measured using the generalized RMSD,[60] which is the average RMSD between all pairs of structures. Another possibility is to compute the RMSD between each structure and the calculated average structure and then average these RMSD values over all structures.[61] A naive approach to this problem is to pick one of the structures and superimpose all structures to this chosen one. A quadratic complexity algorithm to this problem was published by Konagurthu et al.[60] and is used, for example, in Pymol.[49] A more advanced approach is to superimpose all pairs of structures, order the pairs by the quality of the superimposition, and then superimpose the structures to an iteratively computed average.[6] An improved approach to this problem, with nearly linear complexity (in the number of structures), was published by Eidhammer et al.[16] and later generalized by Wang et al.[61] This method is based on iteratively superimposing each structure onto the average model of the structures superimposed in the previous step until a stable configuration is reached.

In this work, we focus on the comparison of large sets of protein structural motifs. Such large sets are generally collected in an automated fashion by querying the primary or secondary structure of proteins and will thus consist mainly of motifs with similar 2D structure. The possibility to perform the multiple superimposition of a large number of protein motifs with similar 2D structure would open the door to innovative thinking. One could find meaningful structural trends or peculiarities that could identify evolutionarily related proteins or could explain and even predict function and activity related features of known or engineered proteins.

To our knowledge, no implementation of such a methodology is available to date, even though many state of the art software packages offer the possibility to superimpose protein structures to various extents. Thus, our goal was to fill in this gap and to develop and implement a methodology for superimposing large sets of protein structural motifs in an efficient, flexible, and intuitive manner, so as to fuel inquisitiveness and creativity in the investigation of protein structure and function. A challenging aspect of protein structural motif superimposition is that the motifs need not refer only to linear protein subsequences but may also consist of the 3D surroundings of residues or sequences, binding sites of metals or sugars, or any other selected parts of protein 3D structure. This means that some of the superimposed motifs may not have any sequence similarity. Our methodology guarantees the best superimposition even in such cases.

METHODS

When performing the superimposition of two protein structural motifs, one faces two challenges. One challenge is to find the best pairing of chemically corresponding atoms from the first and second motif. This pairing establishes which atoms from the first motif should be fitted to which atoms from the second motif in the optimal fitting phase. The other challenge is to calculate the geometrical transformation that optimally fits the structures of the two protein motifs together.

In our methodology, we address the first issue by a systematic approach employing heuristics tailored to proteins (described in detail below) and the second by using a state of the art quaternion algebra approach.[39] A detailed description of how we employ this approach is provided in the Supporting Information. The main mathematical object employed in our methodology is a molecular graph,[62,63] which was adapted for protein structural motifs. The formalized mathematical description of our methodology is available in the Supporting Information.

Pairing. Using the most appropriate atom pairing is a prerequisite for a successful superimposition, and failure to identify the best pairing leads to poor results, as is shown in Figure 1. For superimposing protein structural motifs, we cannot use implicit pairing (i.e., the i-th atom from one motif with the i-th atom from the other motif) because the order of the atoms or amino acid residues in the PDB file of one motif might differ from the order in the PDB file of the other motif. Figure 1 demonstrates that even for the superimposition of two PHE residues there can be a significant difference between the superimposition calculated using implicit pairing and the superimposition calculated using the best possible pairing. Employing sequence alignment is also not applicable because some of the superimposed motifs may not have any sequence similarity. Subgraph matching (i.e., searching for the largest identical subgraph contained in both motifs) is also not suitable because protein motifs can consist of several identical residues and can be very symmetrical, and thus, many relevant subgraphs can be found. We therefore decided to use a systematic approach, which tests all possible pairings.

The disadvantage of the systematic approach is its complexity. When superimposing two motifs with n atoms, there are $n!$ possible pairings (e.g., about 3×10^{40} pairings for 30 atoms). It is thus desirable to reduce the number of tested pairings as much as possible. An initial decrease in the number of pairings can be achieved by looking only at those pairings that are chemically meaningful, such that two atoms will be paired only if

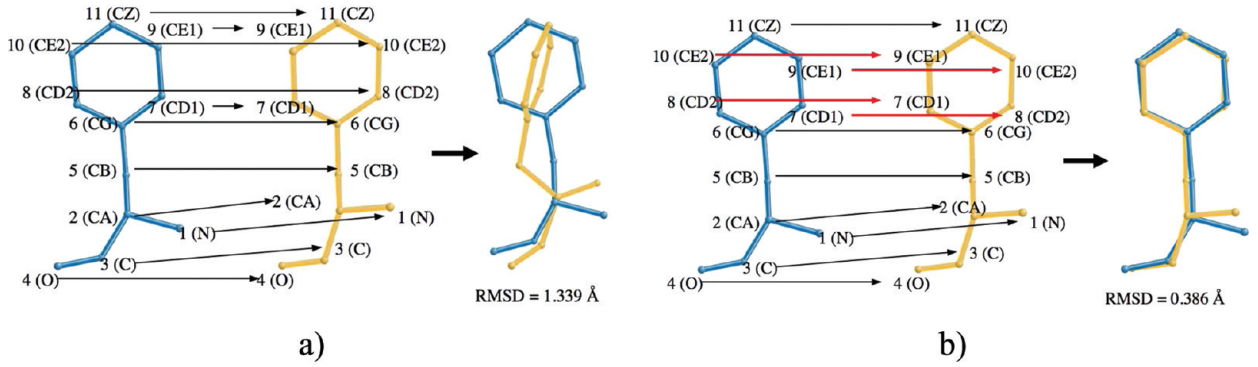


Figure 1. (a) Implicit pairing between residues PHE 83 (blue) and PHE 91 (orange) from the PDB entry 2wh6 and the superimposition calculated by the program VMD, which uses this pairing. (b) The best possible pairing between PHE 83 (blue) and PHE 81 (orange) from 2wh6 and the superimposition calculated by our program SiteBinder, which is able to find this pairing. The differences compared to the implicit pairing are depicted by red arrows. In both (a) and (b), atoms are denoted by their number in the residue, while their PDB name is in brackets.

they are of the same chemical element. A further decrease in the number of tested pairings can be achieved by using the information available in the PDB files. In the PDB file format, each atom is assigned to a residue. Each residue is given a name and a residue identifier (number), which specifies the residue's location in the amino acid sequence. All this information is useful in deciding which atom pairings are worth testing. One can use residue identifiers to make sure that atoms belonging to a single residue in the first motif will be paired only to atoms belonging to a single residue in the second motif and not to atoms belonging to separate residues. Finally, a very effective reduction in the number of possible atom pairings can be achieved if one considers residue names, as this ensures that only atoms belonging to residues with the same names will be paired.

Grouping. We described the basic ideas how to reduce the number of tested atom pairings. To implement these ideas, we need to group the atoms in both motifs into sets and subsets according to the above-mentioned properties. The set containing the atoms from a motif divided into these sets and subsets is denoted as grouping. The groupings help to markedly reduce the number of tested pairings because only the pairings which respect the groupings will be considered. This means that if some atoms from the first grouping are together in a set or subset they can be paired only with atoms from the second grouping that are also together in a relevant set or subset. Conversely, if the respective atoms are not in the same set or subset, they cannot be paired with atoms that are together in a set or subset.

We denote two groupings as compatible if there is at least one pairing (i.e., bijection) that can be created between their atoms. Only compatible groupings can be used in the process of superimposition. We introduce three different types of groupings—residue name, residue identifier, and element symbol grouping.

Residue name grouping assigns atoms to sets according to the name and identifier of the residue they come from. These sets of atoms are further divided into subsets according to their chemical element symbols. For the protein motif in Figure 2, the residue name grouping is $\{\{\{1,3\}^{N}, \{2,4,5\}^{C}\}^{HIS1}, \{\{6,8\}^{N}, \{7,9,10\}^{C}\}^{HIS2}, \{\{11,12\}^{O}, \{13\}^{C}\}^{ASP3}, \{\{14,15\}^{O}, \{16\}^{C}\}^{GLU4}, \{\{17\}^{Zn}\}^{Zn5}\}$. For clarity, the sets and subsets are denoted by the relevant residue name, residue identifier, and element symbol; a similar denotation will be used in further examples of grouping.

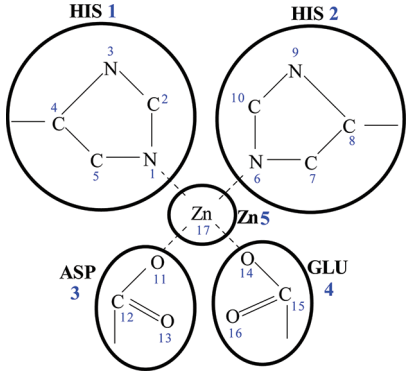


Figure 2. Example of a protein motif.

We use the residue name and identifier jointly for establishing the sets because of two reasons. First, if one uses just the residue names, the atoms from identically named residues will not be separated. For the motif in Figure 2, the grouping would then be $\{\{\{1,3,6,8\}^{N}, \{2,4,5,7,9,10\}^{C}\}^{HIS}, \{\{11,12\}^{O}, \{13\}^{C}\}^{ASP}, \{\{14,15\}^{O}, \{16\}^{C}\}^{GLU}, \{\{17\}^{Zn}\}^{Zn}\}$. Second, if one uses only residue identifiers, the information about the residue name is lost, and it is hard to distinguish for example between the atoms from ASP and GLU in the motif from Figure 2.

Two residue name groupings are compatible if for each set in one grouping there is a set in the other grouping that contains the same number of atoms with the same chemical element symbol that originate from residues with the same name. Thus, using this grouping type is limited (e.g., there are compatible residue name groupings for the dipeptides ALA-GLY and ALA-GLY, but there are no compatible residue name groupings for ALA-GLY and ALA-UNK). On the other hand, the residue name grouping is the most effective grouping type as it reduces the number of tested pairings to a minimum.

Residue identifier grouping assigns atoms to sets according to the identifier of the residue from which they originated. These sets are further divided into subsets according to chemical element symbols. For the protein motif in Figure 2, the residue identifier grouping is $\{\{\{1,3\}^{N}, \{2,4,5\}^{C}\}^{1}, \{\{6,8\}^{N}, \{7,9,10\}^{C}\}^{2}, \{\{11,12\}^{O}, \{13\}^{C}\}^{3}, \{\{14,15\}^{O}, \{16\}^{C}\}^{4}, \{\{17\}^{Zn}\}^{5}\}$. Two residue identifier groupings are compatible if for each set in one grouping there is a set in the other grouping that contains the

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

same number of atoms with the same chemical element symbols. Using this grouping type is also limited (e.g., there can be compatible residue identifier groupings for two dipeptides ALA-GLY and ALA-UNK, but there are no compatible residue identifier groupings for a dipeptide ALA-GLY and a residue UNK). The residue identifier grouping is slightly less effective than the residue name grouping in reducing the number of tested pairings.

Element symbol grouping assigns atoms to sets according to their chemical element symbols. For consistency, we further divide these sets into subsets, but these are also based on chemical element symbols. For the protein motif in Figure 2, the element symbol grouping is $\{\{\{1,3,6,9\}^{N}\}^{N}, \{\{2,4,5,7,8,10,12,14\}^{C}\}^{C}, \{\{11,13,14,16\}^{O}\}^{O}, \{\{17\}^{Zn}\}^{Zn}\}$. Two element symbol groupings are compatible if for each set in one grouping there is a set in the other grouping that contains the same number of atoms that have the same chemical element symbol. This grouping type is very general and can be used in all cases where the superimposed motifs have the same molecular formula. On the other hand, the element symbol grouping has the lowest effectiveness in reducing the number of tested pairings.

Generating Atom Pairings. Before generating all relevant atom pairings that will be tested, it is desirable to find the most effective grouping type that can be used. We first prepare residue name groupings for both motifs and test if these groupings are compatible. If the residue name groupings are compatible, we can employ them. Otherwise, we prepare residue identifier groupings for both motifs and test their compatibility. If the residue identifier groupings are compatible, we can employ them. Otherwise, we prepare element symbol groupings for both motifs. If the element symbol groupings are compatible, we employ these groupings. If no compatible grouping can be found, the motifs cannot be superimposed, and the user needs to change the selection of atoms in at least one of the motifs.

Once we have found compatible groupings for our motifs, we create all possible pairings (i.e., bijections), which respect the groupings (as described above).

Complete Algorithm for Superimposing Two Protein Motifs. To summarize the description given above, we provide a pseudocode of the algorithm for superimposing two motifs.

- Step 1: Prepare the residue name groupings for both motifs. If they are compatible, go to Step 5.
- Step 2: Prepare the residue identifier groupings for both motifs. If they are compatible, go to Step 5.
- Step 3: Prepare the element symbol groupings for both motifs. If they are compatible, go to Step 5.
- Step 4: There is no compatible grouping. Modify the atom selection in at least one motif and go to Step 1.
- Step 5: Use the groupings resulted in the last performed step and generate all possible atom pairings, which respect the groupings.
- Step 6: For each generated pairing do the following: Use quaternion algebra and calculate the transformation that optimally fits one motif to the other. Fit the motifs together using this transformation and calculate the RMSD value.
- Step 7: Find the pairing (among all the generated pairings) that leads to the smallest RMSD.
- Step 8: Superimpose the motifs using the pairing found in Step 7. Return the new coordinates of the motifs (i.e., return the superimposed motifs) and the RMSD value.

Multiple Superimposition of Protein Motifs. Our goal is to provide the most effective solution for this problem that would fit a whole set of protein motifs together as tightly as possible. For this purpose, selecting one of the motifs and superimposing all the others to this one is not a feasible solution as it would only provide an indication of how the rest of the motifs differ from the selected one. Therefore, we designed a multiple superimposition approach that uses the method published by Wang et al.,[61] adapted it to protein motifs and combined it with our algorithm for the superimposition of two motifs. This approach minimizes the RMSD of the whole set of motifs:

$$\mathrm{RMSD}(M) = \sqrt{\binom{m}{2}^{-1} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathrm{RMSD}(M_i, M_j)^2} \quad (1)$$

where M is the set of motifs, and m is the number of motifs in this set.

The multiple superimposition approach works in two steps. First, each motif is superimposed to the first one. This simple superimposition of two motifs is done as described in the pseudocode above, and its purpose is to establish an initial pairing of the atoms and calculate an initial RMSD value. We use this atom pairing and calculate an average motif (M_{avg}) as the arithmetic average of the x, y, and z coordinates of the corresponding atoms. Next, all the motifs in the set are superimposed to the average motif. The new coordinates of all these superimposed motifs are stored, together with the new atom pairing. From these new coordinates, we calculate a new RMSD value (denoted RMSD′). We then calculate the normalized difference (δ) between the original and new RMSD

$$\delta = \frac{\mathrm{RMSD} - \mathrm{RMSD}'}{\mathrm{RMSD}} \quad (2)$$

If $\delta \leq \varepsilon$, where ε is a constant set to 0.005, the process is complete, and the new coordinates are returned. If not, we replace the original coordinates by the new ones, the original pairing by the new one, set the value of the RMSD to RMSD′, and repeat the process. For clarity, we provide also the pseudocode of this approach:

- Step 1: Perform the superimposition of each motif to the first one in order to obtain an initial pairing and calculate an initial value for the RMSD.
- Step 2: Calculate the average motif M_{avg} using the pairing.
- Step 3: Superimpose all motifs to M_{avg} and store the new coordinates and new pairing. Calculate RMSD′ and δ.
- Step 4: If $\delta \leq \varepsilon$, go to Step 6.
- Step 5: Replace the original coordinates of the motifs by the new ones, the original pairing by the new one, set RMSD = RMSD′, and go to Step 2.
- Step 6: The process is complete. Return the new coordinates and RMSD′.

Advantages and Limitations of the Methodology. A great advantage of our methodology is that the accuracy of the superimposition does not depend on the sequence similarity of the superimposed motifs, as all the relevant pairings are tested. This guarantees that the methodology will find the best superimposition (i.e., the superimposition providing minimal RMSD), even when the input motifs do not have any sequence similarity. An example of employing our methodology for the superimposition of motifs that have low sequence similarity (Figure S0 a) and that do not have any sequence similarity

(Figure S0 b) is given in the Supporting Information. On the other hand, the degree of sequence similarity may affect the speed of our approach. Generally, the higher sequence similarity, the fewer pairings need to be tested, and thus, the faster the best pairing will be found. Another advantage of our methodology is that it can very effectively employ information from the PDB files and use this information to decrease the number of tested pairings (i.e., by using groupings). A further significant advantage is that the multiple superimposition does not depend on the order of superimposed motifs. Last but not least, the methodology is able to process any residues in the PDB files, including ligands.

Implementation. We implemented the above-described methodology and developed the Web application SiteBinder, which provides an effective, intuitive, and user-friendly IT solution for the superimposition of multiple protein structural motifs. SiteBinder is implemented in C# using the Microsoft Silverlight platform. Currently, the application can be run in any common Internet browser under Windows and Mac. Full Linux support will be available as soon as the new version of the Moonlight framework plugin (Linux adaptation of Microsoft Silverlight) will be released. The user interface of SiteBinder (depicted in Figure 3) consists of three basic elements: the rendering view, the input panel, and the results panel. The input panel includes the list of motifs and the selection tree.

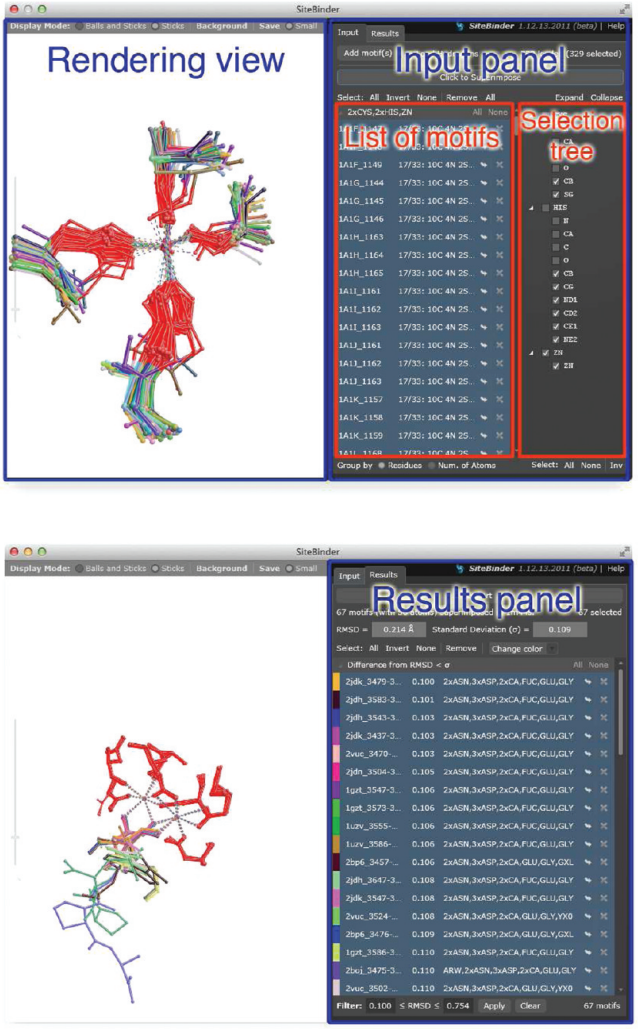


Figure 3. User interface of SiteBinder.

- The rendering view allows the user to view, rotate, and zoom the motifs; change the visualization mode (balls and sticks or sticks); or change the background. Here, the user can also select individual atoms by clicking on them.
- The list of motifs is part of the input panel and shows the loaded motifs grouped by the residues they contain. The user can add or remove motifs from this list and select the particular motifs that will be superimposed at one time.
- The selection tree is also part of the input panel and allows the user to select specific atoms or residues for superimposition.
- The results panel shows the RMSD value of the set of RMSD superimposed motifs. It also provides a list of all superimposed motifs and for each motif its RMSD compared to the average motif ($RMSD_M$). This list of superimposed motifs is sorted according to $RMSD_M$. In addition, the motifs are grouped on the basis of the difference (D_M) between RMSD and $RMSD_M$. There are four groups: $D_M < \sigma$, $\sigma \leq D_M < 2\sigma$, $2\sigma \leq D_M < 3\sigma$, and finally $D_M \geq 3\sigma$, where σ is the standard deviation of the set of D_M values. The RMSD data can be exported into a CSV table and the atomic coordinates into a PDB file. The exported atomic coordinates reflect the superimposition. The structure of the average motif structure can also be written out.

The SiteBinder is a powerful tool but still has some technical limitations. It can superimpose any motifs as long as the atom selections are compatible, meaning that the same number of atoms of the same chemical element need to be selected in each motif. SiteBinder can process at most 7000 to 10000 motifs at a time depending on the computer memory available. For optimum performance, each residue in a superimposed motif should not contain more than 12 atoms of the same element. The reason is that we employ a systematic approach to search for a relevant pairing, which can become significantly slower if each motif contains more than 12 atoms of the same element.

RESULTS AND DISCUSSION

Benchmarking Study—Comparison of Eukaryotic Linear Motifs. Linear motifs (LMs) are short elements embedded within larger protein sequence segments. They operate as regulatory sites and can be found in a wide range of proteins.[64] ELM, the Eukaryotic Linear Motif database,[65,64] is a bioinformatics resource for investigating candidate linear motifs in eukaryotic proteins. ELM currently contains 174 motifs, represented by regular expressions, which describe the occurrence of amino acids in the motif. For example, the regular expression "RF[^P][IV]" indicates that the motif should contain arginine followed by phenylalanine, then any aminoacid except for proline, afterward isoleucine or valine, and finally another amino acid.

This large and heterogeneous resource provides us with a rich area for analysis of protein motifs using SiteBinder. In our investigation, we asked two questions. First, is SiteBinder robust and fast enough to process large sets of low homology linear motifs? Second, do some linear motifs retain conservation at the level of their 3D structure?

In order to address these questions, we first prepared a data set. For each of the 174 linear motifs in ELM (access date: 1.12.2011), we found all its instances in the Protein Data Bank (access date: 1.12.2011). These instances correspond to the ELM regular expressions and may or may not perform the biological function assigned to them in the ELM database.

Table 1. Summary Information about ELM Data Set and Results of Performance and Conservation Study Performed with SiteBinder[a]

information about the motif			performance study			conservation study		
			1000 motifs			1000 motifs	motifs with RMSD < σ	
name	regular expression	no. of res.	no. of compatible atoms in a motif	time (s)	RMSD (Å)	$RMSD_B$ (Å)	no. of motifs	$RMSD_\sigma$ (Å)
LIG_AP2alpha_2	DP[FW]	3	24	10	1.936	0.833	820	0.657
LIG_RGD	RGD	3	23	59	2.603	1.077	883	0.998
LIG_MAPK_2	F.FP	4	33	60	2.693	1.443	833	1.063
LIG_HCF-1_HBM_1	[DE]H.Y	4	32	84	3.238	1.584	816	1.448
LIG_WW_1	PP.Y	4	30	31	2.987	1.601	859	1.519
LIG_EH_1	.NPF.	5	34	50	2.689	1.705	767	1.259
TRG_Cilium_RVxP_2	RV.P.	5	33	80	2.801	1.777	801	1.363
LIG_SPAK-OSR1_1	RF[^P][IV].	5	37	77	3.108	1.962	802	1.428
LIG_TRFH_1	[FY].L.P	5	34	55	3.029	1.869	839	1.525
LIG_APCC_KENbox_2	.KEN.	5	34	79	3.044	1.83	849	1.535
LIG_AP2alpha_1	F.D.F	5	38	79	3.245	1.995	807	1.657
LIG_BIR_III_2	DA.P.	5	28	51	2.641	1.865	853	1.68
LIG_WW_3	.PPR.	5	33	101	2.901	1.962	887	1.714
LIG_BIR_III_4	DA.G.	5	42	30	2.615	1.961	882	1.744
CLV_PCSK_FUR_1	R.[RK]R.	5	37	103	3.65	2.021	819	1.774
LIG_SH3_5	P..DY	5	35	20	3.188	1.963	844	1.856
LIG_EVH1_2	PP..F	5	33	41	3.027	2.096	835	1.944
LIG_PTAP_UEV_1	.P[TS]AP.	6	32	51	2.684	2.121	788	1.895
CLV_PCSK_PC7_1	[R]...[KR]R.	6	41	91	3.884	2.392	791	1.986
LIG_SH3_2	P..P.[KR]	6	33	39	3.033	2.267	883	2.113
LIG_14−3−3_1	R.[^P]([ST])[^P]P	6	35	77	3.257	2.311	861	2.131
LIG_TRAF2_2	P.Q..D	6	36	51	3.337	2.44	854	2.317
LIG_NRBOX	[^P]L[^P][^P]LL[^P]	7	40	73	2.069	1.678	884	0.657
LIG_PP2B_1	.P[^P]I[^P][IV][^P]	7	38	82	2.937	2.45	842	1.924
LIG_SH3_1	[RKY]..P..P	7	36	100	3.079	2.586	870	2.384
LIG_USP7_2	P.E[^P].S[^P]	7	38	42	3.3	2.847	828	2.592
LIG_BRCT_BRCA1_2	.(S)..F.K	7	42	71	3.803	2.928	811	2.607
LIG_RRM_PRI_1	.[ILVM]LG..P.	8	40	110	3.555	3.011	818	2.75
LIG_SH3_4	KP..[QK]...	8	43	92	4.015	3.159	866	2.935
LIG_MDM2	F...W..[LIV]	8	50	211	4.262	3.177	853	2.949
MOD_TYR_ITSM	..T.(Y)..[IV]	8	46	70	3.976	3.31	885	3.134
MOD_PKB_1	R.R..([ST])[^P]..	9	51	181	4.615	3.573	858	3.26

[a]Motifs are sorted first according to their number of residues and then according to $RMSD_\sigma$. Motifs with conserved 3D structure are marked in bold. A brief explanation of the special characters used in the regular expressions can be found on the ELM Help Page.[66]

The files containing the instances of motifs were named *pdbid_index.pdb*, where *pdbid* is the PDB ID of the parent protein, and *index* is the PDB file atom index of the first atom in the motif. Information about the number of instances of each motif and the number of proteins containing at least one instance of each motif is provided in the Supporting Information (Table S1). The program we used for identifying and retrieving ELMs from PDB is also provided in the Supporting Information (program_1). From these 174 linear motifs, we selected 32 as a relevant sample for our benchmarking study. The following criteria were used for this selection. The motif should be frequent enough but not too general (number of instances between 1000 and 30000). The motif should contain at least two identical amino acid residues, and one amino acid residue position defined by a selection of at most four possibilities. In this way, we ensure that it is meaningful to evaluate the structural conservation of the motif. After applying these criteria, we selected the minimum number of motifs so that each of the 20 amino acid residues appears as a firm part of some motif at least once. This procedure provided us with a strong data set for our benchmarking study. The names, regular expressions, and number of residues for the ELMs used in this study are summarized in Table 1.

We then focused on the first question and tested the performance of SiteBinder. For each linear motif in our data set, we selected 1000 instances. Specifically, we went through all P instances of the motif in PDB (sorted alphabetically according to their file names) and took each $(P/1000)^{th}$ instance (e.g., each second instance if the motif appeared 2000 times in PDB). Subsequently, in order to simplify the process of superimposition in SiteBinder, we used a unifying renaming convention for each motif. For instance, the residues in motif "RF[^P][IV]" were renamed as "ARG-PHE-RE1-IL_-RE2". The renaming program (program_2), the unifying residue names for each motif (Table S2), as well as the 1000 renamed instances of each motif are given in the Supporting Information. For each motif, we loaded the 1000 renamed instances into SiteBinder, selected all compatible atoms, and performed the superimposition. By "compatible atoms", we denote all heavy atoms shared by all instances of a particular motif. Table 1 shows the number of atoms used, the duration, and RMSD for each motif. The SiteBinder

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

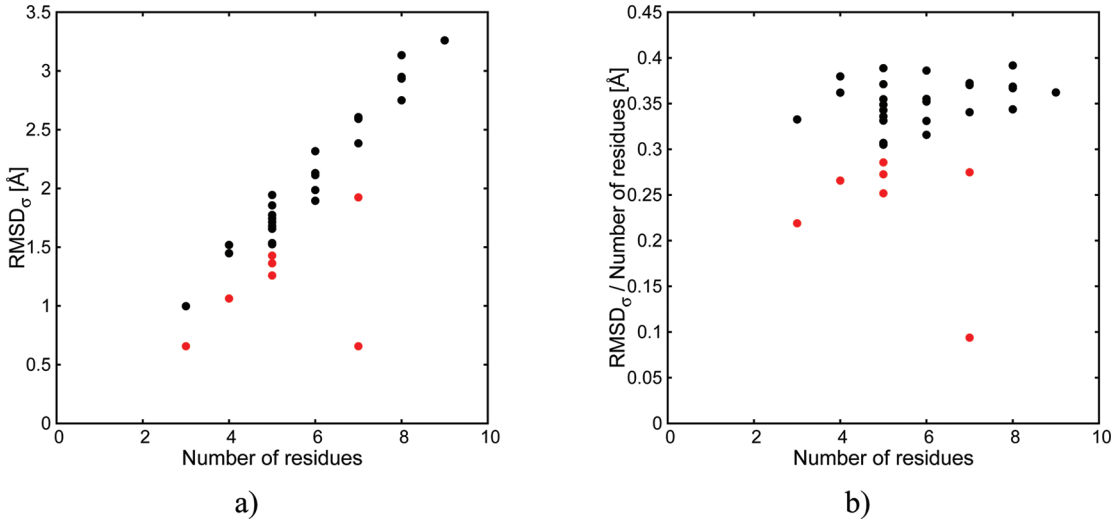


Figure 4. (a) Dependency of RMSD$_\sigma$ on the number of residues in the motif. (b) Dependency of normalized RMSD$_\sigma$ (RMSD$_\sigma$/number of residues) on the number of residues. Motifs with conserved 3D structure are marked red.

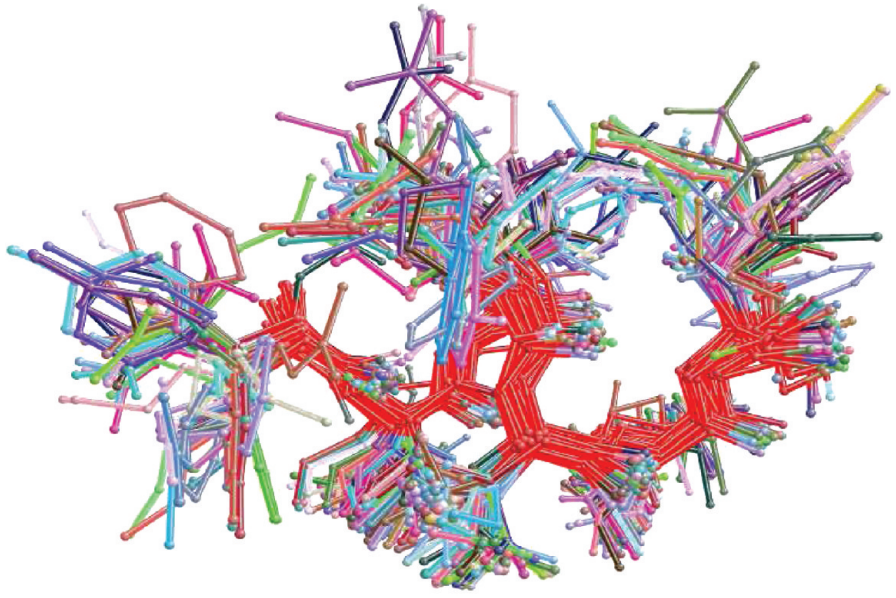


Figure 5. Superimposition of LIG_NRBOX motif instances for which RMSD < σ (only the first 80 instances are shown).

successfully performed the superimposition in all cases, regardless of the number (3 to 9), size (23 to 51 compatible atoms), nature (all 20 amino acids), or degree of conservation of the residues. These results demonstrate the robustness of SiteBinder. The performance test also highlights an exclusive feature of our multiple superimposition methodology, which is that optimal atom pairing can be achieved, and the superimposition can be performed regardless of the degree of amino acid sequence similarity.

We then addressed the second question and investigated whether some linear motifs have a particularly conserved 3D structure. For this stage of the benchmarking, which we denote the "conservation study", we used the same 32 motifs, each with 1000 (renamed) instances, but this time we used only the backbone atoms for the superimposition and obtained RMSD$_B$. To further refine our findings, we performed an additional superimposition for each motif, using only those instances with RMSD$_B$ < σ and thereby obtained RMSD$_\sigma$. The results of the conservation study are also given in Table 1.

The RMSD$_\sigma$ values provide the most relevant information for evaluating the 3D structure conservation of each motif. The RMSD$_\sigma$ grows with the growing number of residues in the motif (Figure 4a), and the dependency is mainly linear. However, seven motifs do not respect this linear trend (marked in red in Figure 4 and in bold in Table 1) and therefore seem much more structurally conserved than the other linear motifs. To clearly identify these motifs, we computed the normalized RMSD$_\sigma$ value by dividing RMSD$_\sigma$ by the number of residues in each motif. We can now clearly visualize the degree of structural conservation. The same seven motifs easily stand out in this analysis, as they have the lowest values of normalized RMSD$_\sigma$ (Figure 4b). The motif LIG_NRBOX seems to be the most structurally conserved by far (Figure 5). Several studies (e.g., Leers et al.,[67] Johansson et al.,[68] Phillips et al.[69]) come to substantiate our finding that LIG_NRBOX is highly conserved. Thus, our analysis was able to easily point out several eukaryotic linear motifs that are conserved at the structural level regardless of the degree of sequence similarity between their

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

parent proteins, and the results of this study are in agreement with published experimental results.

Case Study I—Comparison of Sugar Binding Sites in *Pseudomonas aeruginosa* Lectin II. *Pseudomonas aeruginosa* (PA) is an opportunistic pathogen that can infect almost every human tissue when immunity barriers are lowered.[70] Chronic lung colonization by the bacterium is the major cause of morbidity and mortality in cystic fibrosis patients.[71] *P. aeruginosa* produces the lectin PA-IIL (Pseudomonas lectin II, LecB), which is one of the virulent factors of the pathogen. Each monomer of this lectin contains a sugar binding site that aids the pathogen in host recognition. Knowledge of its structure can lead to better design of new antibacterial—adhesion drugs that minimize the risk of infection. The binding site contains two close calcium cations that mediate the binding of the sugar. These cations are coordinated by seven amino acids, namely, three aspartic acids, two asparagines, one glutamic acid, and one glycine from the adjacent monomer (Figure 6). The sugar is

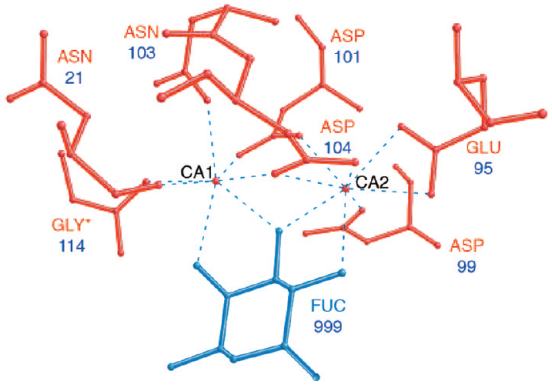


Figure 6. Amino acids coordinating calcium ions in the PA-IIL binding site with α-L-fucose. The depicted binding site originates from the monomer A of the structure with PDB ID 1uzv.

further stabilized by hydrogen bonds with other neighboring amino acids as shown by Mitchell et al.[70]

PA-IIL strongly prefers fucose, but it can also bind other saccharides, albeit with lower affinity. An interesting question that helps to understand the behavior and activity of PA-IIL is whether the structure of this binding site changes when different sugars are bound. We employed SiteBinder to address this question. First, we identified all samples of PA-IIL and its mutants present in the Protein Data Bank (access date: 3.8.2011). We then processed these samples by a program (Supporting Information, program_3) to find and extract the sugar residue, the pair of calcium atoms, and the surrounding seven amino acid residues, as described above and depicted in Figure 6. By this procedure, we obtained 18 structures of PA-IIL and its mutants, which gave us a total of 67 sugar binding sites. Most of these complexes are unique combinations of sugars and the PA-IIL protein or its mutants. There are just three exceptions, i.e., three PDB structures (1gzt, 1oxc, and 1uzv) containing wild-type PA-IIL complexed with α-L-fucose ligands. From these three closely related structures, we kept only the structure with the best resolution (i.e., 1uzv with a resolution of 1 Å) and removed the other two structures. However, we provide a comparison of these three structures in the Supporting Information (Figure S1). It documents the influence of the source organism (1gzt was purified from *P. aeruginosa*, 1ixc and 1uzv were purified from *E. coli*) and the resolution.

We thus obtained a set of 16 PA-IIL structures containing 61 sugar binding sites. These protein structures appear as protein—sugar complexes with nine different sugars. The sugar varies from monosaccharides (i.e., α-L-fucose, α-D-mannose, or α-L-galactose), via their simple derivatives (i.e., methyl-β-D-arabinoside, methyl-α-D-mannoside), to complex synthetic ligands (i.e., 2G0 or LZ0). Basic information about the PA-IIL PDB entries used in this case study can be found in the Supporting Information (Table S3).

In the next step, we used SiteBinder to superimpose the binding sites that bind the same saccharide. The most representative results of this comparison are shown in Figure 7, while the complete set of results can be found in the Supporting Information (Figure S2). These results demonstrate that the binding sites for the same sugar have a very similar structure in different PDB entries (RMSD <0.14 Å), and this feature does not depend on the size of the ligand Figure 7a compared to Figure 7b). The only exception is the binding site of α-methyl-fucoside (RMSD $\leq$ 0.478 Å).

For obtaining a broader overview and in the search for an explanation for the higher RMSD in the case of MFU binding sites, we again employed SiteBinder and superimposed all 61 sugar binding sites. The results of the superimposition are depicted in Figure 8a), and the $RMSD_M$ values are summarized in the Supporting Information (Table S4). This comparison shows that, despite the binding sites originating from different PA-IIL samples (wild types or mutants) and binding different sugars, their structure is very similar (RMSD 0.214 Å). This general comparison also explains the higher RMSD for the binding site of α-methyl-fucoside. The reason is that two of the four binding sites in a mutant of PA-IIL (PDB ID 2jdp) differ from the remaining 59 binding sites (i.e., they have the $RMSD_M$ > 0.7 Å, while the other motifs have the $RMSD_M$ < 0.2 Å). The main difference in these binding sites is that glycine is oriented outward and does not support the binding of the calcium ion (Figure 8b). Nevertheless, this exception does not change the main conclusion, which is that the sugar binding site in PA-IIL is highly conserved.

Our findings that the structure of the sugar binding site in PA-IIL is very similar for nine different sugars could be in direct correlation with the fact that PA is able to infect so many kinds of tissues. In addition, the high level of conservation of this binding site raises the question whether this motif can be used also by other organisms, and because the motif has such a well-defined 3D structure, it can be easily identified. Thus, we used our program_3 to search the complete Protein Data Bank for this motif (access date: 3.8.2011). We searched for two close calcium atoms surrounded by exactly five oxygens from ASP, two oxygens from ASN, two oxygens from GLU, and one oxygen from GLY. From each of the structures found, we obtained the binding site by extracting the sugar residue, the calcium atoms, and the seven surrounding amino acids, as depicted in Figure 6. This way, we collected the 11 sugar binding sites described in Table 2.

These binding sites originate from the proteins *Chromobacterium violaceum* lectin II (CV-IIL) and *Burkholderia cenocepacia* lectin A (BclA). Table 2 shows that the sugar binding sites in these bacteria are very similar as in PA-IIL (RMSD < 0.65 Å). This is in agreement with the fact that the biological activity of BclA[71,72] and CV-IIL[73] is very similar to that of PA-IIL. Moreover, the characteristic propeller assembly of their

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

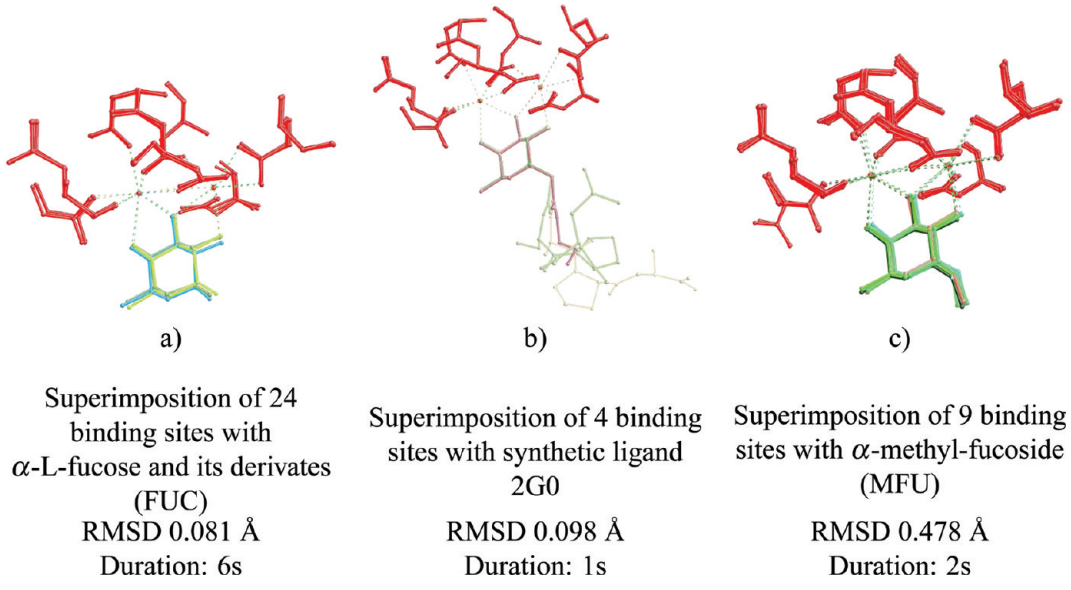


Figure 7. Representative results of the superimposition of PA-IIL binding sites that bind the same sugar-based ligand. Only the atoms in red were used for the superimposition.

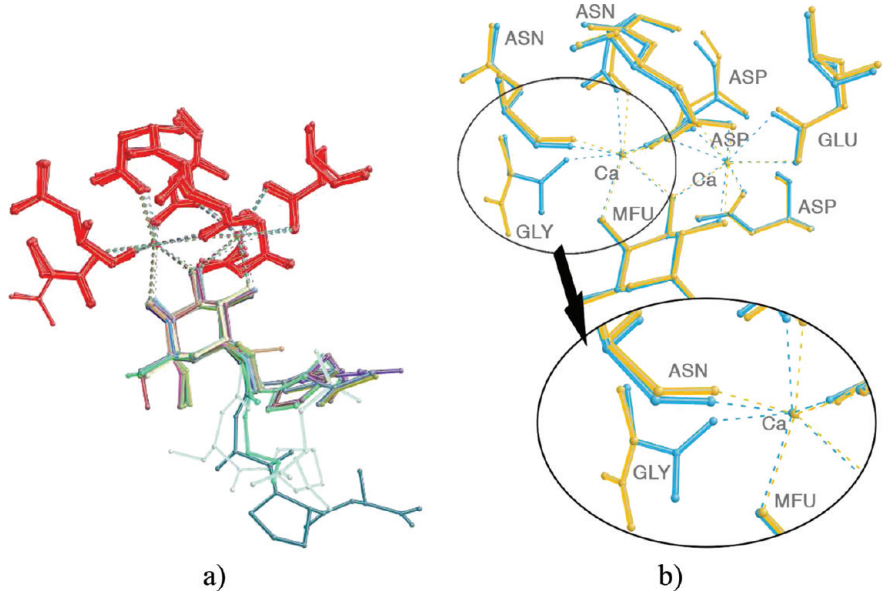


Figure 8. (a) Superimposition of all 61 sugar binding sites, RMSD 0.222 Å, duration 16 s. (b) Comparison of the sugar binding site in the wild type of PA-IIL (PDB ID 2jdm, monomer D, in blue) and in its mutant (PDB ID 2jdp, monomer D, in orange), RMSD 0.754 Å. Part of the calcium binding site in detail.

beta strands classifies these proteins as one family. Our superimposition analysis was able to immediately direct us to identify related family members without any prior knowledge of this fact (i.e., only a calculated model of the sugar binding site in PA-IIL and its mutants was used). One could envision that such analyses could be used to identify related proteins that have been misplaced in different families based on their dominant fold.

Case Study II—Comparison of Zn Binding Sites in Cys_2His_2 Zinc Fingers. Cys_2His_2 zinc fingers are one of the most common structural motifs in eukaryotes.[74,75] Each finger recognizes three to four base pairs of DNA, and several fingers can be linked in tandem to recognize a broad spectrum of DNA sequences with high specificity.[76] There is evidence that some Cys_2His_2 zinc fingers bind RNA and that others may participate in protein−protein interactions, but it appears that their predominant role is in protein−DNA recognition.[74] Individual fingers contain approximately 30 amino acids, and the hallmark of the motif is the presence of two cysteines and two histidines that serve as zinc ligands. The simplest definition of such zinc finger motifs is based on the spacing of the zinc ligands in the amino acid residue sequence. This spacing has the pattern X_2−CYS−X_{2-4}−CYS−X_{12}−HIS−X_{3-5}−HIS,[77] where X represents any amino acid residue. The abundance of this motif, its biological importance, and its simple but apposite description make it an attractive target for research.

We used SiteBinder to determine whether the center of the zinc finger motif (i.e., two CYS, two HIS, and a Zn atom) has a conserved geometry. In order to do this, we went through a few different stages. First, we used a simple program (Supporting

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

Table 2. Sugar Binding Sites with a Very Similar Structure as the PA-IIL Sugar Binding Site

protein name	PDB ID	organism	sugar	monomer	RMSD to the average motif[a] (Å)
CV-IIL	2boi	*Chromobacterium violaceum*	MFU	A	0.136
CV-IIL	2boi	*Chromobacterium violaceum*	MFU	B	0.118
CV-IIL	2bv4	*Chromobacterium violaceum*	MMA	A	0.155
CV-IIL	2bv4	*Chromobacterium violaceum*	MMA	B	0.189
BclA	2vnv	*Burkholderia cenocepacia*	MMA	A	0.621
BclA	2vnv	*Burkholderia cenocepacia*	MMA	B	0.553
BclA	2vnv	*Burkholderia cenocepacia*	MMA	C	0.633
BclA	2vnv	*Burkholderia cenocepacia*	MMA	D	0.567
BclA[b]	2wr9	*Burkholderia cenocepacia*	MAN	A	0.576
BclA	2wr9	*Burkholderia cenocepacia*	MAN	C	0.543
BclA	2wr9	*Burkholderia cenocepacia*	MAN	D	0.521

[a]The average motif was calculated by SiteBinder from all PA-IIL sugar binding sites except those from the mutant 2jdp. [b]The binding site from monomer B of BclA was not included in this study because no sugar was found in the crystal structure at this site.

Information, program_4) and collected all motifs from the Protein Data Bank (access date: 3.8.2011) that fulfill the description of zinc fingers (i.e., Zn coordinated by two CYS and two HIS that are part of a pattern of the type X_2−CYS−X_{2-4}−CYS−X_{12}−HIS−X_{3-5}−HIS). If a protein structure was obtained by NMR, only the motifs from the first model contained in the PDB file were used in our study. We found 329 zinc fingers from 205 different Protein Data Bank entries. For each hit, we extracted the zinc atom and the two HIS and two CYS surrounding this atom. By this procedure, we obtained the zinc finger central motifs and subsequently used these motifs as inputs for SiteBinder. We performed four superimpositions for our set of zinc finger central motifs. These procedures differed in the number of atoms selected for superimposition (displayed in red in Figure 9).

The first superimposition was done using only nine atoms from each motif (zinc, the nitrogens from the imidazole cycle of each HIS, and the sulfur and beta carbon of each CYS), the second superimposition with 15 atoms from each motif (to the previous selection, we added the rest of the imidazole ring atoms of each HIS and the alpha carbon of each CYS), the third with 19 atoms (to the previous selection, we added the beta carbon of each HIS and the carboxylic carbon of each CYS), and the fourth superimposition used all atoms. The superimposed motifs are depicted in Figure 9, which contains also the RMSD values and durations of the superimposition. The values of $RMSD_M$ for each individual motif in all four superimpositions are given in the Supporting Information (Table S5). The RMSD values for the first three superimpositions are similar (between 0.5 and 0.6 Å), which demonstrates that the part of the motif which closely surrounds Zn has a stable structure. Figure 9 demonstrates that the conformation of more distant parts of CYS and HIS may differ.

We further note that, despite the fact that we compared 329 motifs with 9−33 atoms, the superimposition took about 2 min even for the most complex case.

Then we focused on a special group of zinc finger Cys_2His_2 motifs, namely, those known to bind RNA. Superimposing them reveals a very interesting feature. The motifs coming from the PDB entry 1zu1 are markedly different than those in the other RNA binding proteins we investigated (Table 3). This is likely explained by the fact that one of the two HIS residues is facing the binding site with the opposite face of the imidazole ring (Figure 10). What is even more interesting is the biological consequence of this change. Unlike the other zinc finger motifs we discuss here, the motifs contained in 1zu1 have evolved to bind double stranded RNA.[78] The structural peculiarity that we identified by our superimposition analysis without any prior knowledge of RNA binding preference was confirmed by Moller et al.[78] The fact that this structural peculiarity is immediately connected to a functional peculiarity reinforces the structure−function paradigm. This reasoning could be generally applied in order to identify other proteins containing the same functional motif but with slightly different functionality and possibly different behavior toward the same drug molecules.

Case Study III—Comparison of BH3 Domains in Apoptotic Proteins. Apoptosis is a form of cell death that helps to maintain tissue homeostasis and removes malignant cells upon internal and external cellular stress in a biochemically controlled fashion. Apoptosis is downregulated (decreased) in cancer and excessive in neurodegenerative diseases or stroke. The decision whether an initial cellular signal, like a receptor induced stimulus, is tolerated or leads to cell death is controlled by a carefully balanced biochemical cascade of pro-survival or pro-apoptotic proteins of the BCL-2 family.[79,80] The proteins from a pro-apoptotic subgroup of the BCL-2 family, the BH3-only proteins, integrate specific stress signals such as genotoxic stress,[81] serum-deprivation stress,[82] or stress due to the accumulation of unfolded proteins[83] into downstream apoptotic signals. These proteins are called "BH3-only" because they share only the third (of a total of four) BCL-2 homology (BH) domains with the rest of the BCL-2 family. The proteins Bax and Bak, from another pro-apoptotic subgroup, induce the formation of pores into the mitochondrial outer membrane. This phenomenon is a decisive step in apoptosis execution. On the other hand, pro-survival BCL-2 family proteins bind to Bak and Bax, as well as to BH3-only proteins, to prevent unwanted apoptosis. The interaction between pro-survival and pro-apoptotic BCL-2 proteins is mediated by the BH3 domain.[84] The BH3 domains of BH3-only proteins consist of an amphipathic α helix and contain 9−16 amino acids.[85]

A controversy has arisen regarding the role of BH3-only proteins. Originally, it was thought that stress-induced up-regulation (increase) of BH3-only proteins is sufficient to release Bax and Bak from their complexes with pro-survival proteins and thus lead to pore formation. Nonetheless, increasing evidence indicates that an additional step is necessary, namely, the direct activation of Bax and Bak.[86] If such a step is required, two distinct groups of BH3-only proteins are predicted. One group is denoted as "enablers" and comprises the proteins Noxa, Bad, Bmf, Hrk, and Bik. These proteins presumably only bind to pro-survival proteins and thereby release Bax and Bak. The second group of BH3-only proteins, denoted as "activators" are believed to activate Bax and Bak in an explicit activation step. The proteins Bid, Bim, and Puma are examples of activators.[87]

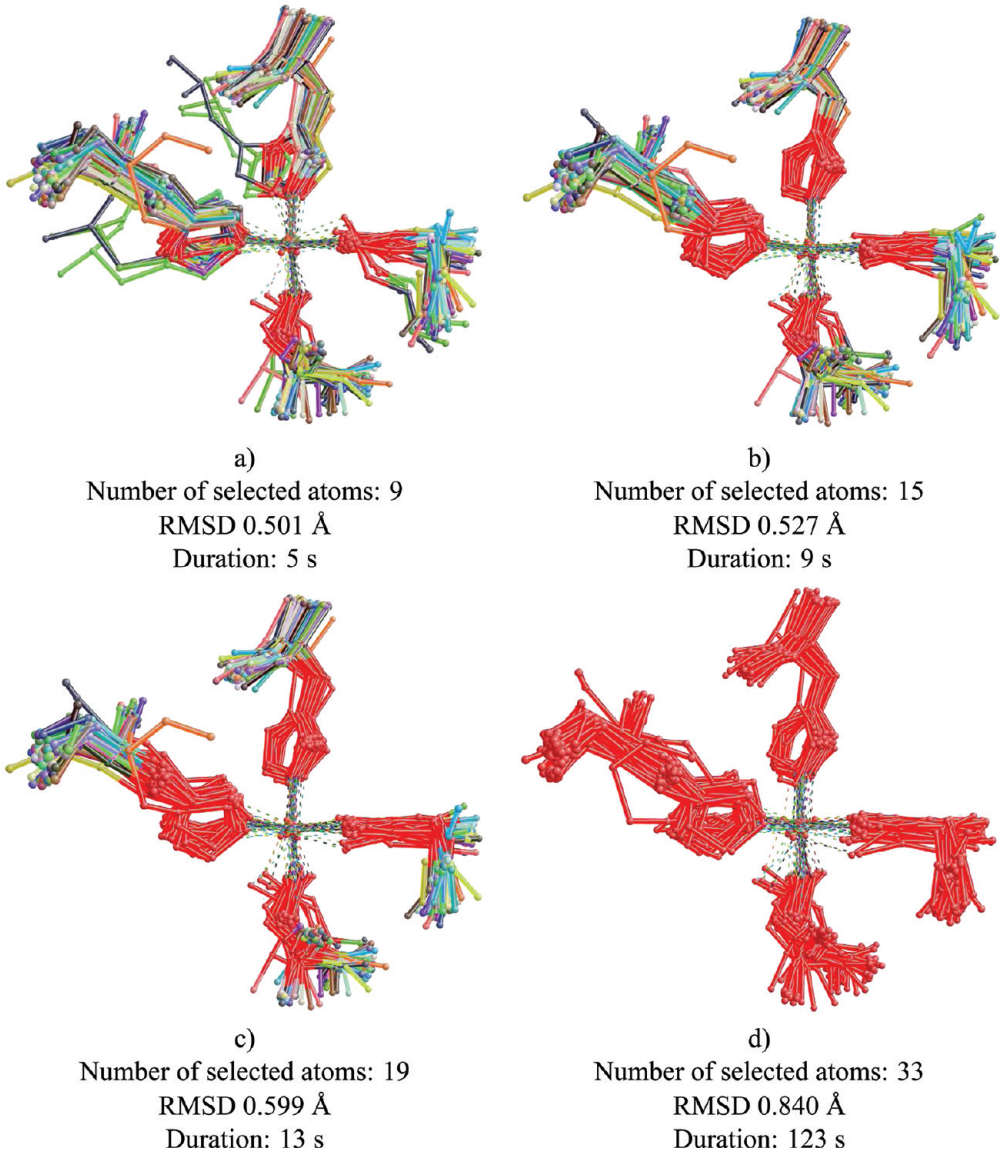


Figure 9. Superimposition of 329 zinc finger central motifs. From (a) to (d), the number of atoms used in the superimposition procedure (displayed in red) increases step by step. For ease of visual interpretation, only the first 80 motifs are displayed.

Table 3. Results of the Superimposition of Zinc Finger Central Motifs of RNA Binding Proteins

protein PDB ID	index of Zn atom	RMSD from the average model (Å)
1un6	4524	0.814
2hgh	3176	0.829
2j7j	717	0.830
1un6	4530	0.880
2hgh	3166	0.881
1un6	4523	0.898
2j7j	718	0.929
1un6	4534	0.956
2ab7	511	0.981
2ab3	494	0.984
2yu5	640	1.108
1zu1	1951	1.821
1zu1	1952	1.834

We used SiteBinder to compare the 3D structures of the BH3 domains of different BH3-only proteins with the goal to investigate whether there is a structural basis of this segregation in activators and enablers. Specifically, we focused on the proteins for which the primary structure of the BH3 domain was described and aligned by Chipuk et al.[79] We obtained the structures of these proteins from the Protein Data Bank, except for the proteins Hrk and Bik, whose structures are not available in this database. The Noxa A protein (PDB ID 2rod) was omitted because its PDB structure was determined by NMR, while the structures of all other BH3-only proteins considered here were determined by X-ray crystallography. The protein names, their PDB identifiers, the BH3-only pro-survival complex from which the structure was derived, and the amino acid sequences are given in Table 4.

As shown in Table 4, the structures of the BH3-only proteins we are using were obtained from larger complexes, in which they are bound to various pro-survival BCL-2 proteins. This complex binding may affect the structural features of the BH3-only proteins. To estimate the influence of these other proteins, we built a reference data set comprising only complexes of the BH3-only protein Bim with all relevant pro-survival BCL-2 proteins (Table 5).

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

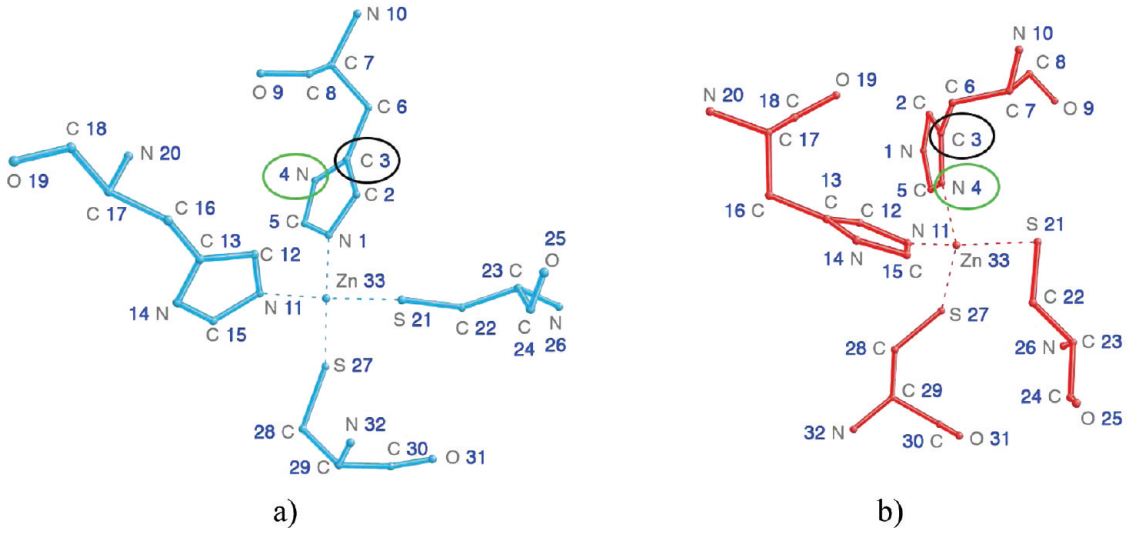


Figure 10. (a) Example of a common structure of the zinc finger central motif in RNA binding proteins (PDB ID 1un6, zinc ion with index 4524). (b) Example of a rare structure of the zinc finger central motif (PDB ID 1zu1, zinc ion with index 1951).

Table 4. Names, PDB Identifiers, and BH3 Domain Amino Acid Sequences of the Activators, and Enablers Used for the Superimposition with SiteBinder[a]

group	PDB ID	BH3-only protein	complexed with	amino acid sequence in BH3 domain[b]
activators	2voi	**Bid**	A1	ILE ALA **ARG** HIS **LEU** **ALA** GLN **ILE** **GLY** **ASP** GLU MET ASP
	2vm6	**Bim**	A1	ILE ALA **GLN** GLU **LEU** **ARG** ARG **ILE** **GLY** **ASP** GLU PHE ASN
	2vof	**Puma**	A1	ILE GLY **ALA** GLN **LEU** **ARG** ARG **ILE** **ALA** **ASP** ASP LEU ASN
enablers	2bzw	**Bad**	BCL-XL	TYR GLY **ARG** GLU **LEU** **ARG** ARG **MET** **SER** **ASP** GLU PHE GLU
	2vog	**Bmf**	A1	ILE ALA **ARG** LYS **LEU** **GLN** CYS **ILE** **ALA** **ASP** GLN PHE HIS
	2nla	**Noxa B**	MCL-1	GLU CYS **ALA** GLN **LEU** **ARG** ARG **ILE** **GLY** **ASP** LYS VAL ASN
SiteBinder denotation				A01 A02 A03 A04 A05 A06 A07 A08 A09 A10 A11 A12 A13

[a]We also mention the pro-survival proteins present in the complexes obtained from PDB. The last row shows our unifying denotation used in the SiteBinder input files. [b]The amino acids that have a degree of conservation higher than 50% for all BH3-only proteins are in bold. Information about the degree of conservation was obtained from the work of Chipuk et al.[79]

Table 5. Summary Information about the Bim Molecules Superimposed Using SiteBinder[a]

PDB ID	2vm6	3fdl	2wh6	2nl9	2pqk	3kj0	3kj1
complexed with	A1	BCL-XL	BHRF1	MCL-1	MCL-1	MCL-1	MCL-1

[a]Entries 3kj0 and 3kj1 contain Bim mutants.

We next extracted the BH3 domains characterized by the amino acid sequence described in Table 4 from the PDB files mentioned in Tables 4 and 5. The amino acid residues that had to be superimposed have different names. In order to simplify their processing by SiteBinder, we introduced a simple unifying denotation for amino acid residue names. Specifically, we re-named the BH3 domain amino acids in the SiteBinder input files according to their position in the sequence (Table 4). This solution was implemented with minimal effort and was feasible because the sequences had already been aligned. The original and modified SiteBinder input files are available in the Supporting Information.

For each of our three groups of motifs (activators, enablers, and Bim samples), we did the following:

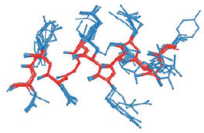
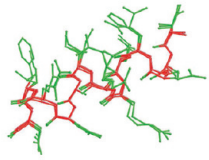
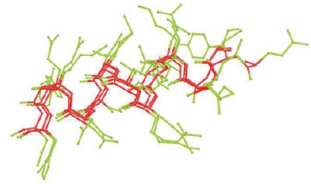
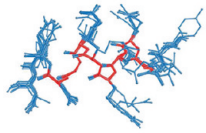
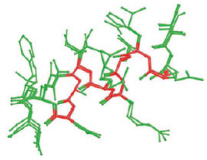
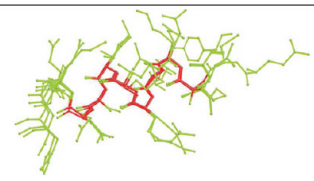
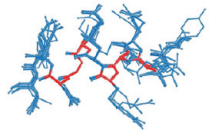
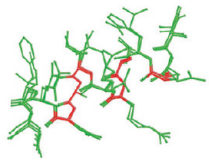
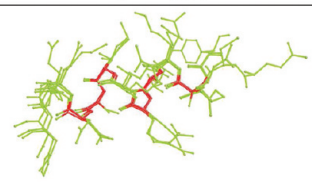
- Superimpose the entire motifs (amino acids A01−A13)
- Superimpose the inner parts of the motifs (amino acids A03−A10)
- Superimpose the conserved parts of the motifs (amino acids A03, A05, A06, A08−A10)

All motifs in a group were employed in the superimposition. Only backbone atoms were used (in red in Table 6), and thus, the RMSD reflects only the backbone geometry conservation.

The results of this superimposition are summarized in Table 6 and in the Supporting Information (Table S6). Each multiple superimposition procedure took less than 5 s. The results in Table 6 indicate that activators have a very conserved BH3 domain (RMSD < 0.25 Å, even when considering the entire motifs). On the contrary, the structure of the BH3 domain in the enablers group shows significant dissimilarity within the group, as well as to the activators group (RMSD > 0.5 Å, even for the inner or most conserved part of the motifs). In addition, comparing Bim motifs extracted from various pro-survival complexes showed smaller RMSD differences than for the activators group in general. This confirms that the pro-survival proteins did not cause significant structural changes upon complex formation. Overall, our results support the hypothesis that activators and enablers may be two functional subgroups of BH3-only proteins. Moreover, they suggest that all activators act in a similar manner to induce cell death. In contrast, the structural

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

Table 6. Superimposition of the BH3 Domains from Several Data Sets (Activators, Enablers, Bim Samples)[a]

	BIM samples	Activators	Enablers
Whole motifs Amino acids: A01–A13 Number of atoms: 39	 RMSD = 0.211 Å	 RMSD = 0.246 Å	 RMSD = 1.438 Å
Middle part of the motifs Amino acids: A03–A10 Number of atoms: 24	 RMSD = 0.147 Å	 RMSD = 0.189 Å	 RMSD = 0.502 Å
Similar parts of the motifs Amino acids: A03, A05, A06, A08–A10 Number of atoms: 18	 RMSD = 0.150 Å	 RMSD = 0.194 Å	 RMSD = 0.534 Å

[a]Only the backbone atoms (in red) were used for the superimposition, and thus, the RMSD values reflect the backbone structural conservation.

heterogeneity of the BH3 domains of different enablers advocate for a specific binding to pro-survival proteins. As different stresses and cells specifically express distinct enablers, this provides a flexible, cell and stress specific, gate-keeping mechanism for enabling or preventing the activation step by the activators.

CONCLUSION

In our work, we focused on the superimposition of very large sets of protein structural motifs. We found the most appropriate state of the art superimposition algorithms available in literature, improved and compiled them, and developed a methodology that is fully tailored to the multiple superimposition of protein structural motifs. This methodology employs the systematic approach for finding the equivalence between atoms and decreases its complexity by using heuristics that consider several types of atom grouping. Fitting the motifs is solved by quaternion algebra. The described superimposition methodology guarantees that the best fit will be found and can be applied even when sequence similarity is low or does not exist at all. Multiple motifs are processed by iteratively superimposing all the structures to an average model until a stable configuration is reached. We have implemented this methodology and have created the Web application SiteBinder. This application is able to process up to thousands of protein structural motifs in a very short time (from a few seconds to a few minutes). Moreover, it provides an intuitive and user-friendly graphical interface, which allows the user to visualize the motifs, select specific atoms or residues for superimposition, export the coordinates of the superimposed structures, as well as the RMSD values, etc.

We have performed a benchmarking analysis by superimposing 1000 experimentally determined structures for each of 32 eukaryotic linear motifs. This analysis shows that our methodology and its implementation are robust, efficient, and versatile. It also demonstrates that SiteBinder can be used for studying general trends in large data sets of low homology protein structural motifs. The applicability of SiteBinder was demonstrated using three case studies that dealt with the comparison of large sets of biochemically important motifs. In the first case study, we compared the structural motifs of 61 PA-IIL sugar binding sites containing nine different sugars. The comparison showed that, despite the binding sites originating from different PA-IIL samples (wild types or mutants) and binding different sugars, their structure is very similar (RMSD 0.222 Å). This finding correlates with the ability of this pathogen to infect many kinds of host cells. In addition, we were able to identify the related proteins CV-IIL and BclA simply by studying the binding site motifs in PA-IIL. This is an example of how a superimposition analysis done with SiteBinder can help in identifying functionally related proteins. The second case study was focused on the analysis of Cys_2His_2 zinc finger structures contained in the Protein Data Bank (more than 300 motifs). We performed four different superimpositions of these motifs, successively increasing the number of superimposed atoms. The results demonstrated that the part of the motifs that closely surrounds Zn has a stable structure (RMSD values are between 0.5 and 0.6 Å). Moreover, we found that a small difference in the structure of RNA binding motifs could be responsible for binding double stranded RNA. In the last case study, we attempted to superimpose 12 BH3 domains from several pro-apoptotic proteins. The results indicated that the activators have a very conserved BH3 domain (RMSD < 0.25 Å, even for the entire motifs). On the contrary, the structure of the BH3 domain in enablers differs across this group of proteins and also differs significantly from the activator group (RMSD > 0.5 Å, even for the most conserved part of the motifs). These results are in agreement with the hypothesis that two functional subgroups of BH3-only proteins, activators and enablers, are present during apoptosis. The three case studies demonstrate the versatility of SiteBinder and show how our

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343–359

software can be used to gain insight into the relationship between protein structure and function. The software is available to the community at http://ncbr.muni.cz/SiteBinder.

■ ASSOCIATED CONTENT

Supporting Information

Superimposition of motifs having low or no sequence similarity (Figure S0), detailed description of the quaternion algebra approach, formalized mathematical description of our superimposition methodology, program to extract the ELMs from PDB (program_1), program for renaming the residues in the ELM PDB files (program_2), summary information about ELMs and their occurrence in PDB (Table S1), unifying residue names for each ELM (Table S2), PDB files of the protein motifs used in benchmarking study, program to extract sugar binding sites (program_3) and zinc fingers (program_4) from PDB, PDB files of the protein motifs used in all case studies, results of the superimposition of PA-IIL sugar binding sites from proteins 1gzt, 1oxc, and 1uzv (Figure S1), results of the superimposition of PA-IIL binding sites which bind the same sugar based ligand (Figure S2), basic information about the PA-IIL PDB entries used in case study I (Table S3), $RMSD_M$ values of the superimposed motifs from case study I (Table S4), case study II (Table S5), and case study III (Table S6). This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: svobodova@chemi.muni.cz (R.S.V.); jaroslav.koca@ceitec.muni.cz (J.K.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (ME08008 to M.W.), the Czech Science foundation (GD301/09/H004 to C.M.I.), the Science Foundation Ireland Grant 08/IN.1/B1949 to H.J.H. and by the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068 to J.K. and R.S.V.) from the European Regional Development Fund. C.M.I. and D.S. thank Brno City Municipality for the financial support provided to them through the program Brno Ph.D. Talent. The access to MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is highly appreciated.

■ REFERENCES

(1) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892−906.

(2) Xie, L.; Xie, L.; Bourne, P. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics* **2009**, *25*, i305−i312.

(3) Thornton, J. M.; Todd, A. E.; Milburn, D.; Borkakoti, N.; Orengo, C. A. O. A. From structure to function: Approaches and limitations. *Nat. Struc. Biol.* **2000**, *7*, 991−994.

(4) Kinoshita, K.; Nakamura, H. Protein informatics towards function identification. *Curr. Opin. Struct. Biol.* **2003**, *13*, 396−400.

(5) Watson, J. D.; Laskowski, R. A.; Thornton, J. M. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* **2005**, *15*, 275−284.

(6) Eidhammer, I.; Jonassen, I.; Taylor, W. R. Structure comparison and structure patterns. *J. Comput. Biol.* **2000**, *7*, 685−716.

(7) Chang, Y. S.; Gelfand, T. I.; Kister, A. E.; Gelfand, I. M. New classication of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. *Proteins* **2007**, *68*, 915−921.

(8) Via, A.; Ferre, F.; Brannetti, B.; Valencia, A.; Helmer-Citterich, M. Three-dimensional view of the surface motif associated with the P-loop structure: *cis* and *trans* cases of convergent evolution. *J. Mol. Biol.* **2000**, *303*, 455−465.

(9) Ausiello, G.; Peluso, D.; Via, A.; Helmer-Citterich, M. Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. *BMC Bioinformatics* **2007**, *8*, S24.

(10) Gherardini, P. F.; Wass, M. N.; Helmer-Citterich, M.; Sternberg, M. J. E. Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.* **2007**, *372*, 817−845.

(11) Gasteiger, J.; Engel, T. *Chemoinformatics: A Textbook*; Wiley-VCH: Weinheim, Germany, 2003.

(12) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(13) Lemmen, C.; Langauer, T.; Klebe, G. FLEXS: a method for fast exible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.

(14) Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573−591.

(15) Baum, D. Multiple semi-flexible 3D superposition of drug-sized molecules. *CompLife* **2005**, *3695*, 198−207.

(16) Eidhammer, I.; Jonassen, I.; Taylor, W. R. *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*; Wiley: Chichester, England, 2004.

(17) Gherardini, P. F.; Helmer-Citterich, M. Structure-based function prediction: Approaches and applications. *Briefings Funct. Genomics Proteomics* **2008**, *7*, 291−302.

(18) Shapiro, J.; Brutlag, D. FoldMiner: Structural motif discovery using an improved superposition algorithm. *Protein Sci.* **2004**, *13*, 278−294.

(19) Taylor, W. R.; Orengo, C. A. Protein structure alignment. *J. Mol. Biol.* **1989**, *208*, 1−22.

(20) Holm, L.; Park, J. DaliLite workbench for protein structure comparison. *Bioinformatics* **2000**, *16*, 566−567.

(21) Michalopoulos, I.; Torrance, G. M.; Gilbert, D. R.; Westhead, D. R. TOPS: An enhanced database of protein structural topology. *Nucleic Acids Res.* **2004**, *32*, D251−D254.

(22) Harrison, A.; Pearl, F.; Sillitoe, I.; Slidel, T.; Mott, R.; Thornton, J.; Orengo, C. Recognizing the fold of a protein structure. *Bioinformatics* **2003**, *19*, 1748−1759.

(23) Madej, T.; Gibrat, J. F.; Bryant, S. H. Threading a database of protein cores. *Proteins* **1995**, *23*, 356−369.

(24) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−747.

(25) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607−633.

(26) Chang, D. T.; Chen, C.; Chung, W.; Oyang, Y.; Juan, H.; Huang, H. ProteMiner-SSM: A web server for efficient analysis of similar protein tertiary substructures. *Nucleic Acids Res.* **2004**, *32*, W76−W82.

(27) Spriggs, R. V.; Artymiuk, P. J.; Willett, P. Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 412−421.

(28) Kinoshita, K.; H., N. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **2003**, *12*, 1589−1595.

(29) Ausiello, G.; Via, A.; M., H.-C. Query3d: A new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinf.* **2005**, *6*, S5.

(30) Barker, J. A.; Thornton, J. M. An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis. *Bioinformatics* **2003**, *19*, 1644−1649.

(31) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137−145.
(32) Gherardini, P. F.; Ausiello, G.; Helmer-Citterich, M.; Hofmann, A. A local structural comparison program that allows for user-defined structure representations. *PloS One* **2010**, *5*, e11988.
(33) Moll, M.; Bryant, D. H.; Kavraki, L. E. The LabelHash algorithm for substructure matching. *BMC Bioinformatics* **2010**, *11*, 555.
(34) Ferre, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M.; Helmer-Citterich, M. Functional annotation by identification of local surface similarities: A novel tool for structural genomics. *BMC Bioinformatics* **2005**, *6*, 194.
(35) Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006**, *65*, 124−135.
(36) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209−220.
(37) Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755−1778.
(38) Bauer, R. A.; Bourne, P. E.; Formella, A.; Frommel, C.; Gille, C.; Goede, A.; Guerler, A.; Guerler, A.; Hoope, A.; Knapp, E. W.; Poschel, T. Others, superimpose: A 3D structural superposition server. *Nucleic Acids Res.* **2008**, *36*, W47.
(39) Coutsias, E. A.; Seok, C.; Dill, K. A. Using quaternions to calculate RMSD. *J. Comput. Chem.* **2004**, *25*, 1849−1857.
(40) Hess, B.; Kutzner, C.; Spoel, D.; Lindahl, E. Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(41) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera: A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(42) *MOE (The Molecular Operating Enviroment)*, version 2005.06; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2009.
(43) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM: A new method for protein modeling and design. Application to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.
(44) Zu-Kang, F.; Sippl, M. J. Optimum superimposition of protein structures: Ambiguities and implications. *Fold. Des.* **1996**, *1*, 123−132.
(45) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.
(46) Raymond, J. W.; Gardiner, E. J.; Willett, P. Rascal: Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.* **2002**, *45*, 631.
(47) Humphrey, W.; Dalke, A.; Schulten, K. VMD: VisualMolecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(48) Laaksonen, L. *gOpenmol*, version 2.0; CSC — IT Center for Science Ltd.: Espoo, Findland, 2001.
(49) *The PyMOL Molecular Graphics System*, version 1.3r1; Schrödinger, LLC: New York, 2010.
(50) Needleman, S.; Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.
(51) Abagyan, R. A.; Batalov, S. Do aligned sequences share the same fold? *J. Mol. Biol.* **1997**, *273*, 355−368.
(52) *Discovery Studio*, version 2.5; Accelrys Software, Inc.: San Diego, CA, 2009.
(53) Chen, B. Y.; Fofanov, V. Y.; Kimmel, D. M.; Lichtarge, O.; Kavraki, L. E. Algorithms for structural comparison and statistical analysis of 3d protein motifs. *Pac. Symp. Biocomput.* **2005**, 334−345.
(54) McLachlan, A. D. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr.* **1972**, *28*, 656−657.
(55) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* **1976**, *32*, 922−923.
(56) Horn, B. K. P. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A* **1987**, *4*, 629−642.
(57) Diamond, R. A note on the rotational superposition problem. *Acta Crystallogr.* **1988**, *44*, 211−216.
(58) Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr.* **1989**, *45*, 208−210.
(59) Karney, C. F. F. Quaternions in molecular modeling. *J. Mol. Graphics Modell.* **2007**, *25*, 1849−1857.
(60) Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. MUSTANG: A multiple structural alignment algorith. *Proteins* **2006**, *64*, 559−574.
(61) Wang, X.; Snoeyink, J. Defining and computing optimum RMSD for gapped and weighted multiple-structure alignment. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2008**, *5*, 525−533.
(62) Koca, J. A mathematical model of the logical structure of chemistry. A bridge between theoretical and experimental chemistry and a general tool for computer-assisted molecular design I. An abstract model. *Theor. Chim. Acta* **1991**, *80*, 29−50.
(63) Koca, J. A mathematical model of realistic constitutional chemistry. A synthon approach. II: The model and organic synthesis. *J. Math. Chem.* **1989**, *3*, 73−89.
(64) Gould, C. M.; et al. ELM: The status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.* **2010**, *38*, D167−D180.
(65) Puntervoll, P.; et al. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **2003**, *31*, 3625−3630.
(66) ELM Help Page. Functional Sites in Proteins, The Eukaryotic Linear Motif. http://elm.eu.org/infos/help.html#regular_expressions (accessed December 1, 2011).
(67) Leers, J.; Treuter, E.; Gustafsson, J. Mechanistic principles in NR box-dependent interaction between nuclear hormone receptors and the coactivator TIF2. *Mol. Cell Biol.* **1998**, *18*, 6001−6013.
(68) Johansson, L.; Bavner, A.; Thomsen, J.; Farnegardh, M.; Gustafsson, J.; Treuter, E. The orphan nuclear receptor SHP utilizes conserved LXXLL-related motifs for interactions with ligand-activated estrogen receptors. *Mol. Cell Biol.* **2000**, *20*, 1124−1133.
(69) Phillips, K. J.; Rosenbaum, D. M.; Liu, D. R. Binding and stability determinants of the PPAR gamma nuclear receptor-coactivator interface as revealed by shotgun alanine scanning and in vivo selection. *J. Am. Chem. Soc.* **2006**, *128*, 11298−11306.
(70) Mitchell, E.; Houles, C.; Sudakevitz, D.; Wimmerova, M.; Gautier, C.; Perez, S.; Wu, A. M.; Gilboa-Garber, N.; Imberty, A. Structural basis foroligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients. *Nat. Struct. Biol.* **2002**, *9*, 918−921.
(71) Govan, J. R. W.; Deretic, V. Microbial pathogenesis in cystic fibrosis: Mucoid *Pseudomonas aeruginosa* and *Burkholderia cepacia*. *Microbiol. Rev.* **1996**, *60*, 539−574.
(72) Lameignere, E.; Malinovska, L.; Slavikova, M.; Duchaud, E.; Mitchell, E. P.; Varrot, A.; Sedo, O.; Imberty, A.; Wimmerova, M. Structural basis for mannose recognition by a lectin from opportunistic bacteria *Burkholderia cenocepacia*. *Biochem. J.* **2008**, *411*, 307−318.
(73) Pokorna, M.; Cioci, G.; Perret, S.; Rebuffet, E.; Kostlanova, N.; Adam, J.; Gilboa-Garber, N.; Mitchell, E. P.; Imberty, A.; Wimmerova, M. Unusual entropy-driven affinity of *Chromobacterium violaceum* lectin CV-IIL toward fucose and mannose. *Biochemistry* **2006**, *45*, 7501−7510.
(74) Pabo, C. O.; Peisach, E.; Grant, R. A. Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.* **2001**, *70*, 313−340.
(75) Krishna, S. S.; Majumdar, I.; Grishin, N. V. Structural classification of zinc fingers: Survey and summary. *Nucleic Acids Res.* **2003**, *31*, 532−550.
(76) Choo, Y.; Sanchez-Garcia, I.; Klug, A. In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature* **1994**, *372*, 642−645.
(77) Brown, R. S.; Sander, C.; Argos, P. The primary structure of transcription factor TFIIIA has 12 consecutive repeats. *FEBS Lett.* **1985**, *186*, 271−274.

(78) Moller, H.; Martinez-Yamout, M.; Dyson, H.; Wright, P. Solution structure of the N-terminal zinc fingers of the *Xenopus laevis* double-stranded RNA-binding protein ZFa. *J. Mol. Biol.* **2005**, *351*, 718−730.
(79) Chipuk, J. E.; Moldoveanu, T.; Llambi, F.; Parsons, M. J.; Green, D. R. The Bcl-2 family reunion. *Mol. Cell* **2010**, *37*, 299−310.
(80) Huber, H. J.; Duessmann, H.; Wenus, J.; Kilbride, S. M.; Prehn, J. H. Mathematical modelling of the mitochondrial apoptosis pathway. *Biochim. Biophys. Acta* **2011**, *1814*, 608−615.
(81) Herr, I.; Debatin, K. M. Cellular stress response and apoptosis in cancer therapy. *Blood* **2001**, *89*, 2603−2614.
(82) Bruns, C. J.; Harbison, M. T.; Davis, D. W.; Portera, C. A.; Tsan, R.; Hicklin, D. J.; Radinsky, R. Epidermal growth factor receptor blockade with C225 plus gemcitabine results in regression of human pancreatic carcinoma growing orthotopically in nude mice by antiangiogenic mechanisms. *Clin. Cancer Res.* **2000**, *6*, 1936−1948.
(83) Ron, D.; Walter, P. Signal integration in the endoplasmic reticulum unfolded protein response. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 519−529.
(84) Han, J.; Flemington, C.; Houghton, A. B.; Gu, Z.; Zambetti, G. P.; Lutz, R. J.; Zhu, L.; Chittenden, T. Expression of bbc3, a pro-apoptotic BH3-only gene, is regulated by diverse cell death and survival signals. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 11318−11323.
(85) Shibue, T.; Taniguchi, T. BH3-only proteins: Integrated control point of apoptosis. *Int. J. Cancer* **2006**, *119*, 2036−2043.
(86) Leber, B.; Lin, J.; Andrews, D. W. Embedded together: the life and death consequences of interaction of the Bcl-2 family with membranes. *Apoptosis* **2007**, *12*, 897−911.
(87) Green, D. R. *Means to an End: Apoptosis and Other Cell Death Mechanisms*; Cold Spring Harbor Laboratory Press: New York, 2011.