JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Protein Structural Statistics with PSS

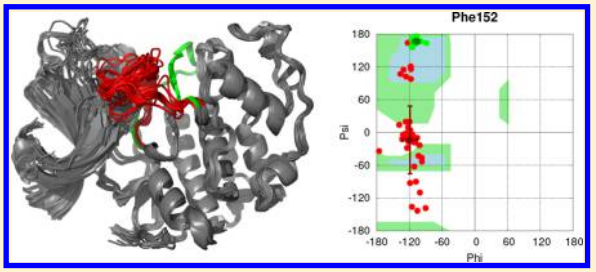
Thomas Gaillard,*[,†] Benjamin B. L. Schwarz,[‡] Yassmine Chebaro,[‡] Roland H. Stote,[‡] and Annick Dejaegere*[,‡]

[†]Laboratoire de Biochimie, UMR 7654 CNRS, Ecole Polytechnique, 91128 Palaiseau Cedex, France

[‡]Biocomputing Group, Integrated Structural Biology Department, IGBMC UMR 7104 CNRS U64 INSERM, 1 rue Laurent Fries, BP 10142 F-67404 Illkirch CEDEX, France

Supporting Information

ABSTRACT: Characterizing the variability within an ensemble of protein structures is a common requirement in structural biology and bioinformatics. With the increasing number of protein structures becoming available, there is a need for new tools capable of automating the structural comparison of large ensemble of structures. We present Protein Structural Statistics (PSS), a command-line program written in Perl for Unix-like environments, dedicated to the calculation of structural statistics for a set of proteins. PSS can perform multiple sequence alignments, structure superpositions, calculate Cartesian and dihedral coordinate statistics, and execute cluster analyses. An HTML report that contains a convenient summary of results with figures, tables, and hyperlinks can also be produced. PSS is a new tool providing an automated way to compare multiple structures. It integrates various types of structural analyses through an user-friendly and flexible interface, facilitating the access to powerful but more specialized programs. PSS is easy to modify and extend and is distributed under a free and open source license. The relevance of PSS is illustrated by examples of application to pertinent biological problems.



INTRODUCTION

With the steadily growing number of structures deposited in the Protein Data Bank (PDB),[1] it is becoming increasingly common in structural biology and bioinformatics to compare large ensembles of protein structures with similar sequence and fold. These analyses are useful in a large variety of applications such as (i) identifying flexible and rigid regions of a protein structure; (ii) selecting structural templates for homology modeling, in particular when there is a large number of structures available and no obvious discriminant filter; (iii) studying the effects of a structural perturbation on a protein, such as ligand binding or mutations; (iv) assessing the existence of distinct conformational states of a protein; (v) comparing a newly obtained protein structure to existing ones; or (vi) analyzing the variations in an ensemble of NMR structures. Despite apparent differences in motivations, a series of common requirements can be identified for such analyses.

A necessary preliminary step in the analysis of an ensemble of structures with similar, but different sequences consists of obtaining a multiple sequence alignment that establishes a correspondence between positions of the different sequences. This, in turn, allows one to define equivalent or aligned residues, which can then be mutually compared. Sequence alignment is an important tool of bioinformatics and molecular biology, used to detect sequence relationships, determine conserved regions or patterns, and make structural, evolutionary, or functional hypotheses.[2] Structure superposition provides additional information to that obtained by sequence comparison alone. Identical sequences can indeed adopt different physically relevant conformations (e.g., active versus inactive kinases), and conversely, different sequences can adopt similar structures (e.g., homologous kinases of different organisms). A first view of structural variability can be obtained by visual inspection of a set of superposed structures. Quantitative information can be derived by the calculation of an average structure and standard deviations of Cartesian coordinates, which can then be averaged by residue to provide a structural fluctuation profile along the protein sequence. The conformation of a protein can also be studied in terms of internal coordinates at the residue level. The local degrees of freedom of a residue are usually described by dihedral angles, measuring either the backbone (ϕ, ψ, ω) or the side chain (χ_1, χ_2, χ_3, χ_4) conformation. The internal nature of dihedral angles make them independent of structure superposition. Dihedral angles are thus often complementary to Cartesian coordinates and, depending on the situation, one or the other can be exploited. Finally, structure comparison can be addressed by cluster analysis. The application of cluster analysis for the classification of protein structures can have many applications, including fold classification,[3] definition of conformational states from molecular dynamics trajectories,[4−7] selection of templates for homology modeling,[8] or selection of receptor conformations for ligand docking.[9]

Received: April 19, 2013
Published: August 19, 2013

A certain number of software products are available to partially address the structural comparison problem. The first and simplest approach consists in the visual comparison of structures provided by molecular visualization applications such as VMD[10] or PyMOL.[11] Multiple sequence alignments and limited structural analyses can be accessible through built-in functions or plugins for these type of programs. However, these programs are not generally suited for the automated comparison of large sets of structures. Programs dedicated to the assessment of structure quality are widely used in structural biology. Such programs, such as Procheck,[12] Aqua,[13] and Molprobity,[14] compute different and detailed structural properties such as dihedral angles, bond distances, or residue contact propensities and compare them to reference values gathered from a curated set of structures. However, these programs are usually limited to the comparison of a single structure to a fixed data set of structures and are not suited for the comparison of multiple structures. Another software class is formed by molecular mechanics programs.[15−18] These programs typically propose a wide variety of flexible structural analyses, implemented in an efficient and scalable manner. However, their application is usually limited to a molecular dynamics trajectory, that is, an ensemble of structures sharing an identical sequence. Numerous programs or platforms are able to perform sequence alignment and structure superposition of multiple proteins. However, in most cases, structural analyses are either limited or not proposed. MODELLER[19] is a program primarily dedicated to homology modeling. It offers several features among which are multiple sequence and structure alignments, as well as functions to assess the quality of a model. Only limited structural statistics are available. In addition, MODELLER requires some expertise and time investment to write Python scripts that are taken as input by the program. Web interfaces are available but are less flexible. MULTISEQ (http://www.scs.illinois.edu/schulten/multiseq) is a unified environment that permits the analysis of both sequence and structure data; it is available as a VMD plugin. As the primary interest of the program is evolutionary analysis, few structural analyses utilities are provided. FRIEND[20] is an application for simultaneous analysis and visualization of multiple structures and sequences. The program is focused on the integration of sequence and structure information. Functionalities to investigate interaction motifs are provided. SuperPose[21] is a web server for multiple protein structure superposition, producing sequence and structure alignments. Structural analyses proposed include the computation of RMSD values between the structures. Another web server called iPBA[22] is dedicated to the comparison of protein structures using a structural alphabet composed of protein blocks. This tool can be used for searching for structural neighbors of one query protein and for performing a structural comparison between two protein structures. As a summary, although there exists tools to compare a set of structures with similar, but different sequences, to our knowledge, none of the available programs can perform automatic, flexible, and detailed structural statistics as that done by the program described here.

We describe here Protein Structural Statistics (PSS), a protein structural statistics program for Unix-like environments dedicated to an automatic, flexible, and detailed statistical analysis of large structural data sets. Written in Perl, it is distributed under the free and open source GNU general public license (GPL). A modular and object-oriented implementation was chosen to facilitate evolution and adaptability. A simple command-line interface makes the program easy to configure and control through various options. The program accepts multiple PDB files as input and, among the available features, produces a multiple sequence alignment, performs structure superposition, calculates Cartesian and dihedral coordinate statistics, and performs cluster analysis, in an automated fashion. In addition, figures are generated for Cartesian and dihedral angle fluctuations by residue, Ramachandran, χ_1/χ_2 plots, individual dihedral angle distributions by residue, and clustering dendrograms. An HTML report providing an overview of results is produced.

A description of the PSS architecture is presented in the Implementation section. The Results and Discussion section illustrates usage through representative examples of structural comparison, followed by a discussion of appropriate employment of PSS in order to obtain optimal results.

IMPLEMENTATION

Here, we present the overall architecture and workflow of PSS, focusing on noteworthy aspects of the implementation. More detailed technical documentation is provided with the program.

Overall Architecture. PSS is a command-line program implemented in Perl as two separate files. The `PSS.pl` script is the main program, while the `PSS.pm` library provides object-oriented classes and methods to describe atoms, residues, molecules, sequences, and alignments. The overall architecture of PSS is illustrated in Figure 1.

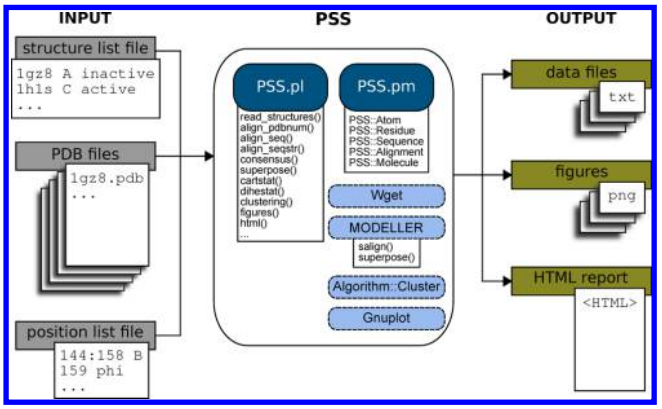


Figure 1. PSS overall architecture. PSS input files are shown in gray boxes on the left, and output files, in khaki boxes on the right. The two Perl components of PSS, `PSS.pl` and `PSS.pm`, are pictured as indigo blue boxes, and their subroutines are listed. PSS dependencies are shown in light blue boxes.

Input data for PSS consists of a set of structure files in PDB format, fetched by default from the execution directory or from another user-specified directory. To have more control of the PDB files read by PSS, a *structure list file* can be provided as argument. Each line of this text file contains a PDB code or file basename, a chain identifier, and an optional user-defined family name. If a requested PDB entry cannot be found locally, it is assumed to be a PDB code and an attempt is made to download it from the PDB server with Wget. If multiple models are present in a PDB file, only the first one is considered. A second text file, the *position list file*, can also be provided to restrict the list of positions and structural variables of interest in the cluster analysis, as well as the positions for which dihedral distribution plots are produced. By default, all positions and variables are considered.

The PSS program is divided into multiple subroutines whose execution depends on requested options. The program workflow is schematized in Figure 2. Initial mandatory steps

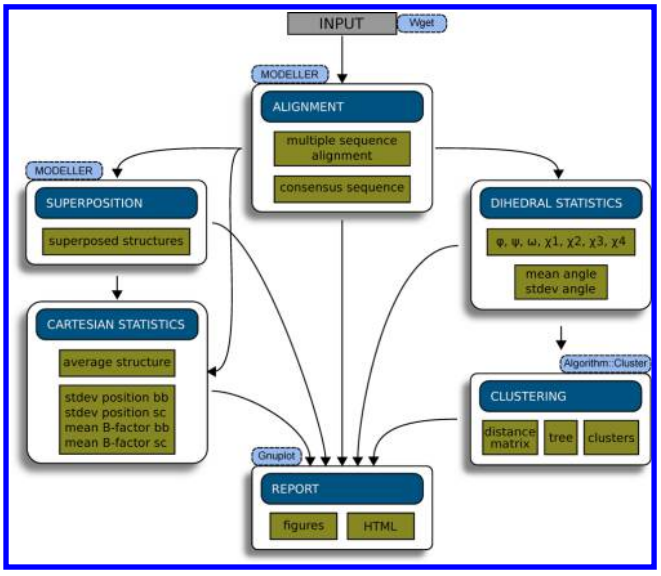


Figure 2. PSS workflow. The color code is the same as that used in Figure 1. Indigo blue boxes represent PSS functionalities. For each functionality, the output is indicated. Black arrows indicate possible workflows.

consist of reading structures and/or sequences and determining a multiple sequence alignment, as well as a consensus sequence. Optional subroutines include structure superposition, Cartesian statistics, dihedral calculation, statistics, clustering, and the production of figures and an HTML report. Iterative use of the program, where a first round of calculations is complemented by a second run, at which point additional analyses can be requested, is possible and facilitated. PSS indeed detects when results are already present and reuses them instead of repeating a calculation. Computationally demanding or complex tasks, such as multiple sequence alignment, structure superposition, and clustering are handled with the help of external programs.

PSS outputs various files into a dedicated directory. Intermediate results are written in space-separated value text files, which is easily accessible and parsable by plotting and statistics programs. Input and output files of external programs are also kept. An HTML report can be produced, facilitating the access to results and analyses through hyperlinks, figures, and tables.

Multiple Sequence Alignment. Three different approaches to generate multiple sequence alignments are proposed in PSS, differing by their complexity and applicability. A fourth possibility is to provide the multiple sequence alignment as an external file in PIR format.

The first and simplest method ("pdbnum") consists of deriving the multiple sequence alignment simply from the PDB numbering. In this case, it is assumed that residues in the ensemble of PDB structures are numbered in a coherent way, and the sequence alignment is straightforwardly determined by residue numbers. The most obvious application is an ensemble of structures of identical sequences, for example, an ensemble of NMR structures. The two other approaches handle more general situations and generate multiple sequence alignments either, based solely on sequence information ("seq"), or using sequence and structure information ("seqstr"). These two latter methods rely on the SALIGN module[23−25] of the MODELLER program.[19] SALIGN is a general alignment tool that exploits several features of protein sequences and structures.

Consensus Sequence. Several operations in PSS require the definition of a reference amino acid type for a given position of the alignment. For this purpose, PSS uses multiple sequence alignment to define a consensus sequence that reflects the most frequent residue type at a given alignment position, gaps excepted. Alternatively, the user can provide an external file containing a consensus sequence determined by an external program.

Superposition of Structures. Superposition of structures on a user-defined reference structure is performed with the `selection.superpose` method of MODELLER. The superposition makes use of the previously determined multiple sequence alignment and employs an iterative least-squares fitting algorithm. Residues that have a RMSD greater than 1.0 Å with respect to the reference structure are removed from consideration in the next iteration. This procedure is carried out until there is no change in the number of equivalent positions. Only backbone atoms are considered in the superposition. PDB files of the superposed structures are produced together with PyMOL and VMD scripts to visualize them.

Cartesian Statistics. Cartesian statistics are computed from atomic coordinates. It is assumed that structures have previously been superposed, either with PSS or with an external program. For each position of the sequence alignment, averages and standard deviations of atomic coordinates, as well as average B-factors are calculated and then averaged separately over backbone and side chain atoms. Only side chains of the same amino acid type as the consensus residue at a given position are considered.

Dihedral Angle Calculations and Statistics. Dihedral angles (ϕ, ψ, ω, χ_1, χ_2, χ_3, and χ_4) are calculated from atomic coordinates, following IUPAC definitions.[26] In order to ensure that chemically equivalent conformations are seen as identical (for example a switch between Oδ1 and Oδ2 atoms in Asp), symmetrization of side chain dihedral χ_2 for residues Asp, Phe, Tyr, and of χ_3 for Glu, is performed, by adding/subtracting 180° to/from their value if it is lower than −90°/higher than 90°.

Dihedral angle statistics are computed for each position of the sequence alignment. Circular averages and their standard deviations, adaptations of equivalent linear quantities to the periodic nature of angles, are calculated. Dihedral angle circular averages are obtained from the components of the mean resultant vector as follows:

$$\bar{\theta} = \text{atan2}\left(\sum_{i=1}^{N} \sin\theta(i), \sum_{i=1}^{N} \cos\theta(i)\right) \quad (1)$$

Dihedral angle circular standard deviations are obtained as

$$S(\theta) = \sqrt{-2\ln R(\theta)} \quad (2)$$

where $R(\theta)$, the module of the mean resultant vector, is computed as

$$R(\theta) = \frac{1}{N}\sqrt{\left(\sum_{i=1}^{N} \sin\theta(i)\right)^2 + \left(\sum_{i=1}^{N} \cos\theta(i)\right)^2} \quad (3)$$

dx.doi.org/10.1021/ci400233j | J. Chem. Inf. Model. 2013, 53, 2471−2482

and N is the number of structures. Side chain dihedral angles are not taken into account in the statistics if they belong to a residue of a different type as the consensus residue. If multiple families have been defined, statistics are calculated for each family, in addition to overall statistics.

Clustering. PSS proposes a hierarchical clustering approach, based on the Algorithm::Cluster Perl module, which is an interface to the C Clustering Library.[27] Structural variables taken into account in clustering can be controlled with the position list file. Any combination of dihedral angle variables can be selected. If no such file is provided, all dihedral angles are considered. Side chain dihedral angles are not taken into account if they belong to a residue of a different type as the consensus residue.

Hierarchical clustering can be seen as a three-step process. The first step consists of the computation of a distance matrix describing pairwise similarities between objects to be clustered. In the second step, a tree is iteratively constructed. At each iteration, a junction is added starting between the two closest objects (i.e., leaves of the tree) and ending when all objects belong to the same cluster. Variants differ by the definition of intercluster or "linkage" distance. The last step consists of trimming the tree, each branch defining a cluster.

In PSS, we define the distance D between two structures i and j as follows:

$$D(i, j) = \frac{1}{\sqrt{N_{\text{var}}}} \sqrt{\sum_{k=1}^{N_{\text{var}}} d(\theta_k(i), \theta_k(j))^2} \tag{4}$$

where N_{var} is the number of dihedral angle variables taken into account in clustering, and d, the circular distance between two angles $\theta(i)$ and $\theta(j)$ of period T, is obtained as

$$d(\theta(i), \theta(j)) = \min\begin{cases} |\theta(i) - \theta(j)| \\ T - |\theta(i) - \theta(j)| \end{cases} \tag{5}$$

The clustering tree is obtained with the `treecluster` method of Algorithm::Cluster. The intercluster distance scheme can be either single- or maximum-linkage. Clusters are derived from the tree by cutting it with the `cut` method of Algorithm::Cluster::Tree, according to a user-controlled radius parameter.

HTML Report and Figures. An HTML report containing a summary of results is produced, including figures generated with Gnuplot.[28] Topping the report is a list of the arguments that were passed to PSS, followed by a table of the structures studied, and then by the multiple sequence alignment and consensus sequence.

Direct links to superposed PDB files are provided, as well as links to PyMOL and VMD scripts that load all superposed structures into the respective programs. Plots of Cartesian coordinate standard deviations and average B-factors for backbone or side chain atoms, as well as standard deviations of dihedral angles, all as a function of the alignment position, are included. These plots are superposable to facilitate comparison and each plot can be hidden/displayed by clicking on the corresponding link. In addition, a sortable and scrollable table containing the raw data is provided below the plots.

Ramachandran (ϕ/ψ), χ_1/χ_2, and side by side individual dihedral angle ("dihedral stripes") distributions are plotted for each alignment position defined in the position list file. Dihedral angle averages and standard deviations can also be visualized on top of the distributions by clicking on the corresponding link. If different families are defined in the structure list file, different colors are used to distinguish points belonging to each family. Side chain dihedral angles are not represented in distribution plots if they belong to a residue of a different type as the consensus residue.

Finally, the clustering results are presented in a table of cluster members. A link to a structure list file with cluster numbers in the family column is also provided, to facilitate the setup of another run of PSS with a definition of families corresponding to identified clusters. At last, a dendrogram plot is included to visualize the clustering tree obtained, with the chosen cutting radius as a vertical line.

RESULTS AND DISCUSSION

Several example applications of PSS are presented that demonstrate features of the program and its relevance for different types of structural analyses. A discussion follows to highlight relevant aspects of PSS usage.

Examples of Application. The interest and use of PSS are illustrated by three examples corresponding to usual problematics of structural biology and bioinformatics. These examples involve proteins of primary biological interest. Nuclear receptor proteins and kinases are indeed implicated in various cell processes and involved in therapeutic strategies. They have been intensively studied, resulting in a large number of experimental structures. In the first example, PSS is used to analyze structural variations in an ensemble of protein structures associated to the same gene of human cyclin-dependent kinase 2 (CDK2), i.e., all structures have the same amino acid sequence up to a few point mutations or gaps. Classification of the structures into two known conformational states is assessed by cluster analysis. In the second example, PSS is used to probe the structural effects of a punctual modification, the phosphorylation of the estrogen receptor β (ERβ). Differences between wild-type and phosphorylated structures are first examined, and, then, the significance of these preliminary observations is evaluated by considering a larger set of structures. In the last example, PSS is used to compare two homologous families of nuclear receptors (NR). Sequence conservation of specific positions among nuclear receptor classes is known and we evaluate the structural conservation of the corresponding residues. Command lines, structure lists, and position lists needed to reproduce the results of the three use cases are given in the Supporting Information.

Conformational States of CDK2. CDK2 is a member of the cyclin-dependent kinase family. It is a catalytic subunit of the cyclin-dependent kinase complex and plays an important role in the regulation of the cell cycle.[29−32] CDK2 adopts a canonical kinase fold (see Figure 3A), comprised of two lobes separated by a hinge region that forms a cleft. The ATP binding site is located within this cleft.[33] The target protein is known to bind to the larger carboxy-terminal lobe, close to the cleft. The activation of CDK2 requires the binding of cyclin, as well as the phosphorylation of a conserved threonine residue (Thr160) located in the activation loop, which is also referred as the T-loop.[34] Upon cyclin binding, important conformational changes occur that include the conversion of the L12 helix preceding the T-loop to a β strand. This, in turn, induces a rearrangement of the T-loop that opens access to the catalytic site. Another change involves the reorientation of the α helix containing a conserved PSTAIRE sequence, leading to an optimal positioning of key amino acids in the active site.[35] CDK2 is an interesting system to illustrate structural variations of a protein

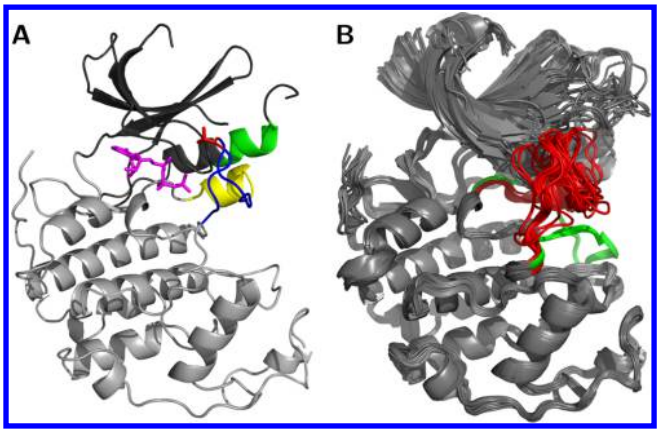


Figure 3. CDK2 structural elements and superposition. (A) The small amino-terminal lobe of CDK2 (1HCK:A) is colored in dark gray, while the large carboxy-terminal lobe is in light gray. The L12 helix is shown in yellow, the T-loop in blue, the Thr160 side chain in red, and the PSTAIRE motif in green. The ATP ligand is pictured as magenta sticks. (B) Superposition of CDK2 structures on 1HCK:A. The L12 helix and T-loop region of major conformational change between inactive and active states is highlighted in red for the inactive state, and green for the active state.

because of the existence of two well-defined inactive and active states and the large amount of structural information available.

A list of 217 human CDK2 structures was retrieved from the P24941 UniProt entry. After exclusion of theoretical models and structures with a resolution above 2.0 Å, the list was reduced to 95 PDB entries. When multiple CDK2 chains were available in a PDB structure, each one was considered separately, yielding a structure list of 104 PDB chains. A reduced structure list of 64 entries was prepared, where chains with missing residues in the L12 helix and T-loop region (residues 144−166) were excluded. These two regions are implicated in the characteristic conformational changes occurring upon CDK2 activation. Each CDK2 chain was labeled as belonging to the families *active* (18 chains) or *inactive* (46 chains), based on the presence or not of a cyclin or CDK inhibitor partner in the structure.

In this example, all structures are essentially derived from the same gene, their sequences differ only by the delimitation of extremities, the presence of gaps where the structure could not be solved, and some point mutations or engineered positions. These small variations in sequence are handled automatically by PSS and the multiple sequence alignment can be straightforwardly derived from PDB numbering with the "pdbnum" alignment method. Structures were then superposed on the A chain of 1HCK, which is the ATP-bound inactive structure with the best resolution (1.9 Å) among selected structures. Superposed structures of CDK2 can be seen in Figure 3B.

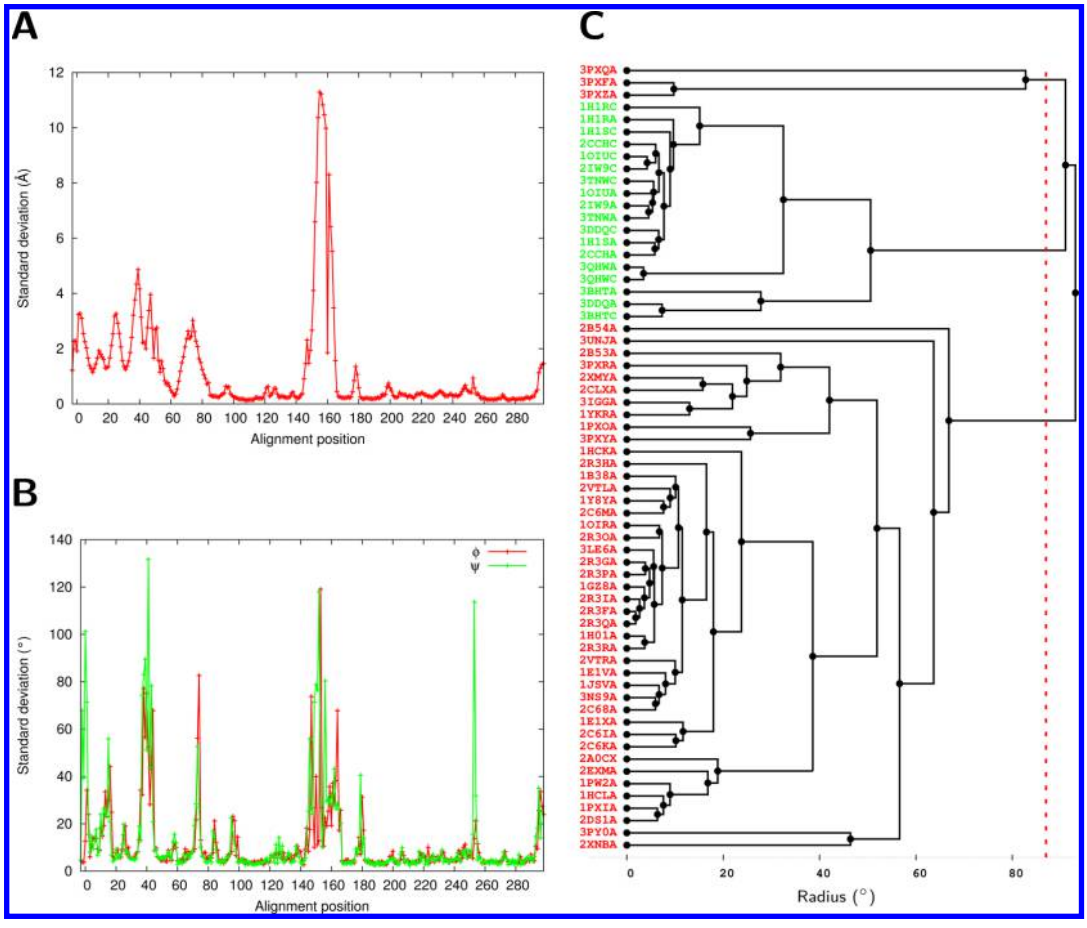


Figure 4. Conformational states of CDK2. (A) Standard deviation of backbone coordinates of CDK2 structures by residue. (B) Standard deviation of ϕ and ψ angles of CDK2 structures by residue. (C) Cluster analysis of CDK2 structures. The tree is obtained from hierarchical clustering with the maximum-linkage method. Variables taken into account are the ϕ and ψ angles of the L12 helix and T-loop region (residues 144−166). Inactive and active structure PDB codes are respectively written in red and green.

dx.doi.org/10.1021/ci400233j | J. Chem. Inf. Model. 2013, 53, 2471−2482

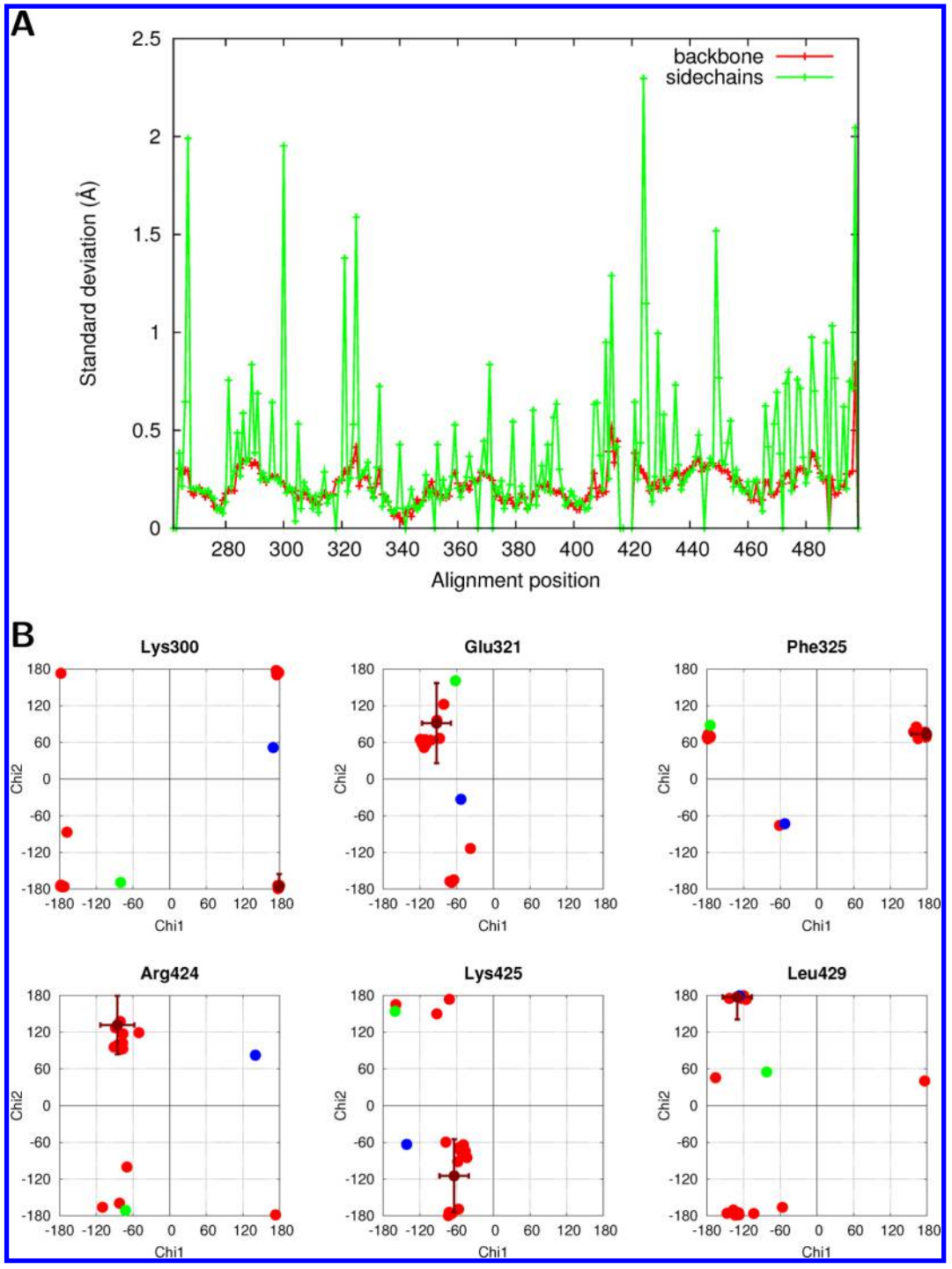


Figure 5. Phosphorylation of ERβ. (A) Backbone (red) and side chain (green) coordinate differences by residue between the phosphorylated (3OLL) and wild-type (3OLS) ERβ structures. (B) χ_1/χ_2 dihedral angle distributions for selected side chains of ERβ structures. The green point corresponds to the phosphorylated structure (3OLL) and the blue point to the wild-type structure (3OLS). Red points correspond to 19 additional wild-type structures, and the dark red point and bars show their average angle values and standard deviations.

Structural fluctuations by alignment position were then calculated with PSS, both in Cartesian (average standard deviation of backbone coordinates, Figure 4A) and dihedral coordinates (average standard deviation of ϕ and ψ angles, Figure 4B). This allows an overall characterization of structural variations in the CDK2 protein data set. Both analyses reveal that the most flexible region of the protein backbone is the 144−166 residue range. Residues 147−153 correspond to the L12 helix, present in the inactive state and known to lose its helical structure in the active state. Residues 154−164 correspond to the T-loop, known to undergo an important conformational change accompanying cyclin binding and Thr160 phosphorylation in the active state. Other significantly variable regions include residues 13−16 of the Gly-rich loop, residues 22−28 of a loop between the strands following the Gly-rich loop, residues 36−44 corresponding to the loop preceding the PSTAIRE helix, and residues 68−77 of a loop between the strands following the PSTAIRE helix. The

dx.doi.org/10.1021/ci400233j | J. Chem. Inf. Model. 2013, 53, 2471−2482

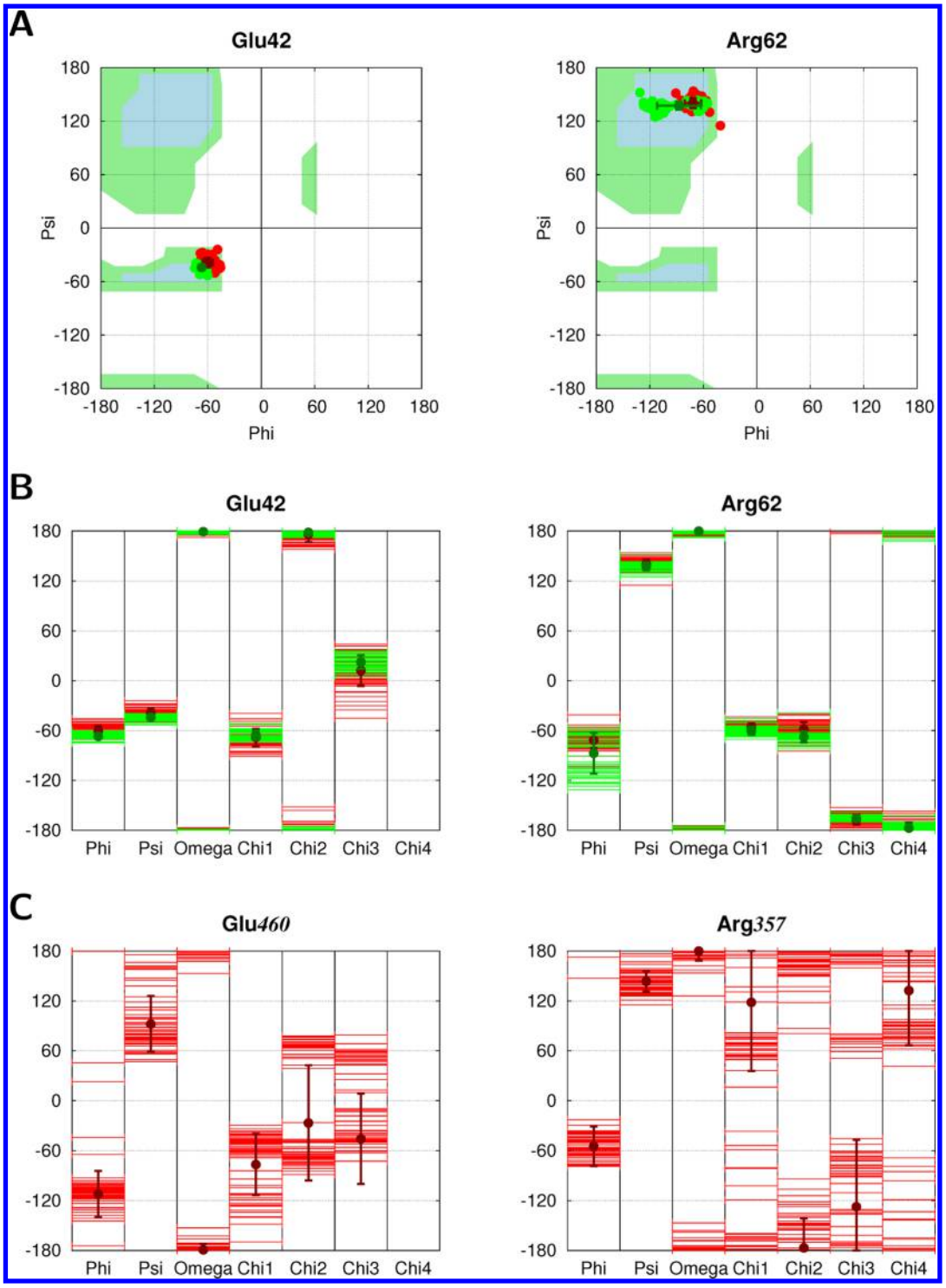


Figure 6. Structural conservation of specific residues in NRs. (A) Ramachandran plots for Glu42 and Arg62 conserved residues (numbering follow the sequence alignment of Brelivet et al.[40]) of PPAR (red) and VDR (green) nuclear receptor structures. (B) Dihedral stripe plots for Glu42 and Arg62 residues of PPAR (red) and VDR (green) structures. (C) Dihedral stripe plots for Glu460 and Arg357 residues (PDB numbering of 1PRG) of PPAR structures.

PSTAIRE motif itself appears in Cartesian but not in dihedral fluctuations, which is indicative of a rigid body displacement.

The CDK2 structures were then submitted to a cluster analysis. We expect to be able to identify two clusters, corresponding to the inactive and active conformational states. Variables taken into account in the clustering were defined in the position list file, as the ϕ and ψ dihedral angles of the L12 helix and T-loop region (residues 144−166). This region is indeed known to be a marker of the difference between inactive and active states and was shown above to be the region with the most important backbone fluctuations in CDK2 structures. Hierarchical clustering was performed using the maximum-linkage method. The clustering tree obtained can be seen in Figure 4C. All active structures and most inactive structures

dx.doi.org/10.1021/ci400233j | J. Chem. Inf. Model. 2013, 53, 2471−2482

were placed in well separated clusters. Three inactive structures (3PXQ:A, 3PXF:A, and 3PXZ:A) unexpectedly appeared in a third cluster, closer to active structures. Examination of these structures revealed that they contain ligands bound to a site other than the ATP binding site, causing a displacement of PSTAIRE and L12 helices, as well as an important deformation of the T-loop. Aside from these three unusual structures, two well separated clusters can be robustly identified from the tree, using a cutting radius ranging from about 70 to 90°. These two clusters correspond exactly to the inactive and active conformational states. The clustering results also show that there is more structural diversity in the inactive than in the active state. This could be explained by the presence of the cyclin partner of CDK2 in the active state, which has an ordering effect on the T-loop located at the binding interface.

In summary, PSS allowed us to characterize the most structurally variable regions of CDK2. A cluster analysis focusing on these elements was able to map the structures into clusters that exactly match the inactive and active states of the protein. Additionally, this analysis allowed us to identify atypical structures, whose particularity could have been missed in a visual inspection of a large number of structures.

Phosphorylation of ERβ. Nuclear receptors are multidomain ligand-dependent transcription factors that control gene activation. The transcriptional activity of these receptors is also subjected to another level of control through post-translational modifications, such as phosphorylation. The estrogen receptor is part of this family of receptors and exists as two subtypes, ERα and ERβ. Experimental studies have shown that phosphorylation of the ERβ estrogen receptor at Tyr488 controls its assembly with the Src kinase.[36,37] More generally, phosphorylation of ERs is known to be important for their localization in the nucleus.[38] The first structure of a phosphorylated nuclear receptor was published in 2010, that of the estrogen receptor ERβ phosphorylated at tyrosine residue 488 (3OLL), issued together with its wild-type (WT) counterpart (3OLS).[39] Structural biology approaches have provided essential information on the molecular mechanism of NR activation by hormones through the determination of a multitude of structures. However, the structural effects linked to ER phosphorylation and its molecular signaling pathways are still not well-understood. In the case of WT and phosphorylated ERβ, the backbone RMSD between the two recently determined structures is small (0.47 Å), and no significant differences were noted by the authors when comparing both structures.[39]

We used PSS to obtain a thorough comparison and systematically identify all structural differences between the phospho-receptor (3OLL) and WT (3OLS) that could be related to functional changes. Structural differences averaged by residue (Figure 5A) reveal that, although only limited changes can be observed in the receptor backbone, some residues exhibit larger differences in their side chains, namely Lys300, Glu321, Phe325, Arg424, Lys425, and Leu429. This allows us to ask the question if these changes in side chain orientations are correlated with phosphorylation. In order to address this question, we performed a second PSS analysis, completing our data set with 19 additional structures of the ERβ wild-type form extracted from the PDB. χ_1/χ_2 dihedral angles (see Figure 5B) were examined for the residues previously identified as having notable side chain differences between wild-type and phosphorylated forms. This analysis revealed that the intrinsic variability of the side chain conformation of these residues in the additional wild-type structures (red) is of similar order as the difference observed between the single structures of wild-type (blue) and phosphorylated (green) forms. One cannot therefore link the sidechain rearrangements observed in the phosphorylated structure (3OLL) as characteristically adopted upon phosphorylation.

In this example, PSS allowed us to easily perform a detailed comparison of the phosphorylated and wild-type forms of ERβ. We were able to easily identify localized structural differences and to assess their significance by comparison to a larger set of WT structures.

Structural Conservation of Specific Residues in NRs. Besides being regulated by ligand binding and phosphorylation, nuclear receptors proteins bind to and are regulated by a variety of other coregulator proteins. An essential aspect is that most NRs are active as dimers. NRs are classified into class I and class II receptors, depending on their ability to respectively form homo- or heterodimers. This distinction has been shown to be present at the sequence level through conservation of specific residues in both classes.[40] In particular, highly conserved salt bridges, that are specific to either homo- or heterodimeric receptors, have been identified,[40] and the functional importance of these salt bridges has been experimentally demonstrated.[41] Using PSS, we can explore whether this specific sequence conservation is also observed at the level of structure, that is, if residues that are highly conserved in sequence have conserved conformations in an ensemble of structures of different nuclear receptors.

To be able to assess the statistical relevance of the observations made in this study, class II nuclear receptors with a sufficient number of available structures were selected. Our data set comprised 112 class II receptor structures: 62 structures of the peroxisome proliferator-activated receptor (PPAR) and 50 structures of the vitamin D receptor (VDR). In heterodimeric receptors (class II), a salt bridge was shown to be conserved[40] between negatively (mostly Glu) and positively charged (mostly Arg) residues, connecting two helices of the ligand binding domain. Dihedral angle distributions were computed with PSS for the 112 receptors structures. The backbone dihedral angles of the salt bridge residues conserved in sequence, Glu42 and Arg62 (numbering follows the sequence alignment of Brelivet et al.[40]), are conserved in the PPAR and VDR structures studied (Figure 6A). The side chain torsion angles (χ_1, χ_2, χ_3, and χ_4) are also remarkably conserved (Figure 6B). It must be noted that conservation of side chain dihedral angles is not systematic, in particular for side chains with the most degrees of freedom. Arg is the most flexible residue (along with Lys) with the highest number of degrees of freedom. A large number of energetically accessible rotamers have been identified for Arg and observed in databases of protein structures (for example, 34 rotamers are defined for Arg by Lovell et al.[42]). Glu has a more restricted conformational space, but distinct rotamers have also been observed (eight rotamers according to Lovell et al.).

For the conserved Arg62 of the PPAR and VDR receptors, the standard deviation of χ_1 is remarkably low (4.5°, while it is of 25.1° for all Arg residues of the structures). The same comparison can be made for Glu42, where the average standard deviation of χ_1 is of 8.0°, while it is of 34.6° for all Glu residues in the structures. For further comparison, the dihedral angles of a salt bridge observed in PPAR structures between residues Glu460 and Arg357 (PDB numbering of 1PRG), but not conserved in other nuclear receptors, have been examined

(Figure 6C). The side chain conformation of these residues is much more variable than for Glu42 and Arg62. The observation that the class-conserved salt bridge Glu42−Arg62 has a narrow distribution of rotameric states thus shows that the sequence conservation is correlated to structural conservation through different receptors.

In this example, PSS allowed us to study local conformational preference of amino acids and to highlight sequence-structure relationships in a large number of nuclear receptor ligand binding domains. The analysis showed that sequence conservation was complemented by structural conservation of side chain rotamers, even for flexible amino acids like arginines.

Usage. The three applications presented above demonstrate the relevance of PSS for various types of structural analyses. We will now provide elements of discussion on selected transverse points of PSS usage that could be useful for the user. As a foreword, we insist on the flexibility and modular nature of PSS that allows a diversity of usages that we cannot exhaustively cover here. In addition, PSS is easy to extend and we encourage users to adapt for example the MODELLER inputs, Gnuplot scripts, or HTML report, if it can better suit their needs. At last, a detailed analysis of PSS time and memory performance is available in the Supporting Information. For a typical run involving 20 structures of 200-residue length, time and memory consumption are about 40 s and 60 Mo on a desktop machine.

Multiple Sequence Alignment. As stated above, obtaining a multiple sequence alignment is a critical preliminary step when comparing structures with similar sequences, conditioning the quality of subsequent analyses derived from it.

For an ensemble of similar structures with a consistent PDB numbering, the method of choice is "pdbnum", where the sequence alignment is simply derived from residue numbers. This method was used in CDK2 and ERβ examples. In this case, the structures usually correspond to the same gene, with only small variations, such as different extremities, point mutations, or engineered positions. Such an ensemble can, for example, be obtained from the PDB codes attached to a UniProt entry. The multiple sequence alignment obtained with the "pdbnum" method is trivial, but in addition to its use by PSS, can be useful for easy identification of missing or mutated residues in the structures.

Whenever residues of the structures are not numbered in a consistent way, it is necessary to perform an actual multiple sequence alignment with the seq or seqstr methods implemented in PSS, which rely on the SALIGN module of the MODELLER program. The seq method was used in the PPAR and VDR nuclear receptor examples. The seqstr method has the highest computational cost, as structure information is used in addition to sequence information. It should be used when the seq method fails to obtain a correct sequence alignment, in particular for families of structures with distant sequences, or regions with large sequence variability. It is indeed known that protein structures are more conserved than sequences,[43] and while it may be difficult to align sequences with low conservation, structures are usually still conserved at lower percentages of identity. Although the three options offered by PSS for sequence alignment should be sufficient to handle most cases, it should be noted that other methods may perform better for extremely divergent sequences.[44,45] To handle such cases, PSS is also able to use as input a multiple sequence alignment computed with an external program or hand curated.

Dihedral vs Cartesian Fluctuations. Dihedral and Cartesian coordinates capture structural fluctuations in a different and complementary manner. Fluctuations are generally pictured in a more diffuse way with Cartesian coordinates, whereas more localized peaks and detailed information in terms of backbone and side chain rotamers are provided by dihedral coordinates. One reason is that we often look at atomic Cartesian fluctuations averaged by residue, which makes them a more collective variable than the individual angles considered in dihedral fluctuations. Another factor is that Cartesian fluctuations are influenced by the positioning of the structures with respect to each others, while dihedral fluctuations are independent of the superposition step.

In particular, when the superposition is global, a subregion moving as a rigid body with respect to the rest of the protein will appear variable in Cartesian coordinates, whereas only residues at the edge of the rigid domain will be highlighted in dihedral coordinates. An example of such situation is the PSTAIRE motif (residues 45−51) of CDK2, which was found in our example to appear in Cartesian but not in dihedral fluctuations. The helix to which the motif belongs indeed has its axis reoriented between inactive and active states, but the internal structure of the helix is maintained. In addition, the magnitude of Cartesian fluctuations in a moving rigid-body subdomain will depend on the distance to the hinge point. For example, some of the loops found as flexible in CDK2 Cartesian fluctuations correspond to the most distant regions of the small lobe with respect to the hinge axis of the interlobe movement.

Clustering. Among the numerous clustering approaches available, we have chosen to use hierarchical clustering in PSS. The advantage of hierarchical methods is that they provide a more global picture of the data and more flexibility to interpret the clustering results as different partitions can be derived from the tree. Hierarchical clustering seems preferable for structural classification problems as, in general, there is no obvious prerequisites on the number of clusters to obtain.

For this reason, the level at which branches of the tree should be cut to define clusters is left for the user of PSS to decide, through the radius option. The radius can be interpreted as the minimum intercluster distance allowed. A radius of zero will thus lead to as many clusters as structures, and increasing the radius will decrease the number of clusters until all structures are in the same cluster. There is no simple way of defining an optimal clustering radius, but the choice can sometimes be guided by examination of the tree when it is clearly divided into branches or can be influenced by preliminary knowledge of the data.

While all variables are considered by default, PSS offers the possibility to precisely select any combination of dihedral angles to take into account for cluster analysis. When information is available to guide the choice of dihedral angles (for example, residues known to be involved in particular interactions, or to undergo particular conformational changes, etc.), limiting the number of variables can contribute to reduce unwanted noise in clustering results.

For example, in our CDK2 application, the cluster analysis was limited to a region known to be a marker of the difference between the two conformational states of interest and previously identified as the most variable part of the protein by the structural fluctuation analysis. The radius choice was then guided by the tree structure, where two main branches were clearly identified.

Table 1. Comparison of Functionalities between PSS and its Closest Competitors

program	different sequences	multiple structures	sequence alignment	structural superposition	Cartesian statistics	dihedral statistics	dihedral clustering	figures	HTML report	comments
PSS	X	X	X	X	X	X	X	X	X	
VMD class	X	X	X	X	X	X		X		not flexible/ automatic
Procheck class					X	X				single structure
CHARMM class		X		X	X	X	X			single sequence
MODELLER	X	X	X	X						
MULTISEQ	X	X	X	X	X			X		
FRIEND	X	X	X	X						limited structural statistics
SuperPose	X	X	X	X	X			X	X	
iPBA	X	X	X	X				X	X	

User-Defined Families. The possibility to define a family attribute for each input structure is a convenient feature of PSS usage. This allows comparative analyses between subsets of structures.

For example, in the CDK2 application, "active" and "inactive" families were defined based on the presence or absence of a cyclin partner in the structures. This classification was useful for interpreting the results. Superposed structures, as well as the leaves of the clustering tree, were indeed colored by family. This allowed an easy verification of the relationship between an a priori classification based on the quaternary structure and the differences in the protein tertiary structures found by PSS.

In the estrogen receptor application, the family attribute was used to distinguish the phosphorylated and wild-type structures. The classification was useful for comparing dihedral angle values between the different groups in dihedral distribution plots, where points were colored by family.

Iterative Use. Iterative use is a powerful feature of PSS. It consists of running the program multiple times for the same project and feeding the results produced in one cycle to the next. This allows an efficient and progressive workflow, where analyses can build one on top of another. In the CDK2 application, for example, different clustering parameters (various variables and clustering radius) could be tested, without regenerating the multiple sequence alignment and recalculating dihedral angles for each iteration.

Comparison to Other Programs. A comparison of PSS to other programs and software classes was presented in the Introduction in terms of general scope and applicability. We found no direct competitor able to perform automatic, flexible, and detailed structural statistics of multiple structures with similar but different sequences. A more detailed comparison in terms of technical features is presented in Table 1. To summarize, visualization programs (VMD, Pymol, etc.) can provide sequence/structure alignment and limited structural statistics but lack flexibility and are not suited for automation; structure quality assessment programs (Procheck, Aqua, Molprobity, etc.) deal with single structures; molecular mechanics programs (CHARMM, AMBER, GROMACS, MDAnalysis, etc.) provide detailed structural statistics for structural ensembles most often associated with a molecular dynamics simulation, that is multiple structures with identical sequence; other programs or web servers (MODELLER, MULTISEQ, FRIEND, Superpose, iPBA, etc.) are able to align two or more structures with different sequences but offer limited structural statistics and figures compared to PSS.

CONCLUSIONS

We presented the program Protein Structural Statistics, or PSS, which is dedicated to the comparative analysis of an ensemble of protein structures with similar sequences. Starting from a set of PDB files, PSS can perform multiple sequence alignment and structure superposition, calculate structural statistics in both Cartesian and dihedral coordinates, and perform cluster analysis. An HTML report can be generated providing an overview of results with figures, tables, and hyperlinks. PSS proposes an integrated approach to structural statistics, facilitating the access to expert tools. It has a flexible command-line interface and is easy to modify and extend, allowing its incorporation to an existing workflow or platform. Applications of such analyses are common and will tend to increase in frequency as the number of available protein structures grows. The relevance of PSS was demonstrated by several example applications to biological problems. Perspectives include setting up a web server interface and integration with other alignment and superposition programs. Extensively employed in our laboratories, we believe PSS will be useful for a broader audience of structural biologists and bioinformaticians.

AVAILABILITY AND REQUIREMENTS

PSS is distributed under the GNU general public license (GPL) and can be downloaded from http://bioc.polytechnique.fr/~gaillard/pss/. External programs are required for some of PSS functionalities. The MODELLER program[19] (version 9 or above) is used for sequence alignment and superposition of structures. Details on obtaining the MODELLER program are available at http://salilab.org/modeller. It is also possible to provide the sequence alignment as an external file, and structures superposed with another program can be given as input to PSS. The Algorithm::Cluster Perl module is used for clustering, and can be obtained from CPAN (http://search.cpan.org/dist/Algorithm-Cluster). At last, Gnuplot (version 4.4 or above), Awk, and Grep programs are used to generate figures, and Wget is used to download structures from the PDB server. Gnuplot, Awk, Grep, and Wget are included in most GNU/Linux distributions.

ASSOCIATED CONTENT

Supporting Information

Command lines, structure lists, and position lists needed to reproduce the results of the three use cases. Analysis of PSS time and memory performance. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: thomas.gaillard@polytechnique.edu.
*E-mail: annick@igbmc.fr.

Author Contributions

A.D. and R.H.S. designed the project. T.G. wrote the program with contributions from B.B.L.S. B.B.L.S., T.G., and Y.C. tested the program, and all authors wrote the article.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Nicolas Muzet for feedback and discussion. This work was supported by funds from the Centre National de Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (INSERM), Université de Strasbourg, and Agence Nationale pour la Recherche (ANR Puzzle-Fit, 09-PIRI-0018-02). Computing resources were provided by the Institut du Développement et des Ressources en Informatique Scientifique (IDRIS), the Centre Informatique National de l'Enseignement Supérieur (CINES), and the Méso-centre de Calcul de l'Université de Strasbourg.

REFERENCES

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

(2) Mount, D. W. *Bioinformatics: Sequence and Genome Analysis*, 2nd ed.; Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2004.

(3) Holm, L.; Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **1997**, *25*, 231−234.

(4) Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry* **1993**, *32*, 412−420.

(5) Stote, R. H.; Dejaegere, A. P.; Lefèvre, J.-F.; Karplus, M. Multiple Conformations of RGDW and dRGDW: A Theoretical Study and Comparison with NMR Results. *J. Phys. Chem. B* **2000**, *104*, 1624−1636.

(6) Lei, H.; Wu, C.; Liu, H.; Duan, Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4925−4930.

(7) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E., III Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3*, 2312−2334.

(8) Fernandez-Fuentes, N.; Rai, B. K.; Madrid-Aliste, C. J.; Fajardo, J. E.; Fiser, A. Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics* **2007**, *23*, 2558−2565.

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

(10) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(11) *The PyMOL Molecular Graphics System*; Schrödinger, LLC: New York, 2013.

(12) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283−291.

(13) Laskowski, R. A.; Rullmannn, J. A.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **1996**, *8*, 477−486.

(14) Chen, V. B.; Arendall, W. B., III; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66*, 12−21.

(15) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(16) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668−1688.

(17) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(18) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *J. Comput. Chem.* **2011**, *32*, 2319−2327.

(19) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.

(20) Abyzov, A.; Errami, M.; Leslin, C. M.; Ilyin, V. A. Friend, an integrated analytical front-end application for bioinformatics. *Bioinformatics* **2005**, *21*, 3677−3678.

(21) Maiti, R.; van Domselaar, G. H.; Zhang, H.; Wishart, D. S. SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res.* **2004**, *32*, W590−W594.

(22) Gelly, J.-C.; Joseph, A. P.; Srinivasan, N.; de Brevern, A. G. iPBA: a tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Res.* **2011**, *39*, W18−W23.

(23) Martí-Renom, M. A.; Madhusudhan, M. S.; Šali, A. Alignment of protein sequences by their profiles. *Protein Sci.* **2004**, *13*, 1071−1087.

(24) Madhusudhan, M. S.; Martí-Renom, M. A.; Sanchez, R.; Šali, A. Variable gap penalty for protein sequence-structure alignment. *Protein Eng., Des. Sel.* **2006**, *19*, 129−133.

(25) Madhusudhan, M. S.; Webb, B. M.; Martí-Renom, M. A.; Eswar, N.; Šali, A. Alignment of multiple protein structures based on sequence and structure features. *Protein Eng., Des. Sel.* **2009**, *22*, 569−574.

(26) McNaught, A. D.; Wilkinson, A. *Compendium of Chemical Terminology*, 2nd ed.; Blackwell Science: New York, 1997.

(27) de Hoon, M. J. L.; Imoto, S.; Nolan, J.; Miyano, S. Open source clustering software. *Bioinformatics* **2004**, *20*, 1453−1454.

(28) Williams, T.; Kelley, C., et al. *Gnuplot: an interactive plotting program*. http://gnuplot.sourceforge.net (accessed August 28, 2013).

(29) Norbury, C.; Nurse, P. Animal Cell Cycles and Their Control. *Annu. Rev. Biochem.* **1992**, *61*, 441−470.

(30) Pines, J. Cyclins and their associated cyclin-dependent kinases in the human cell cycle. *Biochem. Soc. Trans.* **1993**, *21*, 921−925.

(31) Fang, F.; Newport, J. W. Evidence that the G1-S and G2-M transitions are controlled by different cdc2 proteins in higher eukaryotes. *Cell* **1991**, *66*, 731−742.

(32) Pagano, M.; Pepperkok, R.; Lukas, J.; Baldin, V.; Ansorge, W.; Bartek, J.; Draetta, G. Regulation of the Cell Cycle by the cdk2 Protein Kinase in Cultured Human Fibroblasts. *J. Cell Biol.* **1993**, *121*, 101−111.

(33) De Bondt, H. L.; Rosenblatt, J.; Jancarik, J.; Jones, H. D.; Morgan, D. O.; Kim, S.-H. Crystal structure of cyclin-dependent kinase 2. *Nature* **1993**, *363*, 595−602.

(34) Fisher, R. P.; Morgan, D. O. A novel cyclin associates with M015/CDK7 to form the CDK-activating kinase. *Cell* **1994**, *78*, 713−724.

(35) Jeffrey, P. D.; Russo, A. A.; Polyak, K.; Gibbs, E.; Hurwitz, J.; Massagué, J.; Pavletich, N. P. Mechanism of CDK activation revealed

by the structure of a cyclinA-CDK2 complex. *Nature* **1995**, *376*, 313−320.
(36) Auricchio, F.; Migliaccio, A.; Castoria, G. Sex-steroid hormones and EGF signalling in breast and prostate cancer cells: Targeting the association of Src with steroid receptors. *Steroids* **2008**, *73*, 880−884.
(37) Migliaccio, A.; Castoria, G.; Di Domenico, M.; de Falco, A.; Bilancio, A.; Lombardi, M.; Barone, M. V.; Ametrano, D.; Zannini, M. S.; Abbondanza, C.; Auricchio, F. Steroid-induced androgen receptor-oestradiol receptor beta-Src complex triggers prostate cancer cell proliferation. *EMBO J.* **2000**, *19*, 5406−5417.
(38) Castoria, G.; Giovannelli, P.; Lombardi, M.; De Rosa, C.; Giraldi, T.; de Falco, A.; Barone, M. V.; Abbondanza, C.; Migliaccio, A.; Auricchio, F. Tyrosine phosphorylation of estradiol receptor by Src regulates its hormone-dependent nuclear export and cell cycle progression in breast cancer cells. *Oncogene* **2012**, *31*, 4868−4877.
(39) Möcklinghoff, S.; Rose, R.; Carraz, M.; Visser, A.; Ottmann, C.; Brunsveld, L. Synthesis and Crystal Structure of a Phosphorylated Estrogen Receptor Ligand Binding Domain. *ChemBioChem* **2010**, *11*, 2251−2254.
(40) Brelivet, Y.; Kammerer, S.; Rochel, N.; Poch, O.; Moras, D. Signature of the oligomeric behaviour of nuclear receptors at the sequence and structural level. *EMBO Rep.* **2004**, *5*, 423−429.
(41) Gaillard, E.; Bruck, N.; Brelivet, Y.; Bour, G.; Lalevée, S.; Bauer, A.; Poch, O.; Moras, D.; Rochette-Egly, C. Phosphorylation by PKA potentiates retinoic acid receptor alpha activity by means of increasing interaction with and phosphorylation by cyclin H/cdk7. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9548−9553.
(42) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The Penultimate Rotamer Library. *Proteins* **2000**, *40*, 389−408.
(43) Chothia, C.; Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823−826.
(44) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.
(45) Dai, Q.; Yang, Y.; Wang, T. Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* **2008**, *24*, 2296−2302.