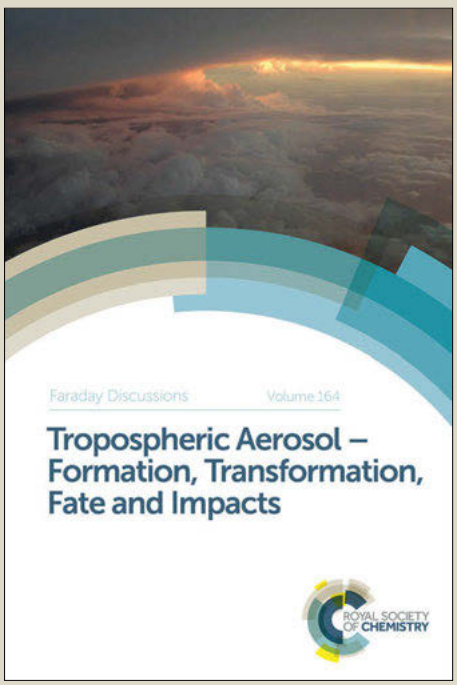
Faraday Discussions

Accepted Manuscript

This manuscript will be presented and discussed at a forthcoming Faraday Discussion meeting. All delegates can contribute to the discussion which will be included in the final volume.

Register now to attend! Full details of all upcoming meetings: http://rsc.li/fd-upcoming-meetings



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/faraday_d

Visualising Intrinsic Disorder and Conformational Variation in Protein Ensembles†

Julian Heinrich[*a,b], **Michael Krone**[a], **Seán I. O'Donoghue**[b,c], **and Daniel Weiskopf**[a]

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/c000000x

Intrinsically disordered regions (IDRs) in proteins are still not well understood, but are increasingly recognised as important in key biological functions, as well as in diseases. IDRs often confound experimental structure determination — however, they are present in many of the available 3D structures, where they exhibit a wide range of conformations, from ill-defined and highly flexible to well-defined upon binding to partner molecules, or upon posttranslational modifications. Analysing such large conformational variations across ensembles of 3D structures can be complex and difficult; our goal in this paper is to improve this situation by augmenting traditional approaches (molecular graphics and principal components) with methods from human-computer interaction and information visualisation, especially parallel coordinates. We present a new tool integrating these approaches, and demonstrate how it can dissect ensembles to reveal functional insights into conformational variation and intrinsic disorder.

1 Introduction

Over the past decade, the role of intrinsically disordered regions (IDRs) in proteins has been increasingly recognised as important, especially in eukaryotes. These regions are now known to play key roles in many biological functions, in regulatory control, and in many diseases[1]. The presence of IDRs in a protein is believed to often confound experimental structure determination, although these regions are present in many of the available 3D structures[2]. Some insights have been gained from examination of structures containing IDRs; in particular, it has become clear that IDRs can exhibit a wide range of structural conformations, from ill-defined and highly flexible to well-defined, upon binding of partner molecules, or upon posttranslational modifications[1]. Overall, however, many aspects of intrinsic disorder in proteins remain poorly understood.

Many structural studies of IDRs have used *homogenous ensembles*, i.e., ensembles comprised of identical molecules that differ only in 3D conformation. This includes ensembles derived from molecular dynamics (MD) simulations,

† Electronic Supplementary Information (ESI) available: http://bit.ly/mega-ensemble

[a] *VISUS, University of Stuttgart, Germany. E-mail: {kroneml | weiskopf}@visus.uni-stuttgart.de*
[b] *CSIRO Computational Informatics, Sydney, Australia. E-mail: julian.heinrich@csiro.au*
[c] *Garvan Institute of Medical Research, Sydney, Australia. E-mail: sean@odonoghuelab.org*

where snapshots are taken at different time points. While MD can be powerful, it is often not feasible to compute sufficiently long trajectories to study key effects on IDRs, such as the binding of partner molecules. A second source of homogenous ensembles that has been used to characterise IDRs are protein structures determined by nuclear magnetic resonance (NMR) studies[3]. These ensembles often exhibit large conformational variations that are widely believed to correlate with the dynamic behaviour of proteins in solution. However, there is good evidence that this belief may be wrong, and that variations observed in NMR ensembles derive primarily from a lack of data to describe the structure fully[4]. In contrast, when structures are derived from X-ray crystallography, any regions lacking sufficient data are simply removed, leaving apparent gaps in the polypeptide chain. A similar approach should probably be taken when using NMR ensembles to study IDRs: regions of the structure with little or no experimental data should often be removed from the analysis. When not done, this may lead to overestimating the conformational variation of IDRs.

In this work, we focus on ensembles that are more heterogeneous, namely ensembles that contain all experimentally-determined structures that are judged to be significantly similar to one 'target' protein sequence, based on a template-based structure prediction method[5]. Currently, such ensembles are readily available for many proteins, often containing information on interactions with other proteins, DNA, RNA, or small molecules — such ensembles are likely to be of increasing significance for molecular biologists, as more structural data becomes available. Examining these ensembles can reveal a wealth of molecular detail on the range of conformations adopted with different binding partners, and can provide insight into IDRs, as these ensembles can capture ranges of conformations across different crystal packing environments, different experimental conditions, and across different molecular complexes. However, these ensembles can be quite complex, with sometimes hundreds, or even thousands of structures.

There are many methods to facilitate the analysis and visualisation of structural ensembles, one of the most widely used being principal components and related methods, which are typically used to find correlated motions[6] either in NMR ensembles or in crystal structures[7]. Such dimension reduction methods are useful for simplifying the resultant visualisations, thus aiding interpretation. However, it remains challenging to augment information about the spatial position of atoms, residues, or secondary structure elements with further attributes such as solvent accessibility, electrostatics, etc. Most methods developed to date focus either on homogenous ensembles, or on ensembles showing structural families[8], which typically include a very diverse range of proteins. The ensembles considered in this work are an intermediate case, and there are few methods for using these ensembles to efficiently gain functional insight into IDRs or other aspects of conformational variation. A key problem with visualising such ensembles is that they are often highly cluttered, particularly in those regions that exhibit high flexibility.

To address these challenges, we propose using parallel coordinates[9] in concert with traditional methods such as principal component analysis (PCA) and molecular graphics for the analysis of intrinsic disorder and conformational flexibility in heterogenous protein ensembles. The use of parallel coordinates allows simultaneous visualisation of high-dimensional data — such as multiple

Table 1 Residue attributes used in the parallel-coordinates view.

Label	Description
2nd	Secondary structure state of residue determined by STRIDE[12].
Contacts	Molecule in contact with current residue.
IUPRED	Predicted disorder score for current residue[13].
Phi	The ϕ backbone angle for current residue.
Psi	The ψ backbone angle for current residue.
Position	Residue position in alignment to the target sequence.
PCA*i* RMSF	Root mean square fluctuation of the Euclidean distance of the C_α atom position to the *i*-th eigenvector.
RMSF	Root mean square fluctuation of the Euclidean distance from the C_α atom to the mean structure.
SAS	Accessible surface area computed with the double cubic lattice method[14] via STRIDE.
Type	The amino acid type (mapped to arbitrary integer).
Chain	PDB[15] chain identifier (mapped to arbitrary integer).
ID	PDB identifier (mapped to arbitrary integer).
PCA*i*	3D coordinates of current structure projected along *i*-th principal component[6].
RMSD	Root mean square deviation of current structure from the top-ranked structure[16].

attributes (see Table 1) from structure ensembles — and is particularly useful to facilitate exploring and finding patterns in the data. In this paper, we construct a multiple view setup[10] that allows residues selected in parallel coordinates to be directly highlighted in a 3D molecular graphics view via brushing-and-linking[11].

2 Related Work

The visualisation of homogenous ensembles is particularly well supported by the VMD[17] molecular graphics tool, as well as other popular tools such as PyMOL[18] or Chimera[19]. Typically, all structures in an ensemble are visualised after being superimposed by minimising the root mean square deviation (RMSD) of the corresponding backbone atoms in the structure, typically using algorithms like those of Kabsch[16] or Coutsias et al.[20]. However, the use of the superimposition approach quickly becomes limited for ensembles containing many structures or those exhibiting large conformational diversity. Therefore, in addition to the generic dimension reduction approaches mentioned above, a range of more tailored approaches have been developed to suit particular cases.

For visualising NMR ensembles, several specialised tools have been developed. One such tool is MOBI[21], which computes a mobility score for the amino acid backbone, based on a combination of Cα interatomic distances and ϕ and ψ angles, then visualises the score using a colour-code mapped onto 3D structure representations. Another tool developed specifically for NMR ensembles is MolMol[22], which offers a 'sausage' visualisation, where a protein's backbone is represented by a tube of variable diameter, scaled according to the mobility of each amino acid.

For visualising MD simulations, specialist systems have been created for studying overall motions, including hierarchical, multiresolution trees[23], as well as interactive linking between alternative visualisations (e.g., DIVE[24]). In addition, methods have been developed for studying even more specialised cases, such as transient cavities[25] or molecular diffusion events[26]. Most MD methods make explicit use of temporal ordering, which is lacking in the ensembles studied in this work.

The above tools typically make use of a range of abstract visualisation methods. One of the first and most popular non-spatial visualisations in structural biology is the Ramachandran plot[27] for the investigation of the distribution of backbone torsion angles with respect to secondary structure elements. Other examples include hydropathy plots, RMSD plots, contact maps — for a recent review, see O'Donoghue et al.[28].

This work focuses on an abstract visualisation method — the parallel-coordinates plot[9] — that has not previously been applied to ensembles of molecular structures, or to intrinsic disorder. This method has been used to visualise high-dimensional data across various application domains, including bioinformatics[29] and systems biology[30,31], where it has been shown to be useful for the analysis of regulatory networks or gene expression (see Heinrich and Weiskopf[32] for a recent survey).

We only found two previous reports using parallel coordinates with protein structure data. The first was from Luke[33] using parallel coordinates to visualise the conformation of the tetrapeptide Met-enkephalin using separate coordinate axes for each rotatable bond in the molecule, similar to the Ramachandran plot. However, this does not scale well as the number of axes increases with the protein size. The second application was from Becker[34], which took a similar approach, but used only main-chain dihedral angles for conformational analysis of proteins. Becker recognised three major advantages of using parallel coordinates for conformational analysis: (i) multiple conformations can be displayed in the same plot, (ii) different types of axes can be mixed in a single plot, and (iii) dynamic clustering and filtering (hiding) can be conducted based on patterns emerging from the plot.

In this work, we further extend these approaches with a richer set of attributes, scalability, interactivity, multiple linked views[10,11], and the integration of statistical methods in the analysis process.

3 Methods

For this work, a plugin to the MegaMol™ framework[35] was implemented to load sets of PDB[15] files, and compute a set of attributes to be used for the parallel-coordinates plot. The system was built using C++ and OpenGL and tested on a Windows workstation with an Intel Core i7, 6 GB RAM and an NVIDIA GeForce GTX 680 (4 GB VRAM).

3.1 Data Preparation

We selected three well-studied human proteins where IDRs and conformational variation were known to influence function (p53, RXR-α, and H2B). In each

case, multiple experimentally determined structures are available — either for that sequence or highly similar sequences — that include a range of residues predicted to be disordered (determined using an IUPRED[13] score ≥ 0.5). In addition, the available structures include many cases with multiple partner molecules. We derived structural ensembles for each target protein sequence with the template-based structure prediction tool HHblits[5], using it to find and align all PDB structures with significantly similar sequence. We included only structures with an expected value of $< 10^{-10}$, a threshold recommended to ensure that all structures are likely to have similar fold to the target protein[36,37].

The three resulting ensembles of PDB structures represent structure variations observed using different experimental methods (NMR, crystallography) across related proteins from several different organisms, and in the presence of a range of binding partners (e.g., DNA or other proteins). For every structure, the HHblits output was used to produce an alignment between each residue in the ATOM records of the PDB file with a corresponding residue in the UniProt[38] sequence of the target protein.

Structures in the resulting ensembles were then clustered based on the region of the match to the full-length protein sequence. We selected one cluster for each sequence — corresponding to one sequence domain — that had a manageable yet sufficiently diverse set of PDB structures (from 47 to 78 cluster members). Structures in the cluster where ranked first by the number of identical residues to the full-length UniProt sequence; in case of matches, PDB structures were then ranked by crystallographic resolution, with NMR structures ranked last. The clustering and ranking were done using the Aquaria resource, currently in development at CSIRO and Garvan (`http://aquaria.ws`).

To prepare for visualisation and further analyses, each structure in an ensemble is superimposed onto the top-ranked structure using the Kabsch algorithm[16], and the respective RMSD is recorded. For NMR structures consisting of multiple models, we used only the first model occurring in the PDB file. For each member of the ensemble, a set of additional attributes were computed. These attributes were selected to reveal different structural aspects that relate to both conformational variation and intrinsic disorder. The attributes are summarised in Table 1, and are further described below:

Secondary structure elements, backbone torsion angles, as well as the solvent accessible area per residue were computed via STRIDE[12].

Intermolecular contacts were defined as follows: for each atom of each residue of the target protein, we searched within a distance of 5Å — if any atoms were found within this distance belonging to another molecule in the PDB structure, this molecule was considered to be in contact with that protein residue.

For the ensemble, we computed a mean backbone structure by averaging the coordinates of superimposed C_α positions for all residues that HHblits matched to the residues in the query UniProt sequence. This mean structure was used to calculate a root mean square fluctuation (RMSF) for each C_α atom, providing a measure of local spatial variation at each residue.

In order to study correlated variations in structure, we applied PCA to each final ensemble using a standard approach[6] developed for analysing MD simulations. Here, a covariance matrix of atom coordinates is calculated and diagonalised to obtain the principal modes that describe most of the spatial variation

within the ensemble. As all structures were superimposed prior to applying PCA, variations caused by the rotation and translation of a whole structure do not affect the computation. Again, only α-carbons are used. For the construction of the covariance matrix, only residues aligned to the target sequence in all members of the ensemble are considered (i.e., gap residues were excluded).

3.2 Visualisation

We constructed a visualisation system using traditional 3D molecular graphics methods[28], to represent ensembles in a spatial context, in concert with a parallel-coordinates view to display additional multidimensional information about the same data set. The system allows users to select residues in parallel coordinates that exhibit certain attributes, with brushing-and-linking allowing the selection to be assessed in the spatial view.

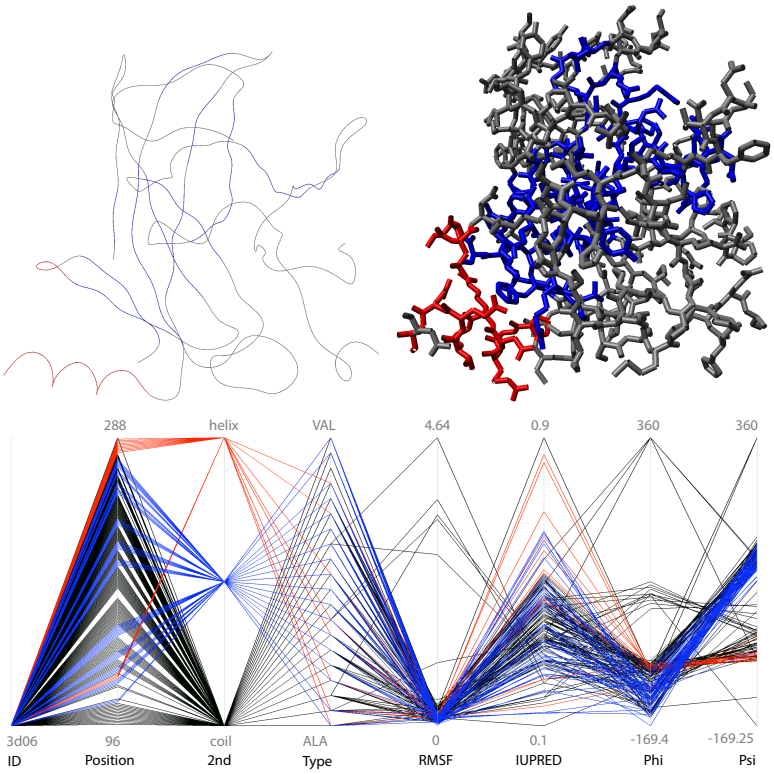


Fig. 1 Mapping residues attributes to poly-lines in parallel coordinates. This figure shows two representations (spline and stick, top) of a single structure from the PDB (`3d06`) and its representation in parallel coordinates (bottom). Each residue is represented as a poly-line (a set of line segments) crossing a set of axes, corresponding to attributes of the residue. Note that for some attributes (such as the PDB 'ID'), all lines of residues from the same structure will cross at the same point on the respective axis. In this view, the '2nd' axis (for secondary structure) was used to brush residues composing α-helices (red) and β-strands (blue).

The 3D view supports most commonly used molecular rendering modes, including ball-and-stick, stick, spacefilling (Van-der-Waals), cartoon[39], solvent excluded surface (SES)[40,41], and Gaussian surfaces[42] (see Figure 1 for examples of a spline and stick rendering). Depending on the type of analysis to be conducted and the question to be answered, the standard practise in molecular graphics is to encode additional attributes (such as secondary structure, electrostatics, hydrophobicity etc.) using colour or glyphs to be visualised together directly with the 3D structure in a spatial context. This approach works well for small numbers of attributes, but can become cumbersome for tasks that require consideration of many different attributes — the standard practise is to switch between attributes or use multiple 3D visualisations. This can become tedious, especially for ensembles, which can impede the discovery of patterns in the data, such as relationships, clusters, dependencies, or outliers.

To facilitate the analysis of disorder in ensembles of structures, we augmented the traditional molecular graphics view with parallel coordinates[9], which allow simultaneous visualisation of a large number of attributes across whole ensembles. In parallel coordinates, multidimensional data is represented by a set of axes arranged in parallel, as opposed to the orthogonal layout of axes in Cartesian coordinates. A data point in multidimensional space is then mapped to a poly-line (a set of line segments) in parallel coordinates, intersecting each axis at its respective coordinates. A point-line-duality between 2D Cartesian and parallel coordinates guarantees a unique mapping of patterns from a 2D scatterplot to a 2D parallel-coordinates plot and vice-versa. This allows us to incorporate well-known statistical plots such as the Ramachandran plot[27] into a parallel-coordinates system of protein ensembles. In addition, parallel coordinates allow us to visualise an arbitrary number of dimensions in a single plot, which can be useful to visually spot multidimensional outliers or clusters in the data and thus provide an analyst with information about protein ensembles that might be difficult or impossible to see using an isolated spatial view.

In our implementation, each poly-line represents a residue in one PDB structure, and each axis represents a residue attribute described in Table 1 (see also Figure 1). In order to map categorical data to axes in parallel-coordinates, we cast non-numerical attributes (such as 'ID' or 'Type') to unique integers with no specific order. As a result of our residue-based representation, lines having an attribute in common will cross at the same point on that attribute axis; for example, all α-helical residues will cross the secondary structure axis at the same point. Axes are rendered as vertical lines with labels for the minimum and maximum of the respective dimension (note that we omit some labels in the figures for the sake of clarity). Since the order of axes is crucial for the determination of patterns in the data, our tool allows the order to be changed interactively.

As is typically done in parallel-coordinates views, our tool allows the user to select lines (representing residues), and to *brush* the selection with a user-defined colour. Selected residues are also immediately highlighted in the spatial view using the same colour. Furthermore, selection can be used to define a set of structures to be removed in both views (called *filtering*, as every structure that contains at least one selected residue is removed from the ensemble) or to hide lines in parallel coordinates; these simple but powerful features enable the user to interactively explore the ensemble based on attributes in parallel coordinates.

To further facilitate interactive exploration, we tailored our system for fast rendering. For the spatial view, we chose to represent the polypeptide backbone using lines or splines (similar to the cartoon model), which was usually effective in providing a cogent visualisation for each of the ensembles used.

Initially, the ensembles used in this study tended to be visually cluttered due to the presence of multiple different molecules in various PDB files. With our tool, a user can easily focus on particular parts of the ensemble by selecting attributes from the parallel-coordinates view. For example, as a first step in our analyses, we used our tool to show only one chain in each PDB structure, namely the chain that aligns onto the target sequence (or the first such chain, in the case of oligomers) — see Figure 2 (top).

We also designed our tool to automatically update the attributes of all parallel-coordinates axes whenever the user filters structures. For instance, this update process completely recalculates the PCA, based only on the currently visible structures and updates the 3D superposition using the top-ranked, non-hidden structure as the target structure.

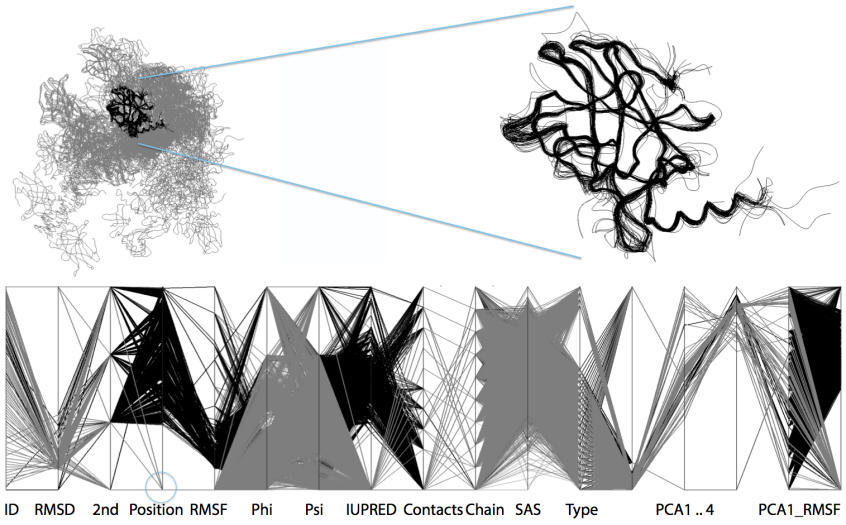


Fig. 2 P53 ensemble of 72 PDB structures, many containing partner proteins and DNA molecules (top left). Our system allows interactive dissection of the ensemble by hiding or revealing structures via selection of attributes from the parallel-coordinates view described in Table 1. The top right view was created from the original ensemble (top left) by a parallel-coordinates selection matching all PDB chains not aligned onto p53 (grey brush in the bottom plot), then filtering all structures that contain brushed residues.

4 Results

In this section, we tested our system by applying it to investigate the three protein ensembles described in Section 3.1. We show how our approach helped to gain insights into the relationship between intrinsic disorder and structural variation for these proteins.

4.1 Cellular Tumour Antigen P53

The ensemble for the human protein *cellular tumor antigen p53* consists of a set of 72 PDB structures that have been aligned to residue positions 94 to 295 of the full-length sequence in UniProt (primary accession `P04637`). This region of p53 is known to bind DNA (e.g. `1TSR`, `2AC0`), as well as partner proteins, such as *p53 binding protein 1* (`1GZH`). Figure 2 shows the initial view, with all PDB structures superimposed (top left) plus a view showing only PDB chains directly aligned onto p53.

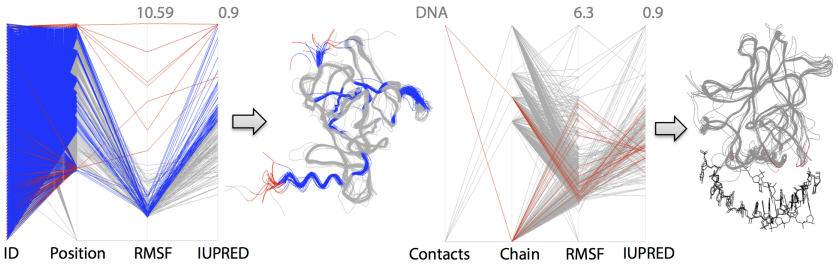


Fig. 3 Disorder and intermolecular contacts in the p53 ensemble. Left: Brushing was used to color red all residues with visually outlying RMSF values (> 3.8Å) — all are predicted to be disordered (i.e., have IUPRED score ≥ 0.5). However, many of the residues predicted to be disordered have low RMSF (blue). Right: Brushing in parallel coordinates allows users to focus on particular partner molecules. In this example, the 'Contacts' axis was used to highlight residues of p53 (red) in direct contact with DNA.

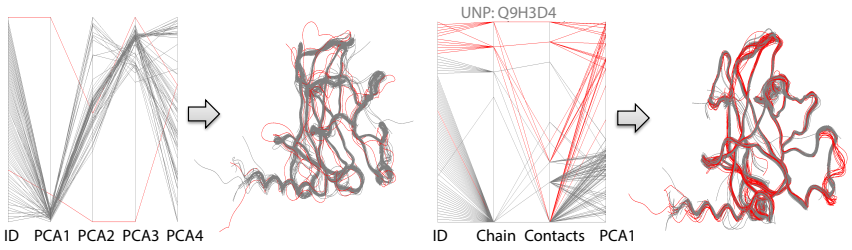


Fig. 4 Finding sub-states in the p53 ensemble. Left: From the set of 'PCA' axes, two prominent outliers in the ensemble are brushed red. Both structures show very different backbones from the ensemble all over the sequence. These outliers were removed for subsequent steps, causing all attributes of the parallel-coordinates plot to be recomputed automatically. Right: Of the remaining structures, selecting from the now updated 'PCA1' axis reveals another subset (red) with distinctly different structure — and somewhat higher apparent disorder — compared to the core ensemble (grey). From the 'Chain' axis, we see that this subset is comprised of molecules p63 and p73 (UniProt accessions `Q9H3D4` and `O15350`), both close relatives of p53.

From the 3D view, it seems that most structures form a rather rigid core, with two outlying regions of high conformational variation — one at the N-terminal α-helix and a second between residues 180 and 190. Applying our tool to this ensemble revealed that residues with very high observed disorder (RMSF) al-

ways had high predicted disorder (IUPRED), while the converse was not true (Figure 3). Figure 4 further illustrates how our tool can be used to successively dissect the p53 ensemble, for example by identifying and removing structures from divergent protein sequences, ultimately deriving a subset of highly similar structures that can be used, e.g., to derive relationships between disorder and secondary structure (Figure 5).

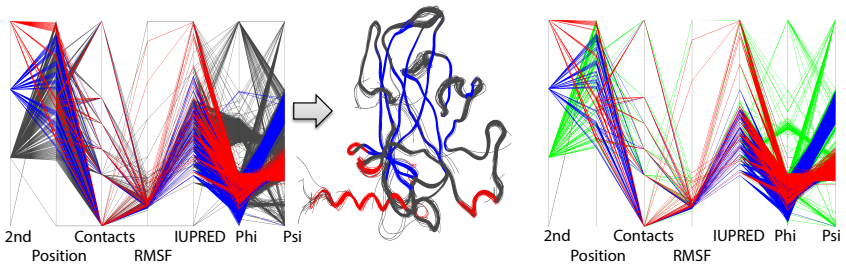


Fig. 5 Relationship between disorder and secondary structure in the p53 ensemble. Left: In this figure, the '2nd' axis (secondary structure) was used to brush α-helices (red) and β-strands (blue). The parallel-coordinates plot shows that helices in this ensemble are more likely to contain disordered residues than β strands, based on both IUPRED score and RMSF. The 'Phi' and 'Psi' axes show the expected configurations for both types of secondary structures, in accordance with the corresponding regions in the Ramachandran plot. Right: Brushing the '2nd' and 'Position' axes reveals residues that adopt different secondary structures across the ensemble. α-helices are shown in red, β-strands in blue, and coils in green. The plot indicates that only a small fraction of residues with ambiguous secondary structure have been predicted to be disordered by IUPRED. Among these, most are associated with low RMSFs. Next steps in the analysis might include filtering by 'Contacts' to investigate the source of the variation in secondary structure.

4.2 Retinoic acid receptor RXR-α

This example encompasses 78 structures that were aligned to residues 132 to 241 of human protein *retinoic acid receptor RXR-α* (P19793). The ensemble initially shows a high degree of variation; one large cluster of similar conformations can be seen, as well as two small clusters (Figure 6). Using the principal component axes in the parallel-coordinates plot (in particular 'PCA3' and 'PCA4'), we were able to quickly identify and brush the smaller cluster. Figure 6 shows two brushing and filtering steps used (from left to right) to filter out these sub-clusters; the remaining cluster, comprising most of the structures, was then examined for disordered regions. Figure 7 compares residues with high predicted disorder (blue) versus those with high observed disorder (red). The blue selection includes many regions with very low observed disorder (e.g. the helices). Upon visual examination of the disordered region (bottom left in the 3D rendering in Figure 7), we found a distinct cluster of structures that could be highlighted using the third principal component (Figure 8). Comparing the spatial views of Figures 7 and 8 further shows that these structural differences are correlated with differences in other disordered regions of the ensemble. After adding the 'Contacts' axis to the parallel-coordinates plot, we can see that these structures are

bound to the same partner protein (PPAR-γ, P37231, see rightmost axis in Figure 8). Obviously, this conformation is not a requirement for binding PPAR-γ, as there are other structures in the ensemble that also bind to PPAR-γ (grey). However, this special conformation seems to prevent the binding of other possible binding partners (blue in selection in Figure 8).

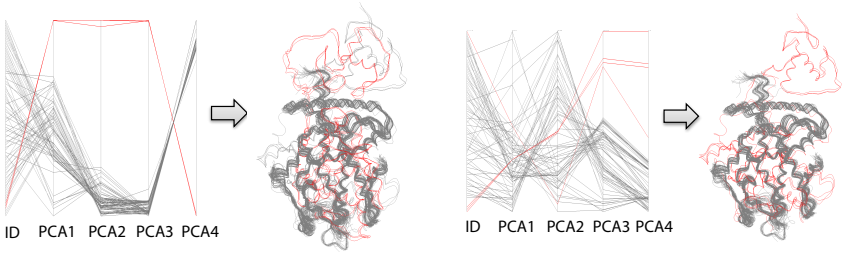


Fig. 6 Successive filtering of sub-states within the RXR-α ensemble. The pattern of lines across the 'PCA' axes clearly shows two distinct clusters (left). Brushing the small cluster (red) allows to remove the corresponding structures from the ensemble, and all attributes to be recomputed. The same procedure can be repeated until only a set of highly similar structures remains (right).

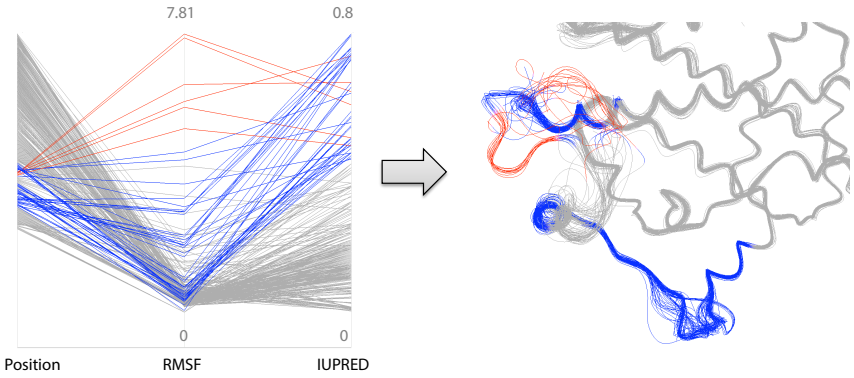


Fig. 7 Predicted vs. observed disorder for the RXR-α ensemble. Blue indicates residues predicted to be disordered (IUPRED score ≥ 0.5), while red indicates residues with visually outlying RMSF values (> 5.2Å). Note in the 3D structure the red coloring was rendered last, and hence conceals some residues coloured blue. The 'Position' axis shows the position of the selected amino acids in the sequence. As with p53, residues with high observed disorder (RMSF) tend to have high predicted disorder (IUPRED), however the converse trend is not as clear.

4.3 Histone H2B

This ensemble consists of 49 structures aligned to residues 30 to 127 of the human protein *histone H2B* (Q96A08). Overall, these structures are highly similar, with only two comparably small regions of disorder at the N- and C-termini (see

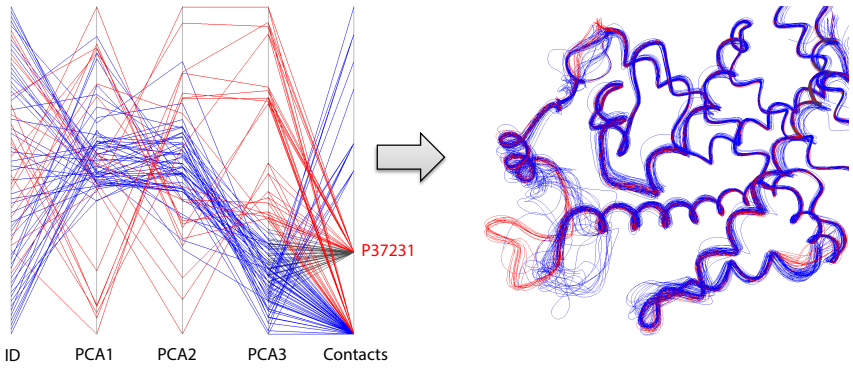


Fig. 8 Disordered region in the RXR-α ensemble. In the disordered region to the right, some structures form a relatively ordered subset, which has been selected (in red) by brushing one of the 'PCAi' axes. Looking at the 'Contacts' axis shows that all members of the ensemble that form the red cluster bind to the same protein (PPAR-γ), whereas none of the members that bind to another protein follows this conformation (blue selection). Thus, binding of PPAR-γ induces order in this region.

Figure 9). As with the previous two ensembles, all residues with highly divergent RMSF values also had high IUPRED scores, while the converse was not true.

5 Discussion

Our tool has many more capabilities than could be presented here, however the cases included in the Results demonstrate that the tool can be useful in dissecting protein ensembles.

Some key trends emerged from the Results. In some cases, we see that binding of partner molecules appears to stabilise regions that are otherwise disordered. This may explain the lack of observed disorder in many residues that were predicted to be disordered, especially for the H2B ensemble, in which all structures have H2B in complex with other histone proteins. The correlation of disorder with secondary structure observed in Figure 5 is also interesting, and may merit further investigation using a larger set of ensembles.

However, the clearest trend to emerge was that in all three ensemble, all residues with high observed disorder (i.e., outlying RMSF values) were predicted to be disordered (IUPRED score ≥ 0.5). Similarly, the converse was consistently not observed, i.e., many residues predicted to be disordered had low RMSF. Like many other methods for predicting disorder, IUPRED is based purely on sequence, and measures the propensity of a sequence region to exhibit disorder, using only amino acid properties. Our results support the suggestion that IUPRED has high recall, but not high precision, for the task of predicting disorder in the heterogenous ensembles used in this study. Stated another way, our results suggest that many of the residues predicted by IUPRED to be disordered are false positives. However, it is important to note that for the ensembles used in this study, the observed RMSF values may differ considerably from the 'true' disorder that occurs when these proteins are alone in solution, with no partner

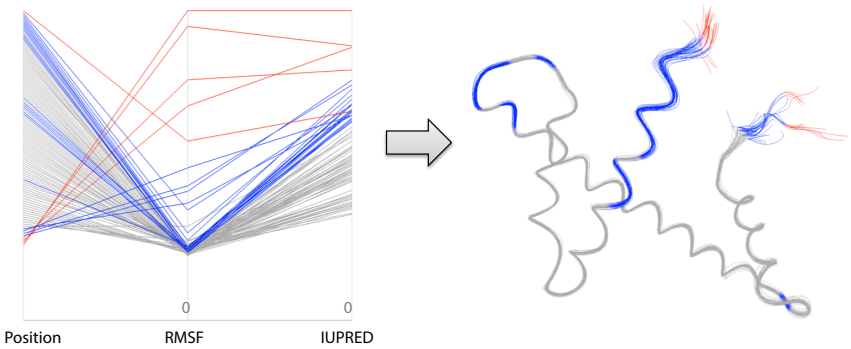


Fig. 9 Predicted vs. observed disorder for the H2B ensemble. Blue indicates residues predicted to be disordered (IUPRED score ≥ 0.5). Red indicates residues with visually outlying RMSF values (≥ 2.2Å), i.e., residues with high observed disorder. High observed disorder always corresponds to high IUPRED scores, but the converse is not true.

molecules — which is the state that IUPRED aims to predict. In contrast, the ensembles we used contained many structures with partner molecules, and almost all were derived from proteins in a crystalline state, not in solution. Nonetheless, such ensembles are a rich and detailed source of experimental data that we believe will be very useful in helping improving our understanding of the functional roles and mechanisms of IDRs.

Overall, the results demonstrate that our tool makes it easy to explore interrelationships in these heterogeneous structural ensembles; in the near future, we intend to use our approach to look at a broader range of cases and, if these correlations stand up, to use them to design statistical tests to further test the trends mentioned above.

While we are using traditional molecular graphics techniques, the parallel-coordinates view greatly facilitates ensemble exploration by showing a large amount of additional information about the ensemble that otherwise would be hidden or difficult to see with a conventional molecular graphics approach. The combination with a spatial 3D view of the molecular structures further enables the analyst to cross-check selected patterns in a well-known environment. The use of parallel coordinates is powerful yet relatively easy to implement, and hence is a good candidate for inclusion in popular molecular graphics tools, such as VMD[17], PyMol[18] or Chimera[19]. Such inclusion would significantly extend the range of attributes, database support, and usability features compared to what is currently available in our implementation, which is a research prototype.

There are some points that need further consideration when using parallel coordinates. One of the most criticised aspect is that patterns depend on the order of axes. There are several approaches to meet the challenge of finding a 'good' axis order: Some authors proposed using an automatic ordering based on various measures such as correlation coefficients or distance metrics (see Heinrich and Weiskopf[32] for an overview), others use manual, interactive reordering of axes (as we did for the system presented in this paper) or show all pairwise correlations in a matrix of parallel-coordinates systems[43].

For very large ensembles and large structures, it may be useful to modify our approach, for example by adding another level of aggregation and compute statistics for whole chains or structures, instead of single residues. This would greatly reduce the number of graphical primitives that may occur, and so improve rendering speed. In part, this has been realised implicitly in our system for axes that depict information on a chain or structure basis (such as the 'Chain' or 'RMSD' axes).

In the future, we plan to extend our tool to achieve a tighter coupling between the parallel-coordinates plot and the 3D visualisation, for example by adding colour maps to an axis, thereby allowing user's to select a parallel coordinate axis and to colour-code the 3D models according to the values on this axis (e.g. using a cool-warm shading). We also plan to add specialised protein ensemble representations (e.g. the 'sausage' visualisation used in MolMol[22]), and to add a range of further protein structure attributes — this will allow us to add further dimensions to the parallel coordinates and potentially find new patterns in the ensembles. Finally, we also plan to investigate the usefulness of this approach for analysing molecular dynamics simulations.

6 Conclusion

Adding views to a system that show different aspects of the data is a well-known and frequently practised approach for a wide range of applications. In this paper, we have shown that parallel coordinates can be a useful add-on to a molecular graphics environment. Applied to the rather complex use-case of analysing protein ensembles, the approach enabled us to dissect these complex datasets, and gain some insight into the correlation between observed and predicted disorder.

Acknowledgements

This work was partially funded by the German Research Foundation (DFG) as part of the Collaborative Research Centre SFB 716 (projects D.4 and D.5). The authors would like to thank Kenneth Sabir for proof reading the manuscript.

References

1 V. N. Uversky and A. K. Dunker, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2010, **1804**, 1231–1264.
2 M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic and A. K. Dunker, *Nucleic Acids Research*, 2007, **35**, D786–D793.
3 A. J. Martin, I. Walsh and S. C. Tosatto, *Bioinformatics*, 2010, **26**, 2916–2917.
4 W. Rieping, M. Habeck and M. Nilges, *Science*, 2005, **309**, 303–306.
5 M. Remmert, A. Biegert, A. Hauser and J. Söding, *Nature Methods*, 2012, **9**, 173–175.
6 A. Amadei, A. B. Linssen and H. J. Berendsen, *Proteins*, 1993, **17**, 412–425.
7 D. van Aalten, D. Conn, B. de Groot, H. Berendsen, J. Findlay and A. Amadei, *Biophysical Journal*, 1997, **73**, 2891–2896.
8 U. Hensen, T. Meyer, J. Haas, R. Rex, G. Vriend and H. Grubmüller, *PLoS ONE*, 2012, **7**, e33931.
9 A. Inselberg, *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*, Springer, NY, USA, 2009.

10 J. Roberts, Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2007, pp. 61–71.
11 M. A. Fisherkeller, J. H. Friedman and J. W. Tukey, Proceedings of ACM Pacific, 1975, pp. 140–145.
12 D. Frishman and P. Argos, *Proteins: Structure, Function, and Bioinformatics*, 1995, **23**, 566–579.
13 Z. Dosztnyi, V. Csizmk, P. Tompa and I. Simon, *Journal of Molecular Biology*, 2005, **347**, 827–839.
14 F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander and M. Scharf, *Journal of Computational Chemistry*, 1995, **16**, 273–284.
15 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Research*, 2000, **28**, 235–242.
16 W. Kabsch, *Acta Crystallographica Section A*, 1976, **32**, 922–923.
17 W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics*, 1996, **14**, 33–38.
18 W. L. DeLano, *The PyMOL Molecular Graphics System*, DeLano Scientific, Palo Alto, CA, USA, 2002.
19 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *Journal of Computational Chemistry*, 2004, **25**, 1605–1612.
20 E. A. Coutsias, C. Seok and K. A. Dill, *Journal of Computational Chemistry*, 2004, **25**, 1849–1857.
21 A. J. M. Martin, I. Walsh and S. C. E. Tosatto, *Bioinformatics*, 2010, **26**, 2916–2917.
22 R. Koradi, M. Billeter and K. Wüthrich, *Journal of Molecular Graphics*, 1996, **14**, 51–55.
23 Y. Zhao, D. Stoffler and M. Sanner, *Bioinformatics*, 2006, **22**, 2768–2774.
24 D. Bromley, S. J. Rysavy, R. Su, T. Toofanny R. D., Schmidlin and V. Daggett, *Bioinformatics*, 2014, In press.
25 M. Petrek, M. Otyepka, P. Banás, P. Kosinová and J. Koca J, Damborský, *BMC Bioinformatics*, 2006, **7**, 316.
26 A. Spaar, C. Dammer, R. R. Gabdoulline, R. C. Wade and V. Helms, *Biophys. J.*, 2006, **90**, 1913–1924.
27 G. Ramachandran, C. Ramakrishnan and V. Sasisekharan, *Journal of Molecular Biology*, 1963, **7**, 95–99.
28 S. I. O'Donoghue, D. S. Goodsell, A. S. Frangakis, F. Jossinet, R. A. Laskowski, M. Nilges, H. R. Saibil, A. Schafferhans, R. C. Wade, E. Westhof and A. J. Olson, *Nature Methods*, 2010, **7**, S42–S55.
29 J. Dietzsch, J. Heinrich, K. Nieselt and D. Bartz, Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, 2009, pp. 179–186.
30 J. Hasenauer, J. Heinrich, M. Doszczak, P. Scheurich, D. Weiskopf and F. Allgöwer, *EURASIP Journal on Bioinformatics and Systems Biology*, 2012, **2012**, 1–13.
31 O. Rübel, G. H. Weber, S. V. E. Keränen, C. C. Fowlkes, C. L. L. Hendriks, L. Simirenko, N. Y. Shah, M. B. Eisen, M. D. Biggin, H. Hagen, D. Sudar, J. Malik, D. W. Knowles and B. Hamann, Proceedings of the Eurographics/ IEEE-VGTC Symposium on Visualization, 2006, pp. 203–210.
32 J. Heinrich and D. Weiskopf, STAR Proceedings of Eurographics, 2013, pp. 95–116.
33 B. T. Luke, *Journal of Chemical Information and Modeling*, 1993, **33**, 135–142.
34 O. M. Becker, *Journal of Computational Chemistry*, 1997, **18**, 1893–1902.
35 S. Grottel, G. Reina, C. Dachsbacher and T. Ertl, *Computer Graphics Forum*, 2010, **29**, 953–962.
36 B. Wellmann, *MSc thesis*, Faculty for Informatics, Technical University of Munich, 2012.
37 M. Kalemanov, *BSc Honours thesis*, Faculty for Informatics, Technical University of Munich, 2012.
38 The UniProt Consortium, *Nucleic Acids Research*, 2013, **41**, D43–D47.
39 M. Krone, K. Bidmon and T. Ertl, Proceedings of Theory and Practice of Computer Graphics, 2008, pp. 115–122.
40 M. Krone, K. Bidmon and T. Ertl, *IEEE Transactions on Visualization and Computer Graphics*, 2009, **15**, 1391–1398.
41 M. Krone, S. Grottel and T. Ertl, Proceedings of IEEE Symposium on Biological Data Visualization, 2011, pp. 17–22.
42 M. Krone, J. E. Stone, T. Ertl and K. Schulten, EuroVis–Short Papers, 2012, pp. 67–71.
43 J. Heinrich, J. Stasko and D. Weiskopf, EuroVis–Short Papers, 2012, pp. 37–41.