**Inferring interface residues from the accessible interaction space defined by distance restraints to improve HADDOCKing models**

\emph{G.C.P. van Zundert and A.M.J.J. Bonvin. To be submitted.}

**\Section{Introduction}**

Uncovering the precise atomic structures of protein complexes is a highly sought-after enterprise. Experimental techniques that provide atomic resolution, mainly X-ray crystallography and NMR spectroscopy, have, unfortunately, so far only revealed a fraction of the whole interactome, the set of all interacting proteins \cite[Stein2011]. Protein-protein docking aims to predict the structure of a complex from its individual proteins to close this knowledge gap \cite[Moreira2010]. However, its success rate using solely first-principles – the so-called *ab initio* docking - is generally low \cite[Huang2015]. Integrating additional information (if reliable) during the docking process can increase the confidence in the resulting models, especially when knowledge about the location of the interface is available \cite[Rodrigues2014].

Mass spectrometry coupled with cross-linking is an upcoming and promising biochemical method that provides inter-residue distance restraints \cite[Leitner2010, Rappsilber2011]. Multiple chemistries are being developed, making the approach more robust and increasing the information content \cite[Leitner2014]. Several software packages have already been developed for visualizing crosslinks, and calculating their path length \cite[Kahraman2011, Holding2013, Kosinski2015]. In the previous chapter we introduced DisVis to quantify and visualize the information content of distance restraints. However, interpreting multiple long-range distance restraints between components from a structural perspective and deducing the interaction surface remains tedious. A simpler interpretation is gained if interface residues can be deduced from the data, as these map directly onto the individual chains and offer a straightforward prediction of the active site. In addition, in the development of protein-protein inhibitors, mainly the protein interfaces are of importance, and less so the precise complex's structure \cite[Sable2015]. Such interface information might be useful in complementing the cross-linking distance restraints, since the allowed distance ranges for the latter can be relatively wide (up to 30Å \cite[Merkley2014]).

Inclusion of cross-linker based distance restraints has already been shown to improve the modeling of both proteins and protein-complexes with the Rosetta software \cite[Kahraman2013], and to heavily decrease the number of accessible conformations of a complex (see \inchapter[chapter:disvis]). Our data-driven docking software HADDOCK is capable of directly incorporating distance restraints during the docking \cite[Dominguez2003, deVries2007]. Currently, Mass Spec Studio provides an advanced software platform for integrative modeling from MS data, such as hydrogen/deuterium exchange and cross-links, with HADDOCK \cite[Rey2014]. However, no thorough benchmark study has been performed to measure the impact and effectiveness of incorporating cross-link based distance restraints in HADDOCK.

Here we introduce a method to infer interface residues when distance restraints are available in addition to models or structures of the components. The method is benchmarked on 90 complexes taken from the Protein-Protein Docking Benchmark 4.0 (PPDB4.0) \cite[Hwang2010] for 3, 5 and 7 cross-links, respectively, with an upper distance restraint of 30Å, comparable to the information content that is provided by disuccinimidyl suberate (DSS) and and bis-sulfosuccinimidyl-suberate (BS3) cross-links \cite[Merkley2014]. Finally, we show how this can be combined with HADDOCK to complement unambiguous distance restraints by derived interface information, benchmarking it on 24 cases of the PPDB4.0.

**\Section{Methods}**

*\Subsection{Inferring interface residues from distance restraints}*

In the previous Chapter, we have introduced the concept of the accessible interaction space, the set of all possible complexes that are consistent with a given number of distance restraints. Indeed, the presence of distance restraints between two interacting macromolecular biomolecules can significantly reduce their accessible interaction space. To infer residues that are likely to be at the interface, we assume that these residues are often found to be interacting in the interaction space consistent with the restraints. Important residues may be determined by performing a full-exhaustive 6 dimensional search of the three translational and three rotational degrees of freedom and counting the number of interactions that each solvent accessible residue forms in complexes consistent with a given number of restraints. We define two residues to be interacting when their \CA\ -- \CA\ distance is smaller than 10Å. We only consider the \CA-atoms of solvent accessible residues of both chains to make the computations more tractable as the number of possible interactions scales with \m{A^2} with \m{A} the number of atoms involved. The average number of interactions per complex (AIC) that a residue \m{i} forms is given by

\placeformula[eq:count-interactions]

\startformula

\overline{N}\_i = \frac{\sum\_R^{\bf P} w\_R \sum\_C^{\boldC\_R}

I\_{C}} {\sum\_R^\boldP w\_R \boldC\_R}

\stopformula

where the first summation is over all rotations \m{\boldP} indexed by \m{R}; \m{w\_R} is a weight factor to correctly average over rotation space; the second summation is over all complexes \m{\boldC\_R} that are formed within a translational scan indexed by \m{C}; and \m{I\_{C}} is the number of interactions that are formed by residue \m{i} in each sampled complex \m{C}.

This approach has been implemented in DisVis (see \inchapter[chapter:disvis], https://github.com/haddocking/disvis), which requires for the interaction analysis an extra input file containing the solvent accessible residue numbers for the fixed and scanning chain. As a result, DisVis outputs a file containing the number of interactions that are formed by each residue for complexes consistent with at least \m{N} restraints.

*\Subsection{Benchmarking interface residue extraction}*

We benchmarked our approach on 90 complexes taken from the protein-protein PPD4.0, of which 58 were classified as Easy, 14 as Medium, and 18 as Difficult. Virtual cross-links were calculated using a local version of the XWalk software \cite[Kahraman2011] on the bound complex. The virtual crosslinks were chosen such that the solvent accessible surface (SAS) distance was shorter or equal than 34Å \cite[Kahraman2013], and the Euclidean distance smaller or equal than 30Å \cite[Merkley2014] and the cross-linked residues should be present in both the bound and unbound proteins. The cross-links were randomly picked from the list of all virtual crosslinks using the SAS-distance dependent probability distribution as was used by \citeauthor{Kahraman2013} to mimic experimental cross-link data: 0 -- 10Å 9\%; 10 -- 15Å 18\%; 15 -- 20Å 34\%; 20 -- 25Å 22\%; and 25 -- 34Å 16\%.

The solvent accessible residues were determined by running {\it naccess} \cite[Hubbard1992] on the two unbound proteins. Residues that had a relative solvent accessibility of the main or side chain of 50\% or higher were used as surface residues. DisVis runs were performed for each complex using 3, 5, and 7 random restraints, respectively, with a 5.27° rotational sampling, and default values for the voxel spacing (1Å), maximum clashing volume (200Å3) and minimum interaction volume (300Å3). The AIC was only calculated from the complexes consistent with all restraints.

Correct interface residues were taken from the experimental structure of the complex using the above definition of interaction, under the restriction that the residue was also present in the unbound proteins. To analyze the predictive capabilities the precision P and recall R were calculated as \placeformula[eq:precision]

\startformula

\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}}

\stopformula

\placeformula[eq:recall]

\startformula

\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}

\stopformula

where TP, FP and FN stand for True Positive, False Positive, and False

Negative, respectively.

*\Subsection{HADDOCKing with virtual cross-links}*

To determine whether the inclusion of DisVis-determined interface residues aids the docking process of HADDOCK, we benchmarked HADDOCK using 24 complexes of the PPDB4.0, of which 16 were the same as those used by \citeauthor{Kahraman2013}. The remaining 8 were randomly picked Easy complexes, as the previous 16 already consisted of 7 Medium and 9 Difficult cases. HADDOCK was benchmarked with 4 different protocols:

* using the restraints directly as unambiguous distance restraints (unambig) with a minimal and maximal Euclidean length of 0 and 30Å, respectively;
* using the unambiguous restraints in combination with center-of-mass restraints \cite[Karaca2013];
* using solely DisVis-based ambiguous interaction restraints (AIRs);
* and using a combination of the unambig restraints and DisVis-based AIRs.

Each protocol was performed with 3, 5 and 7 generated virtual cross-link restraints, respectively, as described in the previous section. The DisVis-based AIRs were determined as follows: a DisVis run was performed as described above using the unbound structures of the complex together with the virtual cross-links; active residues were chosen such that their AIC consistent with all restraints had to be larger than 1; passive residues will be chosen with a to be determined AIC cutoff based following the method described in the previous section. Per HADDOCK run 1000 it0-structures were written to file, using 5 trials per file and 180° sampling, resulting in 10000 sampled solutions (\m{1000 \times 5 \times 2}), of which the 200 best scoring solutions were subjected to the semi-flexible refinement (it1 and itw) (default settings of the server). The solutions were analyzed by calculating the ligand-RMSD (l-RMSD) against the native complex, by first optimally fitting the receptor chain and afterwards calculating the RMSD of backbone atoms of the ligand chain using ProFitV3.1 \cite[Martin2009]. Models with an l-RMSD lower than 10Å were considered acceptable.

**\Section{Results and discussion}**

*\Subsection{Inferring interface residues from distance restraints}*

The analysis of the DisVis benchmark is shown in \infigure[fig:precision-recall] for all complexes and for each difficulty category, by plotting the precision and recall against the AIC cutoff of a residue. This shows for example that for all complexes, using 7 cross-links, for residues that have a AIC ≥ 1.0 the precision is approximately 60\%, i.e. 60\% of all residues with a AIC ≥ 1.0 are true interface residues; the recall at 1.0 AIC is around 40\%, meaning that 40\% of all true interface residues is still retained in the set of residues satisfying the AIC cutoff condition.

\placefigure[top][fig:precision-recall] {\getbuffer[cap:precision-recall]}

{\externalfigure[fig:precision-recall]}

As expected, the precision increases with increasing AIC, while the recall rate drops steadily. Also, both precision and recall rise with the number of available cross-links, reflecting the higher information content of the distance restraints. On average for all complexes, the precision starts at 30\% regardless of the number of available restraints, and rises with increasing cutoff to 50, 70 and 80\% for 3, 5, and 7 restraints, respectively. For Easy complexes this increases even to 60, and 85 and 90\%, while for Medium and Difficult complexes the precision is significantly lower, dropping down to 70\% for Difficult complexes in the presence of 7 restraints. We attribute the noisy behavior of the precision, especially for Medium complexes, to the smaller number of complexes sampled compared to the number of Easy complexes (14 versus 58).

The recall %age averaged over all complexes starts around 90\%, a consequence of how surface residues were chosen: because of small conformational changes between the bound and unbound chains, residues that are regarded as solvent accessible in the bound form might not be accessible in the unbound form. This also explains the lower starting recall rate of Medium and Difficult complexes, with the latter being lower than 80\%, as these more challenging complexes typically exhibit greater conformational change between their bound and unbound form by definition. Not surprisingly, the recall rate decreases steadily with increasing cutoff, as fewer residues will be satisfying the AIC cutoff condition. In contrast to the precision, the recall rate does increase significantly with the inclusion of more restraints for the more challenging categories (Medium + Difficult).

*\Subsection{HADDOCKing with virtual cross-links}*

The HADDOCK benchmark results for the 8 Easy complexes, and the 16 Medium and Difficult complexes are displayed in \infigure[fig:benchmark-easy] and \in[fig:benchmark-rosetta], respectively, in terms of the structure with the lowest l-RMSD after the water-refinement stage, irrespective of its rank. For the DisVis-based AIRs the cutoff AIC for active residues was set to 1, corresponding to a precision of 40 to 60\% and a recall of 20 to 40\% for 3 and 7 cross-links; the AIC cutoff for passive residues was set to 0.1, as the precision increase is steeper approximately until that point, while still keeping a reasonable recall rate of approximately 80\%. For the Easy complexes the unambiguous restraints approach was successful in 50, 37.5 and 75\% of the cases when using 3, 5 and 7 cross-links. Adding the center-of-mass restraint this changed to 62.5, 50, and 62.5\%. Using the DisVis-based AIRs resulted in a success rate of 25, 62.5, and 75\%, and combined with the unambiguous restraints this increased to 50, 75 and 75\% success rate. Interestingly, increasing the number of cross-links does not necessarily result in better structures when using only the unambiguous restraints: the success rate with 5 cross-links is markedly lower than using 3. Also, strangely, the unambiguous approach with center-of-mass restraint is the only method for which no acceptable solutions are generated for the 1QA9 complex, even with 7 restraints included. The results of the DisVis-based approaches, however, are improving with increasing numbers of cross-links.

\placefigure[top][fig:benchmark-easy]

{\getbuffer[cap:benchmark-easy]}

{\externalfigure[fig:benchmark-easy]}

\placefigure[top][fig:benchmark-rosetta]

{\getbuffer[cap:benchmark-rosetta]}

{\externalfigure[fig:benchmark-rosetta]}

In the 16 Medium and Difficult complexes the unambiguous approach is successful in 18.75, 62.5 and 50\% of the cases for 3, 5 and 7 cross-links, respectively; with inclusion of the center-of-mass restraint this becomes 37.5, 56.25, and 75\%. The DisVis-based AIRs result in 12.5, 50 and 50\% successful cases; including the unambiguous restraints this increases to 12.5, 56.25 and 62.5\%.

As with the Easy complexes, the success rate of the unambiguous restraints shows no steady improvement with an increased number of cross-links. Furthermore, the combination of DisVis-based AIRs and unambiguous restraints is superior in general to using only DisVis-based AIRs.

The HADDOCK results using 7 cross-links are compared against Rosetta in \infigure[fig:benchmark-rosetta]{C}. Rosetta was successful in 68.75\% of the cases, a slightly higher %age than HADDOCK using DisVis-based AIRs with unambiguous restraints, but lower than when using unambiguous with center-of-mass restraints. However, in general the results are comparable.

Taking all 24 complexes together we conclude that using only the unambiguous restraints results in a success rate (again defined as generating at least one native-like model in the set of 200 refined models) of 29, 54 and 58\% for 3, 5 and 7 cross-links, respectively; using ambiguous restraints with center-of-mass restraints this increases to 46, 54 and 71\%. Using solely DisVis-based AIRs the success rates are respectively 17, 54 and 58\%; and DisVis-based AIRs combined with unambiguous restraints 25, 63 and 67\%.

Based on these results we conclude that if only 3 cross-links are available, the best protocol to use is to combine unambiguous restraints with center-of-mass restraints. If 5 or more cross-links are available it is best to either again combine the unambiguous restraints with the center-of-mass restraints, or combine them with DisVis-based AIRs.

**\Section{Conclusions}**

In this Chapter we have introduced a method to infer interface residues by enumerating all interactions a residue forms in the interaction space consistent with all restraints and normalizing it against the number of accessible complexes. It was shown that residues with a higher AIC are more likely to be interface residues with precision reaching almost 90\% for rigid complexes. This information can be used to guide future mutagenesis studies and map out the interface, irrespective of the quality of the generated models of the complex. In addition, we benchmarked several protocols within HADDOCK that incorporated cross-link based distance restraints. Based on the analysis, using the cross-links directly as unambiguous restraints is sub-optimal, and instead should be complemented with either center-of-mass restraints or DisVis-based AIRs. Furthermore, we have shown that HADDOCK and Rosetta are very comparable in their docking performance when including the distance restraints. However, it should be noted that we only investigated the quality of the best model. Further analysis should also address the ranking of generated models, as using various types of restraints might significantly affect the scoring functions’ ability to identify near-native models.



The precision and recall rates are plotted against the cutoff used to calculate the average-interactions-per-complex. The results are shown averaged over all 90 benchmarked complexes (58 Easy, 14 Medium, and 18 Difficult).

For each complex the best structure is plotted in terms, of the l-RMSD, for the four procedures tested using (A) 3, (B) 5, and (C) 7 cross-links. The dotted line is the ligand-RMSD cutoff for an acceptable model. Unambig: unambiguous distance restraints; unambig + com: unambiguous restraints combined with center-of-mass restraints; disvis: DisVis-based ambiguous interaction restraints (AIRs); disvis + unambig: unambiguous restraints combined with DisVis-based AIRs.



For each complex the best structure is plotted in terms, of the l-RMSD, for the four procedures tested using (A) 3, (B) 5, and (C) 7 cross-links. The dotted line is the ligand-RMSD cutoff for an acceptable model. Unambig: unambiguous distance restraints; unambig + com: unambiguous restraints combined with center-of-mass restraints; disvis: DisVis-based ambiguous interaction restraints (AIRs); disvis + unambig: unambiguous restraints combined with DisVis-based AIRs. In (C) also the best structures of the Rosetta benchmark are shown \cite[Kahraman2013].