Published in final edited form as:

Methods Mol Biol. 2018 ; 1764: 491–504. doi:10.1007/978-1-4939-7759-8_31.

Selecting Conformational Ensembles Using Residual Electron and Anomalous Density (READ)

Loïc Salmon^{1,*}, Logan S Ahlstrom^{2,3}, James C.A. Bardwell^{2,3}, and Scott Horowitz^{4,*}

¹Institute of Molecular Biology and Biophysics, ETH Zürich, 8093 Zürich, Switzerland ²Howard Hughes Medical Institute, Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA ³Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA ⁴Department of Chemistry & Biochemistry and the Knobel Institute for Healthy Aging, University of Denver, Denver, CO 80208, USA

Abstract

Heterogeneous and dynamic biomolecular complexes play a central role in many cellular processes but are poorly understood due to experimental challenges in characterizing their structural ensembles. To address these difficulties, we developed a hybrid methodology that combines X-ray crystallography with ensemble selections typically used in NMR studies to determine structural ensembles of heterogeneous biomolecular complexes. The method, termed READ, for Residual Electron and Anomalous Density, enables the visualization of heterogeneous conformational ensembles of complexes within crystals. Here we present a detailed protocol for performing the ensemble selections to construct READ ensembles. From a diverse pool of binding poses, a selection scheme is used to determine a subset of conformations that maximizes agreement with the X-ray data. Overall, READ is a general approach for obtaining a high-resolution view of dynamic protein-protein complexes.

Keywords

crystallography; ensemble; conformational dynamics; protein structure; structural biology

1 Introduction

The majority of proteins perform their cellular functions through interactions with other proteins [1]. Thus, obtaining high-resolution structural models of protein-protein interactions yields indispensable insight into biological function and can facilitate the development of therapeutic strategies. However, many protein-protein interactions involve highly dynamic protein-protein association (e.g., intrinsically disordered protein-partner interactions in cell signaling and regulation networks [2]) and thus are particularly challenging for traditional X-ray crystallography and NMR methods. Although recent advances in X-ray

*To whom correspondence should be addressed: loic.salmon@mol.biol.ethz.ch and scott.horowitz@du.edu.

⁵Notes

crystallography [3] and NMR spectroscopy [4–6] allow one to analyze biomolecules in terms of ensembles as opposed to single structures, the information from these techniques is often limited by system size and complexity. X-ray crystallography can provide valuable information on small-scale conformational changes, but observing heterogeneous conformational states falls beyond the reach of current crystallographic techniques. Such studies can be done with NMR, but the systems must exhibit particular properties such as a certain conformational exchange rate, favorable interactions with external alignment media, or sufficient chemical shift dispersion.

To further our ability to obtain high-resolution structural ensembles of dynamic protein-protein interactions, we need to develop new approaches that combine the benefits of experiments and computational strategies in a transferable manner. The elucidation of ensembles of RNA, DNA-protein complexes, and highly flexible proteins by NMR has provided valuable insights into many complex biological problems [7,4–6]. NMR-based approaches can derive highly heterogeneous ensembles to describe complex and flexible interactions, a feat that current crystallographic methods are unable to replicate. Here, we combine crystallographic measurements with an NMR-like ensemble approach [8,9], yielding a general method for observing structural and dynamic properties of proteins and protein complexes [10]. The approach, termed READ, for Residual Electron and Anomalous Density, determines the conformational ensembles of heterogeneous and dynamic proteins by X-ray crystallography. We will first describe the overall approach in general terms and then provide methodological details.

2 Materials

2.1 Residual Electron and Anomalous Densities

1. The foremost requirement for READ are sturdy and highly reproducible crystals of the protein or complex of interest that diffract to ~ 3 Å or higher resolution. Given the ambiguous nature of the electron density, we recommend validating the presence of any components of interest via crystal washing followed by gel or mass spectrometry analysis.
2. Initial attempts at modelling the structure should be made using standard crystallographic model building and refinement procedures, taking the refinement to as close to completion as possible. The resulting structure may later be used as part of a molecular dynamics simulation of the crystal, and the refined electron density will be compressed for use in the selection procedure.
3. Methods for incorporating anomalous scatterers into crystallization components can vary depending on the system. We have successfully used peptide synthesis with unnatural amino acids like iodophenylalanine [10]. Other potential methodologies include amber stop codon suppression [11] to incorporate iodine-containing amino acids, chemical labeling of natural amino acids, such as tyrosine-iodine modifications [12], or binding of heavy metals to cysteine residues. How many anomalous substitutions are required will likely vary

depending on the system. As a general guideline, we recommend beginning by substituting non-conserved residues or using chemically similar substitutions.

2.2 Conformational Pool

1. The second pre-requisite to using the READ protocol is to obtain a large conformational pool of the dynamic component of the crystal. Depending on the size and level of mobility of the substrate, different computational approaches could be used. For cases in which the dynamic fragment is especially large, undergoes high-amplitude conformational changes, or translates considerably within the crystal, we have found a coarse-grained approach to be effective [10,13,14]. Monte-Carlo based approaches such as flexible-meccano [15] could also be useful, or for more computationally amenable systems all-atom simulations [16] could be applicable.
2. The pool of conformations in the crystalline environment should sample all possible conformational space. The crystalline environment can be simulated either through multiple pseudo-crystal environments [10] or by altering the simulation's boundary conditions to match the crystal's dimensions and space group. It will often be more appropriate to favor extensive sampling of space over increased accuracy of local structural details, in which case a coarse-grained approach is especially appropriate.

2.3 Computational Resources

1. The compute time for the selection procedure will depend on various parameters including the amount of experimental data, the size of the conformational pool, and the exact parameters used. For simple cases, a selection should run on a modern desktop computer in a few hours or less. Therefore, standard selection procedures can be easily setup using a single desktop. Extensive testing, multiple cross-validations, or bootstrapping simulations may, however, require a small calculation cluster in order to run in a time-efficient manner (See Note 1). Additionally, the refinement of large conformational ensembles through classical crystallographic approaches will greatly benefit from using a small compute cluster with a crystallographic refinement package such as Phenix [17] or CCP4 [18] installed.
2. To run the READ selection, the computer or cluster must have MATLAB 2015b or later installed. The READ selection program is implemented as a MATLAB toolbox called READ.mltbx. To install the toolbox, simply double click READ.mltbx, which will install the code and add it to your path. The MATLAB toolbox can be downloaded from <https://bitbucket.org/umarsdev/mcdb-bardwell-ga>. Example data for the selection can be downloaded from https://bitbucket.org/umarsdev/mcdb-bardwell-ga_datafiles.

¹Genetic algorithms can easily be encoded to allow for parallel computing, which can allow for the treatment of more complex systems. The MATLAB program provided here takes advantage of this property when running multiple parallel selections.

3 Selection Algorithm

1. The selection of the conformational ensemble can be carried out using various algorithms. Selection procedures can be performed using a Monte Carlo Metropolis approach [9] or using a genetic algorithm [8]. Both approaches can potentially work, but in our experience, a genetic algorithm is computationally more efficient when adequately optimized and was successfully used in our recently published example [10].
2. The overall READ procedure uses the residual electron and anomalous data to select sub-ensembles from a large structural pool (See Note 2). Starting from a pool of conformations of typically 10,000 members, the selection is initiated by randomly generating X ensembles from this pool, where each ensemble contains N conformers. In a typical selection, X is set to 100, and N typically ranges from 1 to 20 [10]. Then, K evolution steps are iteratively performed. K is typically 50,000. Tweaking these parameters will affect the outcome of the selection, and the parameters need to be optimized as part of this protocol. Each evolution step has two phases. The first phase creates new ensembles to diversify the set of possible solutions, and the second phase performs a selection to determine which of the available ensembles best fit the data.
3. New ensembles are generated in the first phase. $4X$ new ensembles are produced by four methods: X by reproduction, X by external mutations, X by internal mutations, and X randomly. Reproduction generates a new ensemble by randomly selecting conformers from two parent ensembles. Mutations are performed by randomly changing one conformer in the ensemble with a new conformer either using conformations already present (internal mutation) or not present (external mutation) in one of the selected ensembles. At the end of this phase, the original X ensembles are combined with the $4X$ new ensembles, providing a total of $5X$ ensembles. This procedure is implemented so that a given conformation cannot appear more than once in any ensemble, and any ensemble cannot be present more than once at each step. This ensures maximal diversity in the sampling of the conformational space.
4. In the second phase, the ensembles are evaluated. The $5X$ ensembles are clustered randomly into T tournaments. Within each tournament, the ensembles are compared to determine which best fits the data. The best fitting ensembles from each tournament are retained, forming a new set of X ensembles. During the course of the selection procedure, the number of tournaments (T) successively decreases, typically from X to 1, increasing the selection pressure. Low selection pressure is important at the beginning of the selection to ensure sufficient diversity, whereas increasing the selection pressure later improves the convergence of the result [19]. In the final rounds of the selection, only a single tournament is used to produce the single best fitting ensemble.

²See Figure 2 in Horowitz et al. (2016) for a diagram of an example procedure.

5. To select the best fitting ensembles, a χ^2 function is used in the tournaments. Best fitted ensembles are those that minimize this target function by reducing the difference between the experimental data and that back-calculated from the ensembles. This χ^2 function is defined as:

$$\chi^2 = \chi_{elec}^2 + \lambda \chi_{anom}^2$$

where χ_{elec}^2 and χ_{anom}^2 are the contributions arising from the electron density and the anomalous positions, respectively, and λ is a scaling factor that allows one to adjust the relative weight of the two data types.

6. For the anomalous data, it is necessary to predict the position of the anomalous atoms from the coordinates. However, the initial conformational sampling may be obtained for the sequence of the non-modified peptide chain and thus might not explicitly include the modifications containing the anomalous atoms. Instead, the anomalous atom positions are inferred from the geometry of the conformers. For example, using iodophenylalanine, the fixed aromatic ring dictates a distance between the C α and the iodine scatterer of 6.5 Å. Thus, the selection searches for iodines as close as possible to 6.5 Å from C α of the mutated residue [10]. This particular approach does not take into account any angular information from the conformer; however, angular information could also be encoded if the experimental information and computational procedure allow for it. Although this approach avoids assumptions regarding backbone angles, it does introduce the problem that anomalous atoms can possibly be placed in positions that cause steric clashes or induce backbone distortions.
7. The anomalous data target function is encoded into the selection using:

$$\chi_{iodo}^2 = c^2 \sum_i \left(1 - \exp \left[- \left(\frac{D_i^{calc} - D_i^{exp}}{\sigma^{anom}} \right)^2 \right] \right)$$

where i represents the different anomalous atom positions, D_i^{calc} is the closest distance between the anomalous atom position and the corresponding C α of any of the conformers in the ensemble, D_i^{exp} is the corresponding expected distance, σ^{anom} is a weight estimating the uncertainty in the anomalous atoms' position, and c is fixed at a canonical value of 2.9846 (See Note 3). The value of D_i^{exp} depends on the nature of the amino acid bearing the anomalous scatter and represents the distance between the anomalous scatter atom and the C α of the

³Using this target function provides a classical χ^2 for data near the target value but reaches a plateau for data points that are very poorly reproduced, and thus better handles outliers. This function is especially useful to avoid the situation in which poor reproduction of a single anomalous atom dominates the target function.

same residue (e.g., 6.5 Å in the case of iodophenylalanine). The value of δ^{anom} in this protocol was the mean squared displacement of the scatterer, derived from the B-factors of the atoms after crystallographic refinement of the anomalous scatterers with a fixed occupancy. In this protocol, occupancy was fixed at 0.5, and crystallographic refinement was performed in Phenix [17].

8. In general, electron density maps contain an enormous amount of data. The number of data points in a typical map is computationally too expensive for practical direct use with this selection procedure. As a result, procedures to compress this information are required. The first compression approach is to only describe the electron density in the space covered by the conformational pool. Importantly, any density outside the initial conformational sampling might be important for understanding the system, but it cannot be recovered by any selection procedure if it is not initially sampled. Removal of data not sampled by the conformational pool dramatically decreases the number of points needed to describe the data by shifting from the entire experimental three-dimensional space of the asymmetric unit to the ensemble of coordinates that account for all the conformers in the pool. Second, the electron density is binned in three-dimensional space to create a three-dimensional electron density histogram. Third, the electron density is filtered for signal intensity and sufficiently averaged to further compress the data. Finally, computing the electron density contribution of each conformer before the selection and then using these pre-computed values as an input in the selection greatly speeds the procedure. Thus, in the selection procedure, only simple operations such as averaging over a limited number of points are required to determine the fit of the ensembles to the electron density.

9. For the electron density evaluation, the following target function is used:

$$\chi_{elec}^2 = \sum_i \left(\frac{\rho_i^{calc} - \rho_i^{exp}}{\delta^{elec}} \right)^2$$

where i spans the different residual electron density data points (discussed in section 4.1), ρ_i^{exp} is the experimental value, ρ_i^{calc} is the predicted value from the considered ensemble, and δ^{elec} represents a constant weight estimating the noise in the residual electron density map (See Note 4).

4 Methods

4.1 Preparing the Electron Densities

1. Prepare individual density maps. Before starting the READ procedure, each conformer in the initial large conformational pool must be crystallographically

⁴Residual electronic density remains between 0 and 1, making the use of the more complex χ^2 unnecessary.

refined, one at a time, together with any fixed atoms in the structure. The refinement procedure should include at least B-factor refinement, but could include other parameters as well. The output Ca atoms of the conformers from this refinement then serve as the input structures for the selection. Depending on the conformational pool, geometric restraints may be necessary during refinement to ensure that the conformers maintain energetically reasonable conformations. This crystallographic refinement can be accomplished with any modern crystallographic software package such as Refmac [20] or Phenix [17], but a computing cluster will likely be required to efficiently refine the entire conformational pool (See Note 5).

2. Average the electron density maps. Taking the refined electron density for each conformer, an average map over the simulation must be produced for subsequent back-prediction of the conformer electron density. This averaged map will allow consistent sampling of the electron density within the crystal space. The average 2mFo-DFc map from all of the refined conformational pool snapshots is calculated using Phenix:

```
Run command phenix.average_map_coeffs file_list=file_list.dat  
labin_mtz="FP=2FOFCWT PHIB=PH2FOFCWT" [output_file=mapfile.mtz]
```

where file_list.dat has spaces or line breaks separating the different mtz file names in the folder containing the refined files. This averaged electron density map then serves as the basis for the electron density used in the selection procedure.

3. Extract map values from the averaged electron density map. After creating the averaged density maps, the contribution of each conformer to the averaged electron density can be estimated. In this procedure, the electron density was estimated for each conformer using the 2mFo-DFc map value of the refined atom using the eight-point interpolation function in cctbx [21]. The script eight_point.py can be used to loop over the conformational pool and output these map values. The name of the model PDB in line 47 eight_point.py must be changed to that of the most up-to-date refined structure of the protein, and the folder name in line 68 must be changed to that containing the individually refined conformational pool. The name of the mtz file containing the electron density data is specified in line 29 of eight_point.py.
4. Bin and filter the electron density. Even after this reduction, the map values will still require typically ~100,000 data points, which remains computationally too expensive for this procedure to be practical. To further compress the information, the map values are binned in three-dimensional space. The binning procedure averages the map values extracted in step 3 to create a three-dimensional histogram of the electron density. The bin size can be adjusted to fine-tune the balance between data compression level and accuracy. The resolution of the binning is an important parameter in the procedure. Testing different bin sizes

⁵The most efficient refinement protocol will likely be case-dependent.

and selecting the most appropriate size can be aided using the cross-validation test described below. To ensure correct addition of the bins, the same binning procedure must be performed for all the conformers using the exact same grid.

Three additional filters were also included to further reduce the data (See Note 6). The first and second filters pertain to the electron density level. High map values corresponding to fully rigid parts of the system that should be handled with standard crystallographic approaches are removed from the selection procedure using the first threshold. Conversely, regions of especially low density are likely noise and can be removed using the second filter. Even removing a region with no density will modify the outcome of the selection, as a region with a zero density level is treated differently than not having the region in the selection. The third filter ensures sufficient sampling. Selecting for a region that is extremely poorly sampled (e.g., only a few conformers within the ~10,000 pool) will not account for the possible structural heterogeneity in the region. If those regions are for some reason judged as essential in the description of the system, the initial sampling should be improved to pass this filter by changing the approach to generating the initial large conformational pool (See Note 7).

To bin the electron density, run `density_histogram.m`. This program gives as outputs the compressed electron density for the target as well as all the conformers. The resulting compressed electron density map will be used as an input in the selection procedure, which will only need to average these values over the conformers selected in a given ensemble. To set the noise level of the electron density maps, set `errorX1k`. The different filters are inputted in the electron density binning procedure using the parameters `densminX1k`, `densmaxX1k` and `popmin`.

4.2 Preparing the Anomalous Data

1. The anomalous atoms are limited in number and therefore can be directly used in the selection procedure. Depending on the space group, the positioning of the anomalous atoms, and the method for generating the conformational pool, you may also need to generate crystallographically identical anomalous atoms in neighboring asymmetric units. This symmetry generation can be easily accomplished in Pymol. After manually merging all the anomalous atoms to the same PDB file using a text editor, open the file in Pymol. Then, click on the PDB object, choose `generate>symmetry mates`, and choose the size required to produce enough duplicate anomalous atoms to cover the required space.
2. Convert the PDB file into `.sym` format using the script `iodine_sym.sh`, changing the filename in the first line of the script to input PDB filename.
3. After the positions of the anomalous scatterers are defined, it is necessary to pre-compute the distance between this position and the corresponding $C\alpha$ of that

⁶It is important to note that these filters must be handled with care, as they explicitly remove experimental information about specific regions.

⁷These filters may or may not be required for different systems, and their precise adjustment will also be system-specific.

residue within each conformer. This can be achieved using the script `iodo_dist.m`, using a text file containing a list of the `.sym` anomalous signal positions in which each line contains the complete path to the `.sym` file, as well as a text file listing the file names for the input `pdb` files from the conformational pool.

4.3 Running the Selection

1. After the experimental data and conformers have been prepared, one can run the selection. Starting with the initial pool, the selection searches for the ensemble that best reproduces the experimental data. A few general parameters will be key in the process and need to be optimized as described below.
2. To run the selection with default parameters, set `chi2scale=1`, `nbGen=50000`, and `nTrials=1`, and run the following commands from the MATLAB command window:

```
outputFile = 'output';
iofile = '/path/to/data/lists/list_anomalous_data';
densfile = '/path/to/data/lists/list_density'

params = processgalist(outputFile, iofile, densfile, 'chi2scale', 1, 'nbGen',
50000, 'nTrials', 1);
```

3. Set the selection ensemble size. The first parameter to consider is the number of conformers in the ensemble (See Note 8). In general, increasing the number of conformers will better reproduce the data, but can lead to overfitting unless controlled for using cross-validation tests as described in section 4.4. Usually, as the number of conformers in the ensemble increases, the data reproduction will reach a plateau value, indicating that adding one or a few extra conformers will not significantly affect the ability to fit the data. Degraded data reproduction may sometimes be observed with an increased number of conformers if the initial pool does not contain enough usable conformers. As an example, if the pool contains only 20 conformations that are close to the correct solution, asking it to select 30 conformations will force it to keep inadequate conformations to fill out the requested ensemble size. The selection should be performed at a variety of ensemble sizes, ranging from one to tens or hundreds of conformers, depending on the system. Validation tests are then used to select the appropriate ensemble size for analysis (see below).

To choose the number of conformations in the ensemble, set `ensSize` to the desired value. Typically, `ensSize` is varied between 1 and 20, and the results are then compared.

4. Set the data weighting. The second key parameter within the selection determines the relative weighting of the residual and anomalous data. In the ideal

⁸Searching for a single conformer corresponds to a static description, and searching for more conformers corresponds to a dynamic description.

case, the weighting should allow for the reproduction of each type of experimental data within experimental error, and with contributions reflecting their information content. It is worth noting that both the number of data points and the estimated errors affect the weight of each data type. If both datasets have similar information content, obtaining a similar final weight in the data reproduction can be reasonable (See Note 9).

The relative weighting of the anomalous positions compared to the electron density data can be set using `chi2scale`.

5. Set the number of selection iterations. The easiest parameter to define is the number of iterations in the selection. This parameter should be large enough to allow for convergence of the algorithm; however, lowering the number will improve computational efficiency. One can estimate this parameter by running a few long initial selections and observing when convergence is reached. Convergence can be judged by monitoring the output files reporting on the χ^2 evolution during the selection. This number can be influenced by the amount of experimental data, the size of the conformational pool, and the other selection parameters.

To set the number of iterations, set the input parameter `nbGen`.

4.4 Testing the Validity and Accuracy of the Selected Ensemble

1. Test the ensemble by replication. After performing the selection procedure, several tests must be used to estimate the validity and accuracy of the selected ensemble. First, the selection should be repeated several times using the exact same conditions, as the selection process is based on stochastic approaches and thus may not always converge on the same results (See Note 10). This step can be used to determine if there is a well-defined single solution or a set of relatively equivalent solutions that are able to reproduce the experimental data to a roughly similar extent. If several solutions exist, they should also present similar conformational properties, such as location in space, secondary structure content, or tertiary contacts. If the solutions do not present similar properties, it probably suggests that the amount of experimental information in the selection is not sufficient to properly define the selected ensemble.

To increase the number of replicated selections, increase the parameter `nTrials` or simply run again the program.

2. Test the approach using simulated data. It is helpful to test that the selection approach can recapitulate known structures pulled from the conformational pool (See Note 11). This can be accomplished by generating a few realistic ensembles of specific characteristics that deviate from the average of the conformational pool using the procedures described below. These ensembles are used to generate

⁹Obtaining a good estimate of experimental uncertainties helps in determining the scaling factor between different data types.

¹⁰Convergence properties may be improved by changing the parameters used in the selection.

¹¹For an example of the results of a simulated data test, refer to Supplementary Figure 5 from Horowitz et al. (2016).

noise-corrupted synthetic datasets that are then used as the data for the selection. The advantage of this approach is that the selected ensemble can then be directly compared to the known target ensemble. The ability to correctly capture the features of the test ensemble can characterize the validity of the procedure (See Note 12).

Simulated electron density maps can be easily generated using the phenix.fmodel tool (<https://www.phenix-online.org/documentation/reference/fmodel.html>). Condensed map generation is then performed as described above, with the inclusion of noise corruption using the density_histogram_target.m script. Similarly, noise-corrupted pseudo-experimental anomalous data can be generated using the script iodine_dist_target.m.

3. Test the ensemble via cross-validation and choose the ensemble size. In a 10-fold hold-out cross-validation procedure, only 90% of the data are used and 10% are randomly removed from the selection. At the end of the selection, the 10% unused data are back-predicted from the final ensemble to test the ensemble's ability to predict the unused experimental data. This process is repeated, typically 10 times, removing a different subset of the data each time (See Note 13.) The exact procedure will depend on the system under consideration, but a safe approach is to use the ensemble with the lowest number of conformers that successfully passes this cross-validation test.

Hold-out cross-validation tests can be seen as similar in essence to the R_{Free} crystallographic test. In fact, assessing the quality of an ensemble on data that are not actively used in the fitting procedure makes cross-validation tests more stringent than a direct fit and greatly helps in identifying potential overfitting or dataset inconsistencies. Thus, cross-validations can be useful in selecting ensemble size, checking data quality and self-consistency, and assessing the predictive power of an ensemble. For example, if the data are overfit or inconsistent, it will not pass this test and the predictive power of the selected ensemble will be worse than the initial pool. Additional external evaluations of the ensemble can also be used in as validations such as using the ensemble in a full crystallographic refinement procedure and computing its R_{free} .

To run the cross-validation, run the cross-validation form of the program using the "processgalist_crossval" function. In this version of the program, the additional parameter "indices" sets the fold number of the cross-validation.

4. Test the approach using bootstrapping. Bootstrapping can be used to estimate the precision of the selection. To perform bootstrapping, a random subset of data, typically on the order of $1/e$ (~37%) of the data [22], are removed from the fit and replaced by an equal number of repeated data points. This bootstrapped dataset is then used for the selection. The procedure is repeated a few hundred

¹²Using simulated data can also be a way to estimate the amount of experimental data required to accurately characterize a given system.

¹³Holding out 10% of the data is a typical value. This percentage may be adjusted depending on the system. The value must be sufficiently high to provide a stringent test but low enough to not completely destabilize the selection procedure.

times and the precision of the selection evaluated using the obtained ensembles. Variations in the selected bootstrapped ensembles can be used to estimate the error in properties of the selected ensemble (See Note 14).

To perform the bootstrapping procedure, generate, for example, 200 new datasets matching the bootstrapping strategy and run “processgalist_bootstraping” function. In this version, the parameter nTrials is used to set the number of bootstrapped datasets.

Acknowledgement

This work was supported by the National Institutes of Health (R01-GM102829 to J.C.A.B. and K99-GM120388 to S.H.). J.C.A.B. is a Howard Hughes Medical Investigator. The authors would like to thank S. Rocchio for comments on the manuscript and M. Mourao for useful discussions. The authors would also like to thank C. Stockbridge and the LSA-IT development team for assistance in coding.

References

1. Keskin O, Gursoy A, Ma B et al. (2008) Principles of protein-protein interactions: What are the preferred ways for proteins to interact? *Chem Rev* 108 (4):1225–1244. doi:10.1021/cr040409x [PubMed: 18355092]
2. Wright PE, Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* 16 (1):18–29. doi:10.1038/nrm3920 [PubMed: 25531225]
3. Fraser JS, van den Bedem H, Samelson AJ et al. (2011) Accessing protein conformational ensembles using room-temperature x-ray crystallography. *Proc Natl Acad Sci U S A* 108 (39): 16247–16252. doi:10.1073/pnas.1111325108 [PubMed: 21918110]
4. Salmon L, Blackledge M (2015) Investigating protein conformational energy landscapes and atomic resolution dynamics from nmr dipolar couplings: A review. *Rep Prog Phys* 78 (12):126601. doi: 10.1088/0034-4885/78/12/126601 [PubMed: 26517337]
5. Salmon L, Yang S, Al-Hashimi HM (2014) Advances in the determination of nucleic acid conformational ensembles. *Annual review of physical chemistry* 65:293–316. doi:10.1146/annurev-physchem-040412-110059
6. Venditti V, Egner TK, Clore GM (2016) Hybrid approaches to structural characterization of conformational ensembles of complex macromolecular systems combining nmr residual dipolar couplings and solution x-ray scattering. *Chem Rev* 116 (11):6305–6322. doi:10.1021/acs.chemrev.5b00592 [PubMed: 26739383]
7. Jensen MR, Zweckstetter M, Huang JR et al. (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using nmr spectroscopy. *Chemical Reviews* 114 (13):6632–6660. doi:10.1021/cr400688u [PubMed: 24725176]
8. Nodet G, Salmon L, Ozenne V et al. (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from nmr residual dipolar couplings. *J Am Chem Soc* 131 (49):17908–17918. doi:10.1021/ja9069024 [PubMed: 19908838]
9. Salmon L, Bascom G, Andricioaei I et al. (2013) A general method for constructing atomic-resolution rna ensembles using nmr residual dipolar couplings: The basis for interhelical motions revealed. *J Am Chem Soc* 135 (14):5457–5466. doi:10.1021/ja400920w [PubMed: 23473378]
10. Horowitz S, Salmon L, Koldewey P et al. (2016) Visualizing chaperone-assisted protein folding. *Nat Struct Mol Biol* 23 (7):691–697. doi:10.1038/nsmb.3237 [PubMed: 27239796]
11. Wals K, Ovaa H (2014) Unnatural amino acid incorporation in e. Coli: Current and future applications in the design of therapeutic proteins. *Front Chem* 2:15. doi:10.3389/fchem.2014.00015 [PubMed: 24790983]

¹⁴For an example of the results of a bootstrapping analysis, refer to Supplementary Figure 7 from Horowitz et al. (2016).

12. Santrucek J, Strohm M, Kadlecik V et al. (2004) Tyrosine residues modification studied by maldi-tof mass spectrometry. *Biochem Biophys Res Commun* 323 (4):1151–1156. doi:10.1016/j.bbrc.2004.08.214 [PubMed: 15451417]
13. Kmiecik S, Gront D, Kolinski M et al. (2016) Coarse-grained protein models and their applications. *Chemical Reviews* 116 (14):7898–7936. doi:10.1021/acs.chemrev.6b00163 [PubMed: 27333362]
14. Salmon L, Ahlstrom LS, Horowitz S et al. (2016) Capturing a dynamic chaperone-substrate interaction using nmr-informed molecular modeling. *J Am Chem Soc* 138 (31):9826–9839. doi:10.1021/jacs.6b02382 [PubMed: 27415450]
15. Ozenne V, Bauer F, Salmon L et al. (2012) Flexible-meccano: A tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 28 (11):1463–1470. doi:10.1093/bioinformatics/bts172 [PubMed: 22613562]
16. Adcock SA, McCammon JA (2006) Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews* 106 (5):1589–1615. doi:10.1021/cr040426m [PubMed: 16683746]
17. Miller BL, Goldberg DE (1995) Genetic algorithms, tournament selection, and the effects of noise. *Complex systems* 9 (3):193–212
18. Afonine PV, Grosse-Kunstleve RW, Echols N et al. (2012) Towards automated crystallographic structure refinement with phenix.Refine. *Acta Crystallogr D* 68:352–367. doi:10.1107/S0907444912001308
19. Winn MD, Ballard CC, Cowtan KD et al. (2011) Overview of the ccp4 suite and current developments. *Acta crystallographica Section D, Biological crystallography* 67 (Pt 4):235–242. doi:10.1107/S0907444910045749 [PubMed: 21460441]
20. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta crystallographica Section D, Biological crystallography* 53 (Pt 3):240–255. doi:10.1107/S0907444996012255 [PubMed: 15299926]
21. Grosse-Kunstleve RW, Sauter NK, Moriarty NW et al. (2002) The computational crystallography toolbox: Crystallographic algorithms in a reusable software framework. *J Appl Crystallogr* 35:126–136. doi:10.1107/S0021889801017824
22. Press WH (2007) *Numerical recipes : The art of scientific computing*. 3rd edn Cambridge University Press, Cambridge, UK ; New York