Towards a Black-Box for Biological EXAFS Data Analysis – I. Identification of Zinc Finger Proteins

Gerd Wellenreuther and Wolfram Meyer-Klaucke

European Molecular Biology Laboratory (EMBL), Hamburg unit, c/o DESY, Notkestr. 85, 22603 Hamburg, Germany

Abstract. EXAFS allows the determination of metal binding motifs in proteins. Here we present an algorithm for the identification of one subgroup, the Zinc finger proteins, by an automated refinement of biological EXAFS data. In combination with scoring criteria inspired by high resolution crystal structures these refinements led to the identification of these motifs, that playing an important role in protein-DNA and protein-RNA interactions. Robust criteria were identified investigating several EXAFS data sets from different zinc binding proteins. For all structural zinc motifs the automated EXAFS refinement led to results consistent with published ones.

Keywords: BioXAS, refinement, metalloproteins, Zinc, structural sites, active sites

PACS: 87.64.Fb, 87.14.Ee, 87.15.By

INTRODUCTION

Metal ions are essential for all living organisms. They are required for many biological processes such as respiration, metabolism, photosynthesis, cell division, muscle contraction, nerve impulse transmission, and gene regulation. The metal can be involved in protein structure stabilization, regulation of gene expression, electron and oxygen transport, as well as enzymatic activity.

Many scientists are aiming at the automation of data treatment for a variety of applications. One of the leading techniques in terms of automation utilizing synchrotron radiation is protein crystallography. Here, the entire pipeline from protein crystallization up to structure refinement is tackled. A similar degree of automation would be desirable for biological x-ray absorption spectroscopy (BioXAS), since this technique is independent of the sample state and can therefore provide information at any stage of a project. Prototype solutions for automated high-throughput data collection[1] and data reduction[2] are available these days or in near future. But the crucial step of the EXAFS refinement remains to be done.

AUTOMATION

Nowadays, despite ample availability of computing power, systematic analysis of EXAFS spectra still is supposed to be done by experienced scientists: First, several models, strongly depending on the scientist’s intuition or background, are tested against the experimental data. Then these results are analyzed, and the most favorite model is identified and in a third step further refined.

Choosing an inappropriate pool of models in the first steps might lead to inadequate sampling of parameter space, resulting in major difficulties during the refinement. Moreover, not every mathematical possible result fulfills the boundary conditions; e.g. is chemically reasonable and does lead to realistic parameters. The importance of these additional factors is the major reason why this job requires profound expertise.

An inexperienced scientist would most likely try to do it the safe way and systematically refine all possible structural models. Obviously, such a tedious task is perfectly suited for automation. However, choosing a model based on the fit-parameter only might not be sufficient; as mentioned above the resulting model could involve unreasonable parameters. Therefore, the challenge in automating the biological EXAFS refinement is the formulation of objective criteria for the quality of a model.

The third step of refinement includes all additional knowledge on the individual sample, and is therefore not very suitable for automation. Thus we focus in this initial project on the classification of metal binding sites, which potentially in most cases results as well in the identification of a good starting model for further refinements.
ZINC BINDING SITES IN METALLOPROTEINS

Although the automated refinement of all biological EXAFS data is our long term dream, we focus in this initial project on one absorber element: zinc. This redox inert ion seems to serve a signaling function in the nanomolar range of concentrations[3]. As well, it can be part as well of the active site of enzymes catalyzing biochemical reactions, or be involved in the stabilization of a secondary structure of proteins, which is necessary frequently necessary for e.g. RNA, DNA, or protein binding. Interestingly, most zinc binding sites are composed of combinations of three different amino acid residues: carboxylate from aspartic acid (Asp) or glutamic acid (Glu), imidazol from histidine (His), sulfur from cysteine (Cys) and OH or water. The major differences in the tuning of zinc affinities and the electronic structure of these metal binding sites originate from number, combination and geometrical arrangement of these ligands: The zinc ligands in zinc-finger proteins, are mainly sulfur (2 to 4), whereas in catalytic active sites there is only rarely up to one sulfur residue. For regulatory sites not much data are available so far. Thus, we leave them out for the moment.

IMPLEMENTATION

The main algorithm was written in Python (www.python.org), which is platform independent and freely available together with including a large number of scientific and numerical add-ons.

The experimental data and all relevant parameters are transferred to the program via a web-interface. The main algorithm is running on a LINUX cluster at EMBL Hamburg, allowing calculating a typical fit including multiple scattering in a few minutes.

Starting Parameters

The user supplies a list of the ligands (e.g. S, His and O) as building blocks for the model, as well as a list of overall coordination numbers (e.g. 4, 5, 6). All permutations of ligands then will be calculated and constitute the pool of models (e.g. [S,His], [S,His] ... [S,His], [S,His]). To incorporate prior knowledge on possible metal environments restraints for the individual ligand types are available; e.g. restraining the His coordination number between 2.0 and 4.0 would reduce the total number of permutations from 18 down to 9, prohibiting amongst others [SHis] and [SHis] from the example above.

For samples containing several metal sites rational coordination numbers are allowed (e.g. 3.0, 3.5, 4.0), and a step size different from 1.0 may be defined (e.g. step size = 0.5) would lead to the addition of models like [S0.5His3.5] to the pool of models shown above.

Moreover, the user further can define the data range to be considered in the fitting process either in energy- or k-units.

Curved Wave Fitting

Curved wave fitting was implemented is realized applying using DL_EXCURV[4], where it provides predefined biological units like e.g. His and Heme. k-weighting is applied, and, by default, intra-unit multiple scattering is considered. The algorithm refines the Fermi-energy, the Debye-Waller factors and the radii of the ligands. To lower the number of free parameters the Debye-Waller factor of all atoms in the first coordination shell is fitted jointly by default. This can be overridden, if individual Debye-Waller factors are justified. e.g. in the case of a bi-metal binding site to fit the metal-metal distance separately.

The radii of all ligands are fitted optimized separately by default, but an option to fit the radii of oxygen-and nitrogen-ligands together jointly is included since they can barely be distinguished for zinc absorber atoms.

DL_EXCURV is minimizing the fit index \( \Phi_{\text{EXAFS}} \), which is the summing up of the square of residuals. It decreases with the fit quality, but also increases when parameters exceed predefined boundaries. In contrast, the reduced \( \chi^2 \) considers the degree of over-determination of the system, based on the number of parameters that have been refined \( N_{\text{pars}} \) and the number of independent data points \( N_{\text{ind}} \)[5]. Thus, it is the basis to select the best statistical fit of the theoretical models used to reproduce the metal geometry coordination

\[
\chi^2 = \frac{1}{N_{\text{pars}}-N_{\text{ind}}} \sum_{i=1}^{N_{\text{ind}}} w_i (\hat{X}_i - \bar{X}_i)^2
\]

where

\[
N_{\text{ind}} = \frac{2 \Delta k \Delta R}{\pi} + 1
\]

\( \Delta k \) is the data range being refined fitted in k space (as determined by the data range given), and \( \Delta R \) is the range of data being given by the distance interval in which backscatters are defined fitted in R space (between \( R_{\text{min}} \) and \( R_{\text{max}} \) ) by the model (between \( R_{\text{min}} \) and \( R_{\text{max}} \) ) where the data are substantial in signal-noise ratio.

Scoring

Selection of the best fit out of the pool of models cannot depend only on the goodness of fit parameters; unreasonable models have to be excluded. Therefore we introduce an additional criterion, \( C_{\text{Bond}} \) based on the deviation of the refined metal ligand distances from ideals. These ideal distances are based on the Hercules work by Marjorie M. Harding[6], who extracted them from the CSD[7] and PDB.

\[
C_{\text{Bond}} = e^{\frac{1}{2} \left( \frac{r_{\text{ref}} - r_{\text{ideal}}} {\sigma^2} \right)^2}
\]
To compose a... The total score is based on a weighted combination of two therefore rely on he Reduced $\chi^2$ and the bond length-criterion. To ensure that each criterion enters the score on the same basis we scaled them to an interval $[0.0, 1.0]$; 0.0 corresponding to complete rejection of the model by failure in this criterion and 1.0 to complete success. The criterions are linked by a logical AND operation, reflected by a multiplication of the scores. Weighting different criteria $C_i$ can be achieved by exponentiations with weights $w_i$.

$$\text{Score} = \prod_i C_i^{w_i} \quad (4)$$

This approach allows including further criterions later. Based on the experience of our initial previous tests we favor a weights of $w_{C_{\chi^2}} = 0.8$ and of our bond-length criterion $w_{C_{\text{Bond}}} = 0.2$.

Moreover, we analyze the Debye-Waller factor $(2\sigma^2)$ of the first coordination shell. In the absence of ab-initio ideal values we limit its range in the individual EXCURV refinement to we reject models outside a fairly large interval of $[0.004, 0.025]$.

**Classification of Zinc Sites**

Using this algorithm we analyzed a couple of BioEXAFS data sets of metalloproteins containing different Zn binding sites. So far, we were successful in identifying the binding motif in (i) the partially folded high-risk human *Papilloma virus 45 oncoprotein E7* [8], (ii) the *Drosophila melanogaster* transcription factor *glia cell missing* (GCM) [9], (iii) the ubiquitin binding domain ZnF-UBP from HDAC6 [10], (iv) the *Pseudomonas aeruginosa* ferric uptake regulator (Fur) [11], the binuclear zinc phosphodiesterase from *Escherichia coli* both (v) wild type and (vi) with the mutation D212C [12] and (vii) a glyoxalase from *Arabidopsis thaliana* [13] as summarized in Table 1.

Thereby the Our algorithm was is thereby able to identify zinc-finger motifs based on the identification of more than 1.6 sulfur ligands per Zn ion in metalloproteins. In addition, it correctly determined the total coordination number and sometimes even yielded the published result as the top-scoring model out of the pool of models. Its main weakness at the current stage is the lacking ability to differentiate between His and O, especially for datasets covering a short k-range (e.g. 3 – 11 Å⁻¹).

**ACKNOWLEDGEMENTS**

We gratefully acknowledge funding of BIOXHIT project by the European Commission under FP6 contract LSHG-CT-2003-503420.

**REFERENCES**


**TABLE 1.** Results of the automated refinement in comparison with published fits for a variety of zinc-proteins. The proteins in bold were identified by the program as zinc-finger proteins.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Highest ranking result</th>
<th>Published Results</th>
<th>Rank of published result</th>
<th>Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Papilloma virus 45 oncoprotein E7</em></td>
<td>S$_4$</td>
<td>S$_4$</td>
<td>1. out of 46</td>
<td>Zinc-finger proteins</td>
</tr>
<tr>
<td><em>D. melanogaster transcription factor GCM</em></td>
<td>S$_3$His$_1$</td>
<td>S$_3$His$_1$</td>
<td>1. out of 399</td>
<td></td>
</tr>
<tr>
<td>ZnF-UBP</td>
<td>S$<em>{1,3,25}$His$</em>{0.66}$</td>
<td>S$<em>{1,3,25}$His$</em>{0.33}$</td>
<td>12. out of 57</td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em> ZipP Fur wild type</td>
<td>His$_{5}O_2$</td>
<td>His$_{5}O_2$</td>
<td>3. out of 40</td>
<td></td>
</tr>
<tr>
<td><em>E. coli</em> ZipP D212C mutant</td>
<td>S$<em>3$His$</em>{4.0}$</td>
<td>S$_{2.5}O_2$</td>
<td>10. out of 253</td>
<td></td>
</tr>
<tr>
<td><em>A. thaliana</em> Glyoxalase II</td>
<td>His$_{5}O_2$</td>
<td>His$_{5}O_2$</td>
<td>43. out of 247</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Catalytic active sites</td>
</tr>
</tbody>
</table>

**REFERENCES**