UNIVERSITÀ DEGLI STUDI DI MILANO

UNIVERSITÀ DEGLI STUDI DI MILANO

PhD Course in Molecular and Cellular Biology

XXXII Cycle

MRSAD-phasing of large macromolecular complexes

Riccardo Pederzoli

PhD Thesis

Scientific tutors: Dr. Thomas R. Schneider

Prof. Martino Bolognesi

Academic year: 2018-2019

Table of Contents

List of Figures

List of Tables

List of abbreviations

ASU: ASymmetric Unit

ATD: AminoTerminal Domain

***At*GLR3.3:** *A. thaliana* GLR isoform 3.3

Bc: *Burkholderia cenocepacia*

Bp: *Burkholderia pseudomallei*

Br-20*S*: Br-soaked human 20*S* proteasome

CDR: Complementarity-Determining Region

CTD: Cytoplasmic Tail

CynS: Cyanate hydratase

dm: density modification

Ep: epitope

GFP: Green Fluorescent Protein

GLR: Plant Glutamate Receptor-Like channels

IgE: Immunoglobulin E

iGluRs: mammalian ionotropic Glutamate Receptors

LBD: Ligand-Binding Domain

LBM: Lanthanide Binding Motif

LBT: Lanthanide Binding Tag

LLG: Log-Likelihood Gain

MAD: Multiple-wavelength Anomalous Diffraction

MAP-CC: Map Correlation Coefficient

MPE: Mean Phase Error (º)

MR: Molecular Replacement

MRSAD: Molecular Replacement in combination with Single-wavelength Anomalous Diffraction

MW: Molecular Weight

NB: NanoBody

NCS: Non-Crystallographic Symmetry

Pal: Peptidoglycan-associated lipoprotein

PDB: Protein Data Bank

r.m.s.d.: root mean square deviation (Å)

RSCC: Real Space Correlation Coefficient (synonym of MAP-CC)

SA: Simulated Annealing

SAD: Single-wavelength Anomalous Diffraction

SAGE: Strategy for Alignment and Grafting of Epitopes

SAXS: Small Angle X-ray Scattering

SeMet: Seleno-Methionine

SRF: Self-Rotation Function

TFZ: Translation Function Z-Score

tNCS: translation Non-Crystallographic Symmetry

V_M: Matthews coefficient

wMPE: weighted Mean Phase Error (º)

ABSTRACT (ITALIAN)

L'obiettivo del presente lavoro di Tesi è lo studio sistematico del metodo MRSAD, un metodo di fasamento cristallografico che sta diventando sempre più un importante strumento nelle mani dei cristallografi, in particolare per quanto riguarda la risoluzione del crescente numero di strutture biologiche a elevato peso molecolare. Il metodo MRSAD è stato testato su diverse proteine a basso e medio peso molecolare e nel caso del proteasoma 20*S* umano, sfruttando la presenza di modelli depositati nel Protein Data Bank (PDB) che hanno reso possibile la comparazione e la valutazione dei risultati. L'applicabilità di una procedura generale per il fasamento e la costruzione di un modello nelle fasi MRSAD è stata studiata, cosí come l'effetto di procedure di "density modification", la completezza dei modelli per Molecular Replacement, la loro accuratezza e la multiplicità dei dati cristallografici. I risultati ottenuti dall'analisi di dati relativi ad un'ampia varietà di sistemi modello permettono di ricavare conclusioni sulle potenzialità e sui limiti del fasamento attraverso MRSAD, e suggeriscono alcune linee guida per la sua applicazione al fine di massimizzare il suo successo. In aggiunta, il fasamento attraverso MRSAD è stato testato positivamente in due casi reali: il primo è rappresentato dall'uso di MRSAD per la risoluzione della struttura del primo recettore del glutammato di pianta. Il secondo riguarda invece l'impiego di MRSAD per il fasamento di antigeni attraverso l'impiego di nano-anticorpi ("nanobodies") ingegnerizzati con una sequenza in grado di legare ioni di lantanidi, sviluppati recentemente all'interno del gruppo. Il lavoro di Tesi ha anche riguardato la determinazione di alcune strutture rimaste a lungo

irrisolte. In questi casi, non è stato possibile ricorrere all'uso di MRSAD a causa della mancanza di dati con segnale anomalo, ed altre strategie di fasamento sono state impiegate. Ognuna delle strutture che è stata risolta rappresenta un caso difficile con le sue proprie peculiarità, ed un ampio spettro di strategie di fasamento e metodi di miglioramento delle fasi è stato impiegato per la loro risoluzione.

ABSTRACT (ENGLISH)

The objective of the Thesis work is a systematic study of MRSAD-phasing, a crystallographic phasing method that is becoming part of the arsenal available to crystallographers and which represents an important tool for phasing of the increasing number of large macromolecular complexes being crystallized. This method has been tested on small and medium size proteins as well as on more challenging human 20*S* proteasome data, taking advantage of the existing deposited models which allow for the comparison and evaluation of the results. The applicability of a general procedure for MRSAD-phasing and model building was investigated, as well as the effect of density modification, MR-search model completeness and accuracy and data multiplicity. The results from the tests on a broad variety of systems allow to draw conclusions on the potentialities and limitations of MRSAD-phasing, suggesting some practical guidelines for its successful application. Moreover, MRSAD-phasing has been tested on two real-life scenarios: the first is represented by the use of MRSAD to solve the structure of the first plant glutamate receptor. The second concerns the use of MRSAD for the phasing of unknown antigens through engineered nanobodies with a lanthanide binding motif recently developed within the group. The Thesis work also dealt with the structure determination of other previously unsolved protein structures, which proved resistant to many attempts at structure solution. In these cases, MRSAD could not be employed because of the unavailability of anomalous signal, and other complex phasing strategies were used. Each structure that was solved represents a difficult case with its own specificities and challenges; in keeping with the Thesis

aims, a broad set of phasing strategies and phase improvement methods were used to tackle such structures.

AIM

The overall objective of the Thesis work is the investigation and the application of methods for the determination of challenging macromolecular structures. This goal has been pursued in two different but complementary ways.

The first one is represented by the systematic study of Molecular Replacement in combination with Single Anomalous Diffraction (MRSAD). It has been shown that this phasing method can lead to structure solution starting from weak anomalous signal and/or poor MR-search models, where both the SAD and MR methods alone would fail. The advent of MRSAD has been triggered by the need to reduce the MR intrinsic model bias, particularly at low resolution, and by the increasing availability of high-resolution structures for components of larger complexes. Despite the structures of several biologically relevant macromolecular complexes could be solved only via MRSAD, its potential has not been fully explored yet. A better characterization of MRSAD would allow to extend the range of tools which crystallographers can use to tackle increasingly challenging systems. This part of the work aims at developing procedures and recommendations for MRSAD-phasing of large systems (molecular weight greater than 500kDa) by performing systematic investigations of its capabilities and limitations. This involves the study of the applicability of a general procedure for MRSAD-phasing and model building and of the effect of MR-search models completeness and accuracy and of data multiplicity. To this aim, different MRSAD-phasing algorithms were tested on a number of systems which differ in terms of size of the asymmetric

unit, data resolution, number and type of the anomalous scatterers and anomalous signal strength. In particular, to test the capabilities of MRSAD-phasing on large macromolecular complexes, the human 20*S* proteasome was chosen as the central model system of this study. The assessment of the performance of MRSAD-phasing is based on a number of indicators and is possible because for each of the test systems a refined model of sufficient quality is available which is used as a reference. To better investigate the potential of MRSAD in a real-life case, this method was applied to two different scenarios. *In primis*, MRSAD was used to solve the structure of the first plant glutamate receptor. The second application concerns the use of MRSAD for the phasing of unknown antigens through engineered nanobodies with a lanthanide binding motif recently developed within the group.

The second way to pursue the overall objective of the Thesis is represented by the application of different phasing methods and phase improvement strategies for the determination of challenging structures. Nowadays, a number of powerful software and automated structure solution pipelines is available which can greatly aid the structure determination process. However, for difficult cases (*i.e.*, for pathologic data and/or cases where, for example, no homologues are available), the expertise of the crystallographer becomes decisive. I therefore embarked on the structure determination of a number of previously unsolved proteins which proved resistant to many structure solution attempts. The following cases were studied: first of all, the abovementioned plant glutamate receptor, for which only twinned native data together with an anomalous data set with weak anomalous signal were available. The second example is represented by the major protein allergen from *Sesamum indicum*. Here, only a native data set

was available and, as for the case of the plant glutamate receptor, all homologues were not sufficiently similar in terms of their structure. The third case is the one of the SAGE1 and SAGE3 proteins, for which only low resolution native data were available (3.4 – 4 Å) and for which experimental validation is ongoing. Lastly, an interesting case of a protein contaminant structure was studied.

1. SYSTEMATIC STUDY OF MRSAD-PHASING PROTOCOLS ON SMALL AND MEDIUM SIZE PROTEINS

1.1. Summary

The aim of the MRSAD project consists in the development of procedures and recommendations for the application of MRSAD-phasing to large macromolecular complexes (having a molecular weight higher than 500kDa). Before studying MRSAD-phasing on large macromolecular complexes, the phasing method was systematically tested on a number of protein models having small to medium size (molecular weight between 15 and 70kDa). The reason was two-fold: in the first place, MRSAD-phasing has not been well studied on small and medium size systems and, in the second place, any application of the method to large and challenging systems requires the knowledge of how it performs on simpler cases. The performance of MRSAD-phasing on simpler cases can be used as a reference point to evaluate the results of the method on more difficult cases. The test systems differ in terms of size of the asymmetric unit (ASU), data resolution, number and type of the anomalous scatterers and strength of the anomalous signal. MR-search models of different nature and completeness were employed. The same test cases where used to test the feasibility of a general procedure for the application of an automated MRSAD-phasing and model building pipeline. In addition, the minimum amount of information required for successful MRSAD-phasing, in terms of size of the MR-search model and diffraction data multiplicity, was determined for

the same test cases. It was found that, for all the test systems, MRSAD-phasing is able to improve the quality of the phases with respect to MR alone. The improvement depends on the accuracy and completeness of the initial search model and on the strength of the anomalous signal. A general pipeline for MRSAD-phasing and model building was successfully applied to systems of small and medium size, but the same protocol was shown to have limitations when the resolution of the data is ~ 2.8 Å or worse, and/or when applied on systems of larger molecular weight. With respect to the amount of data required for successful MRSAD-phasing, it was shown that the completeness of the MR-search model plays a more important role than data multiplicity but, in borderline cases, the effect of the amount of data on the phasing becomes important too.

1.2. Introduction

MRSAD-phasing has been proposed as a strategy to overcome the limitations of MR- and SAD-phasing approaches alone [1], [2]. The advent of MRSAD has been triggered by the increasing availability of high-resolution structures of fragments of large complexes and by the necessity to reduce model bias arising from MR-phasing (perhaps the greatest caveat at low-resolution), since the SAD (or MAD) phases are virtually independent of the MR phases.

The basic working principle of MRSAD-phasing is shown in **Figure 1.1**. The first step is represented by MR with a certain search model, which provides weighted MR-phases. As firstly proposed by Strahs and Kraut in 1968 [3], an anomalous difference Fourier map is used to locate the heavy-atom peaks. Such difference map is computed by combining the MR-

phases 90º shifted and the anomalous differences, so it is a map of the type: $(\Delta \boldsymbol{F}_{obs}, \varphi_{MR} - 90°)$. In general, if phase estimates ϕ_T for the structure are available, anomalous atoms can be found from a Fourier map with amplitudes $\Delta \boldsymbol{F}_{obs} = |\boldsymbol{F}_+| - |\boldsymbol{F}_-|$ and phases $\phi_T - \alpha \simeq \phi_A \rightarrow \phi_T - 90°$. Therefore, Bijvoet differences are used to calculate the so-called Bijvoet-difference Fourier map.

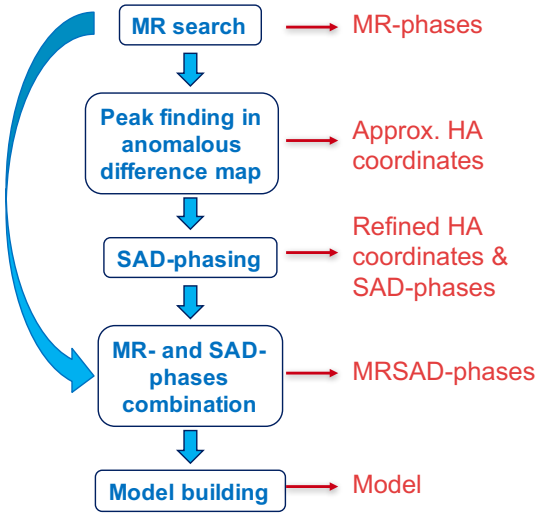


Figure 1.1: Schematic representation of the MRSAD-phasing working principle.

Adding 90° to the phase is a clever trick firstly proposed by Strahs & Kraut in 1968 [3]: it works because $\Delta \boldsymbol{F}$ is large exactly when the phase angle of f^0 for the anomalous scattering atoms happens to be 90° away from that of the rest of the structure. The reflections for which $(\phi_T - 90°)$ is a bad estimate are exactly the reflections for which $\Delta \boldsymbol{F}$ is small and hence does not contribute much to the map. One of the first application of the Bijvoet-difference Fourier map can be found in a paper by Dauter and collaborators in 1999 [4], where a $(\Delta \boldsymbol{F}_{obs}, \varphi_{MR} - 90°)$ map turned out to be useful to

check for the presence of anomalous signal prior to phase determination. The Fourier map is calculated as:

$$\rho_{(x,y,z)} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}(|\boldsymbol{F}_{+}| - |\boldsymbol{F}_{-}|)\, e^{i\alpha_{calc}} e^{-2\pi i(hx+ky+lz)}$$

and the supremacy of this map became clear when compared with a $(\boldsymbol{F}_{obs}, \varphi_{calc})$ map calculated for the same protein region.

After the heavy-atom peaks are identified from such $(\Delta\boldsymbol{F}_{obs}, \varphi_{MR} - 90°)$ map, the usual SAD-phasing steps are followed, meaning that new heavy-atom sites are found, refined and the list of scatterers is updated and this process is iterated (ideally) until substructure completion. Even with a poor MR-model, completion algorithms used by the most powerful programs can succeed in determining the substructure. Once (weighted) MR- and SAD-phases are available, they are combined with correct (maximum-likelihood estimated) relative weights. The MRSAD-phases are then used to build the model in the successive stages, with or without the assistance of density modification. It must be noted that the enantiomorph ambiguity is not an issue in MRSAD, because the chirality of the partial model restricts the sites to the given enantiomer.

MRSAD-phasing has been found to be particularly useful for cases in which the anomalous signal is not sufficiently strong to solve the structure by experimental phasing but is good enough to bootstrap the structure starting from a preliminary MR solution. There are cases in which MRSAD proved to be decisive for structure solution and where it has been carried out starting from weak anomalous signal originating from naturally built-in scatterers such as sulphur and phosphorous atoms [1]. However, for large

macromolecular complexes, in order for the anomalous signal to be of practical use in the phasing procedure, more effective scatterers are required and the use of polynuclear metal nano-clusters has been widely reported in the literature as a source of high scattering power, especially at low-resolution [5], [6]. Among the various possibilities, the tantalum $[Ta_6Br_{12}]^{2+}$ and several W-based clusters are the most widely used compounds [5], [7], [8]. The structures of many large assemblies, ranging from the photosynthetic reaction center to various ribosome and proteasome particles have become accessible only after derivatization and phasing with the abovementioned clusters. Large assemblies are the preferred target of MRSAD-phasing, and they likely represent the cases where its potential could be fully exploited. This is because it is becoming increasingly the case that a significant portion of a large structure is known, but the rest of it is not. At the same time, improvements in the derivatization procedures or in the incorporation of exogenous heavy-atoms into proteins, as well as in the data collection at modern beamlines, are turning the acquisition of anomalous data into a routine procedure. Such large macromolecular complexes usually produce weakly diffracting crystals as a result of the intrinsic flexibility and/or lattice disorder, and this will preclude structure determination at high resolution. Obviously, attempts should be made to improve the resolution at which crystals diffract, but one should not disregard the information that can be obtained from medium-to-low resolution data [9]–[12]. Because of the working MRSAD-principle, medium or low resolution data do not necessarily constitute a problem: in fact, the location of heavy-atoms does not require high-resolution data and, if the substructure can be correctly determined, several techniques can be used to obtain approximate protein phases and to improve them. In

favorable cases, even approximate phases can be significantly improved and extended to higher resolution. The crucial role of MRSAD-phasing in determining the structure of important biological complexes is attested by the literature. Recent examples are represented by the structures of the eukaryotic ribosome [13], the human HOIP/E2-ubiquitin complex [14], the human COP9 signalosome [15], the B and C proteins from the ABC toxin complex of *Yersinia entomophaga* [16] and the Core Mediator Complex from *Schizosaccharomyces pombe* [17].

1.3. Materials & Methods

1.3.1. Sample preparation, crystallization, data collection and processing

The numerous data sets used for the tests have been kindly provided by different people, within and outside the group, with the exception of Br-soaked thaumatin data. In many cases, other types of useful information were also provided as, for instance, sequence files and in-house refined protein models to use for phase comparison. The procedures for sample preparation, crystallization, data collection and processing are described in published papers or in papers to be published soon.

1.3.2. Description of the pipelines

In what follows, the self-written pipelines used to carry out phasing, model building and data analysis in an automated fashion are described. When needed, relevant parameters are automatically calculated and/or extracted from output files. Reference models exist for all systems such that the comparison against the known answer can be made. In particular, for phase comparison, the reference phases were extracted from the reference models through *SFALL* (*CCP4*), and used for the evaluation of the results.

1.3.2.1 Selection and preparation of the models for MR

Truncated models for MR- and MRSAD-phasing were created starting from the reference models and removing waters, ligand molecules and any ions possibly present.

1.3.2.2 MRSAD-phasing and model building

A C-shell script was prepared to automate MRSAD-phasing and model building. It first uses *PHASER* for *MR*- and MRSAD-phasing [18], [19], and then performs (separately) classical model building/density modification with *ARP/wARP* [20], [21], *BUCCANEER* [22], [23] and *PHENIX* [24].

1.3.2.3 Determination of the resolution limits of map interpretation software on the SeMet-FAE data

A Python script was prepared (Python 2.7.10) in order to test the resolution limits of *SHELXE*-MRSAD [25] and of two map interpretation softwares on the SeMet-FAE data, for different resolution and completeness levels of the MR-search models. Five MR-search models at different completeness levels (10%, 30%, 50%, 70% and 90%) were generated from the refined model of SeMet-FAE (structure not deposited). The pipeline starts with MR in *PHASER*, and the resulting MR-model is given to *SHELXE* for MRSAD. The resulting MRSAD map is used by the *PHENIX trace_chain* algorithm for Cα finding; the *trace_chain* model is then used by *pulchra* for backbone extension in order to generate a partial model that can bootstrap successive model building in *ARP/wARP* and *PHENIX*. The same MRSAD map is used for direct model building in *ARP/wARP* and *PHENIX*. This procedure is repeated for each MR-search model and for each resolution level.

1.3.2.4 Determination of the minimum amount of information required for successful MRSAD-phasing

An automated Python script to perform phasing and analysis of the data was prepared to this aim. Depending on the system, the data is simply truncated into smaller wedges or divided into separate turns (1 turn = 360˚ of rotation) which are then progressively combined, one after the other. For the truncation of the data, a program developed-in-house by Fabio Dall'Antonia is used, which cuts the data based on the frame number. The resulting data blocks are used for subsequent phasing (in all cases, scaled and unmerged data were used). For each different MR-search model, MRSAD-phasing is tested at different multiplicity levels with two different softwares (*PHASER* and *SHELXE*). Five MR-search models at different completeness levels (10%, 30%, 50%, 70% and 90%) were generated from the reference models. The pipeline performs *PHASER*-MR and MRSAD, and it then applies *PARROT* density modification (dm) [26] on the MRSAD phases. The pipeline runs *SHELXC* and *SHELXE*-MRSAD on the truncated data, as well.

1.3.3. Evaluation of the results of the pipelines

For all the pipelines, the success of MRSAD-phasing was primarily judged on the basis of the phase-quality, which was assessed through the Mean Phase Error (MPE) against the reference model phases. In order to assess whether MRSAD-phasing was really superior to MR- and SAD-phasing alone, a comparison with them was also made.

Additionally, the real-space correlation coefficients (RSCC) residue-by-residue with respect to the reference electron density map were computed with *phenix.get_cc_mtz_pdb*. The RSCC is a metric of $|\boldsymbol{F}|$ and FOM-weighted phase quality and can be used as a measure of the quality of an

electron density map. To visualize improvements in the MRSAD-maps over MR-maps for the region of the molecule which was not part of the MR-search model, scatter plots of RSCC-MRSAD versus RSCC-MR were generated. Each dot represents a specific residue, the position of which gives information about the improvement of its electron density when performing MRSAD- instead of MR-phasing. Each dot that lies above the diagonal represents a residue for which the RSCC has improved when using MRSAD instead of MR. In addition, the more the dots are shifted to the upper-right corner, the bigger the improvements.

1.3.3.1 Substructure comparison

For substructure comparison, the program *SITCOM* was used [27]. The quality of a given substructure (with respect to the reference substructure extracted from the reference model) was evaluated based on three *SITCOM* indicators: NM, which indicates the number of matches between equivalent sites within 3Å distance, r.m.s.d. (the root mean square deviation) and the score parameter, that gives an indication of site reliability.

1.4. Results & Discussion

Table 1.1 summarizes the systems used to test MRSAD-phasing in the context of the Thesis work. Even though MRSAD-phasing has been tested on all these systems, only the results for three of them will be shown and discussed in this Chapter. This is not only because the test systems are numerous and their discussion would not fit in the context of the present Thesis, but mainly because the results are partly overlapping. Therefore, three representative cases were selected, namely $Cdc23^{NTerm}$, SeMet-FAE and RipA, which summarize what has been observed, in general, for all the other systems. Furthermore, $Cdc23^{NTerm}$, SeMet-FAE and RipA represent

SYSTEM	Acronym	d_{min} (Å)	SG	λ (Å)	No. mol/No. residues ASU	No. and type HA's/ASU	Bijvoet ratio (%)
SeMet-feruloyl esterase	SeMet-FAE	1.2	$P2_12_12_1$	0.9763	2 / 562	16 SeMet	5.1
Br-soaked thaumatin	Br-thau	1.5	$P4_12_12$	0.9197	1 / 207	21 Br	8.1
(GFP+Nb+Tb)-complex	Nanobody complex	Native (1.663 Å), infl (1.773 Å), peak (1.732 Å), hrem (1.689 Å)	$P3_121$	Native (1.0332 Å), infl (1.6505 Å), peak (1.6498 Å), hrem (1.0332 Å)	1 / 369	1 Tb	4.2
***M. tuberculosis* peptidoglycan hydrolase Rv1477**	RipA	1.8	$P2_12_12_1$	1.77	1 / 207	6 Met + 1 Cys	1.0
***M. tuberculosis* peptidoglycan-binding protein Rv1566c**	RipD	1.8	$C121$	1.77	1 / 125	3 Met	0.9
Br-soaked human 20*S* proteasome	20S-Br	2.5	$P2_12_12_1$	0.9198	28 / 6215	57 Br	2.4
Native human 20*S* proteasome	20S-native	2.9	$P2_12_12_1$	2.0664	28 / 6215	322 S (212 Met + 110 Cys), 57 Cl, 6 K	1.6
Subunit of the multimeric anaphase-promoting complex (APC/C)	$Cdc23^{NTerm}$	3.1	$P4_3$	2.69	2 / 564	5 Met + 6 Cys	2.2
80mM Cd-soaked Ferritin	Ferritin	2.7	$F432$	1.0332	1/174	10 Cd	4.2

Table 1.1: List of the test systems on which MRSAD-phasing has been tested. The Bijvoet ratio was computed through the Hendrickson formula at the wavelength of data collection [28]. These are expected Bijvoet ratios and are likely to be lower in reality as the occupancy of heavy-atom sites was assumed to be 100%. The Bijvoet ratios, despite being theoretical, reflect the strength of the anomalous signal in the real crystal.

very different cases and they effectively cover the extremes of the application of MRSAD-phasing to small and medium size systems.

1.4.1. MRSAD-phasing tests on Cdc23^{NTerm}

Cdc23^{NTerm} consists of two molecules in the asymmetric unit with a total of 564 residues, among which 10 are Met and 12 are Cys (there are two disulfide bridges – one per monomer); its structure has been recently determined to 3.1 Å (PDB ID: 5FTP) via S-SAD at 2.69Å wavelength [29]. A monomeric, 1.9 Å resolution structure has been obtained (PDB ID: 3ZN3) via Se-SAD by Zhang *et al.* in 2013 [30]. MRSAD-phasing of Cdc23 has been tested using different search models and the results are summarized in **Table 1.2**. Five search models have been considered: chain A of the 3ZN3 model (3ZN3_A), the polyAla model of 3ZN3 (3ZN3_polyAla), chain A of the 5FTP model (5FTP_A), the polyAla version of chain A of 5FTP (5FTP_A_polyAla) and the full 5FTP model (5FTP_AB).

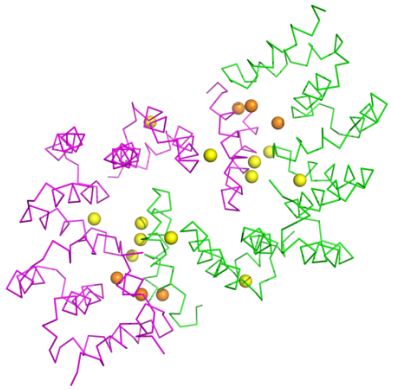


Figure 1.2: **Cdc23^{NTerm} model and the S-substructure.** S-Cys and S-Met are shown as yellow and orange spheres, respectively.

The full, low-resolution model (5FTP_AB) represents the reference case and sets the lower and the upper limits of MPE and MAP-CC achievable via MRSAD on Cdc23, respectively. The mean phase errors obtained with all the other search models are significantly higher and, besides the case of 5FTP_A, they are greater than 50°, which is the limit above which the quality of the phases makes model building considerably more difficult. This threshold is somewhat arbitrary and, in certain cases, maps with a mean phase error ~ 55-60° can still be successfully employed for model building. Despite not being an absolute limit, it is important to set a threshold for the usability of phases to predict whether these phases can be used to build a model or not, which is the final aim of any approach to improve the accuracy of crystallographic phases. Therefore, with the exception of the 5FTP_A and of the 5FTP_AB models, the phases obtained by using the other search models appear to have a borderline mean phase error, even after MRSAD-phasing.

MR-search model	***PHASER-MR***		***PHASER-MRSAD***				
	MAP comparison		**MAP comparison**		***SITCOM* against 5FTP**		
	MPE/°	**MAP-CC**	**MPE/°**	**MAP-CC**	**NM**	**r.m.s.d.**	**Score**
3ZN3_A	59.0	0.495	55.7	0.553	18	0.49	0.838
5FTP_A	44.7	0.585	43.5	0.626	18	0.46	0.847
5FTP_AB	11.6	0.531	23.4	0.726	18	0.42	0.861
3ZN3_polyAla	67.5	0.409	60.9	0.514	18	0.49	0.837
5FTP_A_polyAla	58.7	0.473	54.7	0.559	18	0.46	0.846

Table 1.2: Results of MR- and MRSAD-phasing in terms of MPE and MAP-CC for the Cdc23 test case.

The analysis of the results in **Table 1.2** shows two inconsistencies: *in primis*, it can be observed that, for the reference search model, the MPE on the MRSAD-phases is greater than the one on the MR-phases. This

counterintuitive result can be due to the way the bulk solvent is accounted for when the reference phases for the computation of the mean phase error are computed. This result can also be explained by the fact that the gain, in terms of the MPE between the MR and the MRSAD solution, which can be obtained after MRSAD with the perfect model is necessarily limited and less perfect search models will naturally lead to larger gains. These two explanations are not mutually exclusive. It is also not fully understood why the search model 3ZN3_A does not provide better results, being 3ZN3 the high-resolution model. This could be explained by the different conformation of a number of side-chains between the chain A of 3ZN3 and of 5FTP, which can result in a more difficult MR placement.

Beside the mean phase error of the phases, the success of the MRSAD protocol is evaluated based on the quality of the substructure, which indicates how well the MRSAD protocol has performed given the anomalous signal present in the data. Finding the substructure represents the first step towards the determination of the overall structure: if most of the heavy-atom sites cannot be accurately found, the phasing of the whole protein will be hindered. For this reason, it is important to assess the quality of the substructure. *PHASER* uses the so-called LLG-completion algorithm to find the complete substructure [19]. It is an iterative process in which including the sites that are identified in early rounds should improve the signal for identifying weaker sites in subsequent rounds. The comparison with the program *SITCOM* reveals that all the heavy-atom sites are found reliably in all cases, regardless the completeness and the accuracy of the initial MR-search models. In other terms, even in the cases of lower model completeness, the substructure-completion algorithm is able to accurately locate 18 sites in the structure (this is the number of refined sites in the

deposited model, with the missing four sulfurs not visible because in disordered regions of the structure).

MRSAD-phases were further evaluated by comparison against pure MR- and SAD-phases. The comparison among the three phasing procedures is shown in **Figure 1.3**. As can be appreciated in this figure, the improvements from MR to MRSAD are numerically limited, of the order of few degrees in terms of mean phase error: this is in contrast with what can be observed upon visual inspection of the electron density maps, which clearly suggests MRSAD to be superior. In fact, the electron density of a large fraction of side chains is much better defined in the MRSAD maps as compared with the MR maps (**Figure 1.4** and **Figure 1.5**). The residue-by-residue Real Space Correlation Coefficients computed for the MR and MRSAD maps quantitatively confirm that the electron density for the most of the residues is better defined after MRSAD-phasing, as shown in the scatter plots in **Figure 1.6**. This suggests that the MPE is not the best indicator of the quality of the phases.

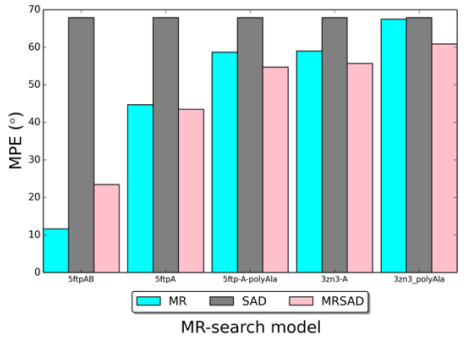


Figure 1.3: Comparison among MR, SAD and MRSAD phasing procedures for the Cdc23 test case. Only phases before any model building/density modification cycle have been considered. The MPE after SAD-phasing is 68.0º.

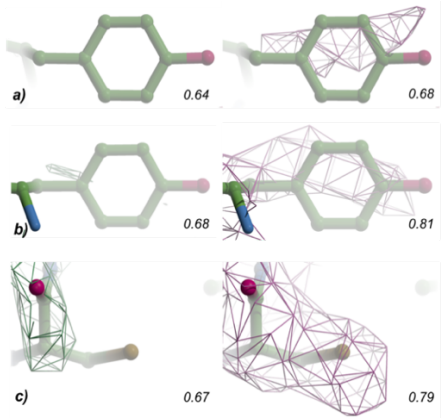


Figure 1.4: Map quality for Cdc23 with 3ZN3-A as a search model. Comparison of the $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ electron density map from MR-phasing (colored green, left) with the same type of map from MRSAD-phasing (magenta, on the right). All maps contoured at 1.5σ level. ***a)*** TYR91-B, ***b)*** TYR174-B and ***c)*** CYS199-B are shown with the associated electron densities and the relative global (main and side chain) RSCC values. The pictures were produced in *Coot* [31].

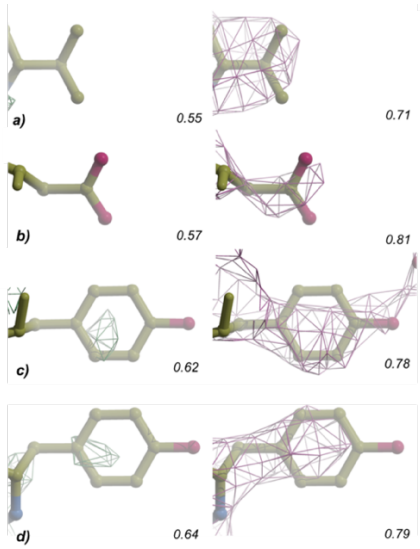


Figure 1.5: Map quality for Cdc23 with 5FTP-A as a search model. Comparison of the $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ electron density map from MR-phasing (colored green, left) with the same type of map from MRSAD-phasing (magenta, on the right). All maps contoured at 1.5σ level. ***a)*** VAL43-B, ***b)*** ASP282-B, ***c)*** TYR91-B and ***d)*** TYR306-B are shown with the associated electron densities and the relative global (main and side chain) RSCC values. The pictures were produced in *Coot*.

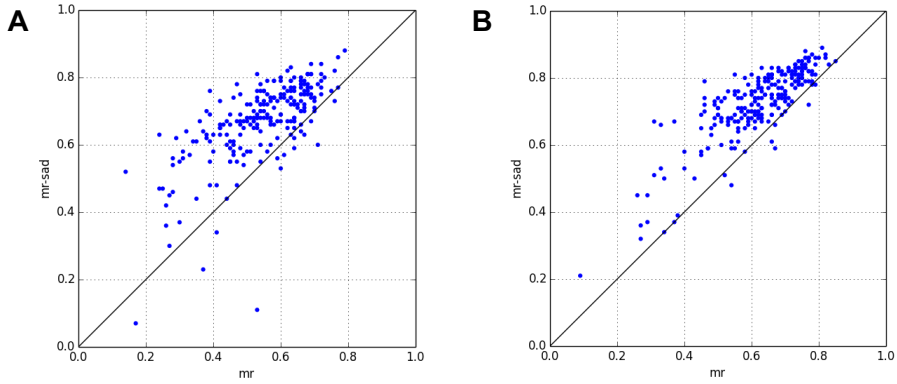


Figure 1.6: Scatter plots for the Cdc23 test case. The plots show the correlation between RSCC-MR and RSCC-*PHASER*-MRSAD for the part of the molecule not used as a search model. **(A)** 3ZN3-A was used as MR-search model (overall RSCC-MR = 0.156, overall RSCC-MRSAD = 0.245), **(B)** 5FTP-A as MR-search model (overall RSCC-MR = 0.288, overall RSCC-MRSAD = 0.399).

The quality of the MRSAD-phases was finally evaluated by assessing the ability of three different map interpretation softwares to build a sensible model with them. In **Table 1.3** the results of model building using the MRSAD phases are shown: among the three software that have been used, *BUCCANEER* shows the best performance, probably owing to the resolution of the data. The SAD- and the MR-solutions, too, have been subjected to model building with the same software (*data not shown*). Based on the auto-traced models it is possible to state that after model building the MRSAD-phases are superior compared to the phases that is possible to obtain with MR and SAD alone.

MR-search model	*ARP/wARP*				*BUCCANEER*				*PHENIX*			
	R	Chains/ aa	MPE /°	MAP-CC	*R*	aa	MPE /°	MAP-CC	R_w/R_{free}	aa built	MPE /°	MAP-CC
3zn3-A	0.27	32/151	65.2	0.360	0.36	480	52.3	0.592	0.29/0.35	358	55.8	0.553
5ftp-A	0.26	27/170	58.5	0.526	0.22	527	28.2	0.724	0.29/0.34	344	43.4	0.626
5ftp-AB	0.22	30/319	49.8	0.613	0.19	521	27.3	0.719	0.22/0.28	439	23.4	0.726
3zn3-polyAla	0.30	26/119	72.2	0.420	0.41	475	57.7	0.543	0.32/0.39	329	60.9	0.514
5ftp-A-polyAla	0.24	32/179	63.6	0.503	0.27	524	45.1	0.650	0.32/0.36	319	54.6	0.559
3zn3-noloops	0.27	27/116	66.4	0.417	0.36	500	52.7	0.588	0.29/0.34	376	57.3	0.536

Table 1.3: Results of MRSAD-phasing after model building/density modification carried out with *ARP/wARP*, *BUCCANEER* and *PHENIX* on the Cdc23 test case.

1.4.2. MRSAD-phasing tests on SeMet-FAE

Feruloyl esterase (FAE) is an enzyme for which different structures have been already solved and deposited. In the present study, unpublished diffraction data from a selenomethionine-derivatized crystal of FAE have been used. The asymmetric unit of the refined model consists of 562 residues, divided into two chains (A and B) of equal length. Out of the 18 selenomethionines theoretically present in the structure, only 16 can be observed since two of them are in the flexible terminus parts. Three low dose MAD-data sets (inflection point, peak and high-energy remote) were collected on a SeMet-FAE crystal around the Se adsorption edge (~ 12.7keV). Data statistics are reported in the *Supplementary Materials & Methods, Appendix A*. The quality of the data allowed straightforward structure solution, model building and refinement. The structure has been solved by MAD, providing *HKL2MAP* [32] with the three data sets and

exploiting the Se anomalous signal only. Both data collection and model building/refinement were carried out by Dr. Anna Polyakova. All the tests were performed using the high-energy remote data set only.

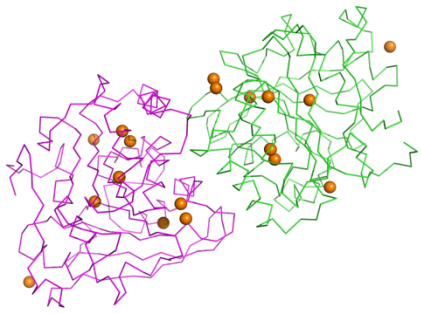


Figure 1.7: SeMet-FAE model and the Se-substructure. The Se atoms are shown as orange spheres.

The following search models were considered: full FAE model, the polyAla version of the full model (FAE_polyAla), chain A (FAE_A), the polyAla version of chain A (FAE_A_polyAla), chain B (FAE_B) and the polyAla version of chain B (FAE_B_polyAla). Among them, the reference case (the full FAE model) sets, respectively, the lower and the upper limits of MPE achievable via MRSAD on SeMet-FAE. The results of MRSAD-phasing on SeMet-FAE by using different MR-search models are summarized in **Table 1.4**. Compared to the Cdc23 case, there is a more significant improvement in the quality of experimental phases after MRSAD, mainly due to the stronger anomalous signal. The LLG-algorithm succeeds in locating all the Se atoms, as confirmed by *SITCOM* comparison.

R-search model	*PHASER*-MR	*PHASER*-MRSAD			
	Map comparison	Map comparison	*SITCOM* against refined model		
	MPE/°	MPE/°	NM	r.m.s.d.	Score
FAE	26.2	13.7	18	0.14	0.952
FAE_polyAla	46.1	28.7	18	0.14	0.953
FAE_A	50.4	35.0	18	0.10	0.967
FAE_A_polyAla	60.5	39.1	18	0.15	0.952
FAE_B	50.8	35.2	18	0.15	0.949
FAE_B_polyAla	60.8	39.3	18	0.17	0.943

Table 1.4: Results of MR- and MRSAD-phasing in terms of MPE and MAP-CC for the SeMet-FAE test case.

The improvements in the phases are reflected in real space: the scatter plots in **Figure 1.8** show that the electron density of all the residues in the region of the molecule which was not used for the MR search is better defined for MRSAD than for MR.

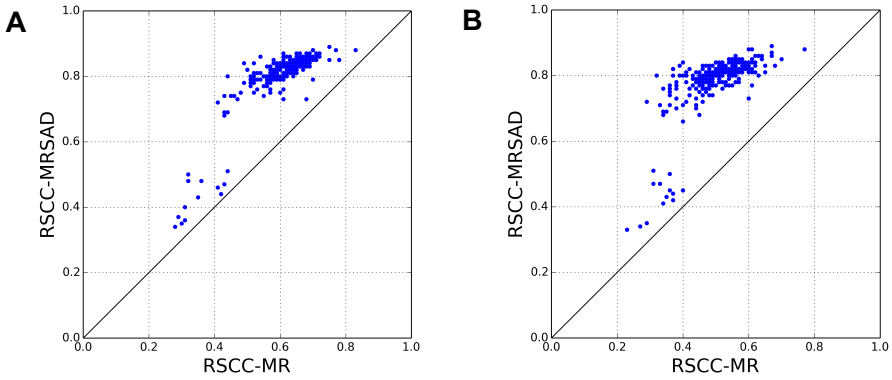


Figure 1.8: Scatter plots for the SeMet-FAE test case. The plots show the correlation between RSCC-MR and RSCC-*PHASER*-MRSAD for the part of the molecule not used as a search model. **(A)** Chain A was used as MR-search model (overall RSCC-MR = 0.267, overall RSCC-MRSAD = 0.519), **(B)** polyAla version of Chain A as MR-search model (overall RSCC-MR = 0.215, overall RSCC-MRSAD = 0.485).

The comparison of the MR, SAD and MRSAD-phasing scenarios confirms the superiority of the MRSAD-phases. As evident from **Figure 1.9**, the

combination of MR- and SAD-phases leads to significantly improved MRSAD-phases for all the MR-search models.

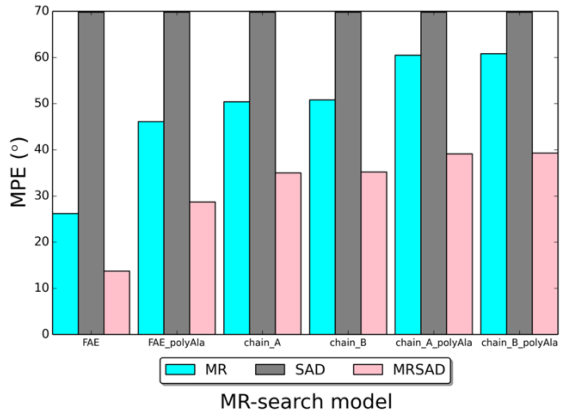


Figure 1.9: Comparison among MR, SAD and MRSAD phasing procedures for the SeMet-FAE test case. Only phases before any model building/density modification cycle have been considered. The MPE after SAD-phasing is 69.8º.

For the case of SeMet-FAE, model building into MRSAD-maps further improves the combined phases. Among the three map interpretation softwares that were used, *BUCCANEER* shows the best performances (**Table 1.5**). However, complete (or almost complete) models could be obtained with all three softwares: this is confirmed by visual inspection and by common model building parameters such as R_{work}, R_{free}, number of residues traced and chains and the MPE of the final model against the reference.

Additional tests on other proteins confirmed that model building on systems of small and medium size further improves the MRSAD-phases. However, the application of the same pipeline for MRSAD-phasing and model building showed limitations on systems of higher molecular weight. For instance, in the case of the Br-soaked human 20*S* proteasome none of the model building softwares tested was able to automatically build a significant part of the proteasome into the MRSAD maps (as will be better

discussed in the following chapter). In the best cases, little more than half of the residues were built, the models lacked backbone continuity and were highly fragmented.

MR-search model	*ARP/wARP*			*BUCCANEER*			*PHENIX*		
	R_w/R_{free}	Chains/ aa	MPE/°	R_w/R_{free}	Chains/ aa	MPE/°	R_w/R_{free}	aa built/ placed	MPE/°
FAE	0.28/0.30	561/2	28.9	0.29/0.30	535/2	24.3	0.28/0.29	563/543	29.6
FAE_polyAla	0.27/0.30	560/2	28.2	0.29/0.30	535/2	25.5	0.28/0.29	562/542	13.7
Chain_A	0.27/0.30	560/2	27.7	0.31/0.33	422/3	30.3	0.28/0.29	551/541	35.0
Chain_A_polyAla	0.27/0.30	560/2	29.1	0.31/0.33	525/2	29.6	0.28/0.30	552/544	39.1
Chain_B	0.27/0.30	560/2	28.4	0.30/0.32	521/2	28.3	0.29/0.30	552/542	35.2
Chain_B_polyAla	0.27/0.30	560/2	27.7	0.30/0.32	507/2	28.5	0.29/0.30	551/541	39.3

Table 1.5: Results of MRSAD-phasing after model building/density modification carried out with *ARP/wARP*, *BUCCANEER* and *PHENIX* on the SeMet-FAE test case.

For example, the MRSAD-solution obtained by using subunits $\alpha 1$ and $\alpha 2$ has a good MPE; however, when the map interpretation programs try to build a model into it, either there is a significant worsening or no improvement at all of the MRSAD-solution.

To determine the limits of the pipeline for MRSAD-phasing and model building, a systematic study was carried out on SeMet-FAE data. The details and the results of this study are discussed in the following section.

1.4.2.1 Limitations to the general applicability of a pipeline for MRSAD-phasing and model building

SeMet-FAE was used as a reference system because of its medium size, the availability of high-resolution data and its dimeric structure: the idea is that any conceived pipeline which does not work on high-resolution data will also be ineffective on lower-resolution data from systems of larger molecular weight. An automated pipeline was developed to test the resolution limits of *SHELXE*-MRSAD and two map interpretation

softwares on the FAE data, for different resolution and completeness levels of the MR-search models. In what follows, the results from the three main parts of the pipeline are discussed separately, in the order in which they are performed:

SHELXE-MRSAD *phasing*

In general, there is no clear drop of any of the *SHELXE* parameters (*Supplementary Materials & Methods, Appendix B*) but rather a gradual decline as a function of resolution. In most of the cases, *SHELXE* succeeds to find the complete, accurate substructure and to provide a good solution regardless of the MR-search model completeness and resolution. Only for the lowest MR-search model completeness level at lowest resolutions a clear distinction can be made between good and bad solutions (between 3.0 and 3.2 Å).

Cα finding and backbone extension with trace_chain and pulchra

The combination of the *trace_chain* and *pulchra* algorithms aims to provide a partial model which is sufficiently good to bootstrap successive model building. However, the partial models so generated did not facilitate the tracing in the last step. In fact, it could be observed that there is no significant difference in the performance of the model building software when providing them with the partial model, or when asking them to build directly into the MRSAD-map.

Assay of individual model building softwares performance

The individual performance of two different model building programs were tested, starting from the MRSAD maps. **Figure 1.10** shows the results of model building with *ARP/wARP* and *phenix.autobuild*. If a MPE between

50° and 60° is taken as the threshold to establish whether the structure is solved or not, it can be said that both softwares display the same resolution limit, around 2.8 Å. This limit happens to be the boundary between sensible models and models which are too poor and fragmented to be plausible. The resolution limit is nearly insensitive to the MR-search model completeness.

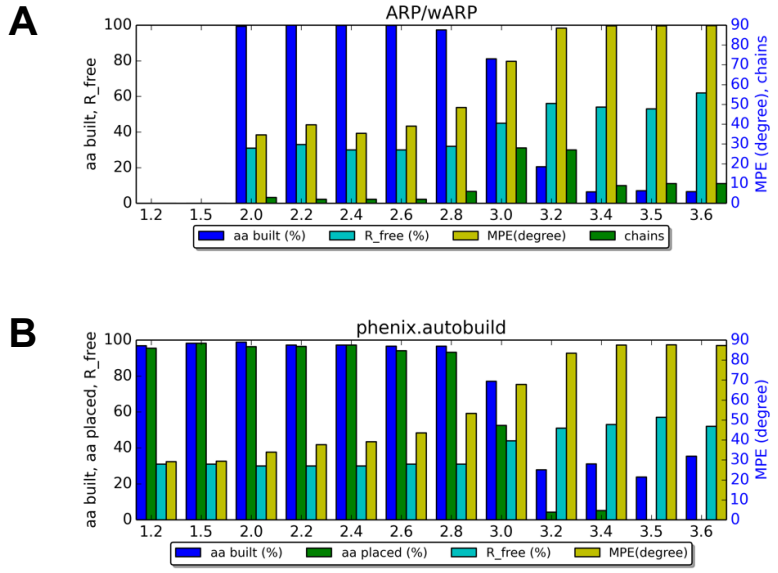


Figure 1.10: Results from model building with *ARP/wARP* and *phenix.autobuild* on FAE-maps truncated at different resolution levels for MR-search model completeness = 90%. The *SHELXE*-MRSAD map at 1.2 Å for MR-search model completeness = 90% (MPE = 19.7˚) was truncated to different resolution levels with *SFTOOLS* (CCP4 [33]). Classical *ARP/wARP* model building was performed, providing the *SHELXE*-MRSAD phases and figure of merit, together with the FAE monomer sequence and the number of residues. Automated building of alpha-helical and beta-stranded fragments ('auto_albe.sh' module) was switched-on. Each of the 5 model building cycles was interspersed with 5 refinement cycles in *REFMAC5* [34]. Standard *phenix.autobuild* was performed providing MRSAD-phases and the monomeric sequence, with keyword 'quick=True'. aa is the number of residues (built or placed).

1.4.3. Determination of the minimum amount of information required for successful MRSAD-phasing

For MRSAD-phasing, it is likely that there is a point where the full structure can no longer be obtained due to the limited size and/or

composition of the search model and/or a the too weak anomalous signal. The systematic study presented in this section aimed to identify whether it is possible to define limits for the minimum information required for successful MRSAD-phasing, in terms of size of the MR-search model and diffraction data multiplicity. This systematic study would allow the definition of the limits but also of the potentialities of MRSAD-phasing. In addition, this information could be useful to guide the planning of the diffraction data collection aimed for MRSAD-phasing. The general scheme of the pipeline is shown in **Figure 1.11**:

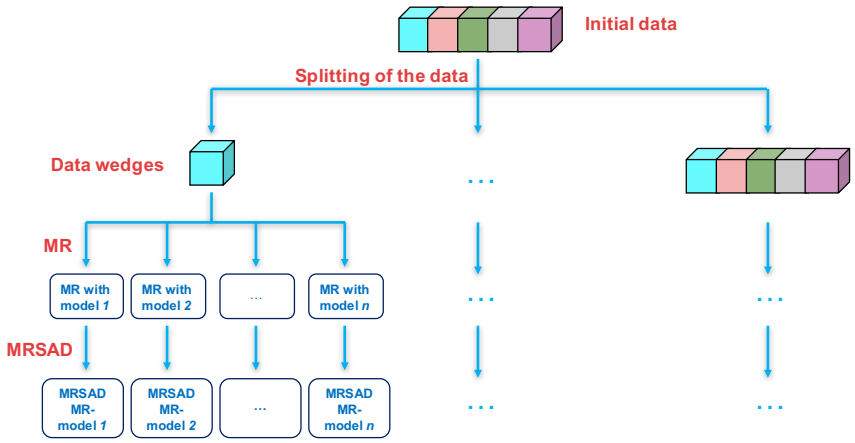


Figure 1.11: General representation of the pipeline used to determine the minimum amount of data needed for successful MRSAD-phasing.

RipA is a small, well-diffracting protein and therefore it represents an appropriate system to first test the pipeline. The anomalous signal comes from the 7 intrinsic sulfurs (the theoretical Bijvoet ratio at the wavelength at which data were collected, 1.77 Å, is ~ 1%). Data statistics are reported in *Supplementary Materials & Methods, Appendix C*. Apart from a few cases in *SHELXE*, it can be observed that the MPE for MRSAD and dm is lower than the MPE for SAD and dm. In addition, the MPE after MRSAD

and dm is lower than the MPE for MR, as expected. In almost all cases, both *PHASER* and *SHELXD* [35] reliably find the complete substructure, regardless the MR-search model completeness. If the information on the starting model is not provided (SAD-phasing), the substructure is always determined reliably with 720° of data or more, but at least 1440° of data are necessary for successful *SHELXE*-SAD-phasing (solving the substructure is a necessary but not a sufficient condition for successful phasing). For the case of RipA, the structure cannot be solved by MRSAD-phasing when the completeness of the MR-search model is 10% or lower (indeed, *ARP/wARP* fails to build any sensible model into the *PHASER*-MRSAD maps after density modification). *SHELXE*-MRSAD phases were also not good enough, as shown by the correlation coefficients between the native structure factors and those calculated from the polyalanine trace (well below the threshold of 25%). The heatmap in **Figure 1.12** shows the variation of the MPE of *PHASER*-MRSAD phases after density modification and of *SHELXE*-MRSAD phases as a function of the completeness levels and of the amount of data used for phasing. Increasing the amount of diffraction data has no effect at 10% search model completeness as it does not lead to the solution of the structure. However, RipA structure is solved as soon as the completeness of the MR-search model is increased to 30% (which is the case for both *PHASER* and *SHELXE*). Multiplicity plays a role only at 30% completeness level (in this case, at least 3 turns of data are needed for successful MRSAD-phasing). As expected, the effect of increasing the amount of data decreases as the MR-search models become more complete.

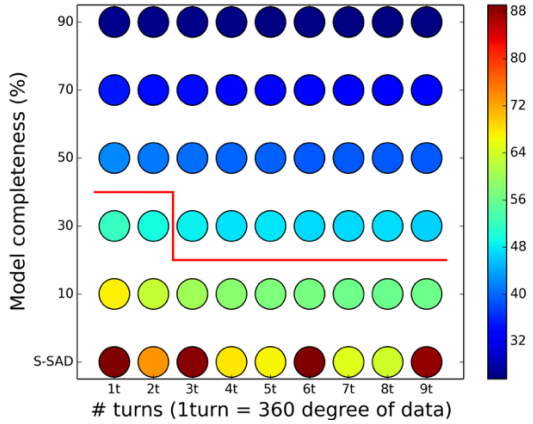


Figure 1.12: Determination of the minimum amount of information for successful MRSAD for the RipA test case. The heatmap shows the variation of the MPE of *PHASER*-MRSAD phases after *PARROT* dm as a function of the completeness level of the MR-search models and of the data turns used for phasing. The baseline of the heatmap is represented by S-SAD phasing, which was carried out in *PHASER* by providing the *SHELXD* substructure. The color legend refers to the MPE. The red line separates the solutions which can be traced by *ARP/wARP* from those who cannot be automatically traced.

The scatter plots qualitatively show the improvement in the MRSAD-phases when the completeness of the MR-search model is increased from 10% to 30%, as in the example reported in **Figure 1.13**. However, because auto-tracing does not work on systems of high molecular weight, for challenging cases the only way to assess whether the structure has been solved or not is to visually inspect the electron density maps. Electron densities for selected residues and/or α-helices at specific MPE values could also be used in another way. In particular, they could represent an indication during the process of solving a new structure, telling how close or far the crystallographer is from solving the structure. Electron density features of α-helices are particularly useful for this purpose: this is because α-helices appear already at 4.0 Å (or worse) resolution in an electron density map, and their characteristic helicoidal shape helps to find them. **Figure 1.14** shows the variation of the electron density of an α-helix for selected MPE values for the RipA case. A MPE ~ 50° is sufficient for the

electron density of the helix and of most of the side chains to become visible; in the maps at lower MPE, the electron density becomes even more defined and additional features appear (e.g., the side chains). However, the map with MPE ~ 50° can be automatically traced since the α-helices are already well distinguishable.

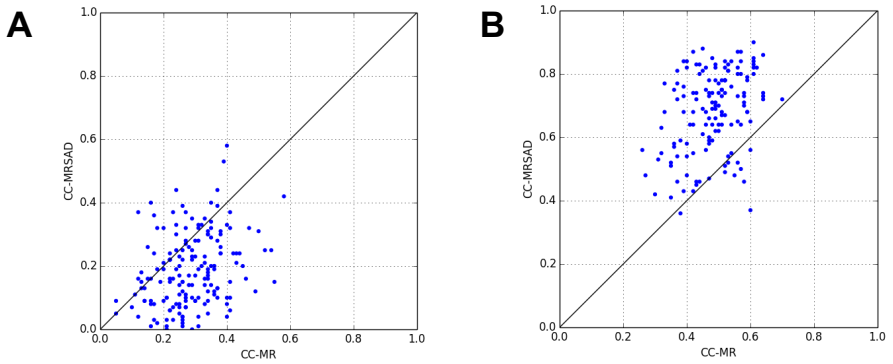


Figure 1.13: Scatter plots for the RipA test case. The plots show the correlation between RSCC-MR and RSCC-*SHELXE*-MRSAD for the part of the molecule not used as a search model. **(A)** Model completeness of MR-search model = 10%, 1 turn of data (360˚). **(B)** Model completeness of MR-search model = 30%, 1 turn of data. A significant improvement in the MRSAD-phases is observed when the model completeness of the MR-search model is increased from 10% to 30%.

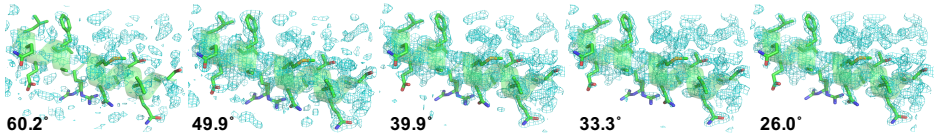


Figure 1.14: Electron density maps for a selected α-helix at different MPE levels for the RipA test case. The figure shows the $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ electron density maps for the α-helix from Ser40 to Gly56 in RipA. The maps are shown for selected MPE values of *PHASER*-MRSAD solutions after *PARROT* density modification, which are indicated at the bottom. Maps contoured at 1.5σ level.

1.5. Conclusions

MRSAD-phasing was tested on a number of systems of small and medium size. These test systems do not represent the best target for the application of MRSAD because their structures could also be solved by MR- or SAD-phasing. However, they are useful and appropriate as most of them are composed of two or more chains (therefore, one can easily pretend that one or more chains are not known) and because in some cases the anomalous signal is not as accurate as it would be in an ideal case. Furthermore, the results obtained by testing MRSAD on these systems can be used as a reference to evaluate the performance of this phasing method on more challenging systems, as described in the next chapter.

Tests on small and medium size systems with decent anomalous signal show the improvements of MRSAD-phasing over MR- and SAD-phasing alone. In the SeMet-FAE case, phases are significantly improved after MRSAD-phasing, both in real (*i.e.*: electron density maps) and reciprocal space (*i.e.*: MPE). Improvements can be observed even on considerably larger systems with a weaker anomalous signal, as for the case of Br-soaked human 20*S* proteasome. It was observed that the mean phase error is not the best metric for phase quality, as it represents an average over all the reciprocal space. On the contrary, the RSCC's reflect local improvements in the electron density and are a better metric to assess the quality of the MRSAD-phases. Scatter plots of RSCC-MRSAD against RSCC-MR are a fast and simple way to assess improvements in the electron density of each residue, and they are particularly informative when used to look at the electron density of the region of the molecule which was not part of the MR-search model.

Building a sensible model into MRSAD-phases is possible only up to a certain resolution limit. This has emerged by testing the performance of popular map interpretation softwares on MRSAD-phases for the SeMet-FAE case. It was shown that it is not possible to build sensible models into ~ 2.8 Å (or worse) resolution MRSAD maps. Because SeMet-FAE is an almost ideal case, the results show that a pipeline for MRSAD-phasing and model building is likely to be ineffective on lower resolution data from systems of larger molecular weight. This is even more evident in a typical challenging system: testing the same pipeline on Br-soaked human 20*S* proteasome has revealed that the three map interpretation softwares cannot automatically build into medium-to-low resolution MRSAD-maps. From these results, it is evident that auto-tracing cannot be used as a metric to establish whether a protein structure has been solved or not. The electron density features of α-helices or of side-chains of selected residues represent an alternative way of evaluating a map and determining whether the map is traceable or not.

The effect of the completeness of the MR-search model and of the data multiplicity on MRSAD-phasing was studied. In general, the completeness of the MR-search model plays a bigger role than data multiplicity, but in borderline cases, as effectively shown by the heatmaps, the role of the amount of data on successful MRSAD-phasing becomes evident. The effect of multiplicity is particularly important in cases as the one represented by RipA, where native phasing with the intrinsic sulfurs is attempted. Tests on RipD, a small protein similar to RipA, where the anomalous signal comes from three intrinsic sulfurs (the theoretical Bijvoet ratio at the wavelength at which data were collected is ~ 0.9%), confirms this conclusion.

Supplementary Materials & Methods

Appendix A

Data statistics for the high-energy remote data set collected on SeMet-FAE (values in parentheses are given for the highest resolution shell):

Beamline	P14, PETRA III
Detector	Pilatus 6M
Transmission (%)	5
Total oscillation (°)	3600 x 0.1
Total exposure time (s)	144
Beam size (V × H, μm^2)	150 x 100
Max dose (MGy)	0.92
Detector resolution at edge (Å)	1.2
Wavelength (Å)	0.9763
$d_{max} - d_{min}$ **(Å)**	56.6–1.20 (1.27–1.20)
Space group	$P2_12_12_1$
Unit cell parameters (Å, °)	a = 58.1, b = 58.1, c = 150.5, $\alpha = \beta = \gamma$ = 90.0
No. of reflections	3,290,087 (508,919)
No. of unique reflections	460,771 (71,857)
Multiplicity	7.0 (7.1)
Completeness (%)	94.6 (91.1)
$\langle I/\sigma(I) \rangle$	19.58 (2.99)
CC(1/2) (%)	99.9 (79.9)
CC_{ano} **(%)**	69 (13)
$R_{r.i.m.}$ **(%)**	5.9 (64.2)
$R_{meas.}$ **(%)**	6.3 (69.3)
SigAno (ΔF/σ)	2.1 (0.8)
Mosaicity (°)	0.056

Appendix B

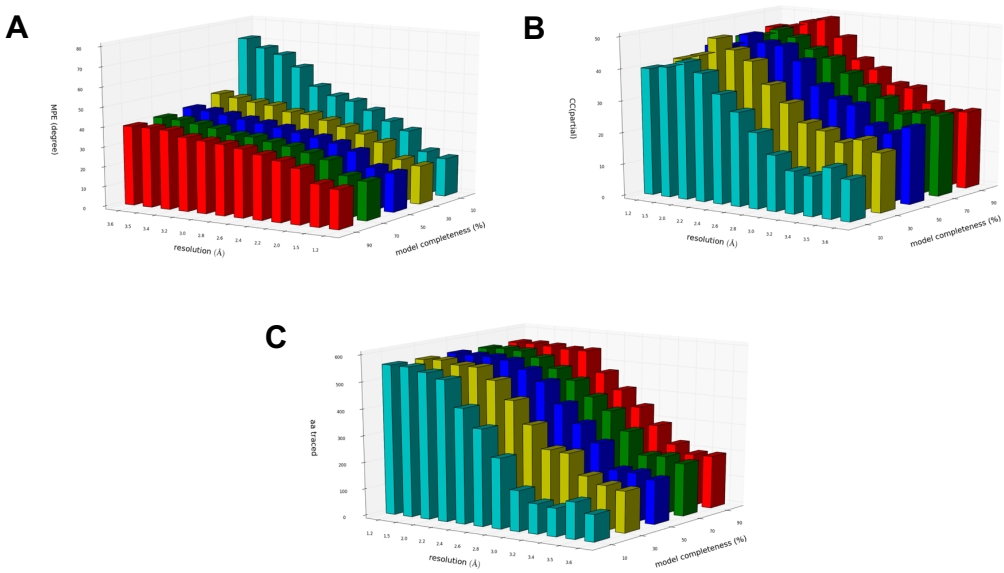


Figure 1.15: Testing the general applicability of a pipeline for MRSDA-phasing and model building on SeMet-FAE data. Variation of **(A)** the *SHELXE*-MRSAD MPE, **(B)** of the correlation coefficient between the native structure factors and those calculated from the polyAla trace and **(C)** of the number of residues traced in the polyAla model as a function of model completeness and resolution. MRSAD has been carried out in *SHELXE* using the partial MR-model, running 10 cycles of density modification and 3 global auto-tracing cycles, with a solvent content of 58.5% (option to keep starting fragments unchanged throughout all cycles has been switched off).

Appendix C

Data statistics for RipA (values in parentheses are given for the highest resolution shell):

Beamline	P14, PETRA III
Detector	Pilatus 6M
Transmission (%)	5
Total oscillation (°)	32400 x 0.1
Total exposure time (s)	1296
Beam size (V × H, μm²)	125 x 167
Max dose (MGy)	1.1
Detector resolution at edge (Å)	1.82
Wavelength (Å)	1.771
$d_{max} - d_{min}$ **(Å)**	67.82–1.78 (1.88–1.78)
Space group	$P2_12_12_1$
Unit cell parameters (Å, °)	$a = 36.7$, $b = 65.6$, $c = 67.8$, $\alpha = \beta = \gamma = 90.0$
No. of reflections	1,387,449 (125,646)
No. of unique reflections	29,943 (4,247)
Multiplicity	46.37 (29.58)
Completeness (%)	98.3 (92.1)
$\langle I/\sigma(I) \rangle$	55.15 (15.29)
CC(1/2) (%)	100.0 (99.4)
CC_{ano} **(%)**	36 (42)
$R_{r.i.m.}$ **(%)**	6.1 (20.8)
$R_{meas.}$ **(%)**	6.2 (21.2)
SigAno (ΔF/σ)	1.16 (0.887)
Mosaicity (°)	0.087

2. MRSAD-PHASING OF THE HUMAN 20*S* PROTEASOME

2.1. Summary

The central model system for the study of MRSAD-phasing of large macromolecular complexes is represented by the human 20*S* proteasome. The 20*S* proteasome was chosen as it represents a possible, real-life target of MRSAD-phasing. This is because of the presence of a high number of molecules in the asymmetric unit and of the naturally built-in sulfurs (or heavy-atoms which can be introduced by derivatization). MRSAD-phasing was tested on two different data sets obtained from native and Br-soaked crystals of the human 20*S* proteasome. For both data sets, the challenge comes from: *i)* the medium-to-low resolution of the data (between 2.5 and 2.9 Å), *ii)* the high molecular weight of the target structure (~ 750kDa) and *iii)* the low completeness of some of the MR-search models. The native 20*S* proteasome data poses an additional challenge because of the data collection wavelength (2.0664 Å) and the source of the anomalous signal, provided only by intrinsic sulfurs. Despite these difficulties, the results show that MRSAD-phasing can improve MR-phases on a large macromolecular complex by using even the anomalous signal of weak scatterers and search-models representing only a small fraction of the target. The most important factors for the success of MRSAD appear to be the accuracy of the collected anomalous differences, the ability of the LLG-algorithms to correctly locate the substructure and the density modification procedure.

2.2. The choice of the human 20*S* proteasome as the central model system

The eukaryotic 20*S* proteasome structures consist of four stacked rings, each one organized in 7 subunits. There are 7 distinct α and 7 distinct β chains making up, respectively, the two outer and the two inner rings (**Figure 2.1**). The eukaryotic proteasome is therefore arranged as $\alpha_{1-7}\beta_{1-7}\beta_{1-7}\alpha_{1-7}$. The asymmetric unit consists of 28 molecules, for a total molecular weight of approximately 750kDa.

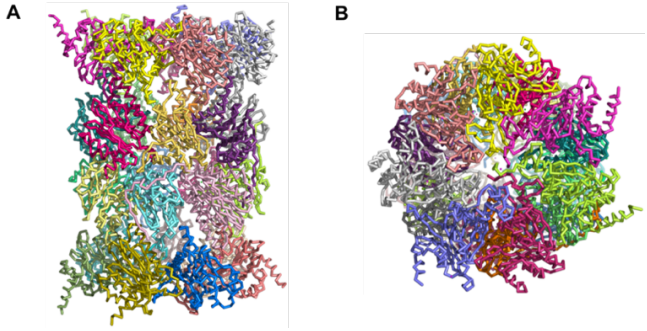


Figure 2.1: Representation of the human 20*S* proteasome model. (A) Side and **(B)** top view of the human 20*S* proteasome structure at 1.8 Å [36] (PDB ID: 5LE5) in ribbon representation. The four stacked rings consist of 7 subunits each; there are 7 distinct α and 7 distinct β chains.

The human and the yeast 20*S* proteasomes have been extensively characterized and described in terms of their structure and function, and a number of models is available in the PDB. Because of all these reasons, the human 20*S* proteasome was chosen as the central model system for this study.

2.3. Materials & Methods

2.3.1. Sample preparation, crystallization, data collection and processing

The details about the analysis of the Br-20*S* data can be found in the previous Chapter "MRSAD-phasing of small and medium size proteins".

The details about the analysis of the native human 20*S* proteasome data collected at 6keV can be found in the manuscript draft in the *Appendix Manuscripts* section.

2.4. Results & Discussion

2.4.1. Tests on Br-soaked human 20*S* proteasome data

The data set was collected at beamline P14 at 13.48keV and processed to 2.5 Å resolution (data statistics in *Supplementary Materials & Methods, Appendix A*). The results of the MRSAD-phasing and model building pipeline are summarized in **Table 2.1** and **Table 2.3**, respectively.

MR-search model	*PHASER*-MR	*PHASER*-MRSAD			
	Phase comparison	Phase comparison	*SITCOM* against refined model		
	MPE	MPE	NM	r.m.s.d.	score
Br-20S	13.1	13.1	57	0.20	0.933
Br-20S_polyAla	38.1	37.9	56	0.25	0.900
α1	60.0	59.1	51	0.38	0.781
α1_polyAla	66.2	64.9	49	0.44	0.734
α1+2	45.4	45.3	56	0.32	0.879
α1+2_polyAla	55.4	54.8	54	0.41	0.817
ß1	61.5	60.3	53	0.37	0.817
ß1_polyAla	67.2	65.8	46	0.46	0.682
ß1+2	47.8	47.2	56	0.26	0.897
ß1+2_polyAla	57.2	56.2	56	0.30	0.883

Table 2.1: Results from MR- and MRSAD-phasing in terms of MPE for the Br-20*S* test case. All the search models have been obtained from a refined model of the human 20*S* proteasome generated from the Br-soaked data set collected at 13.48keV to 2.5Å. *SITCOM* has been used for comparison of substructure sites coming from *PHASER*-MR/MRSAD against the refined model, which contains a total of 57 Br atoms.

A number of search models have been used, all based on the α and ß subunits of which the proteasome is composed. The following search models were used: the full proteasome model (Br-20S), its polyAla version

(Br-20S_polyAla), alpha-1 subunit (α1), beta-1 subunit (ß1), alpha-1 and alpha-2 subunits (α1+2), beta-1 and beta-2 subunits (ß1+2) and their respective polyAla versions (α1_polyAla, ß1_polyAla, α1+2_polyAla, ß1+2_polyAla). The full, low-resolution proteasome model represents the reference case and sets, respectively, the lower and the upper limits of MPE achievable via MR and MRSAD. The mean phase errors obtained with all the other search models are significantly higher than for the full proteasome model (always ~ 45° or above). However, an improvement of the quality of the phases after MRSAD can be observed: this improvement is of the order of ~ 1-2° in terms of MPE but, as already observed for small and medium size proteins, it does not correlate with the quality of the electron density maps (left side of **Figure 2.3**). In fact, after MRSAD, the electron density of the most of the residues that are not part of the initial search model improves, as it can be appreciated in the scatter plots in **Figure 2.2** and **Figure 2.3**. The improvements gained by performing MRSAD are also observed at the level of the substructure, as all or almost all of the Br sites can be found reliably regardless the search models used. As for the cases previously described of $Cdc23^{NTerm}$ and SeMet-FAE, the MRSAD-phasing protocol allows, in all the three cases, to reliably find the substructure, which results in an improvement of the phases, both in real and reciprocal space.

Concerning the real space improvements, as expected, the larger the size of the MR-search model, the higher the increment in the RSCC after MRSAD. For all the MR-search models, the increase in RSCC after MRSAD is significant, ranging from 13 to 27%.

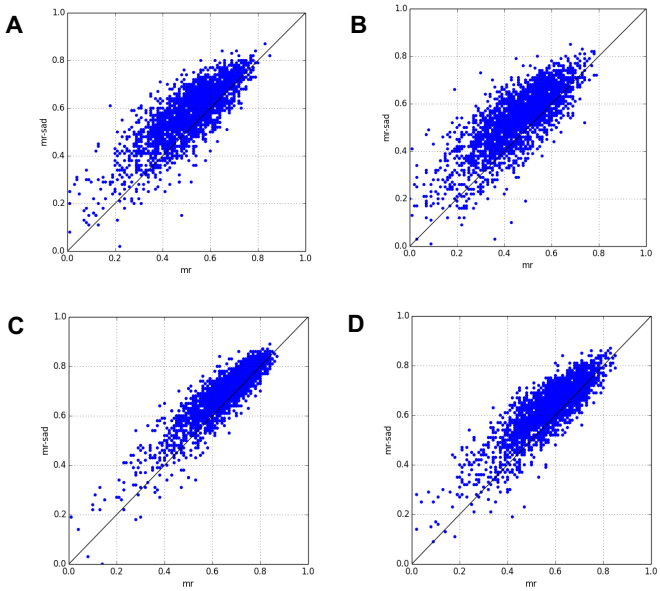


Figure 2.2: Scatter plots for the Br-20*S* test case. Correlation between RSCC-MR and RSCC-MRSAD for the part of the model which was not used for the molecular replacement search. Scatter plots are shown for different search models: **(A)** α1 **(B)** α1_polyAla **(C)** α1+2 **(D)** α1+2_polyAla. Residue-by-residue overall correlation coefficients have been computed with *phenix.get_cc_mtz_pdb*.

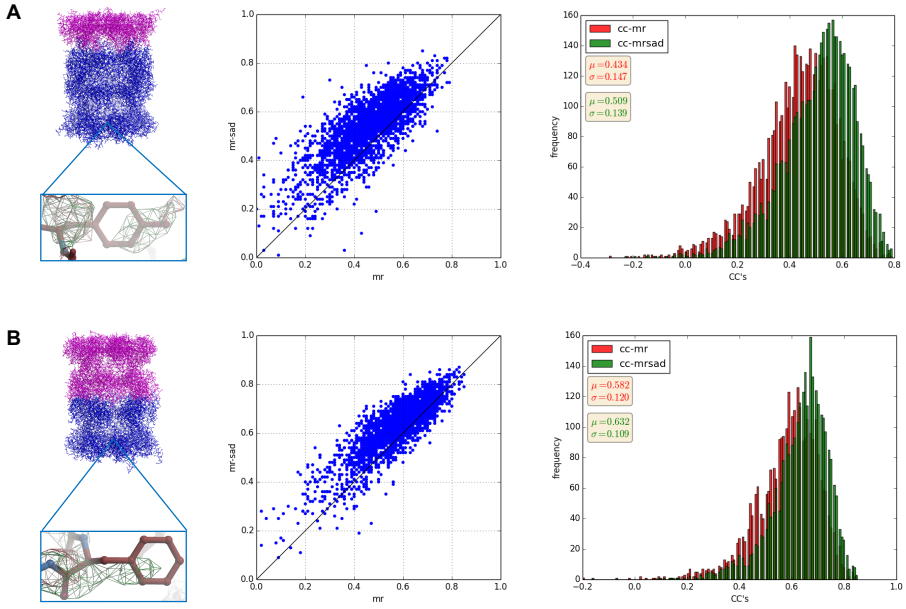


Figure 2.3: Testing MRSAD using different MR-search models for the Br-20*S* test case. (A) α_1_polyAla as MR-search model. (Left) Human 20*S* proteasome model (in magenta is α_1) with a magnification of the electron density of Tyr90 in chain Y (in red: MR-map, in green: MRSAD-

map, all contoured at 1.5σ level). RSCC-MR = 0.40, RSCC-MRSAD = 0.56; (Center) Correlation between RSCC-MR and RSCC-MRSAD for the part of the model which was not used for the MR search; (Right) Distributions of RSCC values: histograms of RSCC-MR and RSCC-MRSAD. **(B)** α_1+2_polyAla as MR-search model. (Left) Human 20*S* proteasome model (in magenta are α_1 and α_2) with a magnification of the electron density of Phe83 in chain N (in red: MR-map, in green: MRSAD-map, all contoured at 1.5σ level). RSCC-MR = 0.71, RSCC-MRSAD = 0.79; (Center) Correlation between RSCC-MR and -MRSAD for the part of the model which was not used for the MR search; (Right) Distributions of RSCC's: histograms of RSCC-MR and -MRSAD.

As it has been observed for $Cdc23^{NTerm}$ and SeMet-FAE, the quantitative improvements as measured by the MPE do not reflect the quality of the electron density, which confirm the RSCC as a better metric to quantify the gains obtained after MRSAD. Scatter plots allow for an easy visual interpretation, but a quantitative measure of the improvement in the RSCC would clearly be a better way to assess it. For quantitative analysis, distributions of overall (main and side chain) RSCC-MR and MRSAD for all test cases were computed and fitted with a normal distribution to determine their means and widths. A selected example is reported in **Figure 2.3**, and in **Table 2.2** the mean and standard deviation are reported for all search models. **Table 2.2** shows that the mean RSCC is shifted towards higher values in the case of MRSAD and that, at the same time, the width of the MRSAD distribution is slightly decreased compared to the MR distribution. Therefore, the variation in the mean of the RSCC-MR and RSCC-MRSAD distributions appears to be a good metric to assess and quantify the real space improvements of MRSAD-phasing. Side-chain-correlation-coefficients in un-built regions are expected to be an even better metric of phase quality. Side-chain-RSCC were computed with the program *Overlapmap* (CCP4) and the resulting distributions were fitted as already described. The mean of the distributions confirmed to be a good metric as was proved by using the overall RSCC (*data not shown*). Having a quantitative measure of the improvement in the RSCC from MR to

MRSAD is of potential importance in setting up an automated pipeline for phasing and model building. For example, by setting a minimum threshold for the real space improvement in the electron density maps, it would be possible to devise a pipeline which automatically tries different phasing scenarios until a certain map quality has been achieved.

MR-search model	MR		MSRAD	
	$\bar{x}$	σ	$\bar{x}$	σ
α1+2	0.651	0.112	0.689	0.103
ß1+2	0.625	0.145	0.662	0.134
α1+2_polyAla	0.582	0.120	0.632	0.109
ß1+2_polyAla	0.544	0..148	0.594	0.137
α1	0.490	0.147	0.552	0.137
ß1	0.487	0.153	0.548	0.146
α1_polyAla	0.434	0.147	0.509	0.139
ß1_polyAla	0.431	0.151	0.502	0.145

Table 2.2: Distribution parameters, mean and standard deviation, for selected MR-search models for the Br-20*S* test case. For the characterization of the distributions of overall (main and side chains) RSCC-MR and RSCC-MRSAD, histograms were fitted with a normal distribution to determine their means and widths. All fits were performed using functions available in the Python NumPy library.

2.4.1.1 Model building into MRSAD-phases

Despite the phase improvement obtained after MRSAD, none of the model building softwares was able to build a significant part of the proteasome into the MRSAD maps (**Table 2.3**). In fact, the reference proteasome structure contains 6215 residues, and in the best cases little more than half of the amino acids were built. In addition, the models lacked backbone continuity and were highly fragmented. For example, the MRSAD-solution obtained by using subunits α1 and α2 has a good MPE; however, when the three map interpretation softwares try to build a model into this map, either there is a significant worsening (*ARP/wARP* and *BUCCANEER*) or no

improvement at all (*PHENIX*) of the MRSAD-solution. Following these findings, a test was carried out in order to understand the capabilities and limitations of the model building programs in a more systematic way.

MR-search model	*ARP/wARP*			*BUCCANEER*			*PHENIX*		
	R_{work}	Chains/ aa	MPE	R_{work}	Chains/aa	MPE	R_{work}/R_{free}	aa	MPE
α1	0.26	352/2251	71.6	0.50	135/1609	74.7	0.43/0.45	2699	59.1
α1_polyAla	0.26	371/2408	75.6	0.49	145/1823	71.3	0.45/0.49	2384	64.9
α1+2	0.26	364/2563	64.6	0.40	194/4417	50.8	0.40/0.44	3735	45.3
α1+2_polyAla	0.28	377/2565	67.9	0.42	184/4410	55.9	0.39/0.44	3711	54.8
ß1	0.26	369/2457	70.8	0.52	203/2222	78.1	0.44/0.46	2494	60.3
ß1_polyAla	0.26	336/2525	75.3	0.49	167/1698	73.4	0.48/0.51	1477	65.8
ß1+2	0.26	308/3048	59.0	0.39	214/3478	50.8	0.40/0.43	3116	47.3
ß1+2_polyAla	0.26	332/3009	64.2	0.40	235/3773	54.5	0.41/0.46	2971	56.2

Table 2.3: Results of MRSAD-phasing after model building/density modifications carried out with *ARP/wARP*, *BUCCANEER* and *PHENIX* on the Br-20*S* data. aa is the number of residues traced in the MRSAD electron density maps.

Here, proteasome models were created by gradually removing one chain at a time. Every model was used for a molecular replacement search in *PHASER*-MR, and the resulting MR-model was used as a starting point for model building with *ARP/wARP* and *BUCCANEER* (*Supplementary Materials & Methods, Appendix B*). Despite differences in the performance, the same trend was observed for both programs. The parameters that were used to monitor the program performances change gradually (*i.e.*: there is not a sudden drop), with the exception of R_{work} which remains almost constant in *ARP/wARP*. Already when the MR-search model is 50% complete, both *ARP/wARP* and *BUCCANEER* can build only half (or less) of the sequence into the electron density maps. Model inspection revealed that model building results in partial and highly fragmented models already after few chains are removed from the full

proteasome structure.

2.4.1.2 Iteration between secondary structure search and MRSAD-phasing

The inability of the model building programs to interpret medium resolution electron density maps inspired the idea of first iterating between (MR)SAD-phasing and α-helices/ß-strands search (**Figure 2.4**).

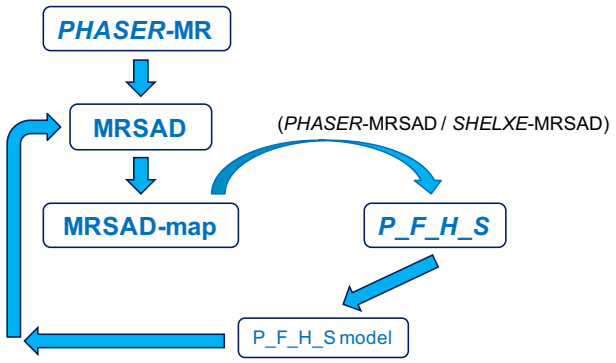


Figure 2.4: Overview of the iterative process between MRSAD-phasing and α-helices/ß-strands search.

At every cycle, new heavy-atom sites should be found, as well as new residues belonging to secondary structure elements. This iteration should gradually improve the substructure and allow more and more secondary structure elements to be found, providing a model which is good enough to bootstrap model building. Two pipelines have been set up which iterate between *PHASER*-(MR)SAD or *SHELXE*-(MR)SAD and *phenix.find_helices_strands* [37], [38] to search for helices and strands. After MR- and MRSAD-phasing, the MRSAD solution is used by *phenix.find_helices_strands* to search for helices and strands. The *phenix.find_helices_strands* model is then used as a starting point for a second iteration cycle, where the (MR)SAD-phasing and helices/strands

search is repeated in the same way. This procedure, where the anomalous (SAD) information is added to the model at each step, is iterated for an arbitrary number of cycles. **Table 2.4** reports the results for the first five iteration cycles when starting from the full proteasome model (ideal case):

Cycle	MRSAD-phasing		*SITCOM*			*phenix.find_helices_strands*
	MPE	**Br sites**	**NM**	**r.m.s.d.**	**Score**	**Total residues**
1	13.1	238	57	0.20	0.933	3287
2	67.6	80	53	0.43	0.796	2901
3	71.2	75	53	0.49	0.779	2621
4	73.4	66	50	0.51	0.728	2450
5	75.1	68	50	0.50	0.730	2277

Table 2.4: Results of the iteration between *PHASER*-(MR)SAD phasing and α-helices/ß-strands search with *phenix.find_helices_strands* for the Br-20*S* test case. The initial molecular replacement search has been carried out using the full proteasome model. (MR)SAD-phasing results are evaluated on the basis of the MPE with respect to the reference model, on the number of Br sites found by *PHASER* and on the substructure comparison between the bromine sites found by *PHASER*-MRSAD and the reference substructure sites. *phenix.find_helices_strands* results are reported in terms of the total number of residues that have been found.

Both pipelines show similar results: in particular, throughout the iterative process there is a gradual worsening of the substructure and of the MRSAD-phases and a reduction of the number of residues found by the *phenix.find_helices_strands* algorithm. This becomes apparent immediately after the first iteration cycle. There are many potential explanations as to why the iteration does not work as expected: *i) PHENIX* does not keep the model fixed after each iteration cycle, *ii)* the anomalous signal might not be sufficiently strong, *iii)* the content of helices and strands in the proteasome might not be sufficiently high and/or *iv)* specific problems in the *phenix.find_helices_strands* algorithm, which prevent it to work successfully on some systems (in fact, the same iteration has been successfully applied to another case). Furthermore, it is important to

consider that the iterative approach does not work in the first iteration cycles because while the initial phase set comes from a full protein, consisting of main and side chains, the second (and all the following) phase sets will come from a (bad) backbone alone. Because of this, an initial worsening is to be expected from the first to the second cycle of iteration, but after it a gradual improvement in the quality of the phases and of the models should be observed.

2.4.2. Tests on native human 20*S* proteasome data at 6 keV

The data was collected using a fine-slicing, low-dose and high-multiplicity data collection strategy at the P14 beamline using the CRL transfocator. Three turns of data (360° * 3) were collected, one after the other, at 6keV from the same position on the same crystal. The processed data extends to 2.9 Å resolution (data statistics in *Supplementary Materials & Methods, Appendix C*).

2.4.2.1 Characterization of the anomalous signal

The source of anomalous signal comes from the 322 S atoms (212 methionines and 102 native or alkylated cysteines) overall present in the structure. Several indicators for the estimation of the anomalous signal have been proposed [39]–[41]. Among them, the Bijvoet ratio, $\langle d''/\sigma \rangle$ and $CC_{anom,1/2}$ are the most reliable. The Bijvoet ratio [28] for all the 322 Sulphur atoms at λ = 2.0664 Å is ~ 1.63% (f_S'' = 0.9479e$^-$, N_{atom} = 48432, N_{ano} = 313). Plots of $\langle d''/\sigma \rangle$ and $CC_{anom,1/2}$ over the resolution range have been analyzed. Taken together, the three indicators show that the anomalous signal at 6 keV is, at the same time, strong and accurate.

2.4.2.2 Tests with human 20*S* proteasome models

Any attempt at solving the structure of the human 20*S* proteasome by SAD-phasing failed, even when the correct and complete substructure was used. The performance of MR as a function of the search-model completeness are shown in **Figure 2.5**.

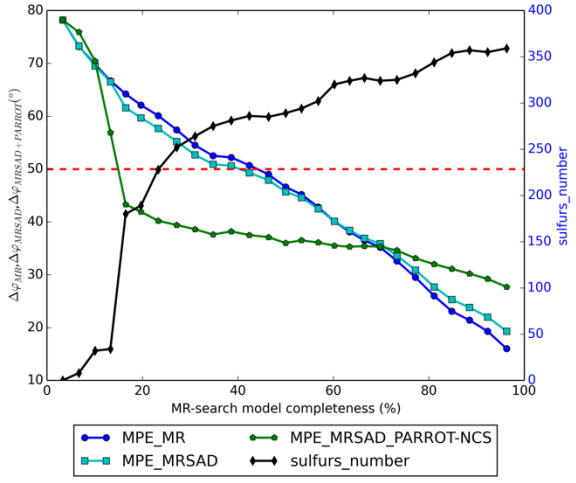


Figure 2.5: Tests with truncated models of the human 20*S* proteasome 5LE5. The plot shows the variation of the MPE for the MR, MRSAD and MRSAD+*PARROT*-NCS solutions and of the number of sites located and refined by the LLG-MRSAD algorithm as a function of the completeness of the MR-search model. The dashed red line at MPE = 50° represents an arbitrary threshold which separates solutions from not solutions.

MR is always able to correctly place the search-models, even when the smallest model (represented by chain b) is used. However, as expected, the MR-search becomes progressively more difficult as more and more chains are removed from the initial and complete model. Visual inspection of the MR-maps in the regions outside the placed models shows that, when the MPE is 50° or above, the electron density of the majority of the main and side chains is not clearly visible and only secondary structure elements (especially α-helices) can be discerned. This means that, when the completeness of the search-models is ~ 50% or lower, the MR-maps are

noisier and reveal less and less details. Density modification was applied to the MR-phases with the aim to improve them. Regardless the application or not of NCS-averaging, in the most of the cases the phases after density modification are comparable to or worse than the initial ones. Therefore, in this case, density modification is not effective in improving the MR-phases. Density modification on MR-phases is particularly important at low resolution and when the target and the template models are not structurally similar and this can explain why, in this case, density modification does not lead to a systematic improvement of the MR-phases.

MR with truncated models of 5LE5 shows that, when the completeness of the search-models is ~ 50% or lower, the interpretation of the MR-maps becomes more difficult. Therefore, there are a number of situations where MR-phases are not optimal and, if the aim is to build a model (which is often the case), they need to be improved. To this aim, MRSAD-phasing was tested for all the truncated models previously used for MR, exploiting only the anomalous signal from the intrinsic sulphur atoms. In the present and in the next section, only the results from the application of MRSAD as implemented in *PHASER* are shown and discussed, but the same procedure was applied with another popular crystallographic program (*autoSHARP*) which gave comparable results.

As it can be seen in **Figure 2.5**, MRSAD improves the MR-phases for most of the search-models, apart when the search-model completeness is high (between ~ 60% and 100%). In these cases, regardless of the density modification scheme, the MRSAD-phases are not better than the MR-phases. This is, however, not the typical real-case scenario: in reality, the homologous models often represent a small fraction of the target structure, and it is in these cases that MRSAD is expected to improve the MR-

solution. Despite improvements can be observed in the most of the cases after MRSAD, they are numerically modest. For example, when only the β2 subunit is used as a search-model for MR, the MPE of the MR-solution is 65.4°. The combination with the anomalous signal lowers the phase error to 62.5°, but this small gain of ~3° is not sufficient to obtain a map that can be easily interpreted.

With the aim to improve the MRSAD-phases, density modification with and without NCS-averaging was applied by using different protocols. Considering the previous example, when density modification is applied to the MRSAD-phases without NCS-averaging, the best protocol leads to a MPE of 55.9°, which already improves the map and reveals previously missing electron density for some of the side chains (**Figure 2.6**).

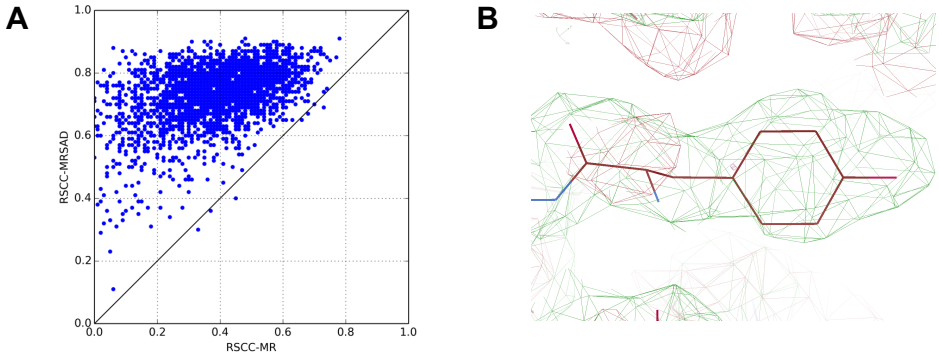


Figure 2.6: Tests with truncated models of the human 20*S* proteasome 5LE5. (A) The scatter plot refers to the case where the β2 ring was used as MR-search model and shows the correlation between RSCC-MRSAD (after *PARROT* density modification) and RSCC-MR for the unknown part (chains A to U). For the MR solution, overall map correlation and map correlation in region of the model are 0.384 and 0.503, respectively. For MRSAD, overall map correlation and map correlation in region of the model are 0.582 and 0.719. **(B)** Map quality comparison for the proteasome model. Comparison of the $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ electron density maps from MR-phasing (red) with the same type of map from MRSAD-phasing and density modification (green) for the same case as in **(A)** (all maps are contoured at 1.5σ level). The maps are shown for Y103D (RSCC-MR = 0.46, RSCC-MRSAD = 0.81). All the electron density images were created in *COOT*.

However, the inclusion of NCS-averaging has an even more significant effect and further improves the MRSAD-phases with respect to the case

where only solvent flattening and histogram matching are used (**Figure 2.7**).

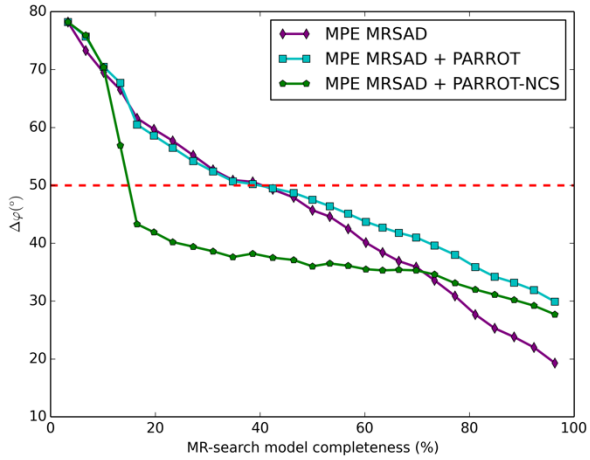


Figure 2.7: Tests with truncated models of the human 20*S* proteasome 5LE5: effect of NCS-averaging. The plot shows the improvement, over a large interval of MR-search model completeness, of the MRSAD-phases when the NCS information from the refined substructure is exploited.

When the NCS-operators are extracted from the sulphur-substructure and used for NCS-averaging, the MPE of the MRSAD-phases drops to 46.4°, with a gain of ~ 16° compared to the pure MRSAD-phases. This improvement in the phases is substantial and it is confirmed by visual inspection of the maps before and after density modification, with and without the use of NCS-averaging. **Figure 2.8** shows, for all the residues not part of the MR-search model, the distributions of RSCC-MRSAD after *PARROT* density modification with and without NCS-averaging and proves that significantly higher RSCC values are obtained when it is included. This demonstrates that the electron density of the most of the residues (in the region of the proteasome that was not used for the MR-search) significantly improves when NCS-averaging is performed. In other terms, density modification procedures in MRSAD-phasing can make the difference between an interpretable and a not interpretable map.

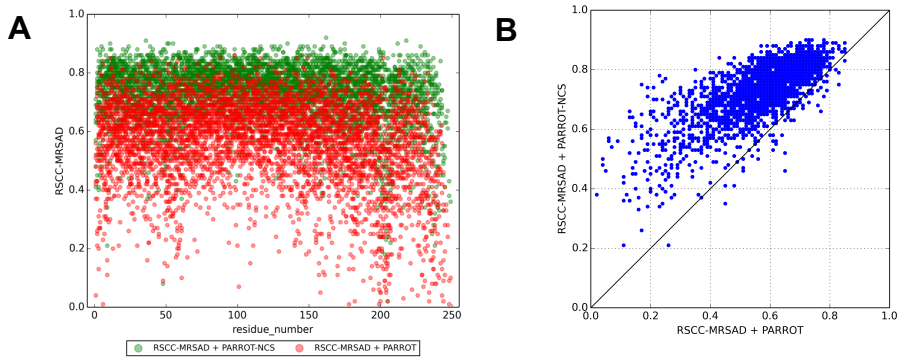


Figure 2.8: Tests with truncated models of the human 20*S* proteasome 5LE5: effect, in real space, of *PARROT* density modification with and without NCS-averaging. (A) The plot refers to the case where the MR-search model contains only chains from W to b. It shows the distribution of RSCC-MRSAD after *PARROT* density modification with (green dots) and without (red dots) NCS-averaging. Electron density comparison was made for the part of the structure which was not used for the MR-search (chain A to V). **(B)** Classical scatter plot referring to the same case.

The lowest model completeness level at which MRSAD (after the application of density modification with NCS-averaging) still improves on MR-phases and provides an interpretable map is around 17%. This means that, by using a truncated 5LE5 model which represents ~ 1/5 only of the final proteasome structure, the combination of MRSAD-phasing and density modification is still able to significantly improve the MR solution, turning a noisy MR-map into one where at least secondary structure features can easily be recognized. Detecting such elements as α-helices and other parts of the structure in a map is important as they could significantly bootstrap subsequent model building. By *SITCOM* comparison of the substructures obtained from the different models with the sulphur sites of the reference 5LE5 structure it can be observed that, in the best cases, as much as 86% of the correct sites are found by the LLG-algorithm. In order to provide an interpretable map (*i.e.*, a map with a MPE of 50° or less), MRSAD needs, at least, ~ 260 correct sites (~ 80% of the total sites in the refined model). The requirements are somehow relaxed when density

modification is considered: in this case, the minimum number of HA-sites becomes ~ 240 when NCS-averaging is not exploited (~ 75%) and it goes down to ~ 150 when included (~ 47%). The results show how powerful the algorithm for HA detection, completion and refinement can be in finding the weak sulphur scatterers, as well as its importance for the success of MRSAD.

2.4.2.3 Tests with yeast 20*S* proteasome models

Figure 2.9 shows that, in analogy with the tests using 5LE5 models, as the size of the search-models decreases, the placement becomes more difficult. However, in this case too, MR is always able to correctly place the search-models, regardless their size. Inspection of the MR-maps shows that the electron density in the region of the model which was used for the search is relatively well-defined for both the main and the side chain residues. However, in the region that was not used as a search-model the electron density is extremely poor: in this case, already after the removal of few chains from the full model, the density for most of the main and side chain residues is missing or very poorly-defined (some of the secondary structure elements can still be recognized).

Density modification improves MR-phases until model completeness ~ 40% (at lower model completeness, density modification does not lead to any improvement). Therefore, density modification is more effective than in the case of the tests with 5LE5 models, mainly because of the structural differences between the human and the yeast proteasome.

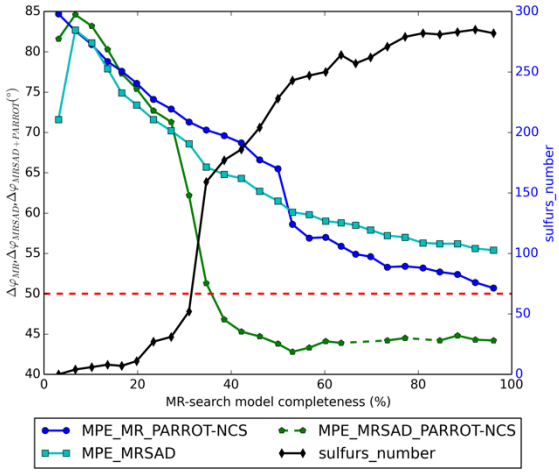


Figure 2.9: Tests with truncated models of the yeast 20*S* proteasome 5CZ4. The plots show the variation of the MPE for the MR+*PARROT*-NCS, MRSAD and MRSAD+*PARROT*-NCS solutions and of the number of sites located and refined by the LLG-MRSAD algorithm as a function of the completeness of the MR-search model.

As compared to the tests with 5LE5 models, the starting MR-maps are of lower quality. This can already be observed at high search-model completeness levels and despite the application of density modification, which improves the phases but is not sufficient to provide an easily interpretable map. As a consequence, in this case it is even more important to improve the MR-phases for later model building.

With the aim to improve the MR-phases, MRSAD was tested in a similar way as described in the previous section for the tests with 5LE5 models. MRSAD improves the MR-phases for most of the search-models and, as opposed to the tests with truncated models of 5LE5, this happens even when the completeness of the search-model is high. This is because the MR placement of a not completely accurate model is more complicated, and the gain which can be obtained by MRSAD is higher. The final MPE of the MRSAD solutions (with or without density modification) is always equal to or higher than the limit of 50° for acceptable solutions.

However, inspection of the MRSAD maps after density modification shows that their quality is still sufficient for the placement of secondary structure elements and for the identification of part of the side chains, even when the completeness of the MR-search model is relatively low (**Figure 2.10**). This shows and confirms that even maps with MPE of 50° or slightly worse can be (at least partly) traced.

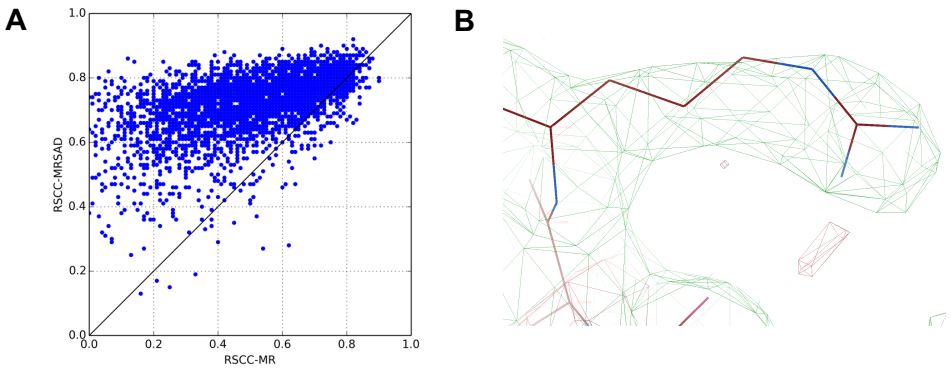


Figure 2.10: Tests with truncated models of the yeast 20*S* proteasome 5CZ4. (A) The scatter plot refers to the case where the MR-search model contains only chains from R to b, and shows the correlation between RSCC-MRSAD (after *PARROT* density modification) and RSCC-MR for the unknown part (chain A to Q). For the MR solution, overall map correlation and map correlation in region of the model are 0.352 and 0.490, respectively. For MRSAD, overall map correlation and map correlation in region of the model are 0.510 and 0.698. **(B)** Map quality comparison for the proteasome model. Comparison of the $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ electron density maps from MR-phasing (red) with the same type of map from MRSAD-phasing (green) for the same case as in **(A)** (all maps are contoured at 1.5σ level). The maps are shown for R8B (RSCC-MR = 0.23, RSCC-MRSAD = 0.75).

As for the tests with 5LE5 models, pure MRSAD-phases are only slightly better (few degrees in terms of MPE) and can be further improved by applying density modification. In the case illustrated in **Figure 2.11**, where a search model ~ 40% complete is used, the MR- and MRSAD-maps have a MPE of 71.1° and 67.7°, respectively. However, following density modification the error on the phases decreases to 62.2°, or 51.3° if NCS-averaging is performed. Again, the improvement gained by including NCS-averaging is substantial and is reflected in real-space (**Figure 2.12**).

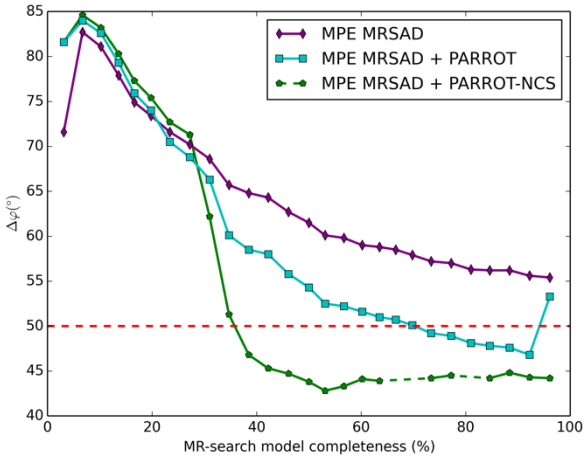


Figure 2.11: Tests with truncated models of the yeast 20*S* proteasome 5CZ4: effect of NCS-averaging. The plot shows the improvement, over a large interval of MR-search model completeness, of the MRSAD-phases when the NCS information from the refined substructure is exploited.

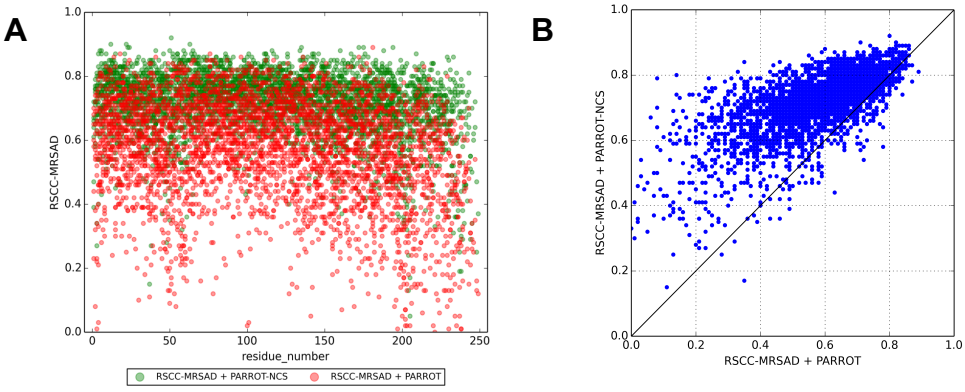


Figure 2.12: Tests with truncated models of the yeast 20*S* proteasome 5CZ4: effect, in real space, of *PARROT* density modification with and without NCS-averaging. The plot refers to the case where the MR-search model contains only chains from R to b. It shows the distribution of RSCC-MRSAD after *PARROT* density modification with (green dots) and without (red dots) NCS-averaging. Electron density comparison was made for the part of the structure which was not used for the MR-search (chain A to Q).

In this case, successful MRSAD-phasing (when combined with density modification and NCS-averaging) is possible down to ~ 30% model completeness (**Figure 2.11**) which, as expected, is higher than the limit observed when using truncated 5LE5 models. This limit does not consider the arbitrary threshold of 50°. However, as discussed above, even maps

with MPE of 50° or slightly worse can be (at least partly) traced. This shows that, by using a homologous model which represents ~ 1/3 of the target structure, and combining the phases after MR-placement with the anomalous signal from the sulphur atoms, it is still possible to improve the initial MR-phases and obtain a map which is considerably easier to interpret. In the best cases (*i.e.*: at high model completeness), the LLG-MRSAD algorithm can find up to 81% of the correct sites. At high model completeness, the number of correct sites is comparable to what is observed for the tests with truncated models of 5LE5. The number of correct sites decreases as soon as the model completeness diminishes, as well as the accuracy of their position. This is confirmed by an analysis of the distribution of the *B*-factor values of the refined substructures, which shows how they increase with decreasing completeness of the MR-models (*data not shown*). However, in this case, even at intermediate or low model completeness, a good number of sites is still found, whose coordinate accuracy is high enough so that they contribute to the improvement over MR-phases.

2.4.2.4 Determination of the minimum amount of information for successful MRSAD-phasing

The minimum amount of information required for successful MRSAD-phasing was determined with truncated ideal models and with truncated models of the homologous yeast structure (5CZ4) to simulate a possible real-case scenario. The discussion that follows is based on the results obtained from the pipeline which employs *autoSHARP*-MRSAD, but analogous results were obtained with the use of *PHASER*-MRSAD [42], [43]. The limits for successful MRSAD-phasing in *autoSHARP* when using truncated 5LE5 models are ~ 23 % and 720° in terms of the size of the MR-

search model and of the rotation range of the anomalous data (corresponding to an anomalous multiplicity, m_{ano}, of 13.5), respectively (**Figure 2.13A**). In terms of the substructure, this translates into 178 correctly determined and refined sulphur sites. In this case, MRSAD-phasing is considered successful when the MPE of the MRSAD-phases after density modification is below the arbitrary limit of 50°.

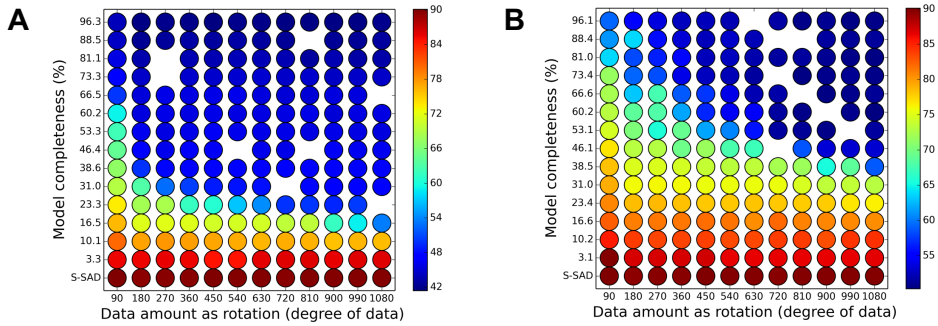


Figure 2.13: Determination of the minimum amount of data for successful MRSAD-phasing starting from truncated (A) 5LE5 and (B) 5CZ4 models. The heatmaps show the variation of the MPE of *autoSHARP*-MRSAD phases after *PARROT* density modification as a function of the completeness level of the MR-search models and of the data used for phasing. The color legend refers to the MPE. The baseline of the heatmaps is represented by S-SAD phasing, which was carried out in *PHASER* by providing the *SHELXD* substructure. Density modification was performed with *PARROT* by exploiting the NCS information present in the substructure determined and refined by *SHARP*. Few data are missing due to technical problems which prevented to run density modification on some cases.

As expected, the limits for successful *autoSHARP* MRSAD-phasing by using truncated 5CZ4 models are lower, meaning that more complete MR-models and higher anomalous multiplicity are required. The limits for successful MRSAD-phasing in *autoSHARP* when using truncated 5CZ4 models are ~ 46 % and 900° in terms of the size of the MR-search model and of the rotation range of the anomalous data (m_{ano} = 15.2), respectively (**Figure 2.13B**). In terms of number of sulphur sites, this translated into 174 correctly determined and refined sites. Comparison of the heatmaps shows immediately that, when truncated 5CZ4 models are used, the shift from

"successful" to "not-successful" cases happens more gradually. The transition is better defined when truncated 5LE5 models are used, meaning that the separation between "successful" and "not-successful" MRSAD cases is sharper. The effect of multiplicity is particularly important in these cases, where native phasing with intrinsic sulphurs is attempted. The completeness of the MR-search model plays a bigger role than data multiplicity, but in borderline cases, as shown by the heatmaps, the effect of data multiplicity on the phasing performance becomes evident.
The program *ANODE* [44] was used to determine the peak heights for the sulfur atoms (for CYS's and MET's sulfurs, separately). As expected, the peak height gradually decreases as the size of the MR-search models and the anomalous multiplicity gets lower (*data not shown*).

2.5. Conclusions

Some general conclusions can be drawn from the analysis of both the proteasome data sets and the other test systems.

As anticipated in the previous Chapter, the application of a general pipeline for MRSAD-phasing and model building showed limitations when applied to systems of higher molecular weight and data at medium-to-low resolution. Three popular map interpretation software with different characteristics were tested but none of them was able to automatically build a significant part of the proteasome into the MRSAD maps. The resolution limit imposed by the map interpretation software poses a limit to the pipeline. Therefore, the auto-tracing results cannot be used as a general metric to establish whether a protein structure has been solved or not, at least when the resolution is equal or worse than 2.8 Å. The numerical improvements after MRSAD, compared to MR, are of the order of few

degrees of the mean phase error. However, visual inspection of the MR and MRSAD-maps clearly shows larger improvements than what the MPE suggests. This is confirmed by the analysis of the RSCC distributions which shows that, after MRSAD-phasing, the electron density improves for the most of the residues. The ability of the LLG-algorithm to find the heavy-atom substructure is critical for the success of MRSAD-phasing. The comparison with the known substructure revealed that the LLG-algorithm, even with less accurate and/or small search models, representing only a limited fraction of the whole target, can still locate a good number of sites that is enough to improve the MR-phases. It was also found that not all the heavy-atom sites are required for successful MRSAD-phasing. Tests to determine the minimum amount of information for successful MRSAD-phasing confirm previous results on smaller test systems, in particular that the addition of even a weak but accurate anomalous signal can greatly improve the MR-phases and that anomalous multiplicity in MRSAD is as much important as in S-SAD. Some similarities were observed between MRSAD and experimental phasing: *i)* not all of the heavy-atom sites are required for successful phasing, *ii)* density modification is critical for the improvement of the phases and *iii)* anomalous multiplicity in MRSAD is as much important as in S-SAD and highlights the necessity to collect, whenever possible, accurate and highly redundant anomalous data.

An exhaustive investigation of the potentialities and limitations of MRSAD-phasing was carried out on the most challenging data set available, *i.e*. the native human 20*S* proteasome data at 6keV. This study allowed to draw some important and specific conclusions about MRSAD-phasing on systems of large molecular weight.

In primis, it was confirmed that density modification plays a crucial role in further improving MRSAD-phases, especially when NCS-averaging is included, and an optimal protocol to improve MRSAD-phases was found, which main advantage lies in its simplicity. In the most favorable cases, the maximum achievable gain (defined as the difference between the MPE of the MRSAD- and MR-phases) is as high as 20°. Furthermore, the maximum gain in terms of MPE is obtained in the region of low model completeness (which was already observed from the tests on small and medium size model systems). The lowest model completeness at which MRSAD (after the application of density modification with NCS-averaging) still improves on the MR-phases and provides interpretable maps was found to be as low as ~ 17 % (*i.e.* representing just ~ 1/5 only of the final proteasome structure). The application of MRSAD-phasing to the human 20*S* proteasome benefits from the high number of equivalent copies in the asymmetric unit, making NCS-averaging a powerful tool to improve the MRSAD-phases of symmetrical complexes. Based on the results collected on a number of different systems, it can be expected that MRSAD is likely to improve the MR-phases even in the cases of systems with a reduced number of symmetrical copies or without internal symmetry. However, the possible benefits coming from the application of MRSAD in systems of reduced internal symmetry would require further tests on real data. Furthermore, the analysis of additional data from other well-known and large macromolecular complexes would allow to draw more general conclusions on the applicability range of MRSAD-phasing and on the best strategies to maximize its success.

Supplementary Materials & Methods

Appendix A

Data statistics for the Br-soaked human 20*S* proteasome data set (values in parentheses are given for the highest resolution shell):

Beamline	P14, PETRA III
Detector	Pilatus 6M
Wavelength (Å)	0.91985
$d_{max} - d_{min}$ (Å)	49.51–2.50 (2.54–2.50)
Space group	$P2_12_12_1$
Unit cell parameters (Å, °)	a = 113.49, b = 202.74, c = 316.06, $\alpha = \beta = \gamma$ = 90.0
No. of reflections	3300517 (43218)
No. of unique reflections	246037 (6513)
Multiplicity	13.4 (6.6)
Completeness (%)	97.6 (52.9)
$\langle I/\sigma(I)\rangle$	12.0 (0.6)
CC(1/2) (%)	99.7 (11.9)
CC_{ano} (%)	10.9 (2.2)
$R_{meas.}$ (%)	22.7 (334)
Mosaicity (°)	0.11

Appendix B

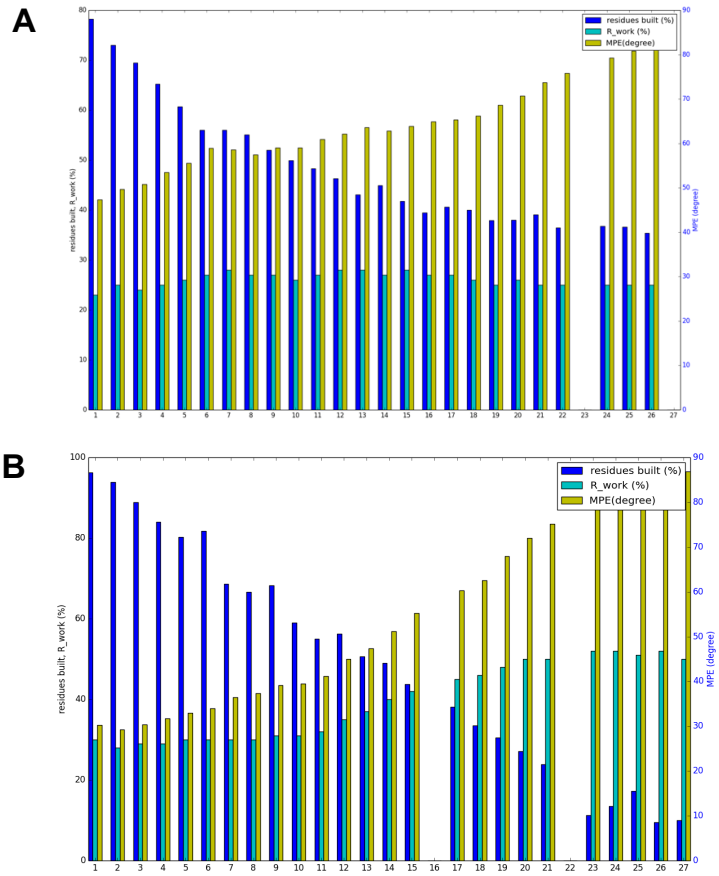


Figure 2.14: Results from (A) *ARP/wARP* and (B) *BUCCANEER* model building on gradually truncated 20*S*-Br models. Models for molecular replacement search have been obtained by gradually removing one chain at a time, starting from chain A of the refined proteasome structure: the proteasome is made up of 28 chains, which means that 27 models were generated. n denotes the number of chains removed every time. Classical *ARP/wARP* model building has been performed, providing the molecular replacement phases and figure of merit, together with the full proteasome sequence. Each of the 5 model building cycles was interspersed with 5 refinement cycles in *REFMAC5*. 10 cycles of auto-building in *BUCCANEER* have been performed each time, providing the molecular replacement phases and figure of merit, together with the full proteasome sequence.

Appendix C

Data statistics for the native human 20*S* proteasome data set (values in parentheses are given for the highest resolution shell):

Beamline	P14, PETRA III
Detector	Pilatus 6M
Wavelength (Å)	2.0664
$d_{max} - d_{min}$ (Å)	170.78–2.87 (2.92–2.87)
Space group	$P2_12_12_1$
Unit cell parameters (Å, °)	$a = 113.39$, $b = 203.25$, $c = 314.94$, $\alpha = \beta = \gamma = 90.0$
No. of reflections	6262578 (72152)
No. of unique reflections	152675 (2859)
Multiplicity	41.0 (25.2)
Completeness (%)	91.7 (35.1)
$\langle I/\sigma(I) \rangle$	41.8 (10.0)
CC(1/2) (%)	100.0 (95.8)
CC_{ano} (%)	37.9 (4.0)
$R_{meas.}$ (%)	9.8 (35.8)
Average mosaicity (°)	0.06

3. MRSAD-PHASING OF A NANOBODY COMPLEX

3.1. Summary

A nanobody-lanthanide complex has been developed in the groups of Dr. Thomas R. Schneider and Dr. Christian Löw with the aim to establish a general tool for crystallographic phasing of large macromolecular complexes. The nanobody is designed to target the protein of interest. While the nanobody should promote the crystallization, the lanthanide atom(s) would provide the anomalous signal for the phasing of the unknown target. To demonstrate the working principle of such nanobody-lanthanide complex, a complex between the Green Fluorescent Protein (GFP) and its nanobody has been engineered. Such complex was successfully crystallized and different crystallographic data sets were collected. The collected data were used to evaluate the applicability of engineered nanobodies (NBs) for crystallographic phasing: to this aim, different phasing scenarios were investigated assuming that the structure of GFP was unknown. In particular, MRSAD-phasing was tested and compared with other classical phasing methods. In addition, since the system had a space group for which several indexing schemes and origins are possible, it was also used to learn how to deal with such cases.

3.2. Introduction

Nanobodies are well known as crystallization chaperons, and they have been extensively used to assist the growth of crystals of challenging systems as membrane proteins, protein complexes and intrinsically

disordered proteins [45]–[47]. General protocols for the generation of NBs to be used as crystallization chaperones have been described [48]. For many different proteins and protein complexes of interest, NBs can be selected from artificial nanobody libraries [49], [50] or after immunization of alpacas or llamas [48]. In both cases, a nanobody specific to the protein of interest (or target protein) is selected and used to favor the crystallization process. The NBs act by stabilizing specific conformations and providing additional crystal contacts [51]–[54]. To solve the structure of several of such protein-nanobody complexes, Molecular Replacement is usually the most commonly employed phasing method [55]–[58]. In such cases, MR is performed by using the structure of the nanobody (or a similar one, or just a part of it) as a search model. However, MR can introduce model bias into the electron density maps, particularly at medium-to-low resolution, which is often the case for large macromolecular systems. The favorite way to reduce (or even avoid) the issue of model bias is by resorting to experimental phasing, which exploits the anomalous signal originated from a wide range of naturally occurring or artificially introduced anomalous scatterers such as sulphur, selenium, cluster compounds and others [59], [60]. In this context, lanthanides have been found to be very powerful phasing agents due to their large anomalous contributions (*f*'' values of ~ 30e^-) at their L_{III} adsorption edges [61], [62] (**Figure 3.1**). However, introducing one of these anomalous scatterers is not always a simple task. In fact, the derivatization of the crystals is often challenging, especially for membrane proteins and large protein complexes, as native crystals from these types of systems can be very fragile and can be easily damaged during manipulation. A long and tedious screening might be required to find the best derivatization condition(s). As a consequence, any method to

introduce anomalous scatterers into protein crystals which does not decrease their diffraction properties would be highly welcomed.

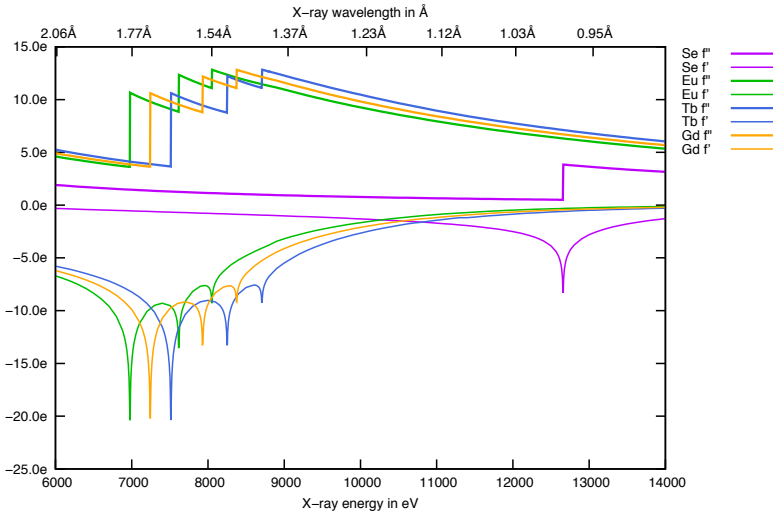


Figure 3.1: L_I-L_{III} X-ray absorption edges of terbium, gadolinium and europium in comparison with selenium *K* X-ray absorption edge. The anomalous signal of the lanthanides at their L_{III} edges (at 1.650 Å for Tb^{3+}, at 1.712 Å for Gd^{3+} and at 1.777 Å for Eu^{3+}) is significantly stronger than the anomalous signal of selenium at its *K* edge (at 0.9795 Å). The plot was generated from http://skuld.bmsc.washington.edu/scatter/.

Based on the promising properties of both nanobodies and lanthanides, the "backpack"-concept has recently been proposed by the Schneider and Löw groups and tested as a strategy to phase challenging proteins. Based on such idea, a lanthanide binding motif (LBM) is engineered into the conserved region of a nanobody fold and this construct is linked to the specific protein antigen, allowing the phasing of the nanobody-antigen complex. Ideally, the nanobody fold is engineered in such a way so that the resulting protein binds the lanthanide ion with high affinity while the complementarity-determining regions (CDRs) interacting with their target protein remain unaffected. In principle, this is a general approach because it can be applied to any nanobody selected against a particular target protein. A GFP*NB-LBM*Tb^{3+} complex has been used to test the "backpack"-principle

(**Figure 3.2**). In this case, the GFP portion acts as the structural probe to test the concept, but in a real case scenario it would be replaced by an unknown target.

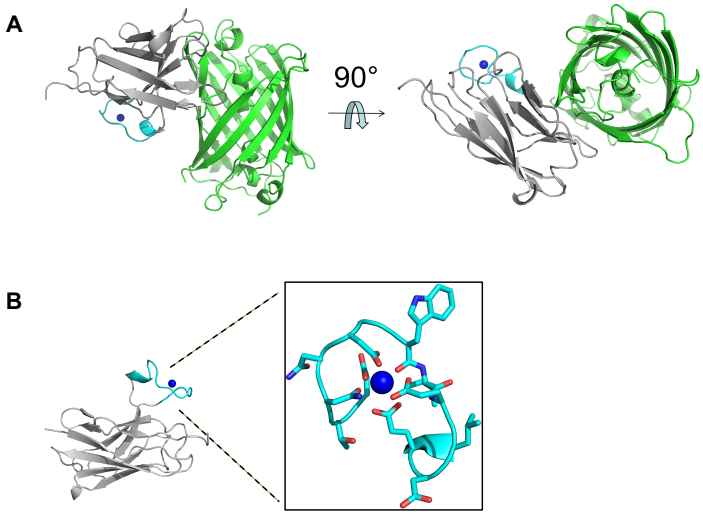


Figure 3.2: Crystallographic structure of the GFP*NB-LBM*Tb^{3+} complex used for the study. The overall structure is shown in **(A)**. GFP protein is in green, the GFP-nanobody is in grey, the LBM in cyan and the Tb^{3+} ion is shown as a blue sphere. **(B)** Magnification of the Tb^{3+} binding site.

3.3. Materials & Methods

3.3.1. Experimental part, structure determination and refinement

The nanobody-GFP complex was designed, produced, purified and crystallized by Sophie Zimmermann. Three low-dose MAD-data sets were collected at beamline P13 by Sophie Zimmermann and Guillaume Pompidor at the peak, inflection point and high-energy remote wavelengths of the terbium L_{III} edge (~1.650 Å). The structure of GFP was solved by Guillaume Pompidor with *Phenix Autosol*, followed by automatic model building with *Autobuild* [24]. Manual model improvement was carried out in *Coot* [31]. Full details about the preparation of the complex, the data

collection and structure determination/refinement can be found in the manuscript draft in the *Appendix Manuscripts* section.

3.3.2. Data processing and analysis

Before any phasing method could be tested, a preliminary issue concerning the MAD data sets had to be addressed. This is because the crystals of the nanobody-GFP complex belong to the trigonal space group $P3_121$ (152), which allows for two possible indexing schemes and two alternate origins, (0,0,0) and (0,0,½). For the purpose of phasing, all the data must be consistent in terms of indexing scheme. However, the same indexing scheme does not ensure that the data will be on the same origin, since the origin of an electron density map is deliberately assigned by any of the phasing software. As a consequence, for the comparison of different phasing scenarios (by means of the MPE, for instance), all the data not only have to be consistently indexed, but they also need to be on a common origin. First of all, the three MAD-data sets were put on a consistent indexing scheme. This was done with *XDS* [63] defining the 'reference' and the 'probe' data sets and using the keyword 'REFERENCE_DATA_SET' in the CORRECT.LP file of the 'probe'. Then, a Python-script was used to carry out the phasing tests. The script automatically performs different phasing scenarios (MR, SAD, MAD, MRSAD) and analyzes the results, ensuring the origin match between the phased data and the reference data which is a *condition sine qua non* for the calculation of the MPE. What follows is a description of the main steps performed by the script for each phasing scenario:

***Pointless* 'Match Index to Reference':** *Pointless* is used to put the phased data (*i.e.*: the data obtained after each phasing scenario) on the same

indexing scheme of the reference file. The latter contains the reference phases, meaning the phases of the final and refined nanobody-GFP model (more information on how the reference phases were calculated can be found in the *Supplementary Materials & Methods*). The data file produced at the end of the phasing is tested against the reference dataset: all the possible alternative indexing schemes are ranked based on the correlation coefficients and the data sets are put on a consistent indexing scheme. The output file will be written in the space group of the reference file.

***CAD* + *CPHASEMATCH*:** this step is required to compute the MPE. *CAD* is used to combine the data file produced at the end of the phasing (on the same indexing scheme of the reference data) with the reference file. The combined file is then given to *CPHASEMATCH* for the computation of the MPE. Two sets of phases can be compared only when they are on the same indexing scheme, which is why the 'Pointless step' is crucial for the calculation of the MPE between phased and reference data.

***phenix.get_cc_mtz_mtz*:** this *PHENIX* command reads in the data file produced at the end of the phasing (on the same indexing scheme of the reference data) and the reference data. The two data files are converted into maps and the origin of one map is adjusted so that the map superimposes on the other map. To do so, *phenix.get_cc_mtz_mtz* finds all the allowed origin shifts compatible with the space group symmetry that maximizes the correlation of the two maps. This shift is applied to the other map (and the correlation of the maps is calculated). Since one of the maps is the reference map, the map calculated from the phasing data will superpose to the reference model. *phenix.get_cc_mtz_mtz* writes out a modified version of the data file produced at the end of the phasing, shifted to match the other.

***FFT*:** this step is used to convert the shifted data file originated in the previous step by *phenix.get_cc_mtz_mtz* to a map, which is used to generate a PyMOL picture of the GFP chromophore.

Scatter plot: additionally, a scatter plot of the residue-by-residue real-space correlation coefficients (RSCC) is generated and used to check (both visually and numerically) the improvements of SAD-, MAD- or MRSAD-phases over MR-phases computed from the invariant part of the NB as search model. The scatter plot is generated for the GFP moiety only, *i.e.* the region of the molecule which was not part of the search model.

PyMOL: finally, a picture (in .png file format) of the electron density for the GFP chromophore is produced and saved.

The pipeline generates various types of outputs. The most important are, for each phasing scenario: the MPE computed against the reference phases, the picture of the electron density around the GFP chromophore and the scatter plot (not for the case of MR-phasing) for the comparison of the specific phasing method with the case where MR with the invariant part of the GFP nanobody is performed. All of them have been used to judge the success of the different phasing scenarios.

3.4. Results & Discussion

After consistent indexing of the MAD data sets, the automated pipeline was used to test different phasing scenarios and to analyze the results, which are summarized in **Table 3.1**. **Figure 3.3** shows an example of scatter plot and the comparison of the GFP chromophore electron density between an MR- and an MRSAD-phasing scenario.

Phasing scenario	MPE vs reference model/º
MR-NB-Tb	56.2
MR-NB-invariant	62.4
Tb-SAD-20dm	35.6
Tb-MAD-0dm	68.3
Tb-MAD-20dm	32.1
Tb-MAD-20dm-a3	31.8
Tb-MAD-20dm-a3-ARP-wARP	24.7
Tb-MRSAD-NB-Tb	30.3
Tb-MRSAD-NB-invariant	30.6

Table 3.1: MPE for selected phasing scenarios on the GFP*NB-LBM*Tb^{3+} data. The scenarios are described as follows: "MR-Nb-Tb": MR by using the Nb model and the Tb atom; "MR-Nb-invariant": MR by using the invariant part of the nanobody (without the Tb atom); "Tb-SAD-20dm": SAD-phasing with 20 density modification (dm) cycles; "Tb-MAD-0dm": MAD-phasing without dm; "Tb-MAD-20dm": MAD-phasing with 20dm cycles; "Tb-MAD-20dm-a3": MAD-phasing with 20dm cycles and 3 auto-tracing cycles; "Tb-MAD-20dm-a3-ARP-WARP": MAD-phasing with 20dm cycles, 3 auto-tracing cycles and *ARP/wARP* model building; "Tb-MRSAD-Nb-Tb": MRSAD with the MR-solution obtained by using the nanobody and the Tb atom as a search model; "Tb-MRSAD-Nb-invariant": MRSAD with the MR-solution obtained by using the invariant part of the nanobody as a search model.

MR was carried out by using two different search models: *(i)* the full GFP nanobody structure and *(ii)* the invariant part of it. The invariant part is a well conserved portion of the nanobody in terms of sequence and structure. As a consequence, the invariant region is a piece of information always available in a real case situation, and can therefore be used either for MR-phasing or to assist experimental phasing. This is the reason why in the scatter plots the MR solution obtained by using the invariant part of the GFP nanobody was used. As expected, the MPE of the MR solution obtained with the full GFP nanobody is lower than the MPE for the solution obtained using the invariant part of the GFP nanobody.

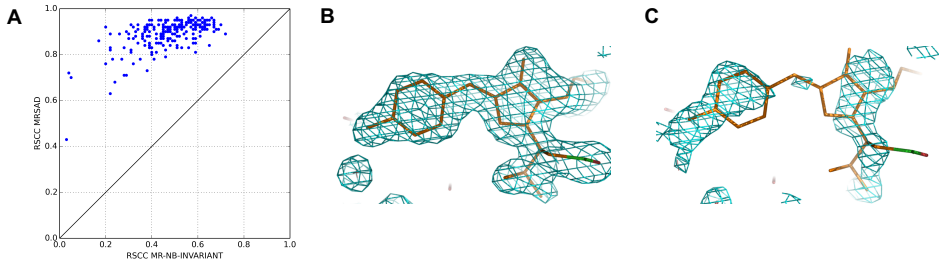


Figure 3.3: Example of scatter plot and the comparison of the chromophore electron density between an MR- and a MRSAD-phasing scenario for the GFP*NB-LBM*Tb^{3+} case. Scatter plot showing the correlation between RSCC-MRSAD-NB-invariant and RSCC-MR-NB-invariant for the GFP molecule. Overall (main and side-chain) correlation coefficients residue-by-residue have been computed with *phenix.get_cc_mtz_pdb*. **(B)** $2|\boldsymbol{F}_o|-|\boldsymbol{F}_c|$ electron density map of the GFP chromophore after MRSAD-phasing using the invariant region of the nanobody as a search model. **(C)** $2|\boldsymbol{F}_o|-|\boldsymbol{F}_c|$ electron density map of the GFP chromophore after MR-phasing using the invariant region of the nanobody as a search model. All electron density maps are contoured at 1.5σ level.

However, in both cases, MR does not allow for the structure of the GFP to be solved. SAD- or MAD-phasing (after density modification) are sufficient for this purpose, as MRSAD-phasing. The modest improvement of MRSAD- over pure SAD- phasing (~ 5° in terms of MPE) is due to the very strong anomalous signal of the terbium atom. This modest improvement might become more important and crucial in other situations. In fact, for the GFP*NB-LBM*Tb^{3+} complex, the size of the target protein is small (compared to what would be a real target) and the anomalous signal is strong, making this an ideal case. However, in a real case scenario, the molecular weight of the target would be higher and the phasing power would be lower (even when using a strong lanthanide scatterer) because of the ratio between the number of atomic scatterers and the number of protein atoms. In such a realistic scenario, MRSAD could be decisive for phasing of the unknown protein.

3.5. Conclusions

The application of the "backpack"-principle to the GFP*NB-LBM*Tb^{3+} complex demonstrates the feasibility of the concept, in particular that it is possible to engineer a nanobody linked to a lanthanide ion and to use this complex for crystallographic phasing of an unknown antigen. The energy of the terbium L_{III} edge (~ 7.5 keV) should allow for successful data collection with such nanobody-Tb complexes on many synchrotron beamlines. Moreover, the strength of the anomalous signal generated from the presence of the terbium (as from other lanthanide ions) should allow for successful experimental phasing. In case the anomalous signal is not sufficiently strong and accurate for successful experimental phasing, MRSAD could represent a viable strategy. In such cases, an MRSAD-phasing approach becomes possible by combining SAD- and MR-phasing with the nanobody as the search model. As the structure of the invariant part of the nanobody (sometimes even the structure of the entire nanobody) is a piece of information always available, MRSAD-phasing represents another concrete route to phase challenging protein structures with the "backpack"-principle. It can be predicted that the "backpack"-concept will allow the determination of more structures of challenging proteins, either by pure experimental phasing or by MRSAD-phasing. Even in those cases where experimental phasing would suffice for structure solution, MRSAD-phasing has the potential to provide more accurate crystallographic models of the antigen protein.

Supplementary Materials & Methods

Preparation of the data file and of the MR-search models

For MR, a scaled and unmerged data file was prepared from the native data set (1.663 Å) from a previous data collection.

For the preparation of the search models for MR, which was carried out in *PHASER*, the invariant region of the GFP nanobody was obtained by removing from the initial sequence all the complementarity determining regions (CDR's) and the lanthanide binding tag. As follows, the invariant regions, the CDR's and the lanthanide binding tag are marked for the GFP nanobody:

invariant region 1: 1 - 25
CDR 1: 26 - 33
invariant region 2: 34-40
lanthanide binding tag: 41 - 53
invariant region 3: 54 - 62
CDR 2: 63 - 73
invariant region 4: 74 - 107
CDR 3: 108 - 115
invariant region 5: 116 - 132

SAD- and MAD-phasing

For SAD- and MAD-phasing the high-energy remote data set was used as native data set because of the higher resolution. For SAD-phasing, the PEAK data was used as anomalous data sets. For MAD-phasing, the

PEAK, INFLECTION and HIGH-ENERGY REMOTE data were used as anomalous data sets. The unit cell parameters are extracted and read by default from the native data, but they were changed to match the ones of the reference model. This is because the comparison is made between the experimental phases and the phases of the reference model, which requires them to lie in the same unit cell.

For SAD, MAD and MRSAD, the *SHELXC/D/E* pipeline was used; in all cases, for sub-structure determination 1 Tb atom was searched at the default anomalous signal resolution (2.3 Å). MRSAD as implemented in *SHELX* was used by performing 20 cycles of density modification ('-m20'), alpha-helices search ('-q') and by refining the sub-structure obtained by *SHELXD* ('-z').

Reference model and phases

In all cases, the MPE was computed against the reference phases, which were obtained from the final and refined model of the nanobody-GFP complex. *SFALL* (*CCP4*) was used to calculate the reference phases starting from the final and refined model. This model, refined at 1.85 Å, was provided by Guillaume Pompidor and has the following cell parameters: $a = b = 69.002$, $c = 169.089$, $\alpha = ß = 90^{\circ}$, $\gamma = 120^{\circ}$.

4. STRUCTURE DETERMINATION OF THE FIRST PLANT GLUTAMATE RECEPTOR BY MRSAD-PHASING

4.1. Summary

The structure of the ligand binding domain (LBD) of plant glutamate receptors (GLR) has remained elusive for decades, despite the availability of a number of homologous structures from all the other kingdoms of life and the many attempts to characterize them from the biochemical and the structural point-of-view. A collaboration with the Structural Biology group of Profs. Martino Bolognesi and Alex Costa (Università degli Studi di Milano) has been established with the aim to obtain the first LBD structure of *A. thaliana* (*At*) GLR isoform 3.3. After significant efforts, the first model of a plant LBD has been obtained and studied. Repeated attempts to solve the structure by molecular replacement with an initially collected and problematic native data set proved unsuccessful, despite the use of rationally edited search models based on the large number of bacterial and eukaryotic GLR LBDs available structures. A data set from a crystal of selenomethionine-substituted GLR3.3 LBD was crucial to solve the phase problem. An MRSAD procedure was cr to solve the structure: approximate experimental phases, obtained by locating some of the selenium atoms and successively improved by density modification, allowed to build a preliminary partial model; the phases extracted from this model were then combined with the initial anomalous phases to produce a more accurate phase set that was used to calculate an electron density map in which a

significantly better model could be built in a semi-automated way. The 3D structure of *At*GLR3.3 obtained by this procedure (in complex with *L*-Glu, at 2.0 Å resolution) was then used as a search model in molecular replacement to obtain models of GLR3.3 LBD in complex with three more different natural ligands (Gly, *L*-Cys and *L*-Met, at resolutions of 1.6, 2.5 and 3.2 Å, respectively). The quality of these final models allowed to gain important biological insights into the function, the mechanism and the evolution of the plant glutamate receptors, compared to their bacterial and eukaryotic counterparts. Moreover, the structure of *At*GLR3.3 LBD now allows to study into more details the physiological role(s) of this and all other plant GLR isoforms, opening up what, until recently, constituted unimaginable research directions and paving the way to new important discoveries in the field of Plant Biology.

4.2. Introduction

Plant Glutamate Receptor-Like (GLR) channels are plant homologs of mammalian ionotropic Glutamate Receptors (iGluRs) [64]. iGluRs are homo- or heterotetrameric cation channels activated by various neurotransmitters (*L*-glutamate, glycine, *D*-serine) released in the synaptic space. They are extensively studied for their central role in neurotransmission, learning and memory [65]. iGluRs homologs have been identified in lower eukaryotes, invertebrates, plants and cyanobacteria, showing the existence of a large family of GLRs across all kingdoms of life.

Past studies have suggested the stoichiometry and arrangement of plant GLRs to be similar to iGluRs (**Figure 4.1**): each subunit hosts an extracellular aminoterminal domain (ATD), an extracellular ligand-binding

domain (LBD), four transmembrane helices (M1 to M4, one of which - M2 - is not fully transmembrane), and a cytoplasmic tail (CTD).

As shown in **Figure 4.1**, the bilobed LBD is made up of lobes L1 and L2 which are composed of segments S1 and S2. L1 residues are mainly contributed by segment S1 and L2 residues are mainly contributed by segment S2. The LBD has a conserved clamshell-shaped architecture and the ligand sits in a cleft between the lobes. In vertebrates, the binding of the ligand induces a variable degree of closure of the clamshell that pulls the transmembrane segments and opens the channel pore [65]. A number of crystallographic structures of soluble LBD domains, some of which at high-resolution, have become available since the 1990s [66]. More recently, cryo-EM structures of tetrameric iGluRs have been obtained and are shedding light on the different steps in the activity cycle of these receptors [67]–[69].

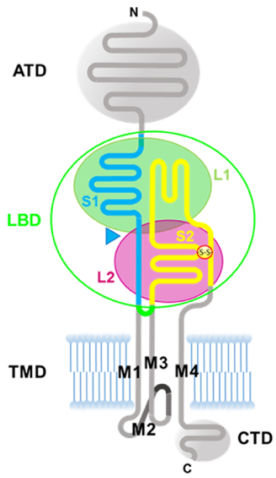


Figure 4.1: General representation of one single eukaryotic iGluR/GLR subunit. Each channel is a homo- or heterotetramer of this subunit. The green boundary encloses the *At*GLR3.3 LBD construct described in this work, with a green arch indicating the site of the linker junction. The disulfide bridge (mostly conserved in eukaryotes) ties the final stretch of S2 to the L2 core.

However, for different reasons, plant GLRs have been difficult to characterize both from the biochemical and structural point-of-view and, as a consequence, only a small wealth of information has been available until recently. The most well-studied plant GLRs are the ones from *Arabidopsis thaliana*. Since the early release of *Arabidopsis* genome sequence, it was evident that plants host a class of putative permeable channels belonging to the class of iGLRs. Specific isoforms have been implicated in a number of physiological processes, such as root growth [70], hypocotyl elongation [71], seed germination [72], long-distance wound signalling [73]–[75], pollen tube growth [76], [77], stomatal aperture [78], [79], as well as Ca^{2+} signalling [80]–[83].

The best characterized *A. thaliana* GLR isoform is 3.3 (*At*GLR3.3). This isoform has been studied for its role in amino acid-induced cytosolic Ca^{2+} increase [80], [84] and recently recognized as a key player in glutamate-mediated defence signalling [74]. These roles are of particular interest: when under stress or attack by external agents, plants react by generating warning signals which propagate rapidly, even to the most distant parts of the plant. One of such signals is represented by the flow of Ca^{2+} ions, which has been linked to the binding of *L*-Glu and other amino acids to GLR3.3. However, despite numerous studies, there is no experimental evidence that any plant GLR isoform can indeed bind glutamate or other ligands. Even a series of studies published in 2018 about the physiological role of GLR3.3 [74], [75], [77] did not provide any biochemical or structural evidence of the role of GLR3.3 as a real amino acid receptor. More so, until recently, for any studied plant GLR there was no direct proof that plant GLRs are able to bind glutamate or other amino acids.

4.3. Materials & Methods

4.3.1. Experimental part, data collection and processing

All the details about the design and cloning of the GLR3.3 LBD construct (native and SeMet-substituted), the over-expression, purification and crystallization screenings can be found in the *Appendix Manuscripts* section. The statistics for data collection, phasing and refinement can also be found in the submitted manuscript.

4.3.2. Structure determination and refinement

The determination of the structure appeared to be significantly challenging and it involved many steps. In fact, the non-optimal quality of the initial data, as well as the lack of suitable search models complicated the structure solution process and different problems had to be solved to reach this goal. What follows is a more detailed description of the structure solution process:

SAD-phasing, density modification and model building: Experimental phasing using the SeMet data set with *CRANK2* (*CCP4i2* suite) was attempted [85]. After several cycles of *BUCCANEER* [23], *REFMAC5* [34] and *PARROT* [26], a model with R_{work}/R_{free} ~ 0.41/0.47 and which contained 412 residues was obtained. Decreasing *R*-factors during model building and refinement, as well as $R_{free} < 0.50$, good electron density and accordance between the position of most of the selenomethionines and the anomalous map led to consider the model as a partial yet promising solution. Despite such good indications, visual inspection revealed regions of the model where the electron density was not good enough for the unequivocal placement of residues: in these regions, the residues were either wrongly placed or completely missing. The model was also fragmented, with two long chains but several short chains as well (mostly

alanines). As a first attempt to get a more complete and accurate model, more cycles of *BUCCANEER* were ran starting from the partial model (and by giving the positions of selenium atoms as a restraint). The best improvement of the initial model was obtained with 100 cycles of model building: the resulting model had R_{work}/R_{free} ~ 0.40/0.46 and 435 residues built into the map. However, this model was still partly incomplete and wrong. *BUSTER* [86] was used to improve the geometry and the quality of the electron density map.

MRSAD-phasing: *CRANK2* (*CCP4i2* suite) was used for MRSAD-phasing in 'rebuild mode' [87] starting from the *BUSTER* model and the SeMet-data. The final model contained 465 residues built, with R_{work}/R_{free} ~ 0.37/0.42. At this point, the model contained only two long chains and all the six SeMet residues found in the 'substructure' step agreed with the map. 10 cycles of *phenix.refine* simulated annealing [24] were used to improve the geometry. At this point, R_{work}/R_{free} ~ 0.36/0.39.

Preparation of the model for MR in the native data: The model after simulated annealing was still partly incomplete and wrong, and manual editing was required to remove parts of the model with low confidence. In particular, residues erroneously included by *BUCCANEER* during model building but not part of the sequence were removed, together with residues at the N- and C-terminus having poor electron density. The edited model was then split into two chains, which were separately used as search models for *MolRep*-MR [88] in the native data set initially processed in C2 at 1.6 Å resolution cut-off. With both search models a clear solution was obtained, but one chain gave a better solution (contrast = 22.00, wRFac =

0.593, Score = 0.471) than the other. After checking for the absence of clashes, the best MR-solution was subjected to 50 cycles of *REFMAC5* restrained refinement, giving a model with R_{work}/R_{free} ~ 0.37/0.41. Visual inspection revealed the presence of regions with little accordance between the model and the map and these regions were consequently removed from the MR refined model. Additionally, all non-mutated residues were converted to the correct ones (with the help of the sequence) and clearly wrong rotamers were adjusted. This procedure gave a fragmented, though more accurate, model composed of three chains. After *REFMAC5* refinement, this model was used for iterative rounds of model completion in *ARP/wARP* [21] and *SHELXE* [89] (**Table 4.1**). *ARP/wARP* and *SHELXE* seem to work in the opposite direction: the first builds as many residues as possible, whereas the second tries to optimize the CC, even if this requires the trimming of some residues from the original model. However, the combination of *ARP/wARP* and *SHELXE* filled in most of the gaps and yielded a model less fragmented, more accurate and more complete. The output model contained 214 residues distributed over two chains, with R_{work}/R_{free} ~ 0.33/0.38. Missing residues belonged to an internal region of the protein, but exposed to a solvent channel, and from the N- and C-terminus. This model was then subjected to several cycles of refinement through *phenix.refine*, using default parameters and including waters, until a plateau in R_{work} and R_{free} was reached (~ 0.26/0.31). Despite the difficulty in decreasing the *R*-factors (higher than what is expected for a structure at 1.6 Å resolution), the model was more than 95% complete, had a good geometry and a well-defined electron density for all the residues.

Parameters	*ARP/wARP* 1	*ARP/wARP* 2	*SHELXE* 1*	*ARP/wARP* 3	*SHELXE* 2*	*ARP/wARP* 4
R_{work}	0.3343	0.3334	-	0.3204	-	0.3345
R_{free}	0.380	0.366	-	0.373	-	0.382
No. of chains	3	3	7	2	6	2
No. of residues built	188	197	187	214	180	214
CC(best)	-	-	23.57	-	23.79	-

Table 4.1: Progress of the model through iterative rounds of *ARP/wARP* and *SHELXE* used to expand the initial GLR model. Some optimization of the *SHELXE* parameters was necessary. Final *SHELXE* arguments were: `xx.pda –s0.51 –m50 –a50 –q –o` ('`–o`' was used to prune the initial model before density modification by eliminating individual residues to optimize the *CC* for the model against the native data). The high *CC*-*SHELXE* values confirmed that a good MR-solution was obtained (for resolution better than 2.5 Å, values above 25% typically indicate a solved structure [89]). After the first and the second *ARP/wARP* cycles, *PARROT* was used for density modification on the *ARP/wARP* map. *ARP/wARP* after the second *SHELXE* run did not improve the model and the model from the fourth cycle of *ARP/wARP* was therefore taken and used in the subsequent steps.

The stable, higher-than-expected *R*-factors suggested problems in the initial assignment of the space group, with the resolution cut-off and/or with the presence of twinning and/or tNCS in the data. It was decided to re-process the native data in P1, and a little improvement in the statistics was noticed. Even if twinning and tNCS were not detected by any program, since they can be easily masked (and tests are not completely reliable), especially with not accurate data, it was decided to re-solve the structure in P1. Solving the structure in P1 temporarily removed the problem of space group assignment without posing significant difficulties because of the presence of only two molecules in the primitive unit cell and the sufficiently high completeness of the data. Moreover, re-analysis of the native data suggested that a more conservative cut-off would be more appropriate. The most complete model (in C2) was used for *MolRep*-MR in the P1 data set truncated at 2.0 Å. The MR-solution (contrast = 32.62, wRFac = 0.579, Score = 0.424) had no clashes with symmetry-related molecules and appeared as a compact dimer with clear 2-fold NCS.

Identification of the ligands and refinement of the model in P1 native data: The model was refined against native data by iterative rounds of *REFMAC5* restrained refinement, *phenix.refine* and manual editing in *Coot* [31]. During refinement, additional positive density observed in both cavities in the $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ and $|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ electron density maps allowed to unambiguously identify the *L*-Glu ligand. The presence of the ligands was confirmed by bias-reduced simulated-annealing OMIT maps generated through the *PHENIX* suite; water molecules were added with *ARP/wARP* (Solvent module) and the final stereochemistry was assessed by *MolProbity* (http://molprobity.biochem.duke.edu/) [90]. Molecular replacement with *MolRep* using the ligand-deprived *L*-Glu structure allowed to obtain the Gly, *L*-Cys and *L*-Met structures. These additional structures were also refined in a similar way as already described. The atomic coordinates and experimental structure factors were deposited in the Protein Data Bank with accession codes 6R85 (GLR3.3 LBD + *L*-Glu), 6R88 (+ Gly), 6R89 (+ *L*-Cys) and 6R8A (+ *L*-Met).

4.4. Results & Discussion

4.4.1. Overall structures of GLR3.3 LBD

All individual chains from the four crystal structures display an excellent structural match in their Cα traces (max r.m.s.d. 0.52 Å); the only significant difference is confined to the C-terminal stretch Lys240-Thr244 (including Cys243, which forms a disulfide bridge with Cys179), whose density has two alternative traces in four of the monomers and is absent in the rest. The GLR3.3 LBD has a bilobed structure which resembles that of prokaryotic and eukaryotic LBDs described in the literature (**Figure 4.2**). The structure has approximate dimensions of 60 x 40 x 40 Å.

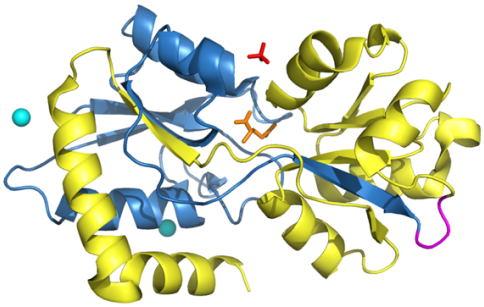


Figure 4.2: Overall structure of *At*GLR3.3 LBD + *L*-Glu. The structure is shown in ribbon representation and colored to highlight the contributions of segments S1 and S2 to lobes L1 and L2. The S1 segment is blue, the S2 segment is yellow, the linker is magenta, *L*-glutamate is in green sticks, the sulfate ion in red stick and the two Na^+ ions are shown as cyan spheres.

The L1 lobe is made up of six α-helices and two β-strands, whereas the L2 lobe is built up by a central five-stranded β-sheet surrounded by five α-helices. The two lobes are connected by a double-stranded hinge and separated by a deep cleft where the binding pocket is located. The DALI server [91] (http://ekhidna2.biocenter.helsinki.fi/dali/) identified the LBDs from a group of vertebrate iGluRs of the kainate subtype (representative PDB ID: 1sd3, r.m.s.d. 2.4 Å) and the rotifer *Adineta vaga* GLR (*Av*GluR, PDB ID: 4io2, r.m.s.d. 2.5 Å) as the most structurally similar PDB entries.

4.4.2. *Post-facto* data analysis

4.4.2.1 *A posteriori* explanation of the failure of the initial attempts to structure solution

Several attempts were made before the structure could eventually be solved:

MR using homologous structures from PDB

The first attempts were made using MR and employing homologous structures identified through standard *BLAST* alignment [92] against the Protein Data Bank. Such known structures of glutamate receptors were used as MR-search models in *PHASER* and *MolRep*, with and without

modifications, according to Schwarzenbacher *et al*. 2004 [93] (*e.g.:* generation of poly-alanine models, removal of solvent-exposed loops, use of either of the two lobes, use of only those residues aligning to the target sequence *etc*...). Screening of the MR solutions was made based on the following indicators: TFZ equivalent and LLG-score for *PHASER* [94], contrast, wRFac and Score for *MolRep*, packing (to verify the absence of clashes) and R_{free} values before and after *REFMAC5* restrained refinement. None of the original or edited models gave a clear MR solution based on these indicators; however, some "promising" MR-models were subjected to cycles of *REFMAC5* refinement followed by *PARROT* density modification and model building (in *ARP/wARP* or *phenix.autobuild*). This strategy just confirmed the bad quality of the original MR-models. Iterating between MR and model building, with the aim of gradually improving the models, did not work too.

Because the *BLAST* search performs only a rough alignment, more sophisticated ways of finding homologous structures were employed, in particular *PSI-BLAST* and *FFAS* [95]. *PSI-BLAST* identified several candidates, among which two with a sequence identity of 30 and 34%, higher than the ones initially found through standard *BLAST* search (where seq. id. varies between 15 and 26%). *FFAS,* too, identified several models, more similar to the target sequence than the ones from the initial *BLAST* search. All these models were screened by performing MR followed by *REFMAC5* restrained refinement and the results were evaluated on the basis of the abovementioned MR-parameters, the presence of clashes and the R_{free} values, but none of the initial models proved to be good.

Because PDB models proved not good enough on their own, ensembles were created by using several programs: the 'align' and 'super' commands in *PyMOL*, *Superpose* and *Gesamt* (CCP4), *phenix.ensembler* and *phenix.sculpt_ensemble* (*PHENIX*) [96]. The basic principle is to apply different *B*-factors weights to the models in order to weigh down unreliable parts and, at the same time, to weigh up those regions which appear to be more conserved. Ensembles of models derived from standard *BLAST* search, *PSI-BLAST* and *FFAS* were created and used as search models for MR, with no significantly better results. *RAPIDO* [97] was used, too, in order to perform a 3D alignment of the different models to find groups of atoms which behave as rigid bodies. Such rigid body groups were then used for MR. However, *RAPIDO* generation of models and their subsequent use for MR did not give significantly better results.

Because the use of the available PDB models proved not successful, different homology modeling programs were used to predict the structure based on the knowledge of the sequence alone. The following servers were employed: *Phyre* [98], *i-TASSER* [99], *MODELLER* [100], *ProtMod* (http://ffas.godziklab.org/protmod/doc.html) and *ROSETTA* [101]. *Phyre* and *i-TASSER* were used to generate models for the full structure as well as for the two individual domains. These homology models (in all the cases predicted with a high level of confidence) were used as MR-search models, with and without modifications as described above, but no promising results were obtained.

A series of other programs, such as *BALBES* [102] *etc*...were employed but proved unsuccessful.

Structural comparison of the refined *At*GLR3.3 LBD model with the closest homologs and with the models obtained through homology modelling shows that *At*GLR3.3 LBD reflects the topological arrangement of known LBDs with a substantial displacement in the Cα trace (**Figure 4.3**). In fact, all the ~ 70 models which were tested in MR (either derived from the PDB or generated through homology modeling) display a Cα r.m.s.d. in the so-called 'twilight zone', or worse. As it has been observed in a vast number of cases, search-models having a sequence identity below 25-30% or, equivalently, a Cα r.m.s.d. higher than 1.5-2.0 Å, are unlikely to work in MR. The Cα r.m.s.d. is more pronounced in the L2 than in the L1 lobe and it is associated with the presence of a number of flexible regions. Despite the overall fold being conserved among plant and all known LBDs, these local differences (in some cases very small while, in others, quite significant) provide a possible *a posteriori* explanation for the failures of the MR-phasing attempts. Model editing generally aims at improving the accuracy at the expenses of completeness. In this case, model editing probably failed to remove all the structurally different regions and/or because the completeness of the edited models was not high enough to produce a significant signal in MR.

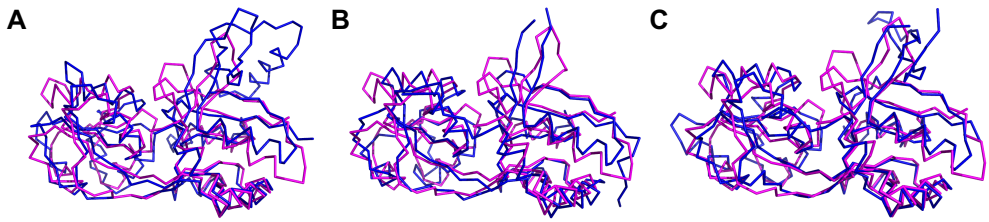


Figure 4.3: Superposition of *At*GLR3.3 LBD model onto three different representative glutamate receptor structures. These are: **(A)** NR3 subtype glutamate receptor from *Rattus norvegicus* (PDB ID: 2RC7, Cα r.m.s.d. 1.94 Å) **(B)** glutamate receptor from *Adineta vaga* (PDB ID: 4IO2, Cα r.m.s.d. 2.22 Å) and **(C)** human structure of a glutamate receptor (PDB ID: 5H8F, Cα r.m.s.d. 1.76 Å). *At*GLR3.3 LBD model in magenta; 2RC7, 4IO2 and 5H8F in blue.

Experimental phasing and MRSAD approaches

The SeMet-data set was first used in the *SHELXC*/*D*/*E*-experimental phasing pipeline [25]. As revealed by *SITCOM* comparison with the substructure in the final model, *SHELXD* was able to find the correct substructure. However, auto-tracing failed to provide a sensible poly-alanine model. Automated structure solution pipelines as *AutoRickshaw* [103] and *autoSHARP* [43] were used for both SAD- and MRSAD-phasing. In the latter case, several MR models generated during the first attempts of solving the structure *via* MR were used. As for *SHELX*, the substructure could easily be found, but none of the combinations of density modification and model building was capable of building a sensible model. *PHASER*- and *SHELXE*-MRSAD were also tried, but no sensible models could be built into the MRSAD-phases. In fact, solving the substructure is a necessary but not a sufficient condition to solve a protein structure [35].

4.4.2.2 *Post-facto* analysis of the raw native data

The program *Zanuda* [104] was used with the aim to assign, *a posteriori*, the correct space group. *Zanuda* performs refinement of the data in the space groups which are compatible with the observed unit cell parameters and, in this way, aims at identifying the most probable space group. The refined model and map obtained from refinement in P1 data were used for refinement in *Zanuda* in both P1 and C2. The best refinement was achieved in C2, which suggested this as the correct symmetry. However, the result from *Zanuda* should not be trusted when dealing with inaccurate data because twinning and tNCS can mask the correct space group. A more in-depth analysis of the raw native data was carried out to identify the presence of pathologies not identified during the initial data processing. *XDS* outputs revealed that only a small fraction of the reflections could be predicted. As

a consequence, the initial processing essentially failed. In addition, the diffraction images showed that the crystal was problematic, as suggested by high-resolution diffraction in lines rather than spots, which is indicative of faulty packing at least in one direction (which means possible anisotropy). A possible explanation for the small fraction of predicted reflections could be twinning. *XDS* does not support integration of multiple lattices, but more recent processing software as *DIALS* [105] do, so the native data set was re-processed in *DIALS* with particular attention to the presence of more than one lattice. Indeed, this showed the unequivocal presence of two lattices (**Figure 4.4**). A third is less certain and others are the tails on the streaks rather than lattices. The two major lattices heavily overlap at high resolution. Integrating the data with two or three lattices and performing a refinement against these data, however, did not significantly improved the model statistics.

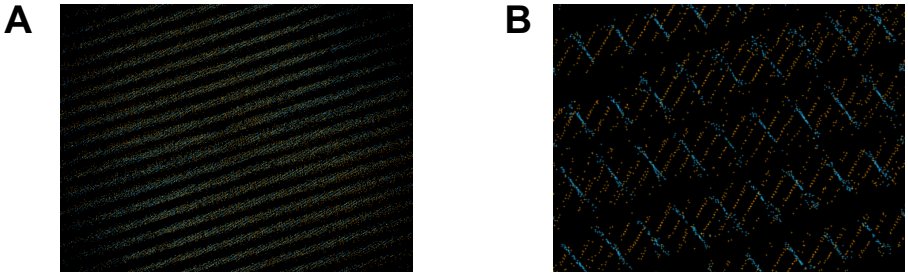


Figure 4.4: Analysis of the diffraction images of the native data set used to solve the structure of *At*GLR3.3 LBD. View of the indexed spots from *dials.reciprocal_lattice_viever*. Multiple lattices are visible as a set of two intersecting lattices. Reflections identified as belonging to distinct lattices are colored differently. **(B)** Closer view of the indexed spots.

4.5. Conclusions

The first three-dimensional and high-resolution atomic model of any plant GLR was obtained. In particular, four different structures of the *At*GLR3.3 LBD with four representative aminoacidic ligands (*L*-glutamate, glycine, *L*-cysteine, *L*-methionine) were solved and refined. Solving the structure represented a considerable challenge because of: *(i)* the sub-optimal quality of the initially available native data set (twinning), *(ii)* the sub-optimal quality of the initially available anomalous data set (weak anomalous signal) and *(iii)* the structural differences between the plant model and all the available GLR homologous structures, which were enough to prevent structure solution by molecular replacement. The phase problem could eventually be solved only by resorting to a new MRSAD algorithm on a far-from-optimal model obtained by SAD-phasing, showing that MRSAD could equally well start from a model obtained through experimental phasing (not necessarily from a MR-model). The MRSAD-phases were not magically correct (up to the point that they did not allow automatic model building) but they were nevertheless sufficiently improved such that manual rebuilding could begin. Building and refining the first structure required significant efforts, and shows the importance that crystallographers still have in solving difficult cases, even in the era of powerful software and resources. Despite the efforts required to solve the structure, the quality of the final models allowed the identification of the natural ligands and of the key residues responsible for the amino acid binding. From the biological perspective, such high-quality models represent a rich source of information. In fact, the plant model confirms the initial hypothesis about the structural similarity and mechanism of action between GLRs and iGluRs and shows, for the first time, that GLRs really

bind glutamate and other small molecules. Furthermore, obtaining the first plant GLR3.3 model virtually gives access to the structure of all the other isoforms (~20 in *A. thaliana*). Such structural knowledge, that adds to the collection of bacterial and animal LBD structures available, on one hand provides a perspective view on the evolution of these ancestral proteins along the plant lineage and, on the other hand, allows to engineer all plant GLR isoforms with the aim to get a deeper understanding of their basic physiology.

5. STRUCTURE DETERMINATION OF *Ses i 2*, THE MAJOR PROTEIN ALLERGEN OF *Sesamum indicum* SEEDS

5.1. Summary

Several studies indicate that food allergies are on the rise worldwide. Of particular importance are the food plant allergies and, among them, the ones caused by the *Sesamum indicum* plant (particularly its seeds), because of the severity of the immune reaction it can elicit in sensitive subjects. Despite these facts, few studies have focused and succeeded on the identification of the major allergenic proteins of the sesame seeds and on their structural characterization. The major allergenic protein in the sesame seeds is *Ses i 2*; its biochemical properties have been recently described but the structure has remained elusive for several years. A collaboration with the groups of Prof. Giuseppe Zanotti and Prof. Vincenzo De Filippis (Università degli Studi di Padova) has been established with the aim to obtain the first structure of *Ses i 2*. After several failed MR-phasing attempts using a number of PDB homologous as well as models obtained by various homology modeling software, the structure of *Ses i 2* was eventually solved by *ab-initio* phasing at 2.0 Å resolution. The initial partial solution, in which few helical fragments were correctly placed, was expanded by iterative cycles of density modification and auto-tracing. The refined model of *Ses i 2* confirms the previous hypothesis about its structure and, more importantly, will allow to study the molecular basis of allergenicity, *i.e.* the structural and conformational characteristics of food

allergens that favor the development of an immune response. As *Ses i 2* is a general model for the study of allergenicity, the information that will be obtained by studying its structure will permit to gain insights into the mechanism of allergic reactions at the molecular level.

5.2. Introduction

5.2.1. *Ses i 2* protein

Food allergy is an abnormal response to a food and it is triggered by the body's immune system. There are several types of immune responses to food, mainly IgE-mediated allergy and IgE-non mediated allergy. IgE-mediated allergy happens when the body produces a specific type of antibody called immunoglobulin E (IgE). The immune response is caused by the binding of IgE to specific molecules present in the food. The response may be mild or severe: in the latter case, a life-threatening reaction called anaphylaxis can occur.

Epidemiological data indicate that food allergy likely affects nearly 5% of adults and 8% of children, with growing evidence of a worldwide increase in prevalence (National Institute of Allergy and Infectious Diseases, July 2012, "Food Allergy An Overview"). A significant number of food allergies is caused by allergens present in the edible plants under various forms (seeds, nuts, beans *etc*…).

The majority of plant food allergens can be classified into families and superfamilies on the basis of their structural and functional properties. One of the most widespread groups of plant proteins is represented by the prolamin superfamily, which includes several important types of allergens found in legumes, tree nuts, cereals, fruits and vegetables, such as proteins

of the '2S albumin' family, the nonspecific lipid transfer proteins, and the cereal α- amylase and protease inhibitors [106].

The existence of the prolamin superfamily was first proposed by Kreis and collaborators in 1989 [107]. This superfamily comprises the important '2S albumin' family. 2S albumins are a major group of seed storage proteins widely distributed in both mono- and di-cotyledonous plants. As storage proteins, they are deposited in the developing seeds and are utilized by the plant as a source of nutrients during subsequent growth steps, but they have also been shown to have other physiological roles. The existence of the superfamily has been proposed based on visual comparisons of amino acid sequences which showed a conserved skeleton of eight cysteine residues, all involved in disulfide bridges, and a similar 3D structure enriched in α-helices. In particular, 2S albumins adopt a common and compact 3D structural scaffold comprising a bundle of five α-helices displayed in different regions and a C-terminal loop folded in a right-handed superhelix stabilized by four conserved disulfide bonds (**Figure 5.1**). The pattern of cysteines appears to be necessary for the maintenance of the tertiary structure. Connecting the α-helices III and IV, there is an exposed and relatively short segment known as "hypervariable region". In addition to the global folding, other structural and biochemical properties are shared by this protein superfamily and have been shown to be implicated in the intrinsic allergenicity of some of their members, including the 2S albumins. In fact, in recent years, some members of this protein family have been described as major food allergens [106]. Many 2S albumins have been classified as major allergens in plant food species, including sesame seeds.

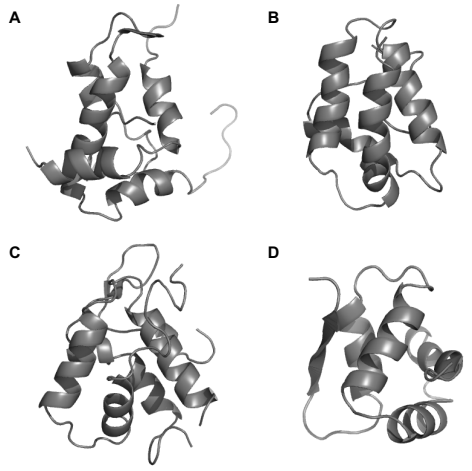


Figure 5.1: Representations highlighting the secondary structure of four members of the prolamin superfamily. (A) rapeseed 2S albumin, PDB ID: 1PNB **(B)** barley Lipid Transfer Protein, PDB ID: 1LIP **(C)** Wheat α-amilase inhibitor, PDB ID: 1HSS **(D)** Soybean hydrophobic seed protein, PDB ID: 1HYP.

Sesame is associated with immunoglobulin E (IgE) mediated food allergy. *Sesamum indicum* is a plant originally from tropical Africa, which is now universally cultivated for its seeds. It is the most important species in the *Sesamum* genus and its annual worldwide production is around 2 million tons. The seeds are used in several food products in different cuisines. Over the past few years the number and severity of reactions to dietary sesame has increased, probably because of the growing use of sesame seeds and sesame oil. In some countries sesame is one of the major causes of food allergy. For example, sesame allergy is common in Eastern countries like Israel, where it is the third most common cause of IgE-mediated food allergy, and is becoming frequent in European countries, too. Sesame seeds represent a potent food allergen and the various allergens are often associated with particularly severe reactions with a high risk of anaphylaxis. Despite the importance of the allergy to sesame, its allergens were only recently identified. In particular, little was still known until

recently about the major allergen of sesame, *Ses i 2*, as testified by the scarce literature on it [108]–[110]. So far, a thorough biochemical and structural characterization has been missing. What is of prominent importance but still lacking is the key information on the structural characteristic of food allergens that favor the development of an immune response. *Ses i 2* represents a suitable model to investigate the structural features that determine the allergenicity of food antigens, which makes this allergen even more important. As a consequence, the structural investigation of the role of *Ses i 2* in allergenicity by means of X-ray Crystallography is important in that it would lead to a better understanding of the molecular basis for allergenicity.

5.2.2. The *ARCIMBOLDO_LITE ab-initio* phasing principle

Compared to MR and experimental phasing, a*b-initio* phasing relies only on the native intensities and does not resort to experimental phase information or previous particular structural knowledge.

ARCIMBOLDO [111] is a program for *ab-initio* phasing of macromolecular structures. It combines the location of model fragments with *PHASER* [94] and density modification [112] and main chain auto-tracing [25], [89], [113] with *SHELXE*. The software receives its name from the Italian painter Giuseppe Arcimboldo, who used to compose portraits utilizing vegetables. Out of the many possible arrangements of such vegetables, only one will truly produce a portrait. In a similar way, only one of all possible placements with small protein fragments will be correct and will allow to get the full 'portrait' of the protein.

Due to the difficulties in discriminating correct but small substructures, many possible fragment locations have to be tested in parallel. Recently,

thanks to the description and the study of the expected value of the LLG [114], [115], the estimation of the difficulty of a problem (given a particular model) has provided an inestimable source of information to guide fragment-based MR. Taken together with the improvements in the Maximum Likelihood MR targets, for some cases the computing requirements have been relaxed. However, on the edge of difficult cases, massive computing power is still required.

Beyond helices, other search fragments can be exploited in an analogous way: libraries of helices with modelled side chains, strands, predictable fragments such as DNA-binding folds [116], fragments selected from distant homologs [117], [118] or libraries of small local folds that are used to enforce nonspecific tertiary structure [119].

5.3. Materials & Methods

5.3.1. Experimental part, data collection and processing

All the details about the extraction and purification of the *Ses i 2* protein from *Sesamum indicum* seeds, as well as protein extraction, purification and crystallization steps can be found in the *Appendix Manuscripts* section. The statistics for data collection, phasing and refinement can also be found in there.

5.3.2. Structure determination and refinement

What follows is a detailed description of the structure solution process:

***ARCIMBOLDO_LITE ab-initio* phasing:** After several failed trials, a final attempt at structure solution was made with *ARCIMBOLDO*. Two main factors inspired optimism: *(i)* the recent extension of the resolution for *ab-initio* phasing up to 2.0 Å and *(ii)* the predicted high content of α-

helices (between 60 and 70 %) suggested by different secondary structure prediction software. After several tests in *ARCIMBOLDO_LITE* by using different input parameters (especially 'fragment_to_search' and 'helix_length'), a promising partial solution was obtained having *CC-SHELXE* ~ 33% [111], [120]. The configuration file (*.bor* format) used to obtain the solution is shown below:

```
[CONNECTION]
distribute_computing: multiprocessing
working_directory= /path/to/sesi.bor

[LOCAL]
path_local_phaser: /path/to/phenix.phaser
path_local_shelxe: /path/to/shelxe

[GENERAL]:
working_directory= /path/to/working_directory/
mtz_path: %(working_directory)s/sesi.mtz
hkl_path: %(working_directory)s/sesi.hkl

[ARCIMBOLDO]
name_job: sesi
molecular_weight: 12453.6
f_label= F_New
sigf_label= SIGF_New
number_of_component: 3
fragment_to_search: 6
helix_length: 15
```

Because of the computational cost, a multiprocessor machine (version: Intel(R) Xeon(R) @ 2.30GHz, with 72 cores and 256 GB RAM) was used to run all *ARCIMBOLDO* jobs. The solution was found within an hour (real time, not CPU time). Such solution was obtained after *SHELXE* expansion and confirmed by the high CC (for resolution better than 2.5 Å, values above 25% typically indicate a solved structure [89], [113]). Visual inspection of the map and of the polyalanine model, consisting of 181 residues, further confirmed the quality of the partial solution. The structure was solved only after the placement of the sixth helical fragment. *SHELXE*

expansion used 15 cycles of density modification interspersed with 8 cycles of auto-tracing. The density sharpening parameter ('-v') was set to 0 and missing reflections were extrapolated up to 1.5 Å. In *PHASER*, the expected r.m.s.d. of the coordinates to the target structure was kept to the default value of 0.2 Å. The resolution for the rotation search, translation search and rigid body refinement were restricted to 1.0 Å. For every rotation or translation search, peaks under 75% of top were rejected. A packing filtering was applied which allowed solutions with significant clashes to be discarded. After the packing check, surviving solutions were subject to rigid body refinement and pruning of duplicates. As can be seen from the configuration file, the successful MR-search was started by using six 15-residues long α-helical polyalanine fragments as initial seeds. The initial guess on the number and the length of the helical fragments was made based on the different homology models predicted with high confidence.

Extension of *ARCIMBOLDO_LITE* partial solution: Because initial solvent content analysis revealed three molecules/ASU, the polyAla trace was still very partial and needed to be extended. After several cycles of model building in *PHENIX-AutoBuild*, manual editing of the model in *Coot* and restrained refinement in *REFMAC5*, a model with R_{work}/R_{free} ~ 0.31/0.37 and containing 240 residues was obtained, in which the electron density was very well defined for chain A, well defined for chain B and less defined for chain C (some regions showing poor or completely absent electron density). After several cycles aimed at improving the model, new density started to appear, which revealed the presence of a fourth molecule. The additional molecule was confirmed after MR (in *PHASER*) by fixing

the coordinates of the trimer and searching with a copy of chain A (LLG = 2479 and TFZ equivalent = 25.4).

Model completion and refinement: After *REFMAC5* refinement of the MR-solution, R_{work}/R_{free} decreased to ~ 0.28/0.32. A single iteration of *PHENIX-AutoBuild* and *REFMAC5* restrained refinement gave a model consisting of 328 residues and R_{work}/R_{free} ~ 0.24/0.27. Serious refinement was started on this model, and it was carried out by iterative cycles of *phenix.refine*, density modification, manual editing in *Coot* and restrained refinement in *REFMAC5*. Density modification (in *PARROT* [26] or *RESOLVE* [121], depending on the case) were important as they improved the electron density and revealed new residues. This procedure allowed to improve the geometry of some residues and to place new ones, previously missing, in the model. At the end of this procedure, the model had R_{work}/R_{free} ~ 0.21/0.25 and contained 20 additional residues. Last refinement stages aimed at improving the poor (sometimes absent) electron density in some traits of chains C and D. Because the electron density in these regions could not be improved, the refinement was ended at this stage.

5.4. Results & Discussion

5.4.1. Overall structure of *Ses i 2*

The final model of *Ses i 2* contains four chains in the asymmetric unit, labelled A, B, C and D. The electron density is not of the same quality for all chains (**Figure 5.2**). Chains A and B are very well defined for residue 19-115, with the exception of residues 43-49. This is because this trait of the protein undergoes proteolytic cleavage, probably in the Golgi. In chain C, the electron density is not visible for residues 89-100 and, in chain D,

for residues 87-103, indicating that these areas are disordered in the crystal, perhaps owing to the small amount of detergent used in the crystallization procedure. In fact, the analysis of the *B*-factors distribution for the different molecules shows the presence of regions, in chains C and D, with significantly higher *B*-factors (**Figure 5.3**).

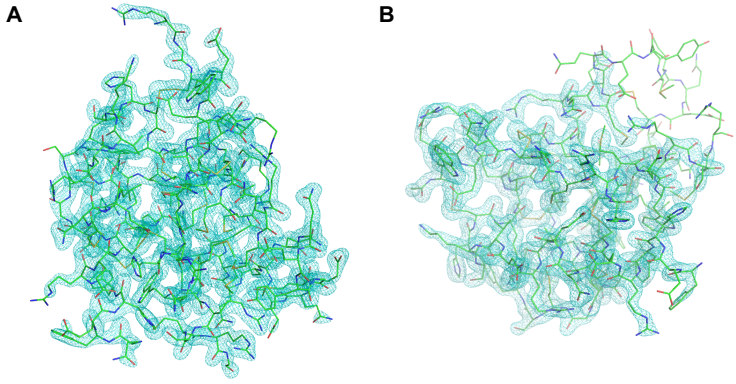


Figure 5.2: Quality of the electron density map for different chains of the *Ses i 2* model. The $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ electron density maps contoured at 1.5σ are shown for **(A)** chain A and **(B)** chain C. For chain C, the electron density is missing for residues 89-100. Missing electron density suggests that this area is disordered in the crystal, probably due to the detergent that was present as an additive in the crystallization procedure.

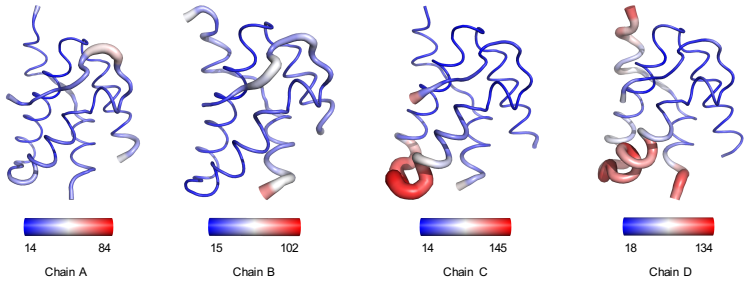


Figure 5.3: *B*-factors distribution for the different molecules in *Ses i 2*. The chains are shown in cartoon putty representation and the B-factor coloring is such that regions with low *B*-factors are colored in blue, whereas areas of high *B*-factors are in red (intermediate values are white).

The Cα r.m.s.d. between different pairs of chains ranges from 0.77Å (superposition of A to B) to 1.01Å (superposition of A to C), making the four chains virtually identical. In the rest of the discussion, the coordinates

of chain A, those better defined, will be used for the description of the molecular structure. The *Ses i 2* structure consists of a globular five-helix motif arranged in a right-handed superhelix with a simple "up and down" topology (**Figure 5.4**), relatively similar, despite significant differences, to that observed in other in 2S albumins and prolamins [122], [123].

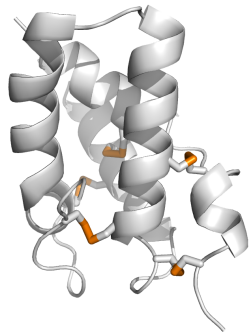


Figure 5.4: Cartoon representation of chain A of *Ses i 2* model with the five disulfide bridges highlighted in orange.

The structure starts with a very short helix (α-helix A) connected by a three-residues stretch to α-helix B (residues 31-41). α-helix C runs from residue 58 to 71, α-helix D from 78 to 92 and the last α-helix, E, from 97 to 112. The overall structure is stabilized by five disulfide bridges. The compact core is very hydrophobic, characterized by the presence of 11 methionine residues (over a total of 13), Trp36 and three leucine residues. All charged residues are exposed on the protein surface, making the surface heavily charged, with prevalence of positively charged residues (13 Lys and Arg, 3 His) compared to 11 negatively charged (Glu and Asp). Of note, the abundance of Cys (10.6%), Arg (13.8%), Gln (17%) and Met (16%) in *Ses i 2* that is much higher than that normally observed in natural proteins: Cys (1.4%), Arg (5.4%), Gln (3.9%) and Met (2.4%) (http//:expasy.org/sprot/relnotes/relstat.html).

5.4.2. *Post-facto* data analysis

5.4.2.1 *A posteriori* explanation of the failure of the initial attempts to structure solution

Several attempts, all based on MR using homologues structures from the PDB or other sources, were made at the beginning. The first MR attempts were made employing: *(i)* homologous structures available in the PDB, identified through standard *BLAST* and *PSI-BLAST* alignment against the Protein Data Bank, *(ii)* homologous structures identified through more sophisticated methods (*FFAS* and *HHpred*) and *(iii)* models derived from homology modelling (*ROSETTA* [101], *Phyre2* [98], *SWISS-MODEL* [124], *MODELLER* [100], *i-TASSER* [99], *QUARK* [125]). Despite some structures were identified which displayed relatively high sequence identities (up to 33% for *PSI-BLAST* and *FFAS*), none of them gave a clear MR solution based on standard indicators. The homology models (in all the cases predicted with a high confidence level) were used as MR-search models but no promising results were obtained. MR failed even when the search models were edited to keep the most conserved core, following [93]. Automated MR- pipelines (*BALBES* [102] and *MrBUMP* [126]) did not succeed, too. Structural comparison of the refined *Ses i 2* model with the closest homologous and with the models obtained through homology modelling reveals a similar overall fold, but shows that there are small yet significant local deviations (**Figure 5.5**). All the models which were used in MR display a Cα r.m.s.d. in the 'twilight zone' (1.5-2.0 Å), or worse. The failures observed when MR was tested show that, even though the target and the homologs share a similar overall fold, local differences are sufficient to prevent structure solution. Even judicious model editing did not sufficiently improve the initial models, which could be attributed to a

failure in removing all the structurally different regions (essentially loops) and/or to the completeness of the edited models that was not high enough to produce a significant signal.

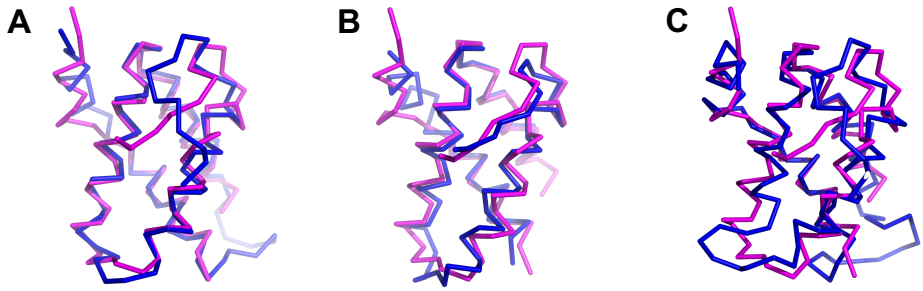


Figure 5.5: Superposition of the *Ses i 2* model to three different representative proteins of the 2S albumin family. (**A**) *Phyre2* predicted model (Cα r.m.s.d.: 1.741 Å) (**B**) 2S albumin from *Moringa oleifera* seeds (PDB ID: 5DOM; Cα r.m.s.d.: 1.730 Å) and (**C**) rproBnIb, a recombinant 2S albumin from *Brassica napus* (rape) seeds (PDB ID: 1SM7; Cα r.m.s.d.: 2.129 Å). *Ses i 2* model always in magenta; *Phyre2*, 5DOM and 1SM7 in blue.

5.4.2.2 Analysis of the variation of wMPE during the structure solution process

The weighted mean phase error (wMPE) for some representative steps of the structure solution process was computed by *CAD* and *CPHASEMATCH* (*CCP4*). *CAD* was used to combine the .MTZ representative of each structure solution step with the reference file containing the phases of the final refined model. The reference file was generated with *SFALL* using the final and refined model. The variation of the wMPE through the different steps is shown in **Figure 5.6**. The initial *SHELXE* solution after density modification has a wMPE of 59.9°: such a value of the wMPE should not appear significantly high. In fact, as it has been proven in a number of cases, promising solutions with a wMPE ~ 60° (or even higher) can still be improved. This can be done by subjecting the partial solution to polyAla tracing and/or to model building cycles which, in favorable conditions (data

resolution, accuracy of the initial model, *etc*…) will expand the initial model and decrease, at the same time, the error on the phases. This is what happens in the case of *Ses i 2*, where the combination of *SHELXE* auto-tracing and *PHENIX* model building hugely improves the initial model.

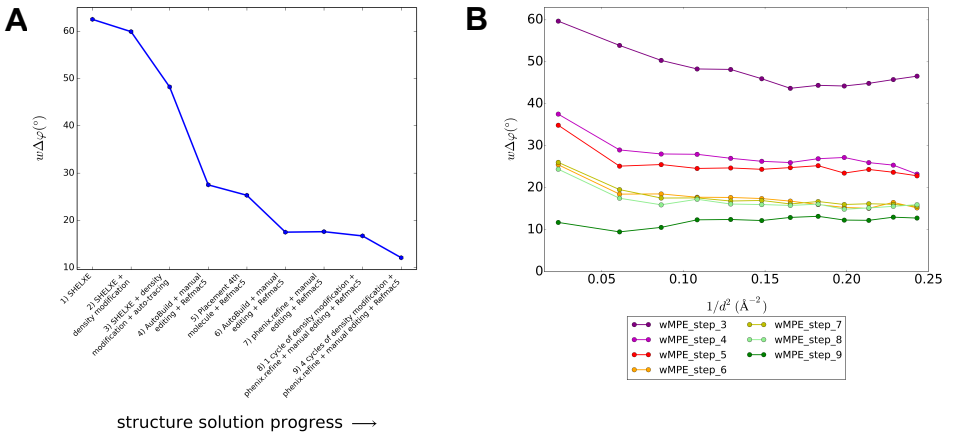


Figure 5.6: Variation of wMPE during the structure solution process. (A) Variation of the wMPE through some representative steps of the structure solution process; **(B)** wMPE across the resolution range at different stages of structure solution. wMPE is computed against calculated phases from the final and *Ses i 2* model.

In particular, *SHELXE* auto-tracing brings the wMPE down to 48.2°, but the most significant drop happens after just a single iteration of *PHENIX-AutoBuild*, manual editing and *REFMAC5* restrained refinement, which lowers the wMPE to 27.5°. In the successive steps, only the placement of the forth molecule will cause an important decrease in the phase error (refinement and density modification steps contribute, too, but only in the order of few degrees).

5.4.2.3 Analysis of the minimal model requirements for structure determination in *ARCIMBOLDO_LITE*

In *ARCIMBOLDO_LITE*, two parameters are known to be particularly important for its success: the length of the α-helices and the number of the α-helices searched for in the electron density map. The minimal model

requirements for the success of *ARCIMBOLDO_LITE* for the *Ses i 2* protein case were tested by running the program for different combinations of the two parameters. A script to automate the procedure was prepared and successfully ran on the powerful 'hyde' computing cluster available at EMBL Hamburg. The number of fragments was varied between 1 and 10, whereas the length of the helices was varied between 5 and 24 residues. The results were evaluated based on the *CC-SHELXE* for the best polyalanine trace. A structure is possibly determined if the *SHELXE* map correlation coefficient exceeds 25% (although the threshold depends on the resolution and on the presence of tNCS). The results are shown in **Figure 5.7** and **Figure 5.8**:

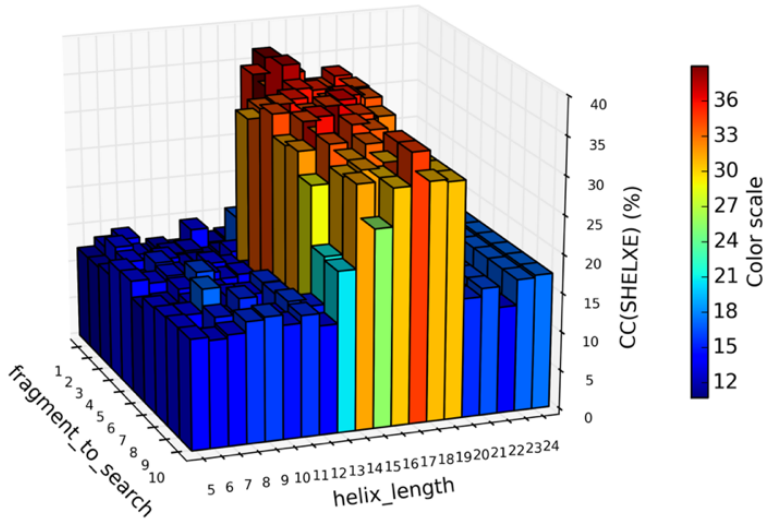


Figure 5.7: Determination of the limits of *ARCIMBOLDO_LITE*. The heatmap shows the variation of CC-*SHELXE* as a function of the 'helix_length' and of the 'fragment_to_search' parameters. The color legend (on the right) refers to the CC-*SHELXE* (%).

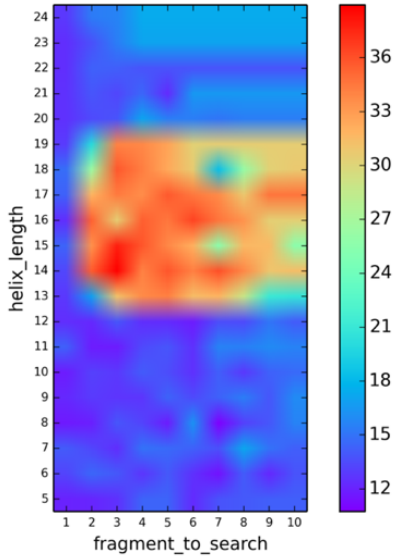


Figure 5.8: Determination of the limits of *ARCIMBOLDO_LITE* (contour plot representation). The contour plot shows the variation of CC-*SHELXE* as a function of the 'helix_length' and of the 'fragment_to_search' parameters. The color legend (on the right) refers to the CC-*SHELXE* (%).

Successful solutions are obtained only when helices with a length between 13 and 19 residues are used; the number of helices appears a less critical parameter. A cluster of very clear solutions can be observed, particularly in the contour plot, which shows that the likelihood of having a solution is higher when the length of the helices is high, even though the number of fragments is low (in some cases, two fragments are enough for the S*HELXE* expansion, which proves the efficacy of density modification and auto-tracing). A clear separation between solutions and not-solutions can be observed, but as soon as two or more helices longer than 12 residues are used for the search (provided they are shorter than 20 residues), the probability to obtain a solution increases considerably. Because searching for a slightly too long or too short helix can result in a failure, it is important to: *(i)* perform a careful examination of the available homologous structures to determine the typical helix length (in particular, the minimum and the

maximum lengths) and *(ii)* to perform different tests by varying (among other parameters) the length of the helices. In fact, in cases similar to the one of *Ses i 2*, parametrization can make the difference between solving or not-solving the structure. In these cases, it becomes important to test as many parameter combinations as possible, which makes the access to powerful computing clusters and the ability to use them a significant advantage.

5.5. Conclusions

The structure of the major allergen of sesame seeds, *Ses i 2*, was determined at 2.0 Å resolution. Despite the availability of close homologous structures from the 2S albumin family, all the attempts to solve the structure of *Ses i 2* by molecular replacement failed. Failure of standard MR happened even though the initial models were subjected to various schemes of model trimming or obtained through sophisticated homology modeling techniques. This example shows that there are cases in which conventional MR does not lead to a successful solution even with models having sequence identity (to the target) better than 30% and with a highly conserved overall fold. In fact, as revealed by a *post-facto* analysis, though the overall fold is the same, small local differences are sufficient to prevent structure solution by conventional MR. Even if homology models were not useful in conventional MR, they could be exploited to solve the structure with *ab-initio* phasing as implemented in *ARCIMBOLDO_LITE*. Based on such homologous, an initial guess on the number and length of the helical fragments was made, which is required by *ARCIMBOLDO_LITE* to locate the initial fragments and to attempt phasing based on these few but

accurately placed helices. The initial guess proved to be, *a posteriori*, a good starting point for the parametrization which eventually led to the solution of the structure. The solution of the structure required extensive parametrization on a multiprocessing machine and was probably complicated, among other factors, by an initially wrong estimation of the ASU composition (three molecules instead of four) which altered the Maximum Likelihood function and the estimation of the expected LLG. Because of this and of the borderline resolution of the native data, *Ses i 2* represents one of the most challenging structures solved so far by *ARCIMBOLDO_LITE*. Finally, systematic tests on the minimal model requirements for structure determination were conducted. This study suggests how, in general, to maximize the chances of finding a solution with *ARCIMBOLDO_LITE*.

6. ACCUMULATING EVIDENCE OF CONTAMINATION FROM EXTERNAL ORGANISMS: THE CASE OF *Serratia* STRAINS

6.1. Summary

In the last decades of protein crystallography, the crystallization of contaminant proteins in place of the proteins of interest, or target proteins, has been reported several times despite the improvements in the expression and purification protocols, the availability of *ad hoc* software for contaminant check and the increasing awareness of crystallographers about this issue. In the vast majority of the cases, the contaminant protein comes from the host expression organism (often *E. coli*) but the possibility of a contamination from other organisms exists and has been reported in few cases. In this Chapter, a case of contamination from a *Serratia* strain is presented, which has remained elusive for several years and resisted numerous attempts at structure determination. *Serratia sp*. is an opportunistic enterobacterium that is most commonly acquired in hospitals but can also be found in the laboratory environment, growing in conditions similar to that of *E. coli*. This case shows that contamination from organisms other than the ones used for expression is not only possible but is probably more common and serious than expected. Furthermore, it suggests that a thorough check for contamination should become an essential step in data analysis prior to any structure determination attempt and it encourages the deposition of contaminant structures to aid the identification of unintended proteins.

6.2. Introduction

It is often the case that, after the structure of a protein has been determined, it becomes interesting to study the structure of the same protein with one or more point mutations. Usually, the aim is to investigate the role of key residues present in the catalytic active site and thought to be involved in the protein activity, or the importance of a region of the sequence in the biological function of the protein. Structure solution of the mutant proteins is usually a straightforward task: this is because the structures of the mutants are similar enough to the one of the wild-type to enable the use of Molecular Replacement with the wild-type protein as a search model. However, there are cases in which this task becomes more difficult or even impossible [127]. Occasionally, even a single-residue mutation can induce a local or a global change in the structure of the mutant, thus altering the conformation to a point where MR becomes very difficult or impossible. Rarely, it is also possible that a contaminant protein is crystallized in place of the protein of interest. In both cases, a significant amount of time and efforts might be invested in unsuccessful MR phasing attempts before it becomes evident that either the conformation has significantly changed or that the nature of the crystal is not the one of the intended target. The case of a contamination is probably worse than the situation when the crystallized protein is still the intended target, but with a significant conformational change in the mutant structure. In fact, contamination is not always easy to spot, and the contamination hypothesis seems unrealistic. This is because: *i)* the search of the PDB for structures with similar cell parameters might not return any hit, *ii)* the molecular weight of the target and of the contaminant are similar enough to make very difficult to reveal them from the SDS-gels and *iii)* only recently, powerful programs have

become available which enable a brute force and thorough check of contaminants. Contamination is an unlikely event, and even more improbable is the contamination of a protein from an organism which is not the one used for the over-expression, *i.e.* from an organism which is in the laboratory environment. There is only a very limited number of reported cases of contamination from organisms other than the one used for over-expression [128], [129].

6.3. Materials & Methods

6.3.1. Experimental part, data collection, data processing and analysis of the unit cell and solvent content

All the details about the experimental protocol, as well as data collection, data processing and the analysis of the unit cell and solvent content can be found in the *Appendix Manuscripts* section.

6.3.2. Structure solution and refinement

All the attempts at solving the structure using Molecular Replacement with a number of search models based on the 4B5C monomer were unsuccessful. As described below, a thorough check for contaminants eventually identified the crystallized protein as a cyanate hydratase of possible bacterial origin. After *ARP/wARP* model building and *REFMAC5* restrained refinement of the MR solution, the origin of the contaminant protein was confirmed to be from a bacterium of the *Serratia* genus. Cycles of refinement with *REFMAC5* and *phenix.refine* (with the application of NCS-restraints) and *Coot* manual editing completed and improved the model. Deposition of the refined model on the PDB is planned.

6.4. Results & Discussion

6.4.1. MR attempts

Initial attempts at solving the structure using Molecular Replacement with the 4B5C monomer were unsuccessful. When the group in Milano first embarked on the project and started to analyze the data, a contaminant search was done by screening the entire PDB for structures having unit cell parameters similar to the ones of the collected data, but no hits were found. As a consequence, when I was given the data and I began to analyze it, I performed more sophisticated MR tests based on the suggested oligomeric assembly deduced from the SRF (**Figure 6.1**):

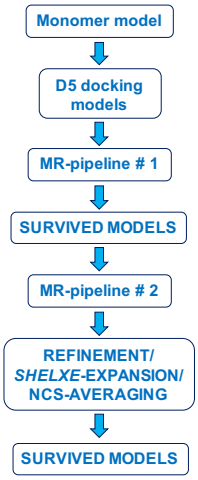


Figure 6.1: Basic scheme of the work-flow which was used to generate the D5 docking models and to test them in MR.

A number of docking software (*SAM* [130], *HSYMDOCK* [131], *ROSETTA SYMMETRY DOCKING* [132], *GalaxyWeb* [133], *SYMMDOCK* [134] and *MZDOCK* [135]) were used to generate ~ 2000 models with D5 symmetry starting from the 4B5C monomer: in a first step, the quality of these models was tested with an automated MR pipeline which employs *PHASER* and *MolRep* with default settings. The most promising solutions, as judged by

the most important MR-indicators (TFZ equivalent, LLG, packing for *PHASER*; Score, contrast, TF/sigma and wRFac for *MolRep*) and the *R*-factors, were used in the second step. Here, the models from the first step were tested in a different automated MR pipeline employing *PHASER* and *MolRep* where critical parameters are varied (data resolution and expected r.m.s.d. for *PHASER*; data resolution, similarity, completeness and number of rotation peaks for *MolRep*). The most promising models (selected with the same criteria used at the end of the first step and described above) were subjected to *REFMAC5* refinement, *SHELXE*-expansion and/or NCS-averaging. However, none of the MR solutions could be successfully refined, expanded or its density improved by any of the methods listed above. This suggested that many of the MR solutions, which were initially considered as promising, were in fact false-positives.

6.4.2. Contaminant search and identification of contaminant origin

At this point, a second, more thorough check for contaminants was carried out using the recently developed program *SIMBAD* [136]. The program quickly identified PDB ID 4Y42 [129] as the likely contaminant. Full MR using 4Y42 as a search model, followed by *REFMAC5* restrained refinement and one cycle of *ARP/wARP* model building confirmed the crystallized protein to be a cyanate hydratase, likely from *Serratia*. To confirm the contaminant origin, the following method was used. A main-chain only model was built with *ARP/wARP*, containing dummy atoms in place of the side-chains electron density. Then, using methods recently described [137], a Position-Specific Scoring Matrix (PSSM) of the sequence was generated. The PSSM essentially describes the probability of each residue along the sequence of being a specific amino-acid, based on the side-chains electron density at that position. The PSSM was used to

query a number of databases with *PSI-BLAST* and *HMMER* [138] in order to find matching known sequences. The best matching sequence found thorough this first iteration (*E*-value of $3.8 \cdot 10^{-41}$, from a strain of *Serratia proteomaculans*) was used to build a full model with *ARP/wARP*. The quality of the model built in this second iteration allowed to unequivocally confirm the initial hypothesis that the contaminant was from *Serratia*. Further cycles of refinement with *REFMAC5* and *phenix.refine* (applying the NCS-restraints) and *Coot* manual editing led to the re-assignment of few residues owing to the better side chains electron density. Alignment with the sequence extracted from the final model shows that the protein comes from an organism of the *Serratia* genus, without showing the exact species.

6.4.3. Hypothesis about the contamination from *Serratia*

In order to better understand the origin of the contamination and, as a consequence, to reduce the possibility of this event to happen in the future, several hypotheses have been considered. Among them, the contamination during the protein expression appears to be the most plausible one. In fact, several antibiotic-resistant *Serratia* strains are known which can grow in the presence of ampicillin [139]. Moreover, it is known that ampicillin is easily degraded, so it could be that at a later point during the expression, the *Serratia* bacteria started to grow too. In addition, *Serratia* is found in the environment and often in the grooves of the floor etc... and it grows in similar conditions to *E. coli*. The combination of these factors (non-fresh or simply easily degradable ampicillin stocks, combined with ubiquitous *Serratia* and non-sterile laboratory environment) is likely to be responsible for the contamination during the expression step. Assuming that the contaminant protein is highly abundant in *Serratia*, it likely co-eluted with

the purified target protein, due to its non-specific binding to the nickel resin. A contamination during the crystallization step seems unlikely, too, but cannot be excluded.

6.4.4. Description of the structure

All individual chains are virtually identical as they display an excellent structural match in their Cα traces, with a maximum Cα r.m.s.d. of 0.203 Å. The protein structure resembles very closely the one of the cyaneses deposited in the PDB. It is composed of ten protomers: each of them consists of two domains: the N-terminal domain forming a 5-helix bundle, and the C-terminal catalytic domain having a unique fold. Pairs of protomers are organized to form dimers through an intricate interaction of two C-terminal cyanase domains, and the dimers assemble into a decamer with 52 point symmetry. The interface between dimers forms a set of 5 symmetrically disposed active sites, where key residues of the catalytic triad (Arg96, Glu99, Ser122 and their NCS-related equivalents at the interface of two adjacent dimers) are responsible for the binding of the substrates. After few refinement cycles, additional positive and symmetrical electron density started to appear from the $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ and $|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ electron-density maps in all the five active sites. Further refinement improved the density and allowed to unambiguously assign the ligand present in the active sites as oxalate ion (**Figure 6.2**). Oxalate, together with other low-molecular-weight dicarboxylic acids and mono-anions, is a known inhibitor of *E. Coli* CynS and can easily be found in the culture media, possibly as the product of bacteria metabolism. Beside the oxalate molecules in the active sites, the structure also contains a number of glycerol molecules, which is not surprising giving its presence in the crystallization condition and in the cryo-solution.

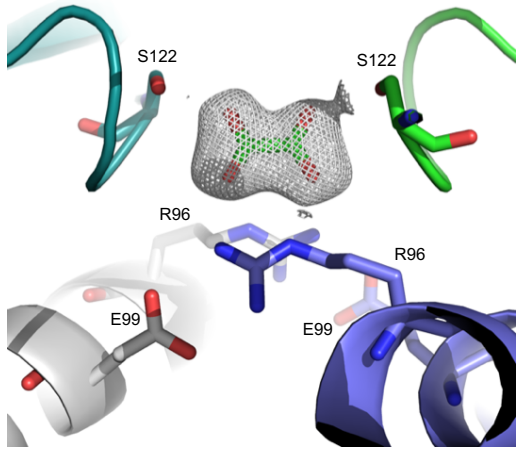


Figure 6.2: Verification of the presence of oxalate in the active sites of the enzyme. $2|\boldsymbol{F}_o| - |\boldsymbol{F}_c|$ composite omit map countered at 1.0σ of the 2.1 Å data of the cyanase crystal structure calculated using *phenix.composite_omit_map* in simulated annealing mode (cartesian; default annealing temperature of 5000K). The SA-composite omit map is shown for one of the oxalate ligands and the key residues of the catalytic triad are labeled.

6.4.4.1 *A posteriori* explanation for the initial failures to find the contaminant and to solve the structure by MR-phasing

6.4.4.2 Failure of the initial contamination check

As discussed in the Introduction, searching for structures with similar cell parameters might fail to return any hit if the space group is not the same and if the error on the cell parameters (despite the space group being the same) is larger than the tolerance set for the search. In this case, the failed identification of a contaminant early on is due to the fact that none of the structures deposited has the same space group. The limitation of this approach has been recently overcome with the implementation of programs for the rapid screening of large databases of structures by MR. The software *SIMBAD*, for instance, implements a brute-force approach in which a rotational search is performed with all structures of known contaminants or, if necessary, with all structures in a non-redundant PDB database. By a rotational search with all structures of known contaminants, *SIMBAD* identified the structure of another cyanate hydratase from *Serratia*

proteamaculans as the likely contaminant. Another reason why a contaminant was ruled out is because contaminants have, usually, smaller size. Despite the fact that known contaminants have a wide range of molecular weights, from 10 to 140 kDa, the most of them weight between 20 and 40 kDa, which also corresponds to the weight range of 70% of all crystallized proteins reported in the PDB [141]. This indeed confirms the unicity of this case, as the MW of the *Serratia* cyanase is ~ 170.2 kDa.

6.4.4.3 Failures to solve the structure with MR-phasing

There are several reasons as to why a contamination was not suspected, even after several failed MR tests. The MR problem was expected to be highly difficult since the beginning. This is because of, at least, the following factors:

a) the many degrees of freedom involved in the generation of D5 models from the monomer structure alone. The variables at play are the following: *i)* relative orientation of each monomer with respect to the others, *ii)* distance between monomers in the pentamer, *iii)* "phase shift" between the two pentamers, *iv)* possibility for the two pentamers to be in "up" or "down" configuration, *v)* distance between the two pentamers, *vi)* distance between the monomers and the 5-fold axis.

b) the well-known high sensitivity of MR to the similarity between the search and the target models,

c) a possible local or global conformational difference in the structure of the monomer induced by the mutation. Concerning the second and the third point, the failure of MR is even more likely to happen for large multimeric proteins, given the higher number of potentially different regions between the target and the search structures.

d) the space group of the data. The MR problem was expected to be difficult because of the space group of the data ($P2_1$). Although it is often true that a TFZ equivalent ~ 8 or higher identifies a true solution, there are difficult cases where a correct solution has a significantly lower TFZ score. Moreover, TFZ scores are generally lower for monoclinic space groups like $P2_1$, at least in the search for the first molecule.

e) Partial proteolysis. Because of the time it took for the crystals to grow, partial proteolysis of the protein was suspected and assumed to be one of the factors complicating the MR problem.

f) False-positive MR solutions. The MR problem was further complicated by the large number of false-positive solutions produced during the first and the second screening step. MR-programs such as *PHASER* and *MolRep* make use of statistics or scoring functions to help identifying true and false solutions. The properties of these parameters have been studied and, based on a high number of tests, it has been concluded that, for example, a TFZ equivalent > 8 and a LLG > 64 are strong indicators of a true solution in *PHASER*. Similarly, a contrast > 2.5 (and/or a Score > 0.3 or a TF/sigma > 8.0) in *MolRep* are considered good indicators of a true solution. This is the reason why these parameters have been used to screen the MR solutions. However, it has been noticed that, in a number of cases, seemingly good MR solutions were simply false-positives either because of the different SRF with respect to the SRF of the data and/or the difficulty to refine or to expand these solutions. A 2D combination of parameters should, in principle, be better than one single parameter to identify a group of distinct MR solutions (for

instance, TFZ and LLG, or Contrast + Score). However, not just single scoring functions showed a weak discriminatory power, but also combinations of them proved to be incapable to separate true and false solutions.

6.5. Conclusions

This interesting case represents the unintentional crystallization of a cyanate hydratase from a bacterium of the *Serratia* genus during efforts to crystallize a different target protein. This result adds to two previous reports of unexpected contamination from *Serratia* [128], [129] and is important under several aspects. First of all, it suggests that the contamination from organisms other than the ones used for over-expression is possible and probably more likely to occur than one might expect. Secondly, it shows that pathogenic organisms can easily grow and proliferate in the laboratory environment and accumulate to the point that they can contaminate machineries, reagents *etc*…Thirdly, the result confirms that the contamination process is serendipitous by its nature and should be expected at any time. Problems with contamination never occurred in the case of 4B5C but manifested for the grafted structure, which confirms that even small changes in the sequence, or in other variables, can have a significant impact during the expression, purification and/or the crystallization steps. Some lessons can be learnt from this and other reports: before anything, these cases highlight the importance of good laboratory practices. The proper and regular cleaning of the laboratory, including all the instrumentation, is of primary importance and is probably the primary and most effective way to reduce (and, hopefully, to avoid) any unintended contamination. This adds to the necessary checks (SDS-page, mass-

spectrometry, chromatographic analysis, …) that must be carried out during the expression and purification of the protein of interest. All these measures will probably not exclude the possibility of a contamination but will certainly reduce it more than any other practice. In the case presented here, several factors together contributed to make more difficult to detect the contaminant as, for instance, the negative results from the screening of the PDB for structures having similar cell parameters, the seemingly positive results from the MR tests and the absence of any contamination problem with the wild-type protein. A check for contaminants should therefore become an essential part of data analysis after data collection and processing and before any attempt at structure solution. The simplicity and the rapidity of such a contamination check should convince crystallographers to include this step into the routine process of data analysis and would most likely reduce future cases as the one reported here. In parallel, crystallographers should report similar cases and should deposit the structures of contaminants, even when they are already known. One of the advantages of increasing the number of deposited contaminant structures is that it will aid the identification of such contaminants by other crystallographers.

7. TOWARDS THE STRUCTURE DETERMINATION OF TWO CHIMERIC ANTIGENS FOR POTENTIAL VACCINE DEVELOPMENT AGAINST MIELOIDOSIS

7.1. Summary

The development of vaccines against life-threatening diseases is of primary importance and is the objective of many efforts by a number of research groups worldwide. One example is represented by the search for an effective vaccine against bacteria from the *Burkholderia* genus, which are responsible for a series of lung infections and, most importantly, for melioidosis, an endemic disease spread in several areas of the world. Identification of conserved protein antigens and epitopes among the different *Burkholderia* spp. (particularly *B. pseudomallei* and *B. cenocepacia*) could be used to design molecules with enhanced immunoreactivity which can serve as the basis to developed more effective diagnostic tools and vaccines. Since the structure and dynamics of antigens or epitopes are fundamental properties for antibody recognition, X-ray Crystallography plays an important role by providing atomic-level information to help the design of improved molecules, more stable and with enhanced immunoreactivity. The Structural Vaccinology Unit directed by Prof. Martino Bolognesi at the University of Milano has made a significant step in this direction through the determination of the X-ray crystal structure of the Peptidoglycan-associated lipoprotein from *B. pseudomallei* (Pal_{Bp}). This information was used to predict and design a promising

epitope peptide (BpEp3). Because of the genomic similarities between the two bacteria, the group is now investigating whether it is possible to design cross-reactive epitopes for the simultaneous diagnosis of *Burkholderia* spp.. To this aim, the BpEp3 epitope was transplanted onto a second antigen (the BPSL2520 acute phase *B. pseudomallei* antigen), using a fast computational tool for epitope grafting called SAGE. Different chimeric constructs containing the foreign BpEp3 were selected and produced for immunological and structural studies. Two of these constructs, named SAGE1 and SAGE3, delivered crystals which were tested for their diffraction properties. However, the crystals diffracted to low resolution (between 3.4 and 4.0 Å), and initial attempts at structure determination were inconclusive. Despite the low resolution of the diffraction data, after some efforts a partial model for both proteins could be obtained. In both cases, a substantial part of the model is missing, probably due to the destabilization introduced by the grafted epitope. Because of this and of the risks of working with low resolution data, the crystallographic models of SAGE1 and SAGE3 require further validation and experiments have been planned to this aim.

7.2. Introduction

Bacteria from the *Burkholderia* genus are Gram-negative pathogens and etiologic agents of the disease melioidosis, which cause potentially fatal lung diseases in infected humans. *Burkholderia* pathogens and melioidosis are a serious endemic problem in South-East Asia, North Australia, South America and in the Indian subcontinent [142]. Present treatment relies on antibiotic administration but problems of multidrug resistance are known and represent a serious obstacle to the effectiveness of this approach [143].

58% mortality rate related to melioidosis means that more than one person out of two that contracts the disease will die. For this reason, an alternative to antibiotic treatment is urgently required. The ideal solution would be a melioidosis vaccine. The best approach to develop a vaccine is represented by the use of epitope peptides [144], [145]. Studies have so far focused on the bacteria *B. pseudomallei* and *B. cenocepacia* and on the analysis of their epitopes. The genomic similarities between the two bacteria suggests the possibility to design cross-reactive epitopes for the simultaneous diagnosis of *Burkholderia* spp. and the development of effective vaccines. In particular, Gourlay and collaborators were recently able to determine the structure of the Peptidoglycan-associated lipoprotein from *B. pseudomallei* [146] (Pal_{Bp}; PDB ID: 4B5C) and used this information to predict and design a promising 20-residue epitope peptide (BpEp3). Using a fast computational tool for epitope grafting called SAGE (Strategy for Alignment and Grafting of Epitopes) [147], they then transplanted the BpEp3 epitope onto a scaffold (BPSL2520; PDB ID: 6F03), with the aim to generate a chimeric antigen useful for the diagnosis of *Burkholderia* spp. and capable of inducing a protective immune response. BPSL2520 is an antigen itself and it was chosen as a suitable scaffold due to its entirely helical nature that is suitable to receive an epitope which is also entirely helical. Moreover, BPSL2520 is one of the most seroreactive antigens conferring resistance to *Burkholderia* infections and its structure is highly stable. Different chimeric antigens were selected after SAGE analysis, each one having the graft sequence inserted in a different position onto the scaffold. The selected grafts were subsequently produced in recombinant form to test their immunological activity (whether this is enhanced or not) and to determine their crystallographic structure. Two of these constructs,

named SAGE1 and SAGE3, delivered crystals which were tested for their diffraction properties. This Chapter describes the efforts which have been made to determine the structures of both SAGE1 and SAGE3.

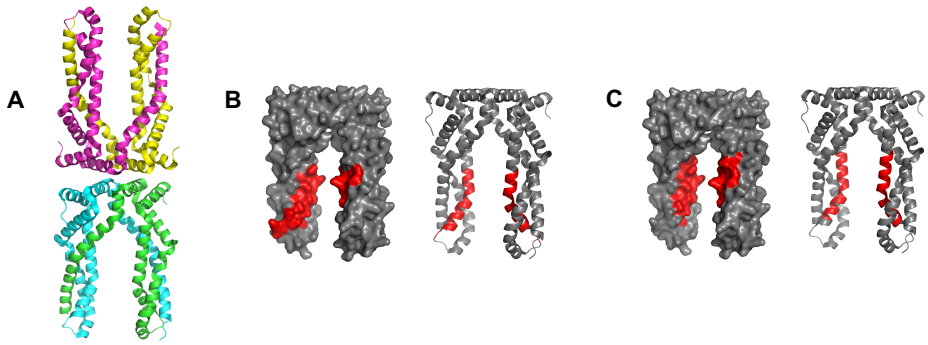


Figure 7.1: Representations of the antigen used as a scaffold and of the two chimeric constructs SAGE1 and SAGE3. (A) Cartoon representation of the structure of the antigen BPSL2520 (PDB ID: 6F03), with each chain colored differently. **(B)** SAGE1 and **(C)** SAGE3 protein surface and secondary structure representation, with highlighted in red the regions corresponding to the graft sequence.

7.3. Materials & Methods

7.3.1. Experimental part, data collection and processing

The Experimental part, as well as the data collection and processing, is part of the Thesis work of Marco Amabili, which was performed under the supervision of Dr. Louise Jane Gourlay at the Department of Biosciences at the University of Milano. All the details about the computational design of the SAGE1 and SAGE3 proteins, their cloning, expression, purification, crystallization, data collection and data processing can be found in his Thesis (Marco Amabili: "A structural vaccinology approach to *Burkholderia pseudomallei* antigen redesign", Master Degree Thesis in Chemical Sciences, University of Milano, 2016/17 Academic Year).

7.3.2. Structure determination and refinement

7.3.2.1 SAGE1

The first attempts were made on SAGE1 data because of the higher resolution (3.4 Å) compared to the SAGE3 case (4.0 Å). What follows is a description of the steps which were made to build a preliminary model:

Identification of the appropriate MR-search model: The wild-type structure of the scaffold, 6F03, represented the obvious starting point as a search-model for a Molecular Replacement search. As the MR-search with the full 6F03 structure did not yield a solution, a number of differently edited search-models, all derived from the wild-type structure, were generated and tested. For the tests, a self-written pipeline was prepared to be run on a multiprocessing machine. The pipeline tests each MR-search model in *PHASER* at different values of the expected r.m.s.d. (which is known to be one of the most critical parameters for the success of MR) and of the number of molecules to search for. The pipeline tests each MR-search model in *MolRep* too by varying the number of molecules to search for and by using the input model as well its polyalanine version. Only one of them gave a convincing MR solution with *PHASER* (TFZ equivalent = 16.4, LLG = 214) with reasonable packing. This solution was obtained by searching for two molecules and using an r.m.s.d. value of 0.4 Å.

Improvements of the initial MR-phases: To improve the initial MR-phases and to (at least) partly remove the model bias, statistical density modification as in *RESOLVE* where applied [121], [148]. This procedure yielded a significantly improved map, less noisy, with clearer boundaries between the protein region and the map, and with clear helical features. As

a consequence, as a first attempt to locate some secondary structure elements, *phenix.find_helices_strands* was used to place helices in the *RESOLVE* map. The partial helical model was refined in *REFMAC5* and this procedure of density modification, helix placement and refinement was iterated until the helical model could not be improved any further.

Location of a second monomer: This helical model was used to bootstrap model building as in *PHENIX AutoBuild*, which found an additional monomer (not initially located by MR).

Attempts to further expand the model: The model obtained from *PHENIX AutoBuild* was refined and attempts were made to expand it. However, the model could only be marginally improved as it was not possible to reveal the missing parts. The final model contains 312 residues and has R_w/R_f = 0.255/0.300. The Ramachandran shows some deviations, but they are in line with what one can expect from a structure at this resolution.

7.3.2.2 SAGE3

A very similar procedure to the one described for SAGE1 was used to obtained a preliminary model for SAGE3. In this case, the search-model which allowed to obtain a solution for SAGE1 was directly used and yielded a convincing solution for SAGE3, too. The initial MR-phases were improved through several iterations of statistical density modification, helix placement and refinement, and the best helical model was given to *PHENIX AutoBuild* which, again, found a second monomer. As for SAGE1, this model could not be significantly expanded. The final model

contains 333 residues and has R_w/R_f = 0.259/0.315. The Ramachandran shows only few outliers.

7.4. Results & Discussion

Initial attempts at solving the structure of SAGE1 and SAGE3 with a number of MR-programs and automated pipelines (*PHASER*, *MolRep*, *Balbes*) using the wild-type monomer, the dimer *etc*…proved unsuccessful. A convincing MR-solution could be found only using a small yet quite compact unit of the wild-type structure (residues 36-65 and 114-150 of chain A together with residues 151-177 of chain B). Varying the expected r.m.s.d. was important to increase the signal-to-noise ratio and therefore to increase the chance of success. Varying the number of molecules of the search-model to look for was necessary, too, as the analysis of solvent content, of the Self-Rotation Function and of the rotation peaks did not give any clear insight about the oligomerization state of the proteins. Because model bias is probably the most important caveat at low resolution, density modification was applied from the beginning with the two-fold aim to reduce the memory from the search-model and to improve the electron density maps. Statistical density modification proved very effective, especially at the very beginning, in that it allowed the helical features to become more visible, therefore making easier the successive placement of the helices. The procedure of iterative density modification, helix placement and refinement was effective in that it improved the initially poor MR-phases and provided a helical model sufficiently good to bootstrap model building. The starting phases for model building were good enough to allow the building of a second molecule, which was not found during the MR tests. During all the abovementioned steps, a decrease

in both R_f and in ($R_w - R_f$) were used to guide the process, as they are robust indicators of better fit to the experimental data and less model bias, respectively. Not just the behaviour of the *R*-factors was used to make decisions on how to proceed, but also the appearance of the electron density maps and the Ramachandran statistics were employed. Refinement at low resolution is notoriously difficult and prone to many errors. The *LORESTR* protocol with the generation of restraints from the homologues 6F03 model was found to be a valuable tool in this sense [149]. In fact, *LORESTR* tries different low resolution refinement protocols, and provides the best model and map based on a series of indicators (*R*-factors, Ramachandran statistics, Clashscore and Molprobity percentile). For SAGE1 and SAGE3, it proved effective in reducing the *R*-factors and in maintaining the ($R_w - R_f$) within an acceptable limit. The overall packing of the final model looks reasonable and the electron density is continuous, which leads to believe that the models are correct. In both cases, the *R*-factors for the final models are lower than what would be expected considering the amount of missing residues, which is around ~ 50% of the total residues based on the sequence. The observed *R*-factors suggest a possible proteolysis: proteases might have cut some of the chains, which would explain why they are not visible in the final models and, at the same time, the *R*-factor values. At the same time, the crystallographic models could also suggest that the epitope grafting have destabilized the constructs, leading to disorder in a significant part of the proteins. The packing of the models in the unit cell shows that there is certainly space for the missing residues, making the hypothesis of proteolysis less unlikely. Grafting a 20-residue epitope is a very ambitious project, so it should not be surprising that a large part of the scaffold structure is destabilized. Given the risks of working at low resolution, the

SAGE1 and SAGE3 models require further validation. To this aim, a native gel analysis will be performed and SAXS measurements of both proteins are planned in order to confirm (or not) the hypothesis that the epitope has destabilized the protein, which could explain the high flexibility of the mutated regions and, therefore, their absence in the electron density.

7.5. Conclusions

A critical prerequisite for low-resolution crystallography is the availability of good experimental phase information as molecular replacement becomes more difficult and dangerous. This is because of the model bias, which is probably the greatest caveat at low resolution, and because the placement of small fragments may not suffice to reveal the missing parts of the model with sufficient clarity. However, obtaining experimental phase information is not always possible. The main factor which makes the generation of a model from low resolution data problematic is the low number of observations used to calculate the electron density map compared to the number of parameters to be defined [9], [12], [150]. This results in maps lacking atomicity, helices appearing as tubes of density, missing side-chains density and so on. All these factors make refinement and model building particularly challenging [150]. Refinement of models against low resolution data requires great care: several low-resolution refinement protocols have to be tested and evaluated, overfitting must be avoided, and additional restraints need to be added in order to maintain a good geometry of the model [149], [151]. Complications arise from the fact that, not always, improvements in the *R*-factors correlate with better model geometry. Model building requires many subjective judgements, especially when other sources of information are not available. The interpretation of

electron density maps is further complicated by the fact that improvements in the model not necessarily result in an improvement of the *R*-factors and many iterations might be required to establish the correctness of the model. As a consequence, building a model in low resolution data might become impossible even for the most skilled crystallographer. All of these problems were encountered during the attempts to determine the structures of SAGE1 and SAGE3. Because of the many traps of working with low resolution data, the crystallographic models of SAGE1 and SAGE3 require further experimental validation. To this aim, a native gel experiment on both proteins is planned to determine the oligomeric state. SAXS measurements will be carried out too, in order to explore whether the missing parts in the crystallographic models are not visible as a result of the destabilization caused by the epitope grafting.

8. CONCLUSIONS AND FUTURE PERSPECTIVES

In the first part of the Thesis work, MRSAD-phasing has been systematically tested on a wide range of protein systems of known structure. The method was first tested on small to medium size proteins. Even though these systems do not represent the ideal target for MRSAD-phasing, they were nevertheless useful to develop and test the pipelines that were used later on for the tests on the central model system. Moreover, the results from the application of MRSAD on small/medium size systems are useful in that they represent a reference against which the results from more challenging cases can be compared and evaluated. By analyzing the results obtained on all the systems, including the human 20*S* proteasome, a number of general conclusions about MRSAD-phasing can be drawn. A general pipeline for MRSAD-phasing and model building cannot be applied to systems of higher molecular weight and/or data at medium or low resolution. The resolution limit after which the map interpretation software cannot build sensible models was found to be ~ 2.8 Å. For all the systems, and for the majority of the MR-search models, MRSAD is able to improve over MR- and SAD-phases alone. The numerical improvements are modest in terms of mean phase error, but visual inspection of the maps clearly shows larger improvements than what the MPE suggests. RSCC was found to be a better metric of phase quality than the error on the phases. The phase improvements observed after MRSAD depend on the ability of the LLG-algorithm to find the heavy-atom substructure. Even with less accurate and/or small search models, the LLG-algorithm can still locate a good number of sites that is enough to improve the MR-phases. As a

consequence, not all the heavy-atom sites are required for successful MRSAD-phasing. Some similarities were observed between MRSAD and experimental phasing: *i)* not all of the heavy-atom sites are required for successful phasing, *ii)* density modification is critical for the improvement of the phases and *iii)* anomalous multiplicity in MRSAD is as much important as in S-SAD and highlights the necessity to collect, whenever possible, accurate and highly redundant anomalous data.

At the same time, conclusions more specific to the proteasome data can be drawn. The results show and confirm that MRSAD-phasing can improve on MR-phases even on large macromolecular complexes by using only the anomalous signal of weak scatterers and search-models representing only a small fraction of the target. It was confirmed that density modification plays a crucial role in further improving MRSAD-phases, especially when NCS-averaging is included, and an optimal protocol to improve MRSAD-phases was found, whose main advantage lies in its simplicity. The analysis of additional data from other well-known and large macromolecular structures would allow to draw more general conclusions on the applicability range of MRSAD-phasing and on the best strategies to maximize its success.

The potential of MRSAD-phasing was also tested on two real-life scenarios. Firstly, MRSAD was used for the structure determination of the first plant glutamate receptor in complex with four natural ligands at medium to high resolution. In this case, MRSAD proved to be crucial for structure solution, as the application of all other methods proved unsuccessful. Therefore, the structure determination of the plant glutamate receptor reinforces, together with growing examples in the literature, the

critical role of MRSAD-phasing in solving the structure of proteins in real-life scenarios. In the second place, the potential of MRSAD for the phasing of unknown antigens with engineered nanobodies with a lanthanide binding motif recently developed within the group was investigated. The results show that MRSAD-phasing represents another concrete route to phase challenging protein structures with the proposed "backpack"-principle. Even in those cases where experimental phasing would suffice for structure solution, MRSAD-phasing has the potential to provide more accurate crystallographic phases and models of the antigen protein.

In the second part of the Thesis work, a number of protein structures has been determined. All the cases presented in the previous chapters had remained unsolved for a long time, despite the significant efforts that had been made by many people in the attempt to solve these structures. Therefore, they all represent challenging cases, each one with its own specificities; for some of them, the difficulty was mainly due to the poor quality of the data, for others to the unavailability of sufficiently close homolog models or anomalous data and, in the third case, to a combination of poor data and poor initial information. Beside the already mentioned case of the plant glutamate receptor, other structures were solved, starting with the case of the major protein allergen of *Sesamum indicum* by *ab-initio* phasing at 2.0 Å resolution. Then, attempts towards the structure determination of two chimeric antigens for potential vaccine development, SAGE1 and SAGE3, were made. Given the low resolution of the data, further experiments will be carried out to validate the latter models. Finally, an interesting case of protein contamination was studied: here, the nature and the origin of the contaminant protein was identified and the structure,

in a new space group, was refined locating a known natural inhibitor present in all the enzyme active sites.

Taken together, all these cases highlight the importance of the role of a competent crystallographer in facing frontier structures. Despite the huge improvements of the last decades in terms of software development, and the availability of a number of automated structure solution pipelines, there are evident cases in which the crystallographer's role is still decisive. The availability of powerful programs and, at the same time, the fact that they are limited to simple or moderately simple cases, probably shows more than ever before the significance of the crystallographer's experience, fundamental skills and determination.

BIBLIOGRAPHY

[1] J. P. Schuermann and J. J. Tanner, "MRSAD: using anomalous dispersion from S atoms collected at Cu Kalpha wavelength in molecular-replacement structure determination" *Acta Crystallogr. Sect. D*, pp. 1731–1736, 2003.

[2] V. Parthasarathy, V. S. Lamzin, S. Manfred, and P. A. Tucker, "On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination" *Acta Crystallogr. Sect. D*, pp. 1089–1097, 2009.

[3] L. Jolla, "Low-resolution Electron-density and Anomalous-scattering- density Maps of Chromatium High-potential Iron Protein" *J. Mol. Biol.*, pp. 503–512, 1968.

[4] Z. Dauter, M. Dauter, E. D. La Fortelle, G. Bricogne, and G. M. Sheldrick, "Can Anomalous Signal of Sulfur Become a Tool for Solving Protein Crystal Structures ?" *J. Mol. Biol.*, pp. 83–92, 1999.

[5] Z. Dauter, "Use of polynuclear metal clusters in protein crystallography" *Comptes Rendus Chim.*, vol. 8, pp. 1808–1814, 2005.

[6] J. Thygesen, S. Weinstein, F. Franceschi, and A. Yonath, "The suitability of multi-metal clusters for phasing in crystallography of large macromolecular assemblies" *Structure*, pp. 513–518, 1996.

[7] J. Knäblein *et al.*, "$Ta_6Br_{12}^{2+}$, a Tool for Phase Determination of Large Biological Assemblies by X-ray Crystallography" *J. Mol. Biol.*, pp. 1–7, 1997.

[8] "Ta_6Br_{14} is a useful cluster compound for isomorphous replacement in protein crystallography" *Acta Crystallogr. Sect. D*, pp. 186–191, 1994.

[9] B. Delabarre and T. Axel, "Considerations for the refinement of low-resolution crystal structures" *Acta Crystallogr. Sect. D*, pp. 923–932, 2006.

[10] F. Dyda, "Developments in low-resolution biological X-ray crystallography" *F1000 Biol Rep.*, vol. 4, no. November, pp. 4–7, 2010.

[11] A. T. Brunger, "Low-Resolution Crystallography Is Coming of Age" *Structure*, vol. 13, pp. 171–172, 2005.

[12] A. T. Brunger, "X-ray structure determination at low resolution" *Acta Crystallogr. Sect. D*, pp. 128–133, 2009.

[13] A. Ben-shem, L. Jenner, G. Yusupova, and M. Yusupov, "Crystal

structure of the eukaryotic ribosome" *Science*, vol. 330, no. November, pp. 1203–1210, 2010.

[14] B. C. Lechtenberg *et al.*, "Structure of the HOIP/E2-ubiquitin complex reveals RBR E3 ligase mechanism and regulation" *Nature*, vol. 529, no. 7587, pp. 546–550, 2016.

[15] R. D. Bunker, "Tackling the crystallographic structure determination of the COP9 signalosome" *Acta Crystallogr. Sect. D*, vol. 1, pp. 326–335, 2016.

[16] J. N. Busby, J. S. Lott, and S. Panjikar, "Combining cross-crystal averaging and MRSAD to phase a 4354-amico-acido structure" *Acta Crystallogr. Sect. D*, pp. 182–191, 2016.

[17] K. Nozawa, T. R. Schneider, and P. Cramer, "Core Mediator structure at 3.4 Å extends model of transcription initiation complex" *Nature*, vol. 545, no. 7653, pp. 248–251, 2017.

[18] A. J. McCoy, L. C. Storoni, and R. J. Read, "Simple algorithm for a maximum-likelihood SAD function" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 60, no. 7, pp. 1220–1228, 2004.

[19] R. J. Read and A. J. McCoy, "Using SAD data in Phaser" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 4, pp. 338–344, 2011.

[20] G. Langer, S. X. Cohen, V. S. Lamzin, and A. Perrakis, "Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7" *Nat. Protoc.*, vol. 3, no. 7, pp. 1171–9, 2008.

[21] A. Perrakis, R. Morris, and V. S. Lamzin, "Automated protein model building combined with iterative structure refinement" *Nat. Struct. Biol.*, vol. 6, no. 5, pp. 458–463, 1999.

[22] K. Cowtan, "Fitting molecular fragments into electron density" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 64, no. 1, pp. 83–89, 2007.

[23] K. Cowtan, "The Buccaneer software for automated model building. 1. Tracing protein chains" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 62, no. 9, pp. 1002–1011, 2006.

[24] P. D. Adams *et al.*, "PHENIX: a comprehensive Python-based system for macromolecular structure solution" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 2, pp. 213–221, 2010.

[25] G. M. Sheldrick, "Experimental phasing with SHELXC/D/E: Combining chain tracing with density modification" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 4, pp. 479–485, 2010.

[26] K. Cowtan, "Recent developments in classical density modification

research papers" vol. 4449, pp. 470–478, 2010.
[27] F. Dall'Antonia and T. R. Schneider, "SITCOM: A program for comparing sites in macromolecular substructures" *J. Appl. Crystallogr.*, vol. 39, no. 4, pp. 618–619, 2006.
[28] J. L. Smith, "Determination of three-dimensional structure by multiwavelength anomalous diffraction" *Curr. Biol.*, pp. 1002–1011, 1991.
[29] M. Cianci, M. R. Groves, D. Barford, and T. R. Schneider, "Data collection with a tailored X-ray beam size at 2.69 Å wavelength (4.6 keV): sulfur SAD phasing of Cdc23 ^{Nterm}" *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 72, no. 3, pp. 403–412, 2016.
[30] Z. Zhang, L. Chang, J. Yang, N. Conin, K. Kulkarni, and D. Barford, "The Four Canonical TPR Subunits of Human APC/C Form Related Homo-Dimeric Structures and Stack in Parallel to Form a TPR Suprahelix" *J. Mol. Biol.*, vol. 425, no. 22, pp. 4236–4248, 2013.
[31] P. Emsley and B. Lohkamp, "Features and development of Coot" *Acta Crystallogr. Sect. D*, pp. 486–501, 2010.
[32] T. Pape and T. R. Schneider, "HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs" *J. Appl. Crystallogr.*, vol. 37, no. 5, pp. 843–844, 2004.
[33] M. D. Winn *et al.*, "Overview of the CCP4 suite and current developments" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 4, pp. 235–242, 2011.
[34] G. N. Murshudov *et al.*, "REFMAC5 for the refinement of macromolecular crystal structures" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 4, pp. 355–367, 2011.
[35] T. R. Schneider and G. M. Sheldrick, "Substructure solution with SHELXD" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 58, no. 10 I, pp. 1772–1779, 2002.
[36] J. Schrader, F. Henneberg, R. A. Mata, K. Tittmann, T. R. Schneider, H. Stark, G. Bourenkov and A. Chari, "The inhibition mechanism of human 20S proteasomes enables next-generation inhibitor design" *Science*, vol. 353, no. 6299, pp. 1–6, 2016.
[37] T. C. Terwilliger, "Rapid model building of alpha-helices in electron-density maps" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 3, pp. 268–275, 2010.
[38] T. C. Terwilliger, "Rapid model building of beta-sheets in electron-density maps" *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 3, pp. 276–284, 2010.
[39] P. H. Zwart, "Anomalous signal indicators in protein

crystallography research papers Anomalous signal indicators in protein crystallography" *Acta Crystallogr. Sect. D*, pp. 1437–1448, 2005.

[40] Z. Dauter, "Estimation of anomalous signal in diffraction data" *Acta Crystallogr. Sect. D*, pp. 867–876, 2006.

[41] P. A. Karplus and K. Diederichs, "ScienceDirect Assessing and maximizing data quality in macromolecular crystallography" *Curr. Opin. Struct. Biol.*, vol. 34, pp. 60–68, 2015.

[42] G. Bricogne, C. Vonrhein, C. Flensburg, M. Schiltz, and W. Paciorek, "Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0" *Acta Crystallogr. Sect. D*, pp. 2023–2030, 2003.

[43] C. Vonrhein, E. Blanc, P. Roversi, and G. Bricogne, "Automated Structure Solution With autoSHARP" *Methods Mol. Biol. Macromol. Crystallogr. Protoc. Vol. 2 Struct. Determ.*, vol. 364, no. 1.

[44] A. Thorn and G. M. Sheldrick, "ANODE: anomalous and heavy-atom density calculation" *J. Appl. Cryst.*, pp. 1285–1287, 2011.

[45] J. Steyaert and B. K. Kobilka, "Nanobody stabilization of G protein-coupled receptor conformational states" *Curr. Opin. Struct. Biol.*, vol. 21, no. 4, pp. 567–572, 2011.

[46] A. Manglik, B. K. Kobilka, and J. Steyaert, "Nanobodies to Study G Protein-Coupled Receptor Structure and Function" *Annu. Rev. Pharmacol. Toxicol.*, 2017.

[47] T. Che *et al.*, "Structure of the Nanobody-Stabilized Active State of the Kappa Opioid Receptor" *Cell*, vol. 172, no. 1–2. pp. 55-67.e15, 2018.

[48] E. Pardon *et al.*, "A general protocol for the generation of Nanobodies for structural biology" *Nat. Protoc.*, vol. 9, no. 3, pp. 674–693, 2014.

[49] I. Zimmermann *et al.*, "Synthetic single domain antibodies for the conformational trapping of membrane proteins" *Elife*, vol. 3, pp. 1–32, 2018.

[50] C. Mcmahon *et al.*, "Yeast surface display platform for rapid discovery of conformationally selective nanobodies" *Nat. Struct. Mol. Biol.*, vol. 25, no. March, 2018.

[51] S. G. F. Rasmussen *et al.*, "Structure of a nanobody-stabilized active state of the b2 adrenoceptor" *Nature*, vol. 469, no. 7329, pp. 175–180, 2011.

[52] E. J. De Genst *et al.*, "Structure and Properties of a Complex of α-

Synuclein and a Single-Domain Camelid Antibody" *J. Mol. Biol.*, vol. 402, no. 2, pp. 326–343, 2010.

[53] D. Monté *et al.*, "Crystal structure of human Mediator subunit MED23" *Nat. Commun.*, no. 2018, pp. 1–7.

[54] A. M. Ring *et al.*, "Adrenaline-activated structure of the β2-adrenoceptor stabilized by an engineered nanobody" *Nature*, vol. 502, no. 7472, pp. 575–579, 2014.

[55] E. Errasti-murugarren *et al.*, "L-amino acid transporter structure and molecular bases for the asymmetry of substrate interaction" *Nat. Commun.*, pp. 1–12, 2019.

[56] A. Koehl *et al.*, "Structural insights into the activation of metabotropic glutamate receptors" *Nature*, 2019.

[57] JJ Ruprecht *et al.*, "The Molecular Mechanism of Transport by the Article The Molecular Mechanism of Transport by the Mitochondrial ADP/ATP Carrier" *Cell*, vol. 176, pp. 435–447, 2019.

[58] X. Jiang *et al.*, "Crystal structure of a LacY – nanobody complex in a periplasmic-open conformation" *PNAS*, vol. 113, no. 44, pp. 1–6, 2016.

[59] W. A. Hendrickson, "Anomalous Diffraction in Crystallographic Phase Evaluation" *Q Rev Biophys*, vol. 47, no. 1, pp. 49–93, 2014.

[60] V. Delft, "An overview of heavy-atom derivatization of protein crystals" *Acta Crystallogr. Sect. D*, vol. 72, pp. 303–318, 2016.

[61] G. Pompidor, O. Maury and R. Kahn "A dipicolinate lanthanide complex for solving protein structures using anomalous diffraction" *Acta Crystallogr. Sect. D*, vol. 66, pp. 762–769, 2010.

[62] Â. Girard and M. Stelter, "A new class of lanthanide complexes to obtain high-phasing-power heavy-atom derivatives for macromolecular crystallography" *Acta Crystallogr. Sect. D*, pp. 1914–1922, 2003.

[63] W. Kabsch, "XDS," *Acta Crystallogr. Sect. D*, vol. 66, pp. 125–132, 2010.

[64] J. Chiu, R. Desalle, H. Lam, L. Meisel, and G. Coruzzi, "Molecular Evolution of Glutamate Receptors: A Primitive Signaling Mechanism that Existed Before Plants and Animals Diverged" *Mol. Biol. Evol. 16(6)826–838.*, vol. 16, no. 6, pp. 826–838, 1998.

[65] S. F. Traynelis *et al.*, "Glutamate Receptor Ion Channels: Structure, Regulation, and Function" *Pharmacol. Rev.*, vol. 62, no. 3, pp. 405–496, 2010.

[66] J. Kumar and M. L. Mayer, "Functional Insights from Glutamate

Receptor Ion Channel Structures" *Annu. Rev. Physiol.*, vol. 75, pp. 313–337, 2013.

[67] M. L. Mayer, "The Challenge of Interpreting Glutamate-Receptor Ion-Channel Structures" *Biophys. J.*, vol. 1, pp. 2143–2151, 2017.

[68] I. H. Greger, J. F. Watson, and S. G. Cull-candy, "Structural and Functional Architecture of AMPA-Type Glutamate Receptors and Their Auxiliary Proteins" *Neuron Rev.*, pp. 713–730, 2017.

[69] K. B. Hansen *et al.*, "Structure, function, and allosteric modulation of NMDA receptors" *J. Gen. Physiol.*, vol. 150, no. 8, 2018.

[70] S. K. Singh, C. Chien, and I. Chang, "The Arabidopsis glutamate receptor-like gene GLR3.6 controls root development by repressing the Kip-related protein gene KRP4" *J. Exp. Bot.*, vol. 67, no. 6, pp. 1853–1869, 2016.

[71] C. Dubos, D. Huggins, G. H. Grant, M. R. Knight, and M. M. Campbell, "A role for glycine in the gating of plant NMDA-like receptors" *Plant J.*, vol. 35, pp. 800–810, 2003.

[72] Y. Cheng, X. Zhang, T. Sun, Q. Tian, and W. Zhang, "Glutamate Receptor Homolog3.4 is Involved in Regulation of Seed Germination Under Salt Stress in Arabidopsis" *Plant Cell Physiol.*, vol. 59, no. February, pp. 978–988, 2018.

[73] S. A. Mousavi, A. Chauvin, F. Pascaud, S. Kellenberger and EE Farmer "GLUTAMATE RECEPTOR-LIKE genes mediate leaf-to-leaf wound signalling" *Nature*, pp. 422–426, 2013.

[74] M. Toyota, D. Spencer, S. Sawai-toyota, W. Jiaqi, and T. Zhang, "Glutamate triggers long-distance, calcium-based plant defense signaling" *Science*, vol. 6, no. September, pp. 1112–1115, 2018.

[75] C. Tam, A. Kurenda, S. Stolz, A. Chételat, and E. E. Farmer, "Identification of cell populations necessary for leaf-to- leaf electrical signaling in a wounded plant" *PNAS*, vol. 115, pp. 10178–10183, 2018.

[76] E. Michard *et al.*, "Glutamate receptor-like genes form Ca^{2+} channels in pollen tubes and are regulated by pistil D-serine" *Science*, vol. 332, pp. 434–437, 2011.

[77] M. M. Wudick *et al.*, "CORNICHON sorting and regulation of GLR channels underlie pollen tube Ca^{2+} homeostasis" *Science*, vol. 536, no. May, pp. 533–536, 2018.

[78] D. Cho *et al.*, "De-regulated expression of the plant glutamate receptor homolog AtGLR3.1 impairs long-term Ca^{2+}-programmed stomatal closure" *Plant J.*, vol. 58, pp. 437–449, 2009.

[79] E. Okuma *et al.*, "L-Met Activates Arabidopsis GLR Ca^{2+} Channels

Upstream of ROS Production and Regulates Stomatal Movement" *Cell Rep.*, vol. 17, pp. 2553–2561, 2016.

[80] Z. Qi, N. R. Stephens, and E. P. Spalding, "Calcium Entry Mediated by GLR3.3, an Arabidopsis Glutamate Receptor with a Broad Agonist Profile 1[W][OA]" *Plant Physiol.*, vol. 142, no. November, pp. 963–971, 2006.

[81] G. L. R. G. W. Oa, N. R. Stephens, Z. Qi, and E. P. Spalding, "Glutamate Receptor Subtypes Evidenced by Differences in Desensitization and Dependence on the GLR3.3 and and GLR3.4 Genes1[W][OA]" *Plant Physiol.*, vol. 146, no. February, pp. 529–538, 2008.

[82] E. D. Vincill, A. M. Bieck, and E. P. Spalding, "Ca^{2+} Conduction by an Amino Acid-Gated Ion Channel Related to Glutamate Receptors" *Plant Physiol.*, vol. 159, no. May, pp. 40–46, 2012.

[83] D. Tapken *et al.*, "A Plant Homolog of Animal Glutamate Receptors Is an Ion Channel Gated by Multiple Hydrophobic Amino Acids" *Sci. Signal.*, vol. 6, no. 279, pp. 1–11, 2013.

[84] I. Responses *et al.*, "Glutamate Receptor-Like Channel3.3 Is Involved in Mediating Glutathione-Triggered Cytosolic Calcium Transients, Transcriptional Changes, and Innate Immunity Responses in Arabidopsis1[W][OA]" *Plant Physiol.*, vol. 162, no. July, pp. 1497–1509, 2013.

[85] P. R. Evans *et al.*, "CCP4i2: the new graphical user interface to the CCP4 program suite" *Acta Crystallogr. Sect. D*, vol. 74, no. August 2017, pp. 68–84, 2018.

[86] E. Blanc, C. Vonrhein, C. Flensburg, S. M. Lea, and G. Bricogne, "Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT" *Acta Crystallogr. Sect. D*, vol. 60, pp. 2210–2221, 2004.

[87] M. W. Bowler, A. R. Correia, S. Larsen, G. A. Leonard, and A. A. Mccarthy, "A new MR-SAD algorithm for the automatic building of protein models from low-resolution X-ray data and a poor starting model" *IUCrJ*, vol. 5, pp. 166–171, 2018.

[88] A. Vagin, A., Teplyakov, "MOLREP: an Automated Program for Molecular Replacement" *J. Appl. Cryst. (1997).*, pp. 1022–1025, 1997.

[89] A. Thorn and M. George, "Extending molecular-replacement solutions with SHELXE" *Acta Crystallogr. Sect. D*, vol. 69, pp. 2251–2256, 2013.

[90] V. B. Chen *et al.*, "MolProbity: all-atom structure validation for

macromolecular crystallography" *Acta Crystallogr. Sect. D*, vol. 66, pp. 12–21, 2010.

[91] L. Holm and L. M. Laakso, "Dali server update" *Nucleic Acids Res.*, vol. 44, no. April, pp. 351–355, 2016.

[92] S. F. Altschul *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs" *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[93] R. Schwarzenbacher, A. Godzik, and K. Slawomir, "The importance of alignment accuracy for molecular replacement" *Acta Crystallogr. Sect. D*, vol. 60, pp. 1229–1236, 2004.

[94] A. J. Mccoy, R. W. Grosse-kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read, "Phaser crystallographic software" *J. Appl. Cryst.*, vol. 40, pp. 658–674, 2007.

[95] L. Rychlewski, L. Jaroszewski, W. Li, and A. Godzik, "Comparison of sequence profiles . Strategies for structural predictions using sequence information" *Protein Sci.*, vol. 9, pp. 232–241, 2000.

[96] G. Bunkózci and R. J. Read, "Improvement of molecular-replacement models with Sculptor" *Acta Crystallogr. Sect. D*, vol. 67, pp. 303–312, 2011.

[97] R. Mosca and T. R. Schneider, "RAPIDO: a web server for the alignment of protein structures in the presence of conformational changes" *Nucleic Acids Res.*, vol. 36, no. May, pp. 42–46, 2008.

[98] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, and M. J. E. Sternberg, "The Phyre2 web portal for protein modeling , prediction and analysis" *Nat. Protoc.*, vol. 10, no. 6, pp. 845–858, 2015.

[99] Y. Zhang, "I-TASSER server for protein 3D structure prediction" *BMC Bioinformatics*, vol. 8, pp. 1–8, 2008.

[100] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints" pp. 779–815, 1993.

[101] T. C. Terwilliger, F. Dimaio, R. J. Read, D. Baker, N. Echols, and Á. R. Á. Phenix, "phenix.mr_rosetta: molecular replacement and model rebuilding with Phenix and Rosetta" *J Struct Funct Genomics*, vol. 13, pp. 81–90, 2012.

[102] F. Long, A. A. Vagin, and G. N. Murshudov, "BALBES: a molecular-replacement pipeline" *Acta Crystallogr. Sect. D*, vol. 64, pp. 125–132, 2008.

[103] T. E. Automated, C. Structure, V. Parthasarathy, V. S. Lamzin, S. Manfred, and P. A. Tucker, "Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment" *Acta Crystallogr.*

Sect. D, vol. 61, pp. 449–457, 2005.
[104] A. A. Lebedev and M. N. Isupov, "Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program Zanuda" *Acta Crystallogr. Sect. D*, vol. 70, pp. 2430–2443, 2014.
[105] G. Winter *et al.*, "DIALS: implementation and evaluation of a new integration package" *Acta Crystallogr. Sect. D*, vol. 74, pp. 85–97, 2018.
[106] F. J. Moreno and A. Clemente, "2S Albumin Storage Proteins : What Makes them Food Allergens ?" *Open Biochem. J.*, pp. 16–28, 2008.
[107] M. Kreis and P. R. Shewry, "Unusual Features of Cereal Seed Protein Structure and Evolution" *BioEssays*, vol. 10, no. 6, pp. 201–207, 1989.
[108] N. Wolff *et al.*, "Identification and characterization of linear B-cell epitopes of b -globulin , a major allergen of sesame seed," 2004.
[109] S. S. K. Tai, T. T. T. Lee, C. C. Y. Tsai, T. Yiu, and J. T. C. Tzen, "Expression pattern and deposition of three storage proteins, 11S globulin, 2S albumin and 7S globulin in maturing sesame seeds" *Plant Physiol. Biochem.*, vol. 39, pp. 981–992, 2001.
[110] K. Beyer, L. Bardina, G. Grishina, H. A. Sampson, and N. York, "Identification of sesame seed allergens by 2-dimensional proteomics and Edman sequencing: Seed storage proteins as common food allergens" *J ALLERGY CLIN IMMUNOL*, vol. 110, no. 1, pp. 154–159, 2002.
[111] D. D. Rodríguez *et al.*, "Crystallographic ab initio protein structure solution below atomic resolution" *Nat. Methods*, vol. 6, no. 9, pp. 651–653, 2009.
[112] G. M. Sheldrick, "Macromolecular phasing with SHELXE" *Z. Krist.*, vol. 217, pp. 644–650, 2002.
[113] I. Usón, "An introduction to experimental phasing of macromolecules illustrated by SHELX; new autotracing features" *Acta Crystallogr. Sect. D*, vol. 74, pp. 106–116, 2018.
[114] R. J. Read, A. J. Mccoy, and I. Usón, "On the application of the expected log-likelihood gain to decision making in molecular replacement" *Acta Crystallogr. Sect. D*, vol. 74, pp. 245–255, 2018.
[115] A. J. Mccoy, "Acknowledging Errors: Advanced Molecular Replacement with Phaser"
[116] K. Pro, M. Sammito, and I. Usón, "Structure solution of DNA-binding proteins and complexes with ARCIMBOLDO libraries" *Acta Crystallogr. Sect. D*, vol. 70, pp. 1743–1757, 2014.

[117] D. Sammito, A. J. Mccoy, C. Milla, A. F. Z. Nascimento, G. Petrillo, and R. D. Oeffner, "Exploiting distant homologues for phasing through the generation of compact fragments, local fold refinement and partial solution combination" *Acta Crystallogr. Sect. D*, vol. 74, pp. 290–304, 2018.
[118] M. Sammito and K. Meindl, "Structure solution with ARCIMBOLDO using fragments derived from distant homology models" *FEBS J.*, vol. 281, pp. 4029–4045, 2014.
[119] M. Sammito *et al.*, "Exploiting tertiary structure through local folds for crystallographic phasing" *Nat. Methods*, vol. 10, no. 11, pp. 1099–1101, 2013.
[120] K. Meindl, G. M. Sheldrick, and I. Usón, "Practical structure solution with ARCIMBOLDO" *Acta Crystallogr. Sect. D*, vol. 68, pp. 336–343, 2012.
[121] T. C. Terwilliger, "Maximum-likelihood density modification" *Acta Crystallogr. Sect. D*, vol. 56, pp. 965–972, 2000.
[122] E. N. C. Millst, W. Trust, and G. Campus, "Plant protein families and their relationshipsto food allergy" *Biochem. Soc. Trans.*, vol. 30, pp. 906–910, 2002.
[123] P. F. Allergens, M. Bruix, J. Santoro, R. Monsalvet, M. Villalbaf, and D. Q. F. Rocasolano, "Solution structure of allergenic 2 S albumins" *Plant Food Allergens*, pp. 919–924, 2002.
[124] A. Waterhouse *et al.*, "SWISS-MODEL: homology modelling of protein structures and complexes" *Nucleic Acids Res.*, vol. 46, no. May, pp. 296–303, 2018.
[125] D. Xu and Y. Zhang, "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field" *Proteins*, no. January, pp. 1715–1735, 2012.
[126] R. M. Keegan and D. Martyn, "MrBUMP: an automated pipeline for molecular replacement research papers" *Acta Crystallogr. Sect. D*, vol. 64, pp. 119–124, 2008.
[127] K. Hatti, Y. K. Mathiharan, N. Srinivasan, and M. R. N. Murthy, "Seeing but not believing : the structure of glycerol dehydrogenase initially assumed to be the structure of a survival protein from Salmonella typhimurium" *Acta Crystallogr. Sect. D*, vol. 73, pp. 609–617, 2017.
[128] P. Musille and E. Ortlund, "Structure of glycerol dehydrogenase from Serratia" *Acta Crystallogr. Sect. F*, vol. 70, pp. 166–172, 2014.
[129] A. Butryn, G. Stoehr, and C. Linke-winnebeck, "Serendipitous crystallization and structure determination of cyanase (CynS) from

Serratia proteamaculans" *Acta Crystallogr. Sect. F*, vol. 71, pp. 471–476, 2015.

[130] D. W. Ritchie and S. Grudinin, "Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry" *J. Appl. Cryst.*, vol. 49, pp. 158–167, 2016.

[131] Y. Yan, H. Tao, and S. Huang, "HSYMDOCK: a docking web server for predicting the structure of protein homo-oligomers with C_n or D_n symmetry" *Nucleic Acids Res.*, vol. 46, no. May, pp. 423–431, 2018.

[132] P. Bradley, C. Wang, and D. Baker, "Prediction of the structure of symmetrical protein assemblies" *PNAS*, vol. 104, pp. 17656–17661, 2007.

[133] J. Ko, H. Park, L. Heo, and C. Seok, "GalaxyWEB server for protein structure prediction and refinement" *Nucleic Acids Res.*, vol. 40, no. May, pp. 294–297, 2012.

[134] D. Schneidman-duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "PatchDock and SymmDock: servers for rigid and symmetric docking" *Nucleic Acids Res.*, vol. 33, pp. 363–367, 2005.

[135] B. G. Pierce, K. Wiehe, H. Hwang, B. Kim, T. Vreven, and Z. Weng, "ZDOCK server: interactive docking prediction of protein – protein complexes and symmetric multimers" *Struct. Bioinforma.*, vol. 30, no. 12, pp. 1771–1773, 2014.

[136] A. J. Simpkin, F. Simkovic, J. M. H. Thomas, and M. Savko, "SIMBAD: a sequence-independent molecular- replacement pipeline" *Acta Crystallogr. Sect. D*, vol. 74, pp. 595–605, 2018.

[137] G. Chojnowski, J. Pereira, and V. S. Lamzin, "Sequence assignment for low-resolution modelling of protein crystal structures" *Acta Crystallogr. Sect. D*, vol. 75, pp. 753–763, 2019.

[138] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching" *Nucleic Acids Res.*, pp. 1–9, 2011.

[139] I. Stock, S. Burak, K. J. Sherwood, T. Grüger, and B. Wiedemann, "Natural antimicrobial susceptibilities of strains of 'unusual' Serratia species: S. ficaria, S. fonticola, S. odorifera , S. plymuthica and S. rubidaea" *J. Antimicrob. Chemother.*, no. March, pp. 865–885, 2003.

[140] R. V. and K. G. Guennadi Kozlov, "Triosephosphate isomerase is a common crystallization contaminant of soluble His-tagged proteins produced in Escherichia coli" *Acta Crystallogr. Sect. F*, vol. 69, pp. 499–502, 2013.

[141] A. Hungler, A. Momin, K. Diederichs and T. Arold, “ContaMiner and ContaBase: a webserver and database for early identification of unwantedly crystallized protein contaminants” *J. Appl. Cryst.*, vol. 49, pp. 2252–2258, 2016.
[142] J. J. Lipuma, “The Changing Microbial Epidemiology in Cystic Fibrosis,” *Clin. Microbiol. Rev.*, vol. 23, no. 2, pp. 299–323, 2010.
[143] A. C. Cheng *et al.*, “Clinical Definitions of Melioidosis” *Am. J. Trop. Med. Hyg.*, vol. 88, no. 3, pp. 411–413, 2013.
[144] C. Peri *et al.*, “Evolving serodiagnostics by rationally designed peptide arrays: the Burkholderia paradigm in Cystic Fibrosis” *Sci. Rep.*, no. August, pp. 1–11, 2016.
[145] J. A. Musson *et al.*, “CD4+ T cell epitopes of FliC conserved between strains of Burkholderia: implications for vaccines against melioidosis and cepacia complex in cystic fibrosis” *J Immunol.*, vol. 193, no. 12, pp. 6041–6049, 2014.
[146] A. Gori *et al.*, “Article Exploiting the Burkholderia pseudomallei Acute Phase Antigen BPSL2765 for Structure-Based Epitope Discovery / Design in Structural Vaccinology” *Chem. Biol.*, vol. 20, pp. 1147–1156, 2013.
[147] R. Capelli, F. Marchetti, G. Tiana, and G. Colombo, “SAGE: A Fast Computational Tool for Linear Epitope Grafting onto a Foreign Protein Scaffold” *J. Chem. Inf. Model.*, vol. 57, pp. 6–10, 2017.
[148] T. C. Terwilliger, “Using prime-and-switch phasing to reduce model bias in molecular replacement” *Acta Crystallogr. Sect. D*, vol. 60, pp. 2144–2149, 2004.
[149] O. Kovalevskiy, R. A. Nicholls, and G. N. Murshudov, “Automated refinement of macromolecular structures at low resolution using prior information” *Acta Crystallogr. Sect. D*, vol. 72, pp. 1149–1161, 2016.
[150] A. M. Karmali and L. Tom, “Model-building strategies for low-resolution X-ray crystallographic data” *Acta Crystallogr. Sect. D*, vol. 65, pp. 121–127, 2009.
[151] P. V Afonine, V. B. Chen, W. Nigel, and D. C. Richardson, “Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution” *Acta Crystallogr. Sect. D*, vol. 68, no. 1991, pp. 381–390, 2012.

ACKNOWLEDGMENTS

First, I would like to thank Dr. Thomas Schneider for the opportunity to work with him on an interesting project, as well as for teaching me to be more critical and independent in my work. This is something that has been and will be extremely useful in my personal growth as a scientist and in my future career.

At the same time, I want to thank Prof. Bolognesi for accepting to be my academic supervisor, for introducing me to his group and for giving me the chance to be involved in a number of interesting projects. Despite the physical distance, I have always felt his support to my work and I have always appreciated his positive attitude, even in difficult times.

I am also thankful to the members of my Thesis Advisory Committee: Dr. Rob Meijers, Dr. Christoph Müller and Dr. Ashwin Chari for their valuable suggestions throughout the project.

In my daily work, I was constantly supported by some people, particularly Dr. Fabio dall'Antonia and Dr. Gleb Bourenkov. Fabio introduced me to Python programming but also shared with me his vast knowledge of Crystallography. I will never forget his patience, dedication, openness and willingness to discuss and to share his expertise. As Fabio, Gleb was useful in many aspects of my work, which reflects his enormous theoretical and practical understanding of Crystallography. Their many qualities, both on the scientific and personal level, are a source of inspiration for me, and they will certainly guide me in the future.

Last but not least, I want to mention some colleagues who contributed to make my PhD a beautiful time but also helped me in my daily work: Dr.

Spyros Chatziefthimiou, Dr. Vivian Pogenberg, Dr. Nabil Hanna, Karen Manalastas, Sandra Kozak, Dr. Amal Hassan, Dr. Philipp Hornburg, Dr. Maria Martinez Molledo, Dr. Thomas Seine, Dr. Xuefan Gao, Andrea Amato, Marco Camerlenghi and Dr. Eduard Avetisyan.

APPENDIX MANUSCRIPTS

"Genetic and structure reveal *Arabidopsis thaliana* GLR3.3 as an amino acid receptor with distinct ligand specificity"

Personal contribution:

In the context of this work I was able to solve and partly refine the structure of the GLR3.3 LBD in complex with *L*-Glu at 2.0 Å resolution. My contribution to the manuscript is important for, at least, three reasons: first of all, this structure represents the first plant glutamate receptor model and its determination marks an important result also in light of the fact that it has resisted many attempts of structure solution for a long time. The determination of this structure was possible only after a *tour de force* phasing procedure which entailed the use of a novel MRSAD algorithm and was further complicated by the non-optimal quality of both the native and anomalous data. In the second place, thanks to the first model of GLR3.3 LBD, three more structures with different natural ligands could be easily solved and more are likely to be obtained in the future. Thirdly, the GLR3.3 LBD model and the structures obtained with different ligands can be used to derive important biological information about, for instance, the evolution and the physiology of glutamate receptors. The data for this work were kindly provided by Dr. Andrea Alfieri who gave me all the crystallographic data sets and other types of information concerning the *At*GLR3.3 construct.

The manuscript is under revision on the *Proceedings of the National Academy of Sciences* (*PNAS*).

Genetics and structure reveal *Arabidopsis thaliana* GLR3.3 as an amino acid receptor with distinct ligand specificity

Andrea Alfieri[a,1,2], Fabrizio G. Doccula[a,3], Riccardo Pederzoli[a,b,3], Matteo Grenzi[a], Maria Cristina Bonza[a], Laura Luoni[a], Alessia Candeo[c], Neli Romano Armada[c,d], Alberto Barbiroli[e], Gianluca Valentini[c], Thomas R. Schneider[b], Andrea Bassi[c], Martino Bolognesi[a,f], Marco Nardini[a], Alex Costa[a,g,1]

[a] Department of Biosciences, University of Milan, via Celoria 26, 20133 Milano, Italy
[b] Hamburg Unit c/o DESY, European Molecular Biology Laboratory, Notkestrasse 85, 22603 Hamburg, Germany
[c] Department of Physics, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy.
[d] INIQUI and Faculty of Engineering, National University of Salta, Av. Bolivia 5150, 4400 Salta, Argentina
[e] Department of Food, Environmental and Nutritional Sciences, University of Milan, via Celoria 2, 20133 Milano, Italy
[f] Pediatric Research Center "Romeo ed Enrica Invernizzi", University of Milan, via Celoria 26, 20133 Milano, Italy
[g] Institute of Biophysics, National Research Council of Italy (CNR), 20133 Milano, Italy

[1] Correspondence to: Andrea Alfieri (andrea.alfieri@unimi.it, phone 00390250314831) and Alex Costa (alex.costa@unimi.it, phone 00390250314831)
[2] Present address: Centro Grandi Strumenti, University of Pavia, via Bassi 21, 27100 Pavia, Italy
[3] These authors contributed equally to this work

E-mails:
AA andrea.alfieri@unimi.it, FGD fabriziogandolfo.doccula@gmail.com, RP rpederzoli@embl-hamburg.de, MG matteo.grenzi@unimi.it, MCB cristina.bonza@unimi.it, LL laura.luoni@unimi.it, ACa alessia.candeo@polimi.it, GV gianluca.valentini@polimi.it, NRA nelir000@gmail.com, ABar alberto.barbiroli@unimi.it, TRS thomas.schneider@embl-hamburg.de, MB martino.bolognesi@unimi.it, ABas andreabassi@polimi.it, MN marco.nardini@unimi.it, ACo alex.costa@unimi.it

ORCID identifiers: Andrea Alfieri 0000-0002-9716-0339; Fabrizio G. Doccula 0000-0002-4499-2297; Riccardo Pederzoli 0000-0002-2286-9538; Maria Cristina Bonza 0000-0001-7096-2967; Laura Luoni 0000-0002-6485-5417; Alessia Candeo 0000-0001-9597-3056; Gianluca Valentini 0000-0002-6340-3021; Andrea Bassi 0000-0002-5017-0775; Martino Bolognesi 0000-0002-9253-5170; Marco Nardini 0000-0002-3718-2165; Alex Costa 0000-0002-2628-1176.

Classification: BIOLOGICAL SCIENCES - Plant biology

Keywords: GLR channels, X-ray crystallography, binding assay, modelling, Ca^{2+} signalling, *in vivo* imaging.

ABSTRACT

A. thaliana Glutamate Receptor-Like (GLR) channels are nonselective cation channels involved in physiological processes including wound signalling, stomatal regulation and pollen tube growth. They are supposedly gated by exogenous amino acid ligands. Using light-sheet fluorescence microscopy and the genetically-encoded Ca^{2+} sensor Cameleon YC3.6, we identified a subpopulation of cells involved in the amino acid-elicited cytosolic Ca^{2+} increase in *Arabidopsis* root tips. Knock-out lines provided genetic evidence of the central role of GLR3.3 in this response, supporting the notion of GLRs as amino acid-gated channels in root tip cells. To elucidate its binding properties, we biochemically reconstituted the GLR3.3 ligand-binding domain (LBD) and analyzed its selectivity profile with binding experiments, which revealed its preference not only for L-Glu but also for sulfur-containing amino acids. Furthermore, we determined the crystal structures of the GLR3.3 LBD in complexes with four different amino acid ligands. Our structural analyses pinpoint the residues involved in ligand binding and lay the grounds for rational mutagenesis. In addition, we analyzed structures of LBDs from non-plant species and generated working models for other GLR isoforms. Our results prove the GLR3.3 receptor potential and provide a structural framework for engineering this and other GLR isoforms to investigate their physiology.

SIGNIFICANCE STATEMENT

GLR channels are plant homologs of glutamate receptors in vertebrate synapses; they are putative calcium-permeable channels involved in root and pollen tube growth, stomatal regulation, wound signalling. This study presents the first crystal structure of a plant GLR ligand-binding domain (LBD) in complex with four different amino acid ligands and identifies the protein residues responsible for amino acid binding. Related binding assays show that those amino acids that trigger GLR-mediated calcium influx in *Arabidopsis thaliana* root tip cells bind the GLR LBD with micromolar affinities.

INTRODUCTION

Plant Glutamate Receptor-Like (GLR) channels are plant homologs of mammalian ionotropic Glutamate Receptors (iGluRs)(1). iGluRs are homo- or heterotetrameric cation channels activated by the neurotransmitters L-glutamate, glycine and D-serine released in the synaptic space. They are extensively studied for their central role in neurotransmission, learning and memory (2).

The identification of iGluRs homologs in other eukaryotes, including invertebrates and plants, and cyanobacteria has outlined the existence of a large family of Glutamate Receptor-Like proteins (GLRs) across all kingdoms of life. In particular, the stoichiometry and architecture of plant GLRs is believed to be similar to iGluRs (3): each subunit hosts an extracellular aminoterminal domain (ATD), an extracellular ligand-binding domain (LBD) composed of segments S1 and S2, four transmembrane helices (M1 to M4, one of which - M2 - is not fully transmembrane), and a cytoplasmic tail (CTD), arranged in the order ATD-S1-M1-M2-M3-S2-M4-CTD (*SI Appendix*, Fig. S1*A* and *B*). The LBD has a conserved clamshell-shaped architecture resembling the periplasmic binding protein-like II superfamily in bacteria (4); in vertebrates, the binding of a ligand/agonist induces a variable degree of closure of the clamshell that pulls the transmembrane segments and opens the channel pore (2).
The 20 *Arabidopsis thaliana* GLR isoforms are grouped in 3 clades (5, 6). Specific isoforms have been implicated in several physiological processes, such as root growth (7), hypocotyl elongation (8), seed germination (9), long-distance wound signalling (10–12), pollen tube growth (13, 14), stomatal aperture (15, 16), as well as Ca^{2+} signalling (17–20); such isoforms are then considered putative Ca^{2+}-permeable channels. In particular, the *A. thaliana* GLR3.3 isoform has been studied for its role in amino acid-induced cytosolic Ca^{2+} ($[Ca^{2+}]_{cyt}$) increases (17, 21), and recently recognized as a key player in glutamate-mediated defense signalling (11). Despite genetic data supporting the role of GLRs as amino acid receptors (11, 16–18, 20–22), there is no experimental evidence that any plant GLR isoform can indeed bind glutamate or other ligands. Furthermore, whereas for iGluRs hundreds of X-ray structures are available for the LBD moiety (23), and an increasing number of cryo-electron microscopy full-length structures are accumulating (24–27)(*SI Appendix*, Fig. S1*B*), no structural information for any plant GLR isoform is available to date.

Therefore, in the present study we set out to investigate the role of *A. thaliana* GLR3.3 in the generation of amino acid-elicited cytosolic Ca^{2+} transients, and reconstituted its ligand-binding domain *in vitro*. The determination of its selectivity profile by binding assays allowed us to identify the transient-eliciting amino acids as high-affinity ligands of the GLR3.3 LBD. Furthermore, we solved the crystal structures of the GLR3.3 LBD in complexes with four representative ligands (L-glutamate, glycine, L-cysteine, L-methionine), providing novel structural information on plant GLR LBDs and a rational explanation for our *in vitro* affinity and *in vivo* functional data. Taken together, the reported results will guide rational mutagenesis *in planta* aimed at interfering with the GLRs binding specificities, to dissect their physiological properties.

RESULTS

Arabidopsis *root tip cells sense exogenous amino acids by GLR3.3*

Glutamate (L-Glu) and other amino acids can elicit $[Ca^{2+}]_{cyt}$ increases and plasma membrane (PM) depolarization in *A. thaliana* and rice seedlings (17, 18, 20, 21, 28), as well as activate currents in the PM of guard cells (16) and pollen tubes (13). Genetic and pharmacological data provide evidence that the GLRs, working as ligand-gated channels, are responsible for the amino acid sensing and for the effects reported above (13, 16–20).

To strengthen the evidence that GLRs link the amino acid perception to the $[Ca^{2+}]_{cyt}$ increase, we studied at high spatial resolution, by means of Light Sheet Fluorescence Microscopy (LSFM), the amino acids-elicited $[Ca^{2+}]_{cyt}$ dynamics in *Arabidopsis* root tip cells of Col-0 plants expressing the genetically-encoded Ca^{2+} sensor NES-YC3.6 (29–31) (Fig. 1 and *SI Appendix*, Fig. S2). The rationale behind the choice of amino acids was based on the current literature (17, 20) and on experiments in which we evaluated, by means of FRET-based wide field fluorescence microscopy analyses, the $[Ca^{2+}]_{cyt}$ transients in *Arabidopsis* root tip in response to the L-enantiomers of all 20 amino acids. In these experiments only seven amino acids, L-Cys, L-Glu, L-Ala, Gly, L-Ser, L-Asn and, to minor extent, L-Met were able to trigger $[Ca^{2+}]_{cyt}$ increases in root tip cells (L-Met evoked a small but clear FRET signal, which was not detected for any of the remaining 13 amino acids) (*SI Appendix*, Fig. S2). Our results confirm those previously obtained in *Arabidopsis* with an earlier Cameleon version and/or aequorin-expressing plants (17, 18). The use of LSFM was here pursued to understand whether the $[Ca^{2+}]_{cyt}$ transients evoked by the seven amino acids were occurring in the same cell types, since this piece of information was missing in previous published works (17, 18, 21). We started our survey with 1 mM L-Glu administration, which triggered a $[Ca^{2+}]_{cyt}$ increase occurring primarily in the lateral cells of the root meristem, then spreading towards the inner cells of the stele (Fig. 1*A*, *B* and *SI Appendix*, Fig. S3). We then tested the other six amino acids, supplied at the same final concentration and in all cases, except for L-Met which failed to show a $[Ca^{2+}]_{cyt}$ increase, the same outer meristematic cells responded (*SI Appendix*, Fig. S3). Although the $[Ca^{2+}]_{cyt}$ increase induced by L-Met was only detectable by wide field fluorescence microscopy (Fig. 1*E* and *F*), these results show that the same meristematic root tip cells of *Arabidopsis* seedlings sense different exogenous amino acids. These cells are the most easily accessible to the applied stimuli, therefore the delayed response of the inner cells (Fig. 1*A* and *B* and SI Appendix, Fig. S3) might depend on a more difficult penetration of the amino acids inside the root tip; however, a non-amino acid-dependent cell-to-cell communication process might be invoked (30, 32). This aspect would require further investigations.

The GLR3.3 was shown to be required for the amino acid-induced $[Ca^{2+}]_{cyt}$ increase and PM depolarization in *Arabidopsis* seedlings (17, 18, 21). We thus analysed the GLR3.3 expression pattern in *Arabidopsis* seedlings root tip cells through confocal analyses of plants expressing the GLR3.3-EGFP fusion protein under the control of the GLR3.3 promoter (19) (Fig. 1*C* and *C'* and *SI Appendix*, Fig. S4). Interestingly, a close inspection of the GLR3.3 subcellular localization showed an apparent accumulation of the protein at the basal and apical membranes, but also intracellular punctate assemblies reminiscent of the endomembrane system (Fig. 1*C'*) which can include the endoplasmic reticulum (ER). Such a hypothesis is supported by the GLR3.3 presence in the ER of phloem sieve elements (12). However our, and previous, results do not rule out the presence of GLR3.3 from the plasma membrane but suggest that it might be subjected to a fine subcellular sorting as reported in pollen (14) and also for animal AMPARs (33). Indeed, the precise regulation of GLR3.3 subcellular localization in root tip cells will require additional research, including the consideration that the GFP tag might affect the *in vivo* subcellular localization of the channel. Nevertheless, a side by side comparison of the GLR3.3-EGFP fluorescence signal with the LSFM Ca^{2+} imaging (Fig. 1*A* and *C*) demonstrated that the GLR3.3 is expressed in those cells where the amino acids-induced Ca^{2+} transients occur. To unequivocally prove that GLR3.3 is involved in the amino acid-induced $[Ca^{2+}]_{cyt}$ increases in root tip cells, we expressed the NES-YC3.6 sensor in two different *GLR3.3* T-DNA lines (*glr3.3-1* and *glr3.3-2*) (17, 18, 21). The comparison of resting Ca^{2+} levels in root tip cells among the wild type and mutant alleles revealed no difference (Fig. 1*D*). Nonetheless, the lack of GLR3.3 completely prevented any amino acid-induced $[Ca^{2+}]_{cyt}$ increase, whereas the response to external ATP was not affected (Fig. 1*E* and *F*). This result confirms previous observations that the GLR3.3 is required for the amino acid response (17, 18, 21) and that might be also directly involved in the generation of the $[Ca^{2+}]_{cyt}$ transients.

To assess GLR3.3 Ca^{2+} permeability *in vivo*, we expressed it in the yeast low-affinity Ca^{2+} uptake-deficient triple mutant K667, which lacks the vacuolar ATPase (PMC1), the vacuolar exchanger (VCX1) and the cytosolic regulatory subunit (CNB1) (34–36). Remarkably, the expression of GLR3.3 in the K667 triple mutant complemented the reduced growth of yeast cells at high external $[Ca^{2+}]$ (*SI Appendix*, Fig. S5), hence supporting its direct role in Ca^{2+} transport, as previously suggested by electrophysiological data obtained in mammalian COS-7 cells (14). All these results prompted us the question whether the GLR3.3 functions as a real receptor and if the different magnitude of $[Ca^{2+}]_{cyt}$ increase observed upon amino acid administration, evaluated as maximum $[Ca^{2+}]_{cyt}$ peak (Fig. 1*F*), reflects different affinities to the receptor LBD or, alternatively, different binding-induced conformational changes of the same domain.

In vitro reconstitution and characterization of GLR3.3 LBD

To investigate the role of GLR3.3 as a receptor and its specificity for amino acid ligands, we engineered a 244-residue fusion protein reproducing the GLR3.3 LBD comprising segments S1 and S2 joined by a Gly-Gly-Thr linker, based on the successful structural determinations of iGluR LBD constructs (37). This sequence is conveniently numbered 1-244 throughout this work (Fig. 2*A* and *SI Appendix*, Fig. S1*A* and Materials and Methods). The boundaries of S1 and S2 were identified by alignment with a number of GLRs/iGluRs sequences from different species (*SI Appendix*, Figs. S6 and S7).

The resulting 27-kDa protein (GLR3.3 LBD), supposed to contain L-Glu as ligand, was purified and characterized (*SI Appendix*, Figs. S8 and S9 and Materials and Methods); interestingly, circular dichroism experiments showed that the apo form of the protein, obtained through extensive dialysis, retains the same secondary structure content as the holo form, but with markedly lower thermal stability; reconstitution of the holo form, by addition of 70-fold excess L-Glu to the apo, restored its stability, thus highlighting (i) the occurrence of a reversible binding event and (ii) the dominant role of the ligand on the structural stability of the holo form of the reconstituted LBD (*SI Appendix*, Fig. S9).

Multiple independent apo GLR3.3 LBD preparations were used to test the affinities of a number of amino acid ligands by microscale thermophoresis, producing consistent results (Fig. 2*B* and Table 1). Microscale thermophoresis monitors the migration of a fluorescently labelled protein across a temperature gradient in the presence of variable ligand concentrations. The panel of amino acid ligands was chosen to match the ones tested *in planta* by external administration. All *in vitro* affinity values were in the micromolar range, with the strongest binding measured for L-Cys and L-Met. Four aminoacidic ligands (L-Glu, L-Ala, L-Asn, L-Ser) cluster in a group of similar affinity and the lowest affinity was measured for Gly. A low but detectable affinity was also recorded for D-Ser. No binding was detected for L-Trp. These data point to a promiscuity of the GLR3.3 LBD binding pocket, with a marked preference for sulfur-containing amino acids (L-Cys, L-Met), a reduction of affinity in the absence of ligand side chain β-atoms (Gly) and complete loss of binding in the case of a bulky side chain (L-Trp).

The above reported scale of *in vitro* affinity data on the isolated GLR3.3 LBD strongly resembles the $[Ca^{2+}]_{cyt}$ increases measured in aequorin-expressing *Arabidopsis* seedlings challenged with different amino acids (18). However, the same scale of *in vitro* affinity data only partially matches our amino acids-induced $[Ca^{2+}]_{cyt}$ increases in root tip cells (*SI Appendix*, Fig. S2). One of the possible reasons for this mismatch is that the measured GLR3.3 LBD binding affinities for amino acids are in the micromolar range, whereas administration to the *Arabidopsis* root tip was at 1 mM. This consideration prompted us to measure the root tip cells $[Ca^{2+}]_{cyt}$ dynamics in response to lower doses of amino acids. We thus tested the *in planta* Ca^{2+} responses against four representative ligands (L-Cys, L-Glu, Gly and L-Met) (Fig. 2*C*) at different concentrations. L-Cys, L-Glu, Gly and L-Met ligands did not trigger any response at 1 and 10 μM and reached the plateau (evaluated in terms of peaks maxima) between 100 and 500 μM. However, at 50 μM L-Cys was more effective than L-Glu and Gly with no response to L-Met. For L-Cys, L-Glu and Gly our results mirror the different *in vitro* affinities also matching the results obtained in *Arabidopsis* seedlings expressing aequorin (panel 5B in (18)). However, L-Met, was unable to trigger a $[Ca^{2+}]_{cyt}$ transient despite binding the GLR3.3 LBD at high affinity.

In conclusion, the different extents of $[Ca^{2+}]_{cyt}$ increases evoked by different amino acids in *Arabidopsis* root tips can be for the most part correlated to the binding properties of the isolated GLR3.3 LBD. Therefore, we set out to obtain the atomic resolution structure of GLR3.3 LBD to identify the determinants underlying its peculiar selectivity profile.

Overall structures of GLR3.3 LBD

The structure of GLR3.3 LBD bound to L-Glu at 2.0-Å resolution was laboriously solved by Molecular Replacement in combination with Single-Wavelength Anomalous Diffraction (MRSAD)(38, 39); it was subsequently used through molecular replacement to solve structures of GLR3.3 LBD in complexes with three different ligands (Gly, L-Cys and L-Met, at resolutions of 1.6, 2.5 and 3.2 Å, respectively). All structures were refined to satisfactory R-factor/R_{free} values with good final stereochemistry (see SI Appendix, Table S1 and Fig. S10 and Materials and Methods for full details on structure solution and refinement).

The GLR3.3 LBD has a bilobed structure of approximately 60x40x40 $Å^3$ resembling the prokaryotic and eukaryotic LBDs described in the literature (Fig. 3*A*). Interrogation of the DALI server (40) (http://ekhidna2.biocenter.helsinki.fi/dali/) identified as the most structurally related Protein Data Bank records the LBDs from a group of vertebrate iGluRs of the kainate subtype (representative PDB ID 1sd3, rmsd 2.4 Å, Z-score 25.0) and the rotifer *Adineta vaga* GLR (AvGluR1, PDB ID 4io2, rmsd 2.5 Å, Z-score 24.9). Lobe 1 (hereafter called domain 1, residues 3-100;201-239) hosts six α-helices and two β-strands, whereas lobe 2 (hereafter called domain 2, residues 101-200) is built up by a central five-stranded β-sheet surrounded by five α-helices. The structural core of each domain is secured by many π interactions between aromatic side chains, produced by the presence of a remarkable number of Tyr and Phe residues (10 and 12, respectively), together accounting for 9% of all residues. The two domains are connected by a double-stranded hinge and separated by a deep cleft where the binding pocket is located (*SI Appendix*, Fig. S10*A-D*). The binding pocket is inaccessible to solvent and has a volume of 196 $Å^3$ as calculated by the CASTp software (*SI Appendix*, Materials and Methods)(http://sts.bioe.uic.edu/castp/) (41). Clear electron density corresponding to the ligand is present in the pocket of all our structures, thus allowing unambiguous positioning of each ligand and identification of their interactions (Fig. 3*B-E* and *SI Appendix*, Fig. S10*A-D*). Two water

molecules are always buried in the pockets, but do not contact the ligand; in the case of Gly, two additional water molecules are trapped at the site where the other amino acid ligands accommodate their side chains. The basic set for anchoring the invariant moiety of any amino acid ligand to the receptor is represented by seven conserved interactions: the guanidino group of the evolutionarily invariant Arg88 chelates the α-carboxy group of the ligand by a bidentate ionic interaction; the same α-carboxy group is hydrogen-bonded with the main chain amides of Ala83 and Phe133; the α-amino group of the ligand is hydrogen-bonded with Asp81 main chain carbonyl and Tyr180 hydroxyl group, and involved in ionic interaction with Glu177 side chain.

In addition to these basic contacts, L-Glu, L-Cys and L-Met share a weak CH/π interaction (42) between their Cβ group (absent in Gly) and the aromatic Tyr63 ring, and additionally develop specific interactions as a consequence of their different side chains: L-Glu with Arg11 (salt bridge), Asn60 (hydrogen bond) and Gln129 (π-stacking); L-Cys with Arg11, Gln129, Tyr180 (hydrogen bonds); L-Met with Arg11 and Gln129 (hydrogen bonds). However, L-Cys and L-Met take advantage of a further binding contribution, as their sulfur atoms are nestled in a remarkable concatenation of sulfur/π interactions taking place between Met66 : Tyr63 : ligand sulfur : Tyr180 (*SI Appendix*, Fig. S11*A* and *B*). The stabilization provided by such architecture provides a structural explanation for L-Cys and L-Met affinities, that are the strongest recorded in our binding assays (Fig. 2*B* and Table 1). Accommodation of a D-Glu molecule in the ligand pocket (by superposing its N-Cα-CO moiety on the same atoms of L-Glu) is expected to be strongly unfavourable, since its γ-carboxy group would fall too close to the negatively charged side chains of Asp176 and Glu177, and possibly lose the salt link to Arg11. The network of hydrogen bonds/ionic interactions extends further away from the ligand molecule, generating a complex outer layer of connections (*SI Appendix*, Fig. S11*C*); however, a superposition of all our structures reveals a striking similarity in the orientation of non-solvent-exposed side chains in the region of the pocket, with the only variability confined to Val18 rotamers (Fig. 3*F*).

In principle, knowledge of the GLR3.3 LBD structure permits to design mutants incapable of binding any or some of the observed ligands, providing a tool for understanding the role of ligand binding in the generation of the downstream cytosolic Ca^{2+} increase in root tip cells. As our complementation assays suggest that *At*GLRs are functional when expressed in yeast (*SI Appendix*, Fig. S5), and given the reported successful expression of functional *At*GLRs in HEK (19, 43) and COS-7 cells (14) and *Xenopus* oocytes (20), we anticipate that a eukaryotic system coexpressing selected full-length mutant GLRs and a Ca^{2+} sensor would be ideal to correlate specific LBD mutations to changes in Ca^{2+} conductance. On these bases, we generated a number of GLR3.3 LBD single or double mutants that were tested in *E. coli* for their level of expression and solubility (*SI Appendix*, Fig. S12 and Table S2). All tested GLR3.3 LBD mutants did not retain sufficient solubility to be scaled up for larger production, with the exception of the S13A-Y14A double mutant, involving neighboring residues not directly in contact with the ligand (*SI Appendix*, Fig. S11*C*). For this double mutant, circular dichroism confirmed retention of the wild-type fold (*SI Appendix*, Fig. S9), but binding assays detected affinities for amino acid ligands comparable to the wild-type protein (*SI Appendix*, Fig. S13), suggesting that the identification of a binding-defective GLR3.3 LBD, through an *in vitro* approach, might not prove to be an easy task.

In conclusion, the X-ray crystal structures of GLR3.3 LBD in complex with different ligands provide a neat atomic explanation of the affinity data recorded (Fig. 2*B* and Table 1) and suggest plausible hypotheses for the differential ability of ligands to evoke $[Ca^{2+}]_{cyt}$ transients (Fig. 1*F* and *SI Appendix*, Fig. S2 and Discussion). Moreover, the crystal structure of a plant GLR LBD not only represents the crucial step along the way to engineer binding-defective receptors, but is also a rational new tool to (i) generate homology models of other *Arabidopsis* GLR isoforms and derive clues about their binding specificities, and (ii) spotlight the peculiarities of GLRs from the plant kingdom through comparison with the known 3D structures of non-plant LBDs.

Homology modelling of* At*GLR isoforms

The availability of a novel experimental structure of an *Arabidopsis* GLR LBD prompted us to create and explore homology models of other *At*GLR isoforms for which information about ligands is available in the literature (GLR1.2 (13), GLR1.4 (20), GLR3.1 and GLR3.5 (16), GLR3.4 (19)) (*SI Appendix*, Table S3 and Materials and Methods). Inspection of these GLR3.3 LBD-based homology models (Fig. 4*A*) and of a sequence alignment of all 20 *At*GLR isoforms LBDs (*SI Appendix*, Fig. S14) shows that the highest structural variability clusters in solvent-exposed regions. We hypothesize that the variability of loop 2 might impact isoform substrate specificity, whereas differences in the αE helix might influence intersubunit contacts (hence gating kinetics). In iGluRs, both substrate selectivity and gating kinetics have been shown to be finely regulated by intersubunit contacts (44, 45).
Although homology modelling cannot equal experimental information from crystal structures, it helps identifying, for a specific GLR isoform, which residues are supposed to be relevant in the ligand pocket. These can be validated by comparisons made with the experimental information on ligand specificity. The GLR3.4 LBD model displays excellent quality statistics and its inspection is particularly interesting, considering that a set of ligands were tested on GLR3.4 homotetramers expressed in HEK cells (19). Despite the overall conservation of the ligand pocket, the binding of L-Glu might be less favoured in GLR3.4 than in GLR3.3 due to the presence of a negative charge (Asp127 replacing the GLR3.3 LBD Val130) at about 6 Å from the L-Glu ligand γ-carboxyl. Moreover, the presence of Leu63 in GLR3.4 LBD in place of the GLR3.3 Met66, and the subsequent shortening of the sulfur/π concatenation described in our GLR3.3 structures, justifies the reported poor agonist effects of L-Cys and L-Ala on GLR3.4 channels (Fig. 4*B*). Interesting hints are also provided by the study of models of LBDs from clade 1 isoforms. *At*GLR1.2 is expressed in pollen; D-Ser and Gly (but not L-Glu) act as agonists in promoting GLR1.2-dependent pollen tube growth (13). Placing a molecule of D-Ser in the GLR1.2 LBD model ligand pocket, by superposing its N-Cα-CO moiety on the Gly ligand, indicates that few crucial residues might underlie the binding of D-Ser (Fig. 4*C*): Thr in place of Ala83 (conferring an additional hydrogen bond to D-Ser hydroxyl group), Leu in place of Phe133 (creating room for the D-enantiomeric conformation), and the pair Met in place of Ser101 and Phe in place of Trp203 (releasing Glu177 hydrogen bonds, which would create room for the D-Ser side chain). Such combination of residues is found in GLR1s only, thus suggesting that their occurrence might be a hallmark of the preference for the D-Ser ligand. The same modelling approach for *At*GLR1.4 appears to justify the binding preference of this isoform for hydrophobic amino acids (Fig. 4*D*) (20).

Comparison of the* At*GLR3.3 LBD structure with non-plant homologous structures

Recent literature extends the evolutionary classification of *A. thaliana* clades to the whole plant kingdom, confirming the late appearance of clade 1 and 2 GLRs in flowering plants (6). Alignments of the GLR3.3 LBD sequence with LBDs of the other 19 *At*GLR isoforms (*SI Appendix*, Fig. S14) and representative plant GLRs (*SI Appendix*, Fig. S15) indicate that sequence conservation within *A. thaliana* clades (≈30% between clades 1 and 3) is lower than intra-clade conservation across different plant species (58-66% sequence identity within the clade 3 sequences of *SI Appendix*, Fig. S15). Therefore, we reckon that the GLR3.3 structure can be viewed as a representative of GLR3s of the whole plant kingdom in a cross-species comparison with non-plant homologs.
The Protein Data Bank hosts a large number of iGluR/GLR LBDs from different species sharing a modest (20 to 25%) sequence identity, with a prevalence for the vertebrate LBDs of the three major types (AMPA, kainate and NMDA). When a comparison of the GLR3.3 LBD structure with a range of representative LBDs (45–50) is run, an overall structural conservation, that is more pronounced in domain 1, is clearly evident (Fig. 5*A* and *B*). However, in the secondary structure arrangement, plant GLR3s operate the peculiar evolutionary choices of: (i) containing the expansion of loop 1 (whose enlargement in NMDA iGluRs affects intersubunit allostery (52)), (ii) expanding the β1-αA and αH-β6 loops, and (iii) drastically rearranging loop 2 (whose first part preceding the conserved Tyr63 is expanded and the second part - bulging outwards in iGluRs - is deleted). Loops 2 and β1-αA host ligand-interacting side chains (Arg11, Asn60), whereas the αH-β6 loop is predicted to face the membrane. Interestingly, none of the above mentioned structural features are predicted to be involved in intersubunit contacts.
Cross-species conservation of specific residues is spread throughout the amino acid sequence and mostly involves Gly or hydrophobic residues contributing to the structural core, including the conserved disulfide Cys189-Cys243 (absent in prokaryotic sequences and plant GLR1s only). In the binding site, the conserved architecture dictates the presence of a ligand-chelating Arg side chain (Arg88) projecting from helix αD, one acidic residue coordinating the α-amino group of the ligand (Glu177) and an aromatic side chain folding on the ligand Cβ (Tyr63) on loop 2. In this area, the only plant-specific conserved residue is Asp81, that is placed at the center of a hydrogen bond network keeping the protein lobes together (*SI Appendix*, Fig. S11*C*).
Finally, we observe that the L-Glu ligand side chain in GLR3.3 LBD binding site maintains its χ_1 dihedral angle in the range observed in vertebrate iGluRs (-73 to -83°), but extends the χ_2 angle to -150° approaching the range observed in prokaryotic GLRs (-174 to -179°; -60 to -77° in iGluRs), locating the side chain half-way between the kinked conformation present in iGluRs and the fully extended conformation of prokaryotic GLRs (Fig. 5*C* and *SI Appendix*, Fig. S16).

DISCUSSION

An increasing body of literature has provided evidence about the numerous physiological roles in which plant GLRs are involved (53), however several pieces of the puzzle are still missing, including the direct link between ligand binding and channel permeation. In this paper, we used a combination of genetics and high-resolution optical microscopy to strengthen the evidence of the primary role played by the GLR3.3 isoform in generating amino acid-evoked $[Ca^{2+}]_{cyt}$ transients in the root tip cells of *Arabidopsis* seedlings. To gain a deeper view of its physiology, we biochemically reconstituted and characterized the GLR3.3 LBD in its binding properties and solved its high-resolution structure. We could thus redefine GLR3.3 as a broad-spectrum amino acid receptor and lay the bases for more precisely dissecting the determinants of plant GLRs physiology.

The ranking of affinities determined by our GLR3.3 binding assays (Table 1) can be rationalized based on the reported crystal structures, since the increase in affinity from Gly to L-Glu to L-Met and L-Cys is explained by the increase in the number of interactions with protein side chains. The amino acid-selective binding site is tuned for acceptance of different ligand residues, in line with previous speculations (18, 53) and in contrast to the selectivity profiles of prokaryotic and other eukaryotic GLRs, where a restricted preference for one or two L-amino acids is usually observed (L-Glu and L-Asp in *Campylobacter* (54); L-Glu in *Nostoc* (55, 56); L-Glu and L-Asp in rotifer *Adineta* (48); Gly in ctenophore *Mnemiopsis* (57); L-Glu in vertebrate AMPA-type and kainate-type iGluRs; Gly, L-Glu and D-Ser in NMDA-type iGluRs (2)). Our affinity values for GLR3.3 (in the sub-micromolar to micromolar range) (Table 1) are in line with the ligand concentrations of our *in vivo* experiments (Fig. 2*C*) and with the values obtained for the animal receptor homologs with the same or different techniques (46, 58).

Interestingly, the LBD of *Av*GluR1 receptor from the rotifer *Adineta vaga* is the only LBD whereby crystal structures are available in complex with a set of amino acid ligands (L-Glu, L-Asp, L-Ser, L-Ala, L-Met, L-Phe) (48); in this receptor, the binding of L-Ser, L-Ala, L-Met is mediated by a chloride ion coordinated by one/two Arg side chains in a position not far from GLR3.3 Arg11. Instead, our crystallographic refinement excluded the presence of ions in the GLR3.3 binding pocket (*SI Appendix*, Fig. S17); moreover, unlike what is observed in AvGluR1 and animal iGluRs, there are no ordered water molecules in direct contact with the ligand (Fig. 5*C*) and no contributions from protein main chain atoms in the recognition of the L-Glu ligand side chain. It is worth noticing that in the binding cavity of animal GLRs, all protein residues interacting with the L-Glu ligand side chain belong to domain 2, with the only exception of the highly conserved equivalent of Tyr63; this suggests that the peculiar expansion of domain 1 loops (loops 2 and β1-αA) observed in GLR3.3 is likely instrumental in broadening the substrate specificity.

One issue that remains unsolved is that the second highest *in vitro* affinity observed for GLR3.3 LBD (L-Met, Table 1) does not correlate with the poor capacity of the same ligand to evoke $[Ca^{2+}]_{cyt}$ increases in root tips (Fig. 1*F* and *SI Appendix*, Fig. S2). This suggests that a simplistic affinity/conductance correlation is true for most but not all ligands, and additional layers of complexity come into play between receptor binding and change in Ca^{2+} conductance. In iGluRs, the early assumption that the extent of the agonist-induced LBD closure correlates with its efficacy (46, 59) was substantially confirmed in full-length structures (60). In the GLR3.3 structures, like in *Av*GluR1 (48), the extent of the LBD clamshell closure is the same for all ligands, despite their different affinities (Table 1) and their different abilities to evoke cytosolic Ca^{2+} increases in root tip cells (Fig. 1*F* and *SI Appendix*, Fig. S2). Therefore, the discrepancy between L-Met affinity and its *in vivo* effect depends on reasons that are not immediately evident from the X-ray structures. A possibility might depend on the amphipathic nature of L-Met (due to the presence of a hydrophobic side chain) which could affect its diffusion within the cell wall. A second hypothesis is that the identity of the GLR isoforms present in the functional tetrameric receptor might modulate its affinity for specific ligands, as previously suggested in plants (18) and observed for animal iGluRs (44, 45); GLR3.3 would be an obligate component of the tetrameric channel, whose ablation prevents any amino acid-induced $[Ca^{2+}]_{cyt}$ increase (Fig. 1*E* and *F*). Future research is indeed needed to shed light on this important aspect.

On the basis of these considerations, homology models of other GLR isoforms based on our structures might prove helpful to gain a clear picture of the GLR response; they confirm ligand selectivity data reported in the literature and predict mutations that impact on ligand binding. Our GLR3.4 model fairly explains why L-Cys and L-Glu are not the best aminoacidic agonists of this isoform, but only binding assays on a reconstituted GLR3.4 LBD would permit specific affinity comparisons between GLR3.3 and 3.4 regarding their common preference for L-Asn, L-Ser and Gly. Our GLR1.2 model posits the response to D-Ser as a feature acquired by clade 1 GLRs; accordingly, our assays on GLR3.3 detected a low affinity for D-Ser (Table 1); however, we cannot exclude that affinity for D-Ser might be additionally finely tuned by residues away from the binding site, as it has been shown for iGluR delta and NMDA receptors (45, 61). The preference of GLR1.4 for amino acid ligands with bulky hydrophobic side chains (L-Met, L-Trp, L-Phe, L-Leu, L-Tyr) (20) is precisely rationalized by our GLR1.4 model, that predicts a hydrophobic environment surrounding the amino acid ligand side chain (Fig. 4*D*). Instead, due to high sequence conservation, the binding pockets in our models of GLR3.1 and GLR3.5 (reported to be specifically activated by L-Met for the regulation of stomatal aperture (16)) are remarkably similar to that of GLR3.3. Actually, both automatically generated homology models publicly available in the SWISS-MODEL Repository (swissmodel.expasy.org/repository) (62) and previous *At*GLR LBD models presented in the literature (8, 16, 20) suffer from problematic alignment with the selected template (generally rat iGluR LBD) and present significant deviations from the experimental structure we present.

In conclusion, this study demonstrates the involvement of GLR3.3 in amino acid response in *A. thaliana* root tip cells, supporting its role as ligand-gated Ca^{2+} channel. Moreover, we present the biochemical and structural characterization of its ligand-binding domain, showing that it works as a broad-spectrum amino acid receptor. Such structural knowledge, that adds to the collection of bacterial and animal LBD structures available, on one hand provides a perspective view on the evolution of these ancestral proteins along the plant lineage and, on the other, represents a working tool to engineer all plant GLR isoforms aiming at a deeper understanding of their basic physiology.

MATERIALS AND METHODS

A. thaliana WT, *glr3.3-1*, *glr3.3-2* plants were in the Col-0 background. Growth conditions, generation of transgenic lines, yeast complementation test, measurement of Ca^{2+} dynamics, description of biochemical and structural methodological assays, statistical methods, protocols used for localization and expression pattern studies and other imaging measurements are reported in the *SI Appendix*, Materials and Methods.

ACCESSION NUMBERS

Sequence data for GLR3.3 (AT1G42540), CCX2 (AT5G17850), can be found in the *Arabidopsis* Araport (https://www.araport.org/) or TAIR (https://www.Arabidopsis.org/) databases. The corresponding GLR3.3 amino acid sequence is UniProtKB Q9C8E7 (UniProt database at https://www.uniprot.org/).
The atomic coordinates and experimental structure factors for the four GLR3.3 LBD structures were deposited in the Protein Data Bank with accession codes 6R85 (L-Glu complex), 6R88 (Gly complex), 6R89 (L-Cys complex), 6R8A (L-Met complex).

ACKNOWLEDGMENTS

The *glr3.3-1* and *glr3.3-2* T-DNA mutant lines were kindly provided by Prof. Zhi Qi (College of Life Sciences, Inner Mongolia University). The vector harboring the GLR3.3 coding sequence was kindly provided by Prof. Josè Feijo and Dr. Michael Wudick (University of Maryland Department of Cell Biology and Molecular Genetics). GLR3.3pro:GLR3.3-EGFP seeds were kindly provided by Prof. Edgar Spalding (Department of Botany, University of Wisconsin). We thank Dr. Antonio Chaves Sanjuán (Department of Biosciences, University of Milan) for helpful discussion and Dr. Delia Tarantino (Department of Biosciences, University of Milan) for help with the microscale thermophoresis analysis. Part of this work was carried out at NOLIMITS, an advanced imaging facility established by the University of Milano.
This work was supported by Ministero dell'Istruzione, dell'Università e della Ricerca Fondo per gli Investimenti della Ricerca di Base (FIRB) 2010 RBFR10S1LJ_001 grant to ACo, by PIANO DI SVILUPPO DI ATENEO 2017 (University of Milan) to ACo, by a post-doctoral fellowship from the Dept. of Biosciences (University of Milan) to AA, by a PhD fellowship from University of Milan to FGD, by a postdoc fellowship from the European Commission within the framework of the SUSTAIN-T Project of the Erasmus Mundus Programme, Action 2—STRAND 1, Lot 7, Latin America to NR, by Laserlab Europe (EU-H2020 654148) to AB.
We acknowledge the ESRF (Grenoble, France) for provision of synchrotron radiation facilities (proposal mx-1894) and the staff of beamline ID29 for in situ assistance, and Diamond Light Source (Didcot, UK) for synchrotron beamtime (proposal MX20221) and the staff of beamline I04 for assistance with remote data collection.

AUTHOR CONTRIBUTIONS

AA and ACo designed and directed the research. AA and MCB generated the constructs of GLR3.3 LBD and the mutated versions. AA expressed, purified and characterized the proteins. ABar performed the circular dichroism experiments. AA performed the MST analyses and the crystallization experiments. AA and RP solved and refined the 3D structures of GLR3.3 LBD; AA analyzed the structures. LL generated the knock-out mutant lines expressing the NES-YC3.6. FGD, MG, ACa, NRA and ABas performed the imaging experiments. MCB performed the yeast complementation tests. AA, FGD, MG, RP, ACa, GV, ABar, MB, ABas, MN and ACo analyzed the data. TRS, MB and MN revised the manuscript. MB and MN provided support and equipment in protein biochemistry and access to synchrotron. AA and ACo generated the figures and the supplemental material. AA and ACo wrote the paper with contributions from MB and MN.

REFERENCES

1. Chiu J, DeSalle R, Lam HM, Meisel L, Coruzzi G (1999) Molecular evolution of glutamate receptors: a primitive signaling mechanism that existed before plants and animals diverged. *Mol Biol Evol* 16(6):826–838.
2. Traynelis SF, et al. (2010) Glutamate Receptor Ion Channels: Structure, Regulation, and Function. *Pharmacol Rev* 62:405–496.
3. Wudick MM, Michard E, Oliveira Nunes C, Feijó JA (2018) Comparing plant and animal glutamate receptors: Common traits but different fates? *J Exp Bot* 69(17):4151–4163.
4. Acher FC, Bertrand HO (2005) Amino acid recognition by venus flytrap domains is encoded in an 8-residue motif. *Biopolym - Pept Sci Sect* 80(2–3):357–366.
5. Davenport R (2002) Glutamate receptors in plants. *Ann Bot* 90(5):549–557.
6. De Bortoli S, Teardo E, Szabò I, Morosinotto T, Alboresi A (2016) Evolutionary insight into the ionotropic glutamate receptor superfamily of photosynthetic organisms. *Biophys Chem* 218:14–26.
7. Singh SK, Chien C Te, Chang IF (2016) The Arabidopsis glutamate receptor-like gene GLR3.6 controls root development by repressing the Kip-related protein gene KRP4. *J Exp Bot* 67(6):1853–1869.
8. Dubos C, Huggins D, Grant GH, Knight MR, Campbell MM (2003) A role for glycine in the gating of plant NMDA-like receptors. *Plant J* 35(6):800–810.
9. Cheng Y, Zhang X, Sun T, Tian Q, Zhang WH (2018) Glutamate Receptor Homolog3.4 is Involved in Regulation of Seed Germination under Salt Stress in Arabidopsis. *Plant Cell Physiol* 59(5):978–988.
10. Mousavi SAR, Chauvin A, Pascaud F, Kellenberger S, Farmer EE (2013) GLUTAMATE RECEPTOR-LIKE genes mediate leaf-to-leaf wound signalling. *Nature* 500(7463):422–426.
11. Toyota M, et al. (2018) Glutamate triggers long-distance, calcium-based plant defense signaling. *Science (80-)* 361(6407):1112–1115.
12. Nguyen CT, Kurenda A, Stolz S, Chételat A, Farmer EE (2018) Identification of cell populations necessary for leaf-to-leaf electrical signaling in a wounded plant. *Proc Natl Acad Sci* 115(40):10178–10183.
13. Michard E, et al. (2011) Glutamate receptor-like genes form Ca2+ channels in pollen tubes and are regulated by pistil D-serine. *Science (80-)* 332(6028):434–437.
14. Wudick MM, et al. (2018) CORNICHON sorting and regulation of GLR channels underlie pollen tube Ca2+homeostasis. *Science (80-)* 360(6388):533–536.
15. Cho D, et al. (2009) De-regulated expression of the plant glutamate receptor homolog AtGLR3.1 impairs long-term Ca2+-programmed stomatal closure. *Plant J* 58(3):437–449.
16. Kong D, et al. (2016) L-Met Activates Arabidopsis GLR Ca2+ Channels Upstream of ROS Production and Regulates Stomatal Movement. *Cell Rep* 17(10):2553–2561.
17. Qi Z, Stephens NR, Spalding EP (2006) Calcium Entry Mediated by GLR3.3, an Arabidopsis Glutamate Receptor with a Broad Agonist Profile. *Plant Physiol* 142(3):963–971.
18. Stephens NR, Qi Z, Spalding EP (2007) Glutamate Receptor Subtypes Evidenced by Differences in Desensitization and Dependence on the GLR3.3 and GLR3.4 Genes. *Plant Physiol* 146(2):529–538.
19. Vincill ED, Bieck AM, Spalding EP (2012) Ca2+ Conduction by an Amino Acid-Gated Ion Channel Related to Glutamate Receptors. *Plant Physiol* 159(1):40–46.
20. Tapken D, et al. (2013) A Plant Homolog of Animal Glutamate Receptors Is an Ion Channel Gated by Multiple Hydrophobic Amino Acids. *Sci Signal* 6(279):ra47.
21. Li F, et al. (2013) Glutamate Receptor-Like Channel3.3 Is Involved in Mediating Glutathione-Triggered Cytosolic Calcium Transients, Transcriptional Changes, and Innate Immunity Responses in Arabidopsis. *Plant Physiol* 162(3):1497–1509.
22. Yoshida R, et al. (2016) Glutamate functions in stomatal closure in Arabidopsis and fava bean. *J Plant Res* 129:39–49.
23. Kumar J, Mayer ML (2013) Functional Insights from Glutamate Receptor Ion Channel Structures. *Annu Rev Physiol* 75(1):313–337.
24. Mayer ML (2017) The Challenge of Interpreting Glutamate-Receptor Ion-Channel Structures. *Biophys J* 113(10):2143–2151.
25. Greger IH, Watson JF, Cull-Candy SG (2017) Structural and Functional Architecture of AMPA-Type Glutamate Receptors and Their Auxiliary Proteins. *Neuron* 94(4):713–730.
26. Hansen KB, et al. (2018) Structure, function, and allosteric modulation of NMDA receptors. *J Gen Physiol* 150(8):1081–1105.
27. Zhu S, Gouaux E (2017) Structure and symmetry inform gating principles of ionotropic glutamate receptors. *Neuropharmacology* 112:11–15.
28. Behera S, et al. (2015) Analyses of Ca2+ dynamics using a ubiquitin-10 promoter-driven Yellow Cameleon 3.6 indicator reveal reliable transgene expression and differences in cytoplasmic Ca2+ responses in Arabidopsis and rice (Oryza sativa) roots. *New Phytol* 206(2):751–760.
29. Krebs M, et al. (2012) FRET-based genetically encoded sensors allow high-resolution live cell imaging of Ca 2+ dynamics. *Plant J* 69(1):181–192.
30. Costa A, Candeo A, Fieramonti L, Valentini G, Bassi A (2013) Calcium Dynamics in Root Cells of Arabidopsis thaliana Visualized with Selective Plane Illumination Microscopy. *PLoS One* 8(10):e75646.
31. Candeo A, Doccula FG, Valentini G, Bassi A, Costa A (2017) Light Sheet Fluorescence Microscopy Quantifies Calcium Oscillations in Root Hairs of Arabidopsis thaliana. *Plant Cell Physiol* 58(7):1161–1172.
32. Hander T, et al. (2019) Damage on plants activates Ca 2+ -dependent metacaspases for release of immunomodulatory peptides. 7486. doi:10.1126/science.aar7486.
33. Brockie PJ, et al. (2013) Article Cornichons Control ER Export of AMPA Receptors to Regulate Synaptic Excitability. 129–142.
34. Cunningham KW, Fink GR (2015) Calcineurin inhibits VCX1-dependent H+/Ca2+ exchange and induces Ca2+ ATPases in Saccharomyces cerevisiae. *Mol Cell Biol* 16(5):2226–2237.
35. Yadav AK, et al. (2015) A rice tonoplastic calcium exchanger, OsCCX2 mediates Ca2+/cation transport in yeast. *Sci Rep* 5:1–15.
36. Corso M, Doccula FG, de Melo JRF, Costa A, Verbruggen N (2018) Endoplasmic reticulum-localized CCX2 is required for osmotolerance by regulating ER and cytosolic Ca $^{2+}$ dynamics in *Arabidopsis*. *Proc Natl Acad Sci* 115(15):3966–3971.
37. Madden DR (2002) Ion Channel Structurethe Structure and Function of Glutamate Receptor Ion Channels. *Nat Rev Neurosci* 3(2):91–101.
38. Schuermann JP, Tanner JJ (2003) MRSAD: Using anomalous dispersion from S atoms collected at Cu Ka wavelength in molecular-replacement structure determination. *Acta Crystallogr - Sect D Biol Crystallogr* 59(10):1731–1736.
39. Panjikar S, Parthasarathy V, Lamzin VS, Weiss MS, Tucker PA (2009) On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination. *Acta Crystallogr Sect D Biol Crystallogr* 65(10):1089–1097.
40. Holm L, Laakso LM (2016) Dali server update. *Nucleic Acids Res* 44(W1):W351–W355.
41. Tian W, Chen C, Lei X, Zhao J, Liang J (2018) CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res* 46(W1):W363–W367.
42. Nishio M, Umezawa Y, Fantini J, Weiss MS, Chakrabarti P (2014) CH-π hydrogen bonds in biological macromolecules. *Phys Chem Chem Phys* 16(25):12648–12683.
43. Vincill ED, Clarin AE, Molenda JN, Spalding EP (2013) Interacting Glutamate Receptor-Like Proteins in Phloem Regulate Lateral Root Initiation in Arabidopsis. 25(April):1304–1313.
44. Furukawa H, Singh SK, Mancusso R, Gouaux E (2005) Subunit arrangement and function in NMDA receptors. *Nature* 438(7065):185–192.
45. Yao Y, Harrison CB, Freddolino PL, Schulten K, Mayer ML (2008) Molecular mechanism of ligand recognition by NR3 subtype glutamate receptors. *EMBO J* 27(15):2158–2170.
46. Armstrong N, Gouaux E (2000) Mechanisms for Activation and Antagonism of an AMPA-Sensitive Glutamate Receptor: Crystal Structures of the GluR2 Ligand Binding Core. *Neuron* 28(1):165–181.
47. Unno M, et al. (2011) Binding and selectivity of the marine toxin neodysiherbaine A and its synthetic analogues to GluK1 and GluK2 kainate receptors. *J Mol Biol* 413(3):667–683.
48. Lomash S, Chittori S, Brown P, Mayer ML (2013) Anions mediate ligand binding in adineta vaga glutamate receptor ion channels. *Structure* 21(3):414–425.
49. Li Y, et al. (2016) Novel Functional Properties of Drosophila CNS Glutamate Receptors Novel Functional Properties of Drosophila CNS Glutamate Receptors. *Neuron* 92(5):1036–1048.
50. Mayer ML, Olson R, Gouaux E (2001) Mechanisms for ligand binding to GluR0 ion channels: Crystal structures of the glutamate and serine complexes and a closed apo state. *J Mol Biol* 311(4):815–836.
51. Hackos DH, et al. (2016) Positive Allosteric Modulators of GluN2A-Containing NMDARs with Distinct Modes of Action and Impacts on Circuit Function. *Neuron* 89(5):983–999.
52. Regalado MP, Villarroel A, Lerma J (2001) Intersubunit cooperativity in the NMDA receptor. *Neuron* 32(6):1085–1096.
53. Forde BG, Roberts MR (2014) Glutamate receptor-like channels in plants: a role as amino acid sensors in plant defence? *F1000Prime Rep* 6. doi:10.12703/p6-37.
54. Müller A, et al. (2007) A Bacterial Virulence Factor with a Dual Role as an Adhesin and a Solute-binding Protein: The Crystal Structure at 1.5 Å Resolution of the PEB1a Protein from the Food-borne Human Pathogen Campylobacter jejuni. *J Mol Biol* 372(1):160–171.
55. Lee JH, et al. (2005) Crystallization and preliminary X-ray crystallographic analysis of the GluR0 ligand-binding core from Nostoc punctiforme. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 61(11):1020–1022.
56. Lee JH, et al. (2008) Crystal Structure of the GluR0 Ligand-Binding Core from Nostoc punctiforme in Complex with l-Glutamate: Structural Dissection of the Ligand Interaction and

Subunit Interface. *J Mol Biol* 376(2):308–316.

57. Alberstein R, Grey R, Zimmet A, Simmons DK, Mayer ML (2015) Glycine activated ion channel subunits encoded by ctenophore glutamate receptor genes. *Proc Natl Acad Sci* 112(44):E6048–E6057.

58. Seidel SAI, et al. (2012) Label-free microscale thermophoresis discriminates sites and affinity of protein-ligand binding. *Angew Chemie - Int Ed* 51(42):10656–10659.

59. Jin R, Banke TG, Mayer ML, Traynelis SF, Gouaux E (2003) Structural basis for partial agonist action at ionotropic glutamate receptors. *Nat Neurosci* 6(8):803–810.

60. Dürr KL, et al. (2014) Structure and dynamics of AMPA receptor GluA2 in resting, pre-open, and desensitized states. *Cell* 158(4):778–792.

61. Tapken D, et al. (2017) The low binding affinity of D-serine at the ionotropic glutamate receptor GluD2 can be attributed to the hinge region. *Sci Rep* 7:46145.

62. Bienert S, et al. (2017) The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res* 45(D1):D313–D319.

FIGURE LEGENDS

Fig. 1. Amino acid-induced $[Ca^{2+}]_{cyt}$ increase in *A. thaliana* root tip cells depends on GLR3.3 activity. (*A*) Ratiometric purple-color images superimposed to cpVenus images from a representative time series of *Arabidopsis* Col-0 root tips expressing NES-YC3.6 treated with 1 mM of L-Glu visualized by LSFM. The number in the images indicates the time passed after acquisition start in seconds. Scale bar = 25 µm. (*B*) Kymograph analysis (performed on the yellow line of 1 pixel-width) showing the progression of the L-Glu-induced $[Ca^{2+}]_{cyt}$ increase which shows signal percolation from lateral root cells to the stele. (C) Confocal image of a representative root tip meristem of an *Arabidopsis* seedling expressing the GLR3.3-GFP (green color in the image) chimeric protein driven by the GLR3.3 promoter. Scale bar = 25 µm. (*C'*) Magnification of root meristem cells shown in *C*. Scale bar = 5 µm. (*D*) Steady state cpVenus/CFP ratios of the Region Of Interest (ROI) (corresponding to the area indicated within the black dashed line in the schematic drawing at the right bottom of the figure) in root tip cells imaged under continuous perfusion preceding (averaged over 50 sec time window) amino acids treatments of Col-0 (light blue), *glr3.3-1* (green) and *glr3.3-2* (yellow) knock out alleles; n ≥ 8; ns: not statistically significant. (*E*) Root tips of seedlings expressing NES-YC3.6 in Col-0, *glr3.3-1* and *glr3.3-2* imaged as in *D* treated with 1 mM of L-Cys, L-Glu, L-Ala, Gly, L-Ser, L-Asn, L-Met and 0.1 mM of external ATP. The same ROI as in *D* in the root tip meristematic zone was analyzed and plotted over time for the averaged cpVenus/CFP ratio ± SD. The black line above the graphs indicates the duration of amino acid or ATP exposure (for 150 seconds followed by washout). (*F*) Maximal relative amplitude of cpVenus/CFP ratio as $\Delta R/R_0$ increase triggered by amino acids and ATP administration in the three analyzed genotypes. Inset: magnification of 1mM L-Met maximum response; n ≥ 4; error bars ± SD; ** p<0.005, *** p<0.0005; (Student t test); ns: not statistically significant.

Fig. 2. Design of the *At*GLR3.3 construct and characterization of its binding properties. (*A*) Design of the GLR3.3 LBD construct from the full sequence; arrows indicate the position of the cloning primers, that introduce a short Gly-Gly-Thr linker (magenta) between segments S1 and S2. (*B*) Fitting of the binding curves of L-Cys, L-Met, L-Glu and Gly to GLR3.3 LBD from the microscale thermophoresis experiments, based on the equation reported in *SI Appendix*, Materials and Methods; the graph reports the concentration of the ligand in logarithmic scale vs the thermophoretic signal normalized as fraction bound. (*C*) Maximal relative amplitude of cpVenus/CFP ratio as $\Delta R/R_0$ increase triggered by different amino acids concentrations (dose-dependent amino acid response). n ≥ 3; error bars ± SD; * p<0.05, ** p<0.005, *** p<0.0005; (Student t test). For the 50, 100, 500 and 1000 µM concentrations, differences in $\Delta R_{max}/R_0$ between incremental concentrations for the same ligand are statistically non-significant, unless indicated.

Fig. 3. Structures of *At*GLR3.3 LBD bound to different ligands. (*A*) Overall structure of *At*GLR3.3 LBD (+ L-Glu) in ribbon representation, colored to highlight the contributions of segments S1 (green) and S2 (magenta) to domains 1 (D1) and 2 (D2). The linker is colored cyan, L-Glu is in cyan sticks. The C-terminal stretch (dashed) has a defined electron density in 4 out of the 14 protein chains present in the different crystal forms. The structure is oriented in such a way that the N-terminus (i.e. the prosecution of the polypeptide chain after the ATD domain) is at the top and the linker (replacing the transmembrane segments M1 to M3) is at the bottom. The traditional secondary structure nomenclature used for animal iGluR LBDs has been maintained as reference (including the names loop1 and loop 2 for the αA-αB and β2-αC loops, respectively). Aside is shown a 2D-diagram of the secondary structure of AtGLR3.3 LBD with the same color code for S1 and S2 segments; cylinders, arrows and lines represent α-helices, β-strands and loops, respectively; blue stars indicate the positions of ligand-interacting residues; the position of ATD and transmembrane domains in the topology of the protein is shown. (*B-E*) Close-up view of the ligand binding pocket in the crystal structures of GLR3.3 LBD + L-Glu (*B*), + Gly (*C*), + L-Cys (*D*) and + L-Met (*E*). The $2|F|_o$-$|F|_c$ electron density omit maps contoured at 1.5 σ are shown for the ligand molecules (cyan sticks) and two additional water molecules of the Gly-bound structure (see *SI Appendix*, Fig. S10*A-D* for the corresponding $|F|_o$-$|F|_c$ maps). The residues or groups of atoms relevant for binding are indicated and represented as sticks, with nitrogen atoms blue and oxygen atoms red; protein carbon atoms are either green (if belong to segment S1) or magenta (if belong to segment S2). The hydrogen bonds are drawn as black dashes; not all interactions are shown for the sake of clarity. (*F*) Stereo view of the ligand binding pocket in a superposition of *At*GLR3.3 LBD structures from the four datasets. The domains 1 from each structure were superimposed. All the side chains (lines) and main chains (tubes) surrounding the ligands are shown, except Tyr63 for clarity. The L-Glu-bound structure is blue, Gly magenta, L-Cys yellow and L-Met orange. Note that in all structures except the one containing L-Met, a water molecule (sphere) resides in the ligand cavity; the two additional water molecules in the Gly-bound structure that are shown in *C* are not represented here for clarity.

Fig. 4. Homology modelling of other *At*GLR isoforms. (*A*) Structural superposition of the GLR3.3 LBD structure (this work, green) with GLR3.3 LBD-based homology models of GLR1.2 (31% sequence identity, yellow), GLR1.4 (32%, orange), GLR3.1 (65%, cyan), GLR3.4 (62%, dark green), GLR3.5 (60%, blue) LBDs, all in ribbon representations. The structurally coincident parts are shown in grey and only the divergent parts are coloured. The L-Glu molecule in the GLR3.3 LBD structure is shown in green sticks. Note the absence of the αE helix in the GLR1.2 and GLR1.4 models. UniProtKB primary accession numbers are: GLR1.2 Q9LV72, GLR1.4 Q8LGN1, GLR3.1 Q7XJL2, GLR3.3 Q9C8E7, GLR3.4 Q8GXJ4, GLR3.5 Q9SW97. (*B*) Model of the binding pocket of GLR3.4 (orange) superposed to the GLR3.3 LBD structure on which the model is based (transparent green). The L-Glu ligand of GLR3.3 is shown in cyan sticks. See *SI Appendix*, Materials and Methods for the numbering of GLR3.4. (*C*) Model of a D-Ser ligand molecule (cyan sticks) in the binding pocket of the GLR1.2 LBD homology model (orange), strictly resembling the pose observed in PDB-deposited structures of D-Ser-containing LBDs (PDB IDs 1pb8, 2rc8, 2rcb, 2v3u, 4ykk). Relevant residues of the GLR1.2 LBD model (orange) and GLR3.3 LBD structure on which the model is based (transparent green) are shown. Relevant hydrogen bonds are represented as dashes (green for GLR3.3 LBD) and the position of a bound L-Glu molecule is indicated in transparency for reference. Note that the Glu177 side chain in GLR3.3 LBD is kept in place by two hydrogen bonds that are lost in the GLR1.2 LBD model. See *SI Appendix*, Materials and Methods for the numbering of GLR1.2. (*D*) Model of the binding pocket of GLR1.4 (orange) superposed to the GLR3.3 LBD structure on which the model is based (transparent green). The L-Glu ligand of GLR3.3 is shown in cyan sticks, with green dashes indicating relevant hydrogen bonds for GLR3.3. See *SI Appendix*, Materials and Methods for the numbering of GLR1.4.

Fig. 5. Comparison of the *At*GLR3.3 LBD structure with non-plant homologous structures. (*A*) Overall superposition of the GLR3.3 LBD structure (+ L-Glu, green) with X-ray structures of LBDs from rat AMPA-subtype GluA2 (RnGluA2, PDB ID 1ftj, purple) (46), human kainate-subtype GluK1 (HsGluK1, PDB ID 2zns, yellow) (47), rat NMDA-subtype GluN3A (RnGluN3A, PDB ID 2rc7, orange) (45), rotifer AvGluR1 (PDB ID 4io2, cyan) (48), fruit fly GluR1A (DmGluR1A, PDB ID 5dt6, blue) (49), and cyanobacterial GluR0 (SsGluR0, PDB ID 1ii5, pink) (50), with the GLR3.3 LBD L-Glu ligand shown as green sticks. The traits that are roughly structurally coincident are shown as grey wires connecting Cαs; only the parts that display relevant structural divergence from the other compared proteins are shown as colored ribbons. Note the large rearrangement of loop 2 in the GLR3.3 structure. Table aside: for the same proteins, % sequence identities with GLR3.3 LBD and Cα trace rmsd (Å) from GLR3.3 LBD are given; for rmsd, values are provided for both the whole LBDs and domains 1 only and calculated excluding the protruding loop 1 that is highly variable in sequence and structure. (*B*) Least squares superposition of the Cα traces of the *At*GLR3.3 LBD structure (+ L-Glu, green) with the crystallographic structures of LBDs from the same proteins shown in *A* (same color codes). This figure differs from *A* by the fact that the structures were superimposed by the domains 1 selectively (excluding the variable loops 1 and 2, not shown); the corresponding rmsd values are tabled in *A*. (*C*) Superposition of the L-Glu ligand molecules (in stick representation) from different LBDs onto the L-Glu molecule of this study (*At*GLR3.3, green): rat AMPA-subtype GluA2 (PDB ID 1ftj, purple), human kainate-subtype GluK1 (PDB ID 2zns, yellow), human NMDA-subtype GluN2A (PDB ID 5h8f_A, orange) (51), rotifer AvGluR1 (PDB ID 4io2, cyan), fruit fly GluR1A (PDB ID 5dt6, blue) and cyanobacterial GluR0 (PDB ID 1ii5, pink). Relevant side chains, main chains and waters coordinating the ligands are shown. The highly conserved structural equivalents of GLR3.3 Tyr63 and Arg88 are indicated. The L-Glu molecules from rat GluA2 and fruit fly GluR1A almost perfectly overlap. The coordination of the L-Glu ligand γ-carboxy group is diversely achieved in different species.

SI Appendix

MATERIALS AND METHODS

Plant material and growth conditions. All *A. thaliana* plants were of the ecotype Columbia 0 (Col-0). Plants were grown on soil under short day conditions (12 h light /12 h dark, 100 µE m^{-2} s^{-1} of Cool White Neon lamps) at 22 °C and 75% relative humidity. Seeds were surface-sterilized by vapor-phase sterilization (1) and plated on half-strength MS medium (2) (Duchefa) supplemented with 0.1% sucrose, 0.05% MES, pH 5.8, and 0.8% plant agar (Duchefa). After stratification at 4 °C in the dark for 2 days, plates were transferred to the growth chamber under long day conditions (16 h light/8 h dark, 100 µE m^{-2} s^{-1} of Cool White Neon lamps) at 22 °C. For wide field imaging the plates were kept vertically and the seedlings were used 6-7 days after germination. For Light Sheet Fluorescence Microscopy (LSFM) imaging the plates were kept horizontally for 36 hours and the germinated seeds transferred to the Fluorinated Ethylene Propylene tubes (FEP, Adtech FT2x3) as reported in (3).

Generation of transgenic plants. Plant transformation of *glr3.3-1* and *glr3.3-2* T-DNA homozygous mutant alleles (4) with NES-YC3.6 (5) was carried out using *Agrobacterium tumefaciens* GV3101 cells by floral-dip (1). At least two independent transgenic lines for both alleles were selected based on the presence of Cameleon fluorescence using a stereo microscope equipped with a GFP filter. To confirm the presence of T-DNA insertions in homozygosity in the *glr3.3-1* x NES-YC3.6 and *glr3.3-2* x NES-YC3.6 we followed the genotyping strategy reported in (4).

Confocal laser scanning microscopy. Confocal microscopy analyses were performed using a Nikon Eclipse Ti2 inverted microscope, equipped with a Nikon A1R+ laser scanning device (Nikon). EGFP was excited with the 488 nm laser and the emission was collected at 525-550 nm. Images were acquired by a CFI Apo LWD 40x WI (N.A. 1.25) or CFI Super Fluor LWD 4x Dry (N.A. 0.20) for large imaging. The stitched image shown in *SI Appendix*, Fig. S4 was obtained using the NIS-Elements™ (Nikon) software. Images were analyzed using FIJI software (https://fiji.sc/).

Wide field fluorescence microscopy. For wide field Ca^{2+} imaging analyses in *Arabidopsis* root tip cells, an inverted fluorescence Nikon microscope (Ti-E) with a 20x N.A 0.75 was used. Excitation light was produced by a fluorescent lamp (Prior Lumen 200 PRO, Prior Scientific) set to 20% with 440 nm (436/20 nm) excitation for the Cameleon (YC3.6) sensor. Images were collected with a Hamamatsu Dual CCD camera (ORCA-D2). The FRET CFP/YFP optical block A11400-03 (emission 1, 483/32 nm for CFP; emission 2, 542/27 nm for FRET) with a dichroic 510-nm mirror (Hamamatsu) was used for the simultaneous CFP and cpVenus acquisitions. Camera binning was set to 2 x 2 and exposure times (from 100 to 200 ms) were adjusted depending on the sensor line. Images where acquired every 5 s. Filters and dichroic mirrors were purchased from Chroma Technology. NIS-Elements™ (Nikon) was used as a platform to control the microscope, illuminator, and camera. Images were analyzed using FIJI.

Root tip seedling wide field fluorescence Ca^{2+} imaging. Seven-day-old seedlings were used for root Ca^{2+} imaging. Seedlings were kept in the growth chamber until the experiment, then were gently removed from the plate according to (6), placed in the dedicated chambers and overlaid with cotton wool soaked in imaging solution (5 mM KCl, 10 mM MES, 10 mM $CaCl_2$ pH 5.8 adjusted with TRIS). The root was continuously perfused with imaging solution while the shoot was not submerged. Treatments were carried out by supplementing the imaging solution with 1 mM of different amino acids (or with lower concentrations where otherwise indicated) or 0.1 mM Na_2ATP (sodium adenosine triphosphate) (from a 200 mM stock solution buffered at pH 7.4 with NaOH) and administered for 3 min under running perfusion.

Light Sheet Fluorescence Microscopy imaging of root tip. For LSFM Ca^{2+} imaging analyses in *Arabidopsis* root tip cells a custom-made setup was used (3, 7). The optical path starts with a single-mode fibre, coupled to a laser emitting at 442 nm (MDL-III-442, CNI), collimated and focalized through a cylindrical lens (f_{CL} = 50 mm) in a horizontal plane. A 1× telescope ($f_1 = f_2$ = 50 mm,

Thorlabs) conjugates the focal plane of the cylindrical lens to the back focal plane of a 10× water-dipping microscope objective (NA=0.3, UMPLFLN 10xW, Olympus), which creates a a vertical light-sheet at the sample level. The light sheet is matched to the field of view of a detection objective (N.A. = 0.5, UMPLFLN 20xW, Olympus) held orthogonally to the excitation axis. For the detection of the FRET cpVenus/CFP ratio, a two-wavelength detection is required. Thanks to the vertical geometry of plant roots, it is practical to record two images with different spectral content on the same detector by splitting the detection path in two spectral channels. To this end, the detection objective is followed by a 1x relay lens system ($f_3 = f_4 = 100$ mm, Thorlabs). A vertical slit is placed in the intermediate image plane with 400 µm horizontal size, which corresponds to half of the field of view. A dichroic filter at 505 nm (DMLP505, Thorlabs) creates two-colour replicas of the sample image, which are then formed on the detector through two band-pass filters (MF479-40 and MF535-22 emission filters, Thorlabs), two broadband mirrors (BBSQ1-E02, Thorlabs) and a tube lens (U-TLU-1-2, Olympus). These create the images of the CFP and the cpVenus fluorescent signals on the two sides of the CMOS sensor (Neo 5.5 sCMOS, 2560 × 2160 pixels, ANDOR). The laser power was set to 20 µW on the sample, which proved not to give relevant photobleaching during the experiment. To minimize the light dose on the sample, an automatic shutter opens the laser beam only when the camera is in acquisition mode. A white LED illuminator is used for trans-illumination, for sample alignment. The sample is held vertically in a custom-made 3D-printed chamber, filled with the imaging solution. The camera acquisition, sample translation stage and shutter are synchronized via a custom-made LabVIEW software. This software permits the observation of the two channels, to visualise their ratio in real time and to record the data. Camera binning was set to 1 x 1 and exposure times to 100 ms. Images were acquired every 5 s and at every time point a Z-stack of 30 planes spaced of 3 µm was acquired. Images were processed using FIJI by analyzing a single plane of the time series. To generate the images shown in Fig. 1 and *SI Appendix*, Fig. S3 the cpVenus/CFP calculated ratio (magenta) was superimposed to the first cpVenus emission image of the time series.

Root tip seedling LSFM Ca^{2+} imaging. Fluorinated ethylene propylene (FEP, Adtech FT2x3) tubes with an internal diameter of 0.8 mm and manually cut in 3 cm long pieces using a razor blade, were cleaned first with 1M NaOH, then with a diluted NaOH solution (0.5 M) and finally with 70% of ethanol (7). After washing with 1M NaOH, a 10 min sonication was performed at each cleaning step. The tubes were then rinsed with MilliQ water and coupled with the head of a 10 µl pipette tips (manually cut), placed into cleaned pipette tip boxes and afterwards autoclaved at 121 °C for 20 min. The FEP tubes were then filled with the MS/2 medium used for the seed germination but in this case jellified with 0.5% Phytagel™ (w/v) (Duchefa) instead of plant agar (3, 7, 8). The tubes were filled from the bottom of the tubes using a P200 micropipette. To prevent the evaporation of water from the Phytagel™-based medium the top of the tubes was covered with a plant agar-based-medium plug, thus creating a small cap. After solidification, a sterilized scalpel was used to remove the exceeding cap medium. After seedlings germination and fluorescence inspection with a stereo microscope, the fluorescent seeds were quickly moved from the plate to the top of the tubes to avoid root drying, using sterilized pliers and without clamping them. Seedlings were placed over the top of the tubes, so the plantlets could grow inside the filled tubes. The tubes were transferred to a tip box that was finally filled with MS/2 liquid medium without sucrose and sealed to avoid contamination. To mount the tubes with the plant in the imaging chamber, we used a custom-made holder (3, 7) consisting of a hollow aluminium tube in which a pipette tip can be attached. The seedlings were let grow until the root tip emerged from the FEP tubes (7/8-day-old). When plants were ready to be imaged, we plugged the pipette tip with the tube into the hollow tube, and quickly moved the whole holder to the imaging chamber of the LSFM setup filled with imaging solution (5 mM KCl, 10 mM MES, 10 mM $CaCl_2$ pH 5.8 adjusted with TRIS), fixing it on a rotation and translation stage for the sample positioning. This procedure prevents any kind of damage or major stress to the root and maintains the seedling vertical. For the analysis of spatiotemporal dynamics of the $[Ca^{2+}]$ variation, a volume of 120 µL (100X) for each tested amino acid was directly added to one corner of the imaging chamber (filled with 12 mL of imaging solution). The final concentration of the stimuli was 1 mM. The ratio images are representative of n = 3 experiments.

Quantitative imaging analysis. Fluorescence intensity was determined over regions of interest (ROIs), which corresponded to the meristematic cells of the root tip. cpVenus and CFP emissions were used for the ratio (R) calculations (cpVenus/CFP) and, where suitable, normalized to the initial ratio (R_0) and plotted versus time ($\Delta R/R_0$). For wide field imaging background subtraction was performed independently for both channels before calculating the ratio. Kymograph was generated with the FIJI plugin using the yellow line reported in the Fig. 1*A*.

Yeast growth complementation assay. *Saccharomyces cerevisiae* strain K667 (vcx1/Δ, cnb1::LEU2, pmc1::TRP1) (9) was transformed with pYES2-URA empty vector (Invitrogen) or pYES2-URA harboring the *GLR3.3* coding sequence (10) in the BamHI/EcoRI sites. The same vector harboring the *Arabidopsis CCX2* (cation/Ca^{2+} exchanger 2) (11) was used as positive control. Transformants were selected for uracil prototrophy as reported in (12). For complementation studies, single URA-plus colonies were grown in SC-URA medium containing 2% (w/v) glucose (SD), pelleted, washed twice with sterile water and diluted to an OD_{600} of 0.1. Three µl of a 10-fold dilution were spotted onto SC-URA plates containing 2% (w/v) galactose (SG) supplemented with 1, 300 or 500 mM $CaCl_2$ and incubated at 30 °C for 3-5 days. All media were supplemented with 50 mM succinic acid/Tris (pH 5.5), 0.7% (w/v) yeast nitrogen base without ammonium sulfate, to prevent precipitation of Ca^{2+}, and 5g/L NH_4Cl.

Cloning of the GLR3.3 LBD construct. The DNA portions codifying for *At*GLR3.3 S1 (residues 463-570) and S2 (residues 681-813) segments were amplified, joined by overlapping PCR (with the concomitant introduction of a Gly-Gly-Thr interspacing linker) and cloned into a pETM-14 vector (EMBL, Heidelberg, Germany) to produce an N-terminally histidine-tagged construct for expression in *E. coli*. Although a 2-residue linker is regularly reported in the literature for LBD constructs, a 3-residue Gly-Gly-Thr linker was designed because expected to be better accommodated in the crystal packing, in the light of *in silico* predictions of *At*GLR3.3 LBD secondary structure and careful alignment of the *At*GLR3.3 sequence with sequences from deposited LBD structures (using *Jalview*) (13).

Production and purification of the native protein (GLR3.3 LBD). Rosetta strain *E. coli* cells (Novagen, Merck Biosciences) were transformed with the above described pETM-14: *At*GLR3.3 LBD plasmid and grown at 37 °C in LB medium (supplemented with kanamycin and chloramphenicol) up to OD_{600} of 0.6-0.8. After cooling down the cultures at room temperature for 20 min, isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 0.1 mM and the growth continued at 20 °C for 16 h. The pelleted cells were resuspended in a buffer containing 50 mM TrisHCl pH 8.0, 500 mM NaCl, 15 mM imidazole, 2 mM β-mercaptoethanol, cOmplete™ EDTA-free protease inhibitor cocktail (Roche), 0.5 mM L-Glu or Gly. All buffers were supplemented with 0.5 mM L-Glu or Gly throughout purification to ensure stabilization of the construct. After sonication and centrifugation, the resulting supernatant was applied onto a nickel column (HisTrap FF, GE Healthcare). The imidazole-eluted sample was mixed with home-made His-tagged human rhinovirus 3C protease (protease:target protein molar ratio ≈1:600) and dialyzed overnight in a dialysis tube (SpectrumLabs, cutoff 4 kDa) to allow for tag cleavage and imidazole removal. The following day, a second passage through the nickel column was performed to separate the sample from both the tag and the protease and in the final size-exclusion chromatography column (Superdex200 10/300 GL, GE Healthcare) the 27kDa protein eluted as a symmetric peak compatible with either a monomer or a dimer (*SI Appendix*, Fig. S8*A*). A dynamic light scattering experiment showed that >99% of the protein in solution is monomeric (*SI Appendix*, Fig. S8*C*). Typical yields were of about 25 mg per liter of culture. The sample was monitored throughout purification by SDS-polyacrylamide gel electrophoresis and spectrophotometry. The final protein construct included 3 post-cleavage N-terminal residues (Gly-Pro-Met) immediately followed by Gly1 (see *SI Appendix*, Figs. S6 and S7 for construct sequence numbering) and was stored in the final size-exclusion chromatography buffer (10 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 0.5 mM L-Glu).

Production and purification of selenomethionine-substituted GLR3.3 LBD. Rosetta cells transformed with the same plasmid described above were grown in minimal M9 medium to OD_{600} of 0.3, supplemented with a cocktail of amino acids including L-selenomethionine, induced by 0.2 mM IPTG 15 min later and grown at 25 °C for 30 h, according to a metabolic inhibition protocol (14). Purification procedures were identical to the ones used for the native protein, except for the inclusion of 20 mM β-mercaptoethanol in all buffers. The incorporation of selenium was assessed on crystals right before data collection by analysis of the X-ray fluorescence emission spectra.

Dynamic light scattering. Measurement was performed on a Punk instrument (Unchained Labs) on GLR3.3 LBD + L-Glu at 1 mg/mL (37 µM) in 10 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 0.5 mM L-Glu, at 20 °C.

Dialysis to produce apo protein. The final protein was expected to contain the amino acid ligand supplemented during purification (L-Glu or Gly) and for this reason was subjected to extensive dialysis against the storage buffer to force the complete release of the ligand (1:150 sample dilution in 8x 6h-passages, giving a final dilution of 10^{17}); the protein sample obtained was used in the binding assays. Turbidity of the sample and heavy precipitation reproducibly occurring after 3-4 dialysis steps strongly suggested a holo to apo transition; apo *At*GLR3.3 LBD is invariably more unstable than the holo (L-Glu or Gly) form, displaying lower solubility and shorter storage life.

Circular dichroism. Circular dichroism experiments were carried out on a J-810 spectropolarimeter (JASCO Corp.) equipped with a Peltier system for temperature control. All data were collected on 0.2 mg/mL (7 µM) protein solutions in 10 mM HEPES pH 7.5, 150 mM NaCl and 0.5 mM EDTA (± 0.5 mM L-Glu), placed in a cuvette with a path length of 0.1 cm. Spectra were recorded from 260 to 200 nm. Temperature ramps were monitored at 220 nm while temperature was increased from 20 to 95 °C at 1 °C/min. T_m was calculated as the first-derivative maximum of the temperature ramps.

Binding assays by microscale thermophoresis. The assays were performed on a Monolith NT.115 instrument (NanoTemper Technologies). To prepare the experiment, GLR3.3 LBD or GLR3.3 LBD S13A,Y14A in their apo form were conjugated to a fluorophore targeting surface lysines (Monolith Protein Labeling Kit RED-NHS, NanoTemper Technologies) and separated from the dye excess using desalting columns. The GLR3.3 LBD construct possesses 14 surface lysines, resulting in a satisfactory and reproducible conjugation process. Each curve was produced at 24 °C by the thermophoretic signal of 16 capillaries (MST power 40%) containing a fixed concentration of labelled protein (100 nM) and increasing concentrations of ligand, in 10 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 0.05% Tween20. All measurements recorded strong thermophoretic signals (response amplitudes between 4 and 15, signal to noise ratio between 8 and 13, in line with what is expected for unambiguous results with this technique), generating well-defined sigmoidal curves reaching plateau. Data were averaged and fit by the instrument software MO. Affinity Analysis (NanoTemper) according to the following formula:

$$F = U + \frac{(B-U)\cdot\left([ligand]+[protein]+K_d-\sqrt{([ligand]+[protein]+K_d)^2-4\cdot[ligand]\cdot[protein]}\right)}{2\cdot[protein]}$$

where F is the fraction of protein bound to the ligand, while U and B represent the response values (normalized fluorescence) of the unbound and bound states, respectively. Application of this technique to LBDs of iGluRs is reported in the literature (15).

Crystallizations. Crystallization screens were performed on an Oryx robot (Douglas Instruments) by the sitting drop vapor diffusion method and manually refined by the hanging drop technique. GLR3.3 LBD (native or SeMet-substituted) purified in the presence of L-Glu or extensively dialyzed GLR3.3 LBD supplemented with 3 mM L-Cys or Gly or L-Met was mixed 2:1 at initial concentration of 12 mg/mL (445 µM) with commercial solutions (Hampton Research) in Greiner Bio-One plates and incubated at 20 °C. Hits showed up within 7 days. GLR3.3 LBD + L-Glu crystals were subsequently optimized to the final reservoir condition 100 mM sodium acetate pH 4.6, 240 mM ammonium sulfate, 30% (w/v) PEG monomethyl ether 2,000. For GLR3.3 LBD + Gly or L-Cys or L-Met, initial crystals were directly used for data collection and were obtained in the following conditions: (+ Gly) 100 mM sodium citrate tribasic pH 5.6, 2 M ammonium sulfate, 200 mM potassium sodium tartrate; (+ L-Cys or + L-Met) 100 mM HEPES pH 7.5, sodium citrate tribasic 1.4 M. SeMet-substituted GLR3.3 LBD gave crystals in 100 mM MES pH 6.5, 200 mM ammonium sulfate, 20% (w/v) PEG 8,000. All cryoprotectants were prepared by adding 25% (v/v) glycerol to the reservoir solution.

Data collections and structure solution. Statistics for data collection, phasing and refinement are summarized in *SI Appendix*, Table S1. For both the native GLR3.3 LBD + L-Glu and the SeMet GLR3.3 LBD + L-Glu datasets, diffraction data were collected at 100K on the ID29 beamline (16) at the European Synchrotron Radiation Facility, Grenoble (France) using the Pilatus 6M-F pixel detector (Dectris). The native data set was collected at a wavelength of 1.000 Å and initially indexed in space group C2 with a resolution of 2.0 Å; the anomalous data set was collected close to the Se *K*-edge at 0.979 Å and showed a tetragonal space group ($P4_32_12$) with a resolution of 2.4 Å. The datasets for GLR3.3 LBD + Gly, L-Cys and L-Met (all in the orthorhombic space group $P2_12_12_1$, with resolutions of 1.6 Å, 2.5 Å and 3.1 Å, respectively) were collected at 100K at the Diamond Light Source (Didcot, UK) on the Eiger2 X 16M detector (Dectris) of beamline I04 at a wavelength of 0.9795Å. Several applications from the *CCP4* suite were used throughout processing (17) (*SI Appendix*, Table S1). Diffraction data were processed using *XDS* (18) and scaled and merged with *AIMLESS* (19); the high-resolution data cut-off was based on the statistical indicators $CC_{1/2}$ and CC* (20). Molecular replacement (MR) was initially attempted with no success on the native L-Glu dataset with standard software, using search models identified through *BLAST*, *PSI-BLAST* (21), *FFAS* (22) or rationally edited LBD models based on the large number of bacterial and eukaryotic GLR LBDs structures available from the Protein Data Bank; pruning of the solvent-exposed loops and sequential use of either of the two lobes of known LBDs were tested in MR, producing in some cases partial solutions that did not improve after subsequent manipulation. The observation that the *At*GLR3.3 LBD reflects the topological arrangement of known LBDs with a substantial displacement in the Cα trace (more pronounced in the D2 domain) provides a possible *a posteriori* explanation for these failures. Experimental phasing on the SeMet dataset was first attempted with various standard software with no success, at least in our hands. The phase problem was finally solved by MRSAD phasing (23, 24), by which approximate experimental phases, obtained by locating some of the selenium atoms and successively improved by density modification, allowed to build a partial model; phases extracted from this model were then combined with the initial anomalous phases to produce a more accurate set of phases and an improved electron density map, as described more in detail below. A highly fragmented and partially wrong model was firstly obtained by the *CRANK2* experimental phasing pipeline (25, 26), using the SeMet dataset as both native and anomalous input: decreasing R-factor during the initial model building and refinement, as well as $R_{free} < 0.50$, good electron density for some parts of the structure and accordance between the position of some of the SeMet residues and the anomalous map led to consider the model as a partial, promising solution, albeit fragmented and incorrect in several parts. The model was then gradually improved by extensive model building with the *BUCCANEER* software (27) and the geometry of the model and the quality of the electron density map were improved with *BUSTER* (28, 29). Subsequent MRSAD-phasing in 'rebuild mode' in *CRANK2* (26) using the *BUSTER* model and the anomalous dataset improved the map; the number of substructure improvement iterations and the number of model building cycles were increased to 5 and 15, respectively (compared to the default *CRANK2* values); all expected SeMet in the model were in agreement with the map. Simulated annealing refinement by *phenix.refine* (30) was then used to improve the geometry and the obtained model was placed into the unit cell of the native data (L-Glu dataset) by MR with *MOLREP* (31). For the subsequent model building of the native L-Glu dataset, the P1 space group was chosen because of better data statistics compared to the alternative monoclinic

C2 assignment, good overall completeness of the data and the presence of only two molecules in the unit cell. The model was refined against native data by iterative rounds of *REFMAC5* restrained refinement (32), *phenix.refine* and manual editing in *Coot* (33). During refinement, additional positive density observed in both cavities in the 2|*F*|$_o$-|*F*|$_c$ and |*F*|$_o$-|*F*|$_c$ electron density maps allowed to unambiguously identify the L-Glu ligand (Fig. 3*B-E* and *SI Appendix*, Fig. S10*A-D*). The presence of the ligands was confirmed by bias-reduced simulated-annealing OMIT maps generated through the *PHENIX* suite (30); water molecules were added with *ARP/wARP* (Solvent module) (34) and the final stereochemistry was assessed by *MolProbity* (http://molprobity.biochem.duke.edu/) (35). MR by *MolRep* (31) using the ligand-deprived L-Glu structure allowed to obtain the Gly, L-Cys and L-Met structures. All 14 individual chains from the four crystal structures display an excellent structural match in their Cα traces (max rmsd 0.55 Å); the only significant difference is confined to the C-terminal stretch Lys240-Thr244 (including Cys243, which forms a disulfide bridge with Cys179), whose density has two alternative traces in 4 out of 14 chains and is absent in the rest: however, in almost all cases the density for the disulfide bridge is detectable. For the L-Met dataset, a moderate degree of anistropy was detected and therefore the reflection data were subjected to ellipsoidal truncation and anisotropic scaling through the UCLA Diffraction Anisotropy Server (http://services.mbi.ucla.edu/anisoscale/) (36); moreover, 29 residues in the chain B of the L-Met-bound structure (all comprised in the D2 domain) displayed missing or very poor density and therefore were not modelled.

Preparation of figures. All structural representations and superpositions were prepared with *PyMOL* (The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC).

Site-directed mutagenesis. All mutations listed in *SI Appendix*, Table S2 were obtained using QuikChange Site-Directed Mutagenesis Kit (Stratagene) on the above described pETM-14 : *At*GLR3.3 LBD plasmid and sequence-verified.

Expression tests of GLR3.3 LBD mutants. Each of the plasmids bearing a mutant version of GLR3.3 LBD was transformed into Rosetta strain *E. coli* cells. Small-scale (10 mL) cultures were grown at 37 °C in LB medium (supplemented with kanamycin and chloramphenicol) up to OD_{600} of 0.6-0.8. After that, they were subjected to either expression condition 1 (induction by 1 mM IPTG followed by shaking at 37 °C for 3 h) or 2 (induction by 0.1 mM IPTG followed by shaking at 20 °C for 16 h). The pelleted cells were then subjected to a shortened small-scale purification protocol limited to sonication and centrifugation, and samples for SDS-PAGE analysis were taken. A wild-type version of GLR3.3 LBD was included in all tests as positive control.

Homology modeling. All models were generated using the online server SWISS-MODEL (swissmodel.expasy.org) (37) providing the GLR3.3 LBD + L-Glu structure as input. Final model quality was assessed by the *MolProbity* score and QMEAN Z-score included in SWISS-MODEL calculations (see *SI Appendix*, Table S3 for details). To generate the GLR3.3 LBD-based models, the following residues from separate S1 and S2 segments (interspaced with the GGT linker) were used: GLR1.2, residues 441-547, 655-776 (numbered 1-232 in Fig. 4*C*); GLR1.4, residues 445-555, 663-785 (numbered 1-237 in Fig. 4*D*); GLR3.1, residues 469-575, 686-808; GLR3.4, residues 493-597, 708-836 (numbered 1-237 in Fig. 4*B*); GLR3.5, residues 487-590, 701-828. For the UniProt KB accession numbers of *At*GLR isoforms, see 'Sequence alignments' in this Appendix.

Cavity volume calculations. The CASTp software (http://sts.bioe.uic.edu/castp/) (38) was used to calculate the Connolly's solvent-excluded volume of the binding pocket, corresponding to the volume of the cavity contained within the contact molecular surface. The calculations were performed on the two datasets with best resolution, producing similar results: 196 Å^3 for the GLR3.3 + L-Glu pocket and 189 Å^3 for the GLR3.3 + Gly pocket.

Sequence alignments. Protein sequence alignments were performed with ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/) (39). However, all alignments were manually corrected after careful inspection of the superimposed structures. All final figures of alignments were prepared with ESPript (http://espript.ibcp.fr) (40). UniProtKB primary accession numbers (https://www.uniprot.org/) of all protein sequences used in the alignments are: *At*GLR1.1 Q9M8W7, *At*GLR1.2 Q9LV72, *At*GLR1.3 Q9FH75, *At*GLR1.4 Q8LGN1, *At*GLR2.1 O04660, *At*GLR2.2 Q9SHV1, *At*GLR2.3 Q9SHV2, *At*GLR2.4 O81776, *At*GLR2.5 Q9LFN5, *At*GLR2.6 Q9LFN8, *At*GLR2.7 Q8LGN0, *At*GLR2.8 Q9C5V5, *At*GLR2.9 O81078, *At*GLR3.1 Q7XJL2, *At*GLR3.2 Q93YT1, *At*GLR3.3 Q9C8E7, *At*GLR3.4 Q8GXJ4, *At*GLR3.5 Q9SW97, *At*GLR3.6 Q84W41, *At*GLR3.7 Q9SDQ4; *Dm*GluR1A Q03445; *Av*GluR1 E9P5T5; *Rn*GluA2 P19491; *Hs*GluK1 P39086; *Hs*GluN2A Q12879; *Ec*GlnBP P0AEQ3; *Ss*GluR0 P73797; *Os*GLR3.1 Q7XP59. Of those sequences for which a UniProtKB record is not available, the entry in the NCBI Protein database (https://www.ncbi.nlm.nih.gov/protein) is: *Pp*GLR1 XP_024390787.1; *Br*GLR3.4 XP_009118614.1. For Gin_bil2 (*Ginkgo biloba* putative GLR2) the sequence of isoform 8 (locus 13956) is taken from (41).

Statistical analysis. All the data are representative of at least ≥ 3 experiments. Reported traces are averages of traces from all single experiments used for the statistical analyses. Results are reported as averages ± standard deviations (SD). Statistical significance was assayed by Student t test and validated using One-way ANOVA (ANalysis Of VAriance) and with post-hoc Tukey HSD (Honestly Significant Difference) tests.

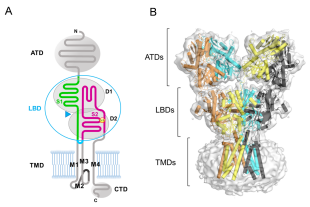


Fig. S1. Structure of iGluR/GLR channels. (*A*) General representation of one single eukaryotic iGluR/GLR subunit. Each functional channel is a homo- or heterotetramer of this subunit. Segment S1 is represented in green, S2 in magenta. The bilobed ligand-binding domain (LBD) is made up of domains 1 and 2 (D1 and D2); D1 residues are mainly contributed by segment S1 and D2 residues are mainly contributed by segment S2. The ligand (blue triangle) sits in a cleft between D1 and D2. The blue boundary encloses the *At*GLR3.3 LBD construct described in this work, with an arch indicating the site of the linker junction. The disulfide bridge (mostly conserved in eukaryotes) ties the final stretch of S2 to the D2 core. ATD, aminoterminal domain; M1 to M4: transmembrane segments of the transmembrane domain (TMD); CTD, C-terminal domain. (*B*) View of homotetrameric GluA2 (rat AMPA-subtype iGluR; PDB ID 5kbv, EMD ID 8232) (42). The 6.8-Å resolution cryo-EM map is shown as transparent surface and the four subunits of the model are shown in different colors with cylinder representation of α-helices. *A* is the general scheme of each one of these four subunits. Figure produced with *PyMOL* (The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC) from publicly available data.

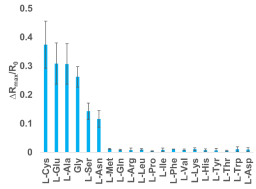


Fig. S2. Maximal relative amplitude of cpVenus/CFP ratio as $\Delta R/R_0$ increase triggered by the 20 amino acid L-enantiomers (1 mM) in the Col-0 wild-type seedling root tip expressing the NES-YC3.6 calcium sensor.

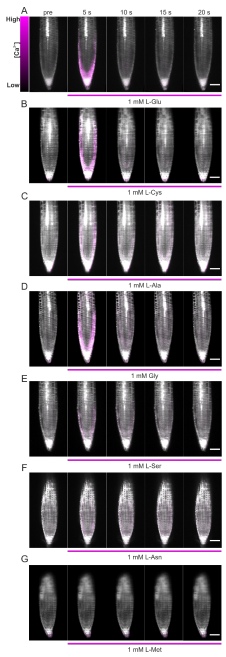


Fig. S3. Ratiometric purple-color images superimposed to cpVenus images from a representative time series visualized by LSFM of *Arabidopsis* Col-0 root tips expressing NES-YC3.6 treated with the 7 different amino acids used for the experiments shown in Fig. 1*E*. The different time series show cpVenus/CFP ratio changes in response to 1 mM L-Glu (*A*), 1 mM L-Cys (*B*), 1 mM L-Ala (*C*), 1 mM Gly (*D*), 1 mM L-Ser (*E*), 1 mM L-Asn (*F*), 1 mM L-Met (*G*). Numbers in the images indicate the time passed after acquisition start in seconds. Scale bar = 55 µm; n = 3.

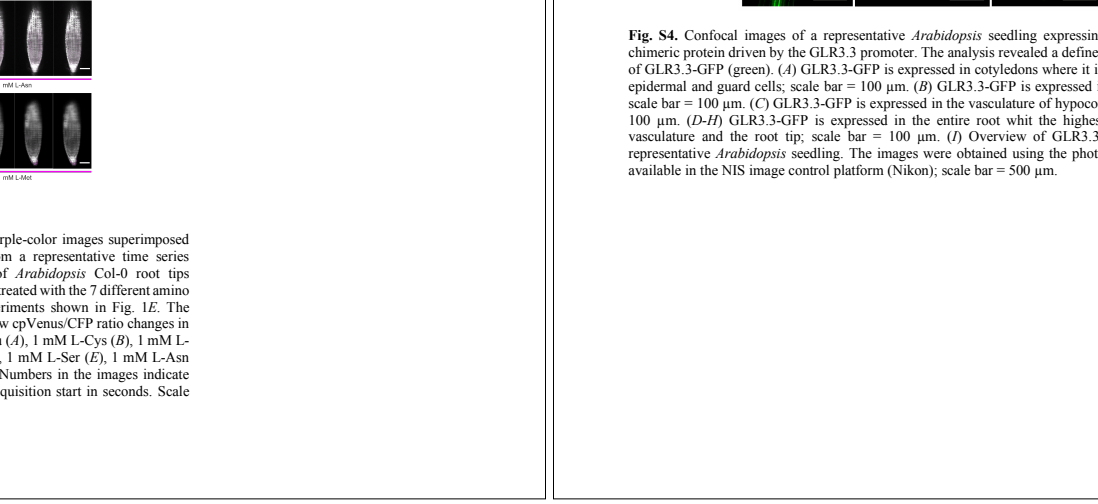


Fig. S4. Confocal images of a representative *Arabidopsis* seedling expressing the GLR3.3-GFP chimeric protein driven by the GLR3.3 promoter. The analysis revealed a defined expression pattern of GLR3.3-GFP (green). (*A*) GLR3.3-GFP is expressed in cotyledons where it is clearly detected in epidermal and guard cells; scale bar = 100 µm. (*B*) GLR3.3-GFP is expressed in the first true leaf; scale bar = 100 µm. (*C*) GLR3.3-GFP is expressed in the vasculature of hypocotyl cells; scale bar = 100 µm. (*D-H*) GLR3.3-GFP is expressed in the entire root whit the highest expression in the vasculature and the root tip; scale bar = 100 µm. (*I*) Overview of GLR3.3-GFP signal in the representative *Arabidopsis* seedling. The images were obtained using the photo stitching software available in the NIS image control platform (Nikon); scale bar = 500 µm.

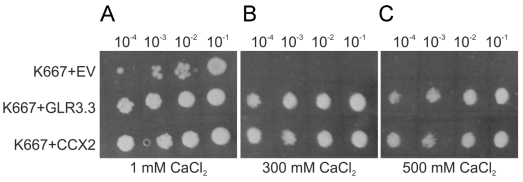


Fig. S5. Growth complementation assay of *S. cerevisiae* K667 transformed with pYES2-URA empty vector (EV), pYES2-URA harboring GLR3.3 or the *Arabidopsis* cation/Ca^{2+} exchanger CCX2 as positive control (11). Yeast cells were grown to OD_{600} of at least 1 and then 3 µl of serial dilutions were spotted onto SG-URA plates supplemented with 1 mM (*A*, control plate), 300 mM or 500 mM $CaCl_2$ (*B* and *C*, selective plates). The experiment is representative of two independent biological replicates showing similar results.

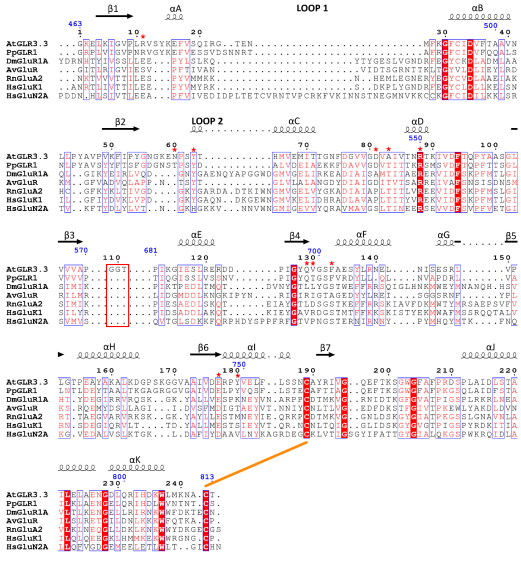


Fig. S6. Structure-based sequence alignment of LBDs (S1+S2 segments) from L-Glu-binding iGluRs/GLRs of different species. *At*GLR3.3: *Arabidopsis thaliana* GLR3.3 (this work); *Pp*GLR1: moss *Physcomitrella patens* GLR1; *Dm*GluR1A: *Drosophila melanogaster* GluR1A (PDB ID 5dt6); *Av*GluR1: rotifer *Adineta vaga* GluR1 (4io2); *Rn*GluA2: *Rattus norvegicus* AMPA-subtype GluA2 (1ftj); *Hs*GluK1: *Homo sapiens* kainate-subtype GluK1 (2zns); *Hs*GluN2A: *Homo sapiens* NMDA-subtype GluN2A (5h8f_A). At the top of the alignment, the *At*GLR3.3 secondary structure (α-helices as coils, β-strands as arrows), full-length numbering (blue) and numbering of the construct used in this paper (black) are shown. Location of the intervening M1-M2-M3 sequence (replaced by the GGT linker in the *At*GLR3.3 construct of this work) is indicated by a red box. The two Cys residues forming the disulfide bridge are connected by an orange line. Residues involved in ligand binding in the *At*GLR3.3 LBD structure are marked with red stars. See *SI Appendix*, Materials and Methods for the production of this alignment.

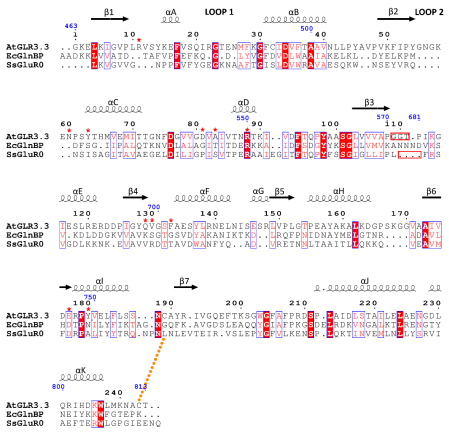


Fig. S7. Structure-based sequence alignment of LBDs (S1+S2 segments) of *At*GLR3.3 with prokaryotic homologous proteins. *Ec*GlnBP: *Escherichia coli* glutamine-binding protein (PDB ID 1wdn); *Ss*GluR0: cyanobacterium *Synechocystis sp.* GluR0 (1ii5). At the top of the alignment, the *At*GLR3.3 secondary structure (α-helices as coils, β-strands as arrows), full-length numbering (blue) and numbering of the construct used in this paper (black) are shown. Location of the intervening M1-M2-M3 sequence (replaced by the GGT linker in the *At*GLR3.3 construct of this work) is indicated by red boxes. Note that *Ss*GluR0 possesses transmembrane segments like *At*GLR3.3, whereas *Ec*GlnBP is a soluble clamshell-shaped periplasmic protein. The position of the disulfide bond in *At*GLR3.3 is indicated by orange dots. Residues involved in ligand binding in the *At*GLR3.3 LBD structure are marked with red stars. See *SI Appendix*, Materials and Methods for the production of this alignment.

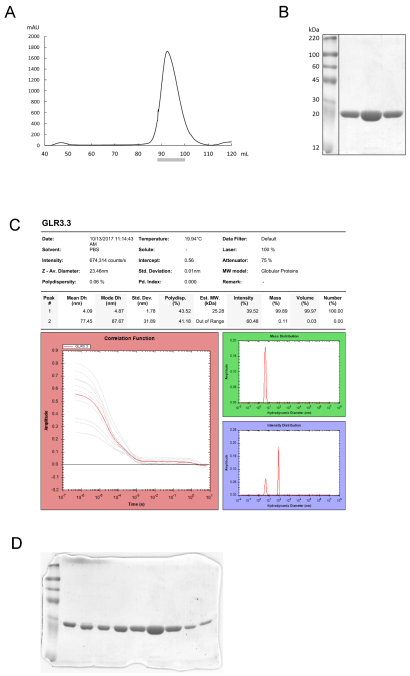


Fig. S8. Purification and characterization of GLR3.3 LBD. (*A*) Elution profile of GLR3.3 LBD from a preparative size-exclusion chromatography column (Superdex200 16/60, GE Healthcare) equilibrated with 10 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 0.5 mM L-Glu. The central fractions of the peak (marked) represent the final sample. (*B*) SDS polyacrylamide gel electrophoresis of three fractions (eluted from *A*) of purified GLR3.3 LBD (final sample), molecular weight ≈ 27kDa. Molecular weight marker: Blue Prestained Protein Standard, Broad Range (New England Biolabs). Gel: ExpressPlus PAGE (GenScript). (*C*) Report from a dynamic light scattering experiment on purified GLR3.3 LBD (1 mg/mL) loaded with L-Glu. The estimated molecular weight (25.28 kDa) is in agreement with the expected one for a monomeric sample (26.92 kDa). Instrument: Punk (Unchained Labs). (*D*) Original scan used to prepare the image in *B*.

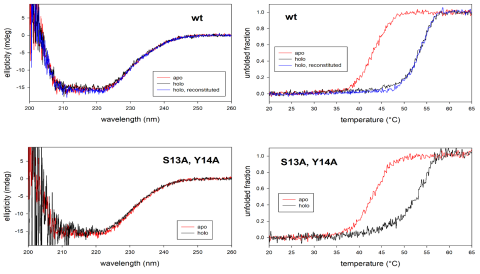


Fig. S9. Circular dichroism characterization of GLR3.3 LBD (wt, top, and mutant S13A-Y14A, bottom). Left panels: far-UV CD spectra; right panels: temperature ramps (the change in ellipticity at 220 nm was normalized as unfolded fraction). Black traces: holo (L-Glu-loaded); red traces: apo; blue traces: reconstituted holo (the reconstituted holo was obtained by addition of L-Glu to the apo). T_m values from the wt GLR3.3 LBD temperature ramps are 53.7 °C (holo), 42.9 °C (apo) and 53.8 °C (reconstituted holo), whereas T_m values from the GLR3.3 LBD S13A-Y14A are 53.7 °C (holo) and 43.2 °C (apo).

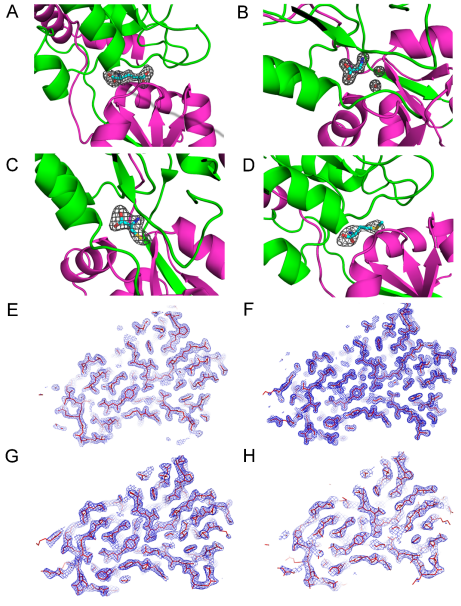


Fig. S10. Quality of the electron density maps. (*A-D*) The 1$\boldsymbol{F}$l$_o$-l$\boldsymbol{F}$l$_c$ electron density omit maps contoured at 3.0 σ are shown for L-Glu (*A*), Gly and the two associated waters (*B*), L-Cys (*C*) and L-Met (*D*). In the early rounds of refinement protein models lacking any ligand molecule produced maps with clear 1$\boldsymbol{F}$l$_o$-l$\boldsymbol{F}$l$_c$ electron densities for the ligands in the pockets. The ligand molecules were then added in the following rounds of refinement. The color code is the same used in Fig. 3*A-E*. See Fig. 3*B-E* for the 2l$\boldsymbol{F}$l$_o$-l$\boldsymbol{F}$l$_c$ omit maps of the ligands. (*E-H*) Representative 2l$\boldsymbol{F}$l$_o$-l$\boldsymbol{F}$l$_c$ electron density maps contoured at 1.5 σ at the end of refinements for the L-Glu- (*E*), Gly- (*F*), L-Cys- (*G*) and L-Met- (*H*) containing datasets.

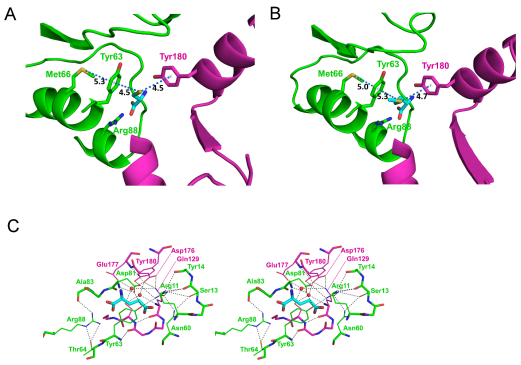


Fig. S11. Structural details of selected binding pockets in GLR3.3 LBD structures. (*A-B*) Close-up view of the ligand binding pocket in the crystal structures of GLR3.3 LBD + L-Cys (*A*) and L-Met (*B*). The ligands (cyan) and relevant side chains are in stick representation. Protein atoms from the S1 segment are green, from the S2 segment magenta. Oxygen is red, nitrogen blue, sulfur yellow. The orientation highlights the array of sulfur/π interactions (blue dashes) generated by the presence of the L-Cys and L-Met ligands. An almost straight line connects Met66 sulfur, the center of Tyr63 ring and the ligand sulfur. Distances between sulfur atoms and centers of the aromatic rings are indicated in Å. (*C*) View of the surroundings of the ligand-binding pocket of GLR3.3 LBD + L-Glu, with the same color codes as in *A-B*, showing the intricate network of interactions immediately outside the residues of Fig. 3*B*. Hydrogen or ionic bonds are shown as dashes; the ligand interactions shown in Fig. 3*B* have been omitted for clarity.

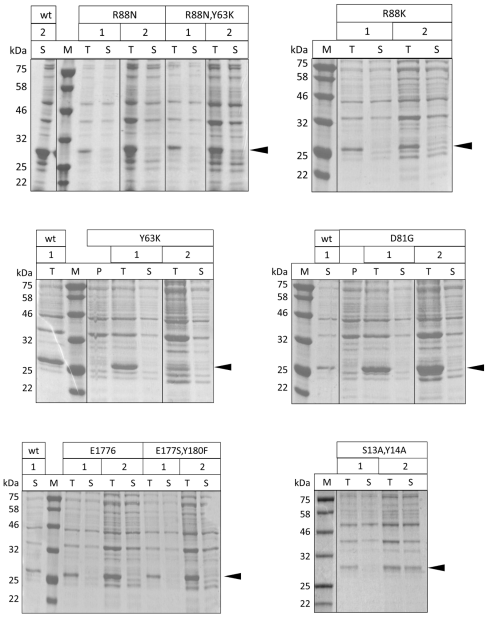


Fig. S12. SDS-polyacrylamide gel electrophoresis of fractions from small-scale expression tests of the GLR3.3 LBD mutants in *E. coli*. 1/2: condition 1 (induction by 1 mM IPTG followed by shaking at 37 °C for 3 h) or 2 (induction by 0.1 mM IPTG followed by shaking at 20 °C for 16 h); T: total cell lysate; S: soluble fraction; P: pre-induction sample; M: Blue Prestained Protein Standard, Broad Range molecular weight marker (New England Biolabs). Wt samples are included for comparison; His-tagged wt and mutant constructs have an approximate molecular weight of 29 kDa.

A

PROTEIN	K_d (μM)	*n*
3.3 LBD wt	2.2 ±0.5	5
3.3 LBD S13A-Y14A	2.1 ±0.4	2

B

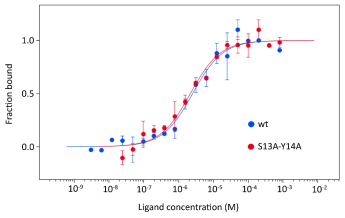


Fig.S13. Characterization of the binding properties of GLR3.3 LBD S13A-Y14A. (*A*) Values of the dissociation constants (K_d) ± SD for the binding of L-Glu to GLR3.3 LBD wt and S13A-Y14A, as determined by microscale thermophoresis; the values reported are averages from *n* repeats. (*B*) Fitting of the binding curves of L-Glu to GLR3.3 LBD wt and S13A-Y14A from the microscale thermophoresis experiments, based on the equation reported in *SI Appendix*, Materials and Methods; the graph reports the concentration of the ligand in logarithmic scale vs the thermophoretic signal normalized as fraction bound.

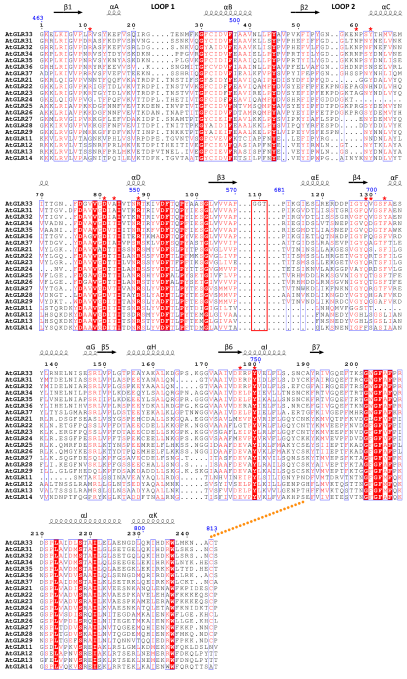


Fig. S14. Sequence alignment of the LBDs (S1+S2 segments) of all *A. thaliana* GLR isoforms, with the GLR3.3 LBD sequence (this work) at the top and the other sequences grouped by clade. Above the alignment, the GLR3.3 secondary structure (α-helices as coils, β-strands as arrows), full-length numbering (blue) and numbering of the construct used in this paper (black) are shown. Location of the intervening M1-M2-M3 sequence (replaced by the GGT linker in the GLR3.3 construct of this work) is indicated by a red box. The position of the disulfide bond in GLR3.3 is indicated by orange dots. Residues involved in ligand binding in the GLR3.3 LBD structure are marked with red stars. See *SI Appendix*, Materials and Methods for the production of this alignment.

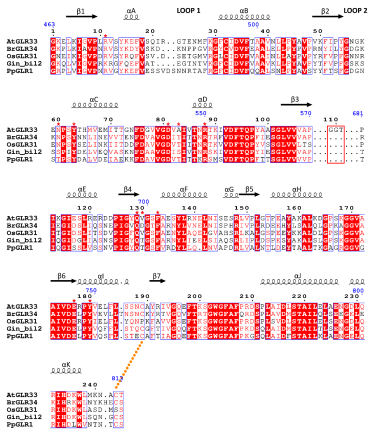
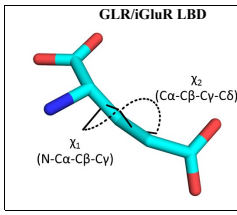


Fig. S15. Sequence alignment of LBDs (S1+S2 segments) of clade 3 GLRs from different plant species. *At*GLR3.3: *Arabidopsis thaliana* GLR3.3 (this work); *Br*GLR3.4: *Brassica rapa* GLR3.4; *Os*GLR3.1: *Oryza sativa* GLR3.1; Gin_bil2: *Ginkgo biloba* putative GLR2; *Pp*GLR1: moss *Physcomitrella patens* GLR1. At the top of the alignment, the *At*GLR3.3 secondary structure (α-helices as coils, β-strands as arrows), full-length numbering (blue) and numbering of the construct used in this paper (black) are shown. Location of the intervening M1-M2-M3 sequence (replaced by the GGT linker in the *At*GLR3.3 construct of this work) is indicated by a red box. The position of the disulfide bond in *At*GLR3.3 is indicated by orange dots. Residues involved in ligand binding in the *At*GLR3.3 LBD structure are marked with red stars. See *SI Appendix*, Materials and Methods for the production of this alignment.



	PDB ID	chain	χ_1 (°)	χ_2
Arabidopsis thaliana GLR3.3 (this work)	6r85	A	-81	-149
	6r85	B	-75	-151
Eukaryotes (non-plant)				
Rattus norvegicus GluA2 (AMPA-type)	1ftj	A	-78	-72
	1ftj	B	-76	-74
	1ftj	C	-73	-73
	3rn8	A	-74	-75
	3rn8	B	-76	-71
	3rn8	C	-75	-73
Rattus norvegicus GluA3 (AMPA-type)	3dln	A	-75	-75
Rattus norvegicus GluA4 (AMPA-type)	3epe	A	-75	-77
	3epe	B	-76	-71
Homo sapiens GluK1 (kainate-type)	2zns	A	-83	-67
Rattus norvegicus GluK2 (kainate-type)	1s50	A	-79	-67
Homo sapiens GluN2A (NMDA-type)	5h8f	A	-83	-60
Adineta vaga AvGluR1	4io2	A	-57	-70
	4io2	B	-59	-71
Drosophila melanogaster GluRIIB	4wxj	A	-73	-81
	4wxj	B	-70	-80
Drosophila melanogaster GluR1A	5dt6	A	-81	-70
Prokaryotes				
Synechocystis sp. GluR0	1ii5	A	-57	-179
Thermus thermophilus GluR0	1us5	A	-61	-177
Nostoc punctiforme GluR0	2pyy	A	-174	-175
	2pyy	B	-175	-174
	2pyy	C	-176	-178

Fig. S16. Indication of the χ_1 and χ_2 dihedral angles (°) of the L-Glu side chain, with a table reporting the values of χ_1 and χ_2 for a number of deposited structures of glutamate-bound GLR/iGluR ligand-binding domains.

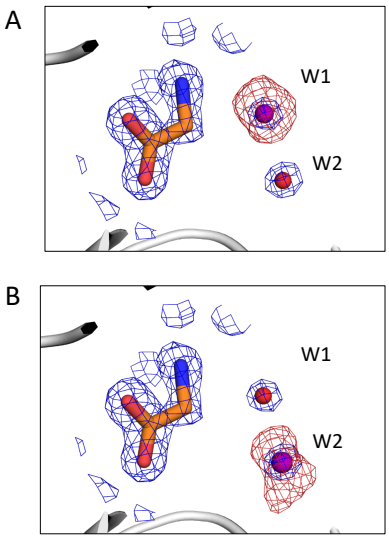


Fig. S17. Alternative refinement of GLR3.3 + Gly dataset. A Cl^- ion was placed in the position of either of the two additional water molecules in the Gly dataset ligand pocket and 5 cycles of restrained refinement were performed by the software *REFMAC5* (32); a clear peak of negative density in the $1|F|_o$-$1|F|_c$ electron density map (red mesh in the figure, showed at 3.0 σ contour level) appeared and the corresponding B-factors increased from 20.2 to 36.6 (W1 position) and from 21.9 to 37.3 (W2 position), ruling out the possibility that the spherical densities may correspond to ions rather than water molecules. Extending this operation to the four water molecules of the pocket in all chains of the Gly dataset invariantly causes increases of B-factors from a range of around 15-20 to a range of 37-60. The blue mesh corresponds to the $2|F|_o$-$1|F|_c$ electron density map at 1.5 σ contour level.

	SeMet +L-Glu	native +L-Glu	native +Gly	native +L-Cys	native +L-Met
Data collection					
Space group	$P4_12_12$	P1*	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell parameters (*a*,*b*,*c*, Å)	98.4, 98.4, 113.9	35.9, 61.3, 64.1	97.0, 98.2, 114.3	97.7, 98.5, 114.1	95.5, 96.8, 114.8
Cell parameters (α,β,γ, °)	90.0	75.2, 75.5, 90.0	90.0	90.0	90.0
Resolution (Å)	50.0-2.4	60.0-2.0	50.0-1.6	50.0-2.5	50.0-3.2
No. of monomers / asymm. unit	2	2	4	4	4
Observations	974572	105514	940331	260518	76068
Unique reflections	22542	33332	144028	38874	17844
R_{merge} [a]	0.28 (4.4)	0.10 (0.44)	0.07 (0.74)	0.25 (1.43)	0.42 (2.32)
Mean *I*/σ(*I*)	16.0 (1.2)	4.9 (1.6)	12.9 (1.7)	6.0 (1.1)	4.5 (1.4)
Completeness (%)	100.0 (99.8)	96.7 (94.7)	100.0 (100.0)	100.0 (100.0)	98.7 (99.6)
Multiplicity	43.2 (43.9)	3.2 (3.2)	6.5 (4.9)	6.7 (6.9)	4.3 (4.6)
Phasing					
Anomalous completeness (%)	100.0 (99.9)	-	-	-	-
Anomalous multiplicity	23.0 (22.8)	-	-	-	-
Overall FOM (centric/acentric) [b]	0.21 (0.09/0.23)	-	-	-	-
Refinement					
R-factor/R_{free} [c]	0.207/0.262	0.151/0.181	0.188/0.228	0.242/0.307	
No. of protein residues / monomer	238;239	242;238;238;243	238;239;239;238	239;209;237;237**	
Average B-factor (Å^2) [d]	24.1	21.2	38.1	39.7	
No. of ligand molecules	2	4	4	4	
Average B-factor (Å^2) [d]	17.2	16.5	37.7	32.1	
No. of ions	4	8	3	2	
Average B-factor (Å^2) [d]	34.4	46.2	46.9	17.5	
No. of water molecules	315	1048	387	1	
Average B-factor (Å^2) [d]	31.1	35.7	32.3	7.0	
rmsd bond lengths (Å) [e]	0.007	0.013	0.007	0.006	
rmsd bond angles (°) [e]	1.48	1.81	1.47	1.38	
Ramachandran plot					
in preferred regions (%)	97.4	97.6	97.5	98.0	
in allowed regions (%)	2.6	2.4	2.5	2.0	
outliers (%)	0.0	0.0	0.0	0.0	
MolProbity Score [f]	1.96 (76th percentile)	1.27 (97th percentile)	1.26 (100th percentile)	1.97 (99th percentile)	
PDB ID	6R85	6R88	6R89	6R8A	

(previous page)

Table S1. Crystallographic statistics. Values in parentheses are for the highest-resolution shell.

* Data from native crystals +L-Glu solved in space group C2 (a=124.2, b=35.9, c=61.3; α=90.0, β=105.4, γ=90.0) showed slightly worse statistics; all statistics reported here for native crystals +L-Glu refer to data solved in space group P1.
** In the chain B of the L-Met dataset, a total of 29 internal residues have not been included in the final PDB due to missing or very poor electron density.
[a] R-merge = $\Sigma_{hkl}\Sigma_j \mid I_{hkl,j} - \langle I_{hkl} \rangle \mid / \Sigma_{hkl}\Sigma_j I_{hkl,j}$.
The high R_{merge} value observed for the selenomethionine dataset and the L-Cys and L-Met datasets was due to the considerable redundancy of the dataset and/or a partial decay of the crystal during data collection. Maps calculated including all data were of higher quality than those calculated by including a largely redundant but more restricted subset of reflections with lower resolution and lower R_{merge}.
[b] Overall figure of merit (and for centric and acentric reflections) calculated by the program *Phaser* (43).
[c] R-factor = $\Sigma_{hkl} |Fobs_{hkl} - Fcalc_{hkl}| / \Sigma_{hkl} |Fobs_{hkl}|$ where Fobs and Fcalc are the observed and calculated structure factor amplitudes, respectively. R_{free} is the R-factor value for 5% of the reflections excluded from the refinement.
[d] Average B-factors calculated with the program Baverage from the *CCP4* suite (17).
[e] Root mean square deviations from ideal values calculated with *REFMAC5* (32).
[f] combines the clashscore, rotamer and Ramachandran evaluations giving one number that reflects the crystallographic resolution at which those values would be expected; from the server *MolProbity* (http://molprobity.biochem.duke.edu/) (35).

MUTANT	DESCRIPTION	CONDITION	EXPRESSION	SOLUBILITY	SCALE-UP
R88N	ligand-contacting residue	1	very good	very poor	no
		2	very good	very poor	no
R88N,Y63K	ligand-contacting residues	1	very good	very poor	no
		2	very good	very poor	no
R88K	ligand-contacting residue	1	very good	none	no
		2	very good	very poor	no
Y63K	ligand-contacting residue	1	very good	none	no
		2	good with degradation	very poor	no
D81G	residue from the outer network	1	very good	none	no
		2	very good	none	no
E177S	ligand-contacting residue	1	very good	none	no
		2	very good	none	no
E177S,Y180F	ligand-contacting residues	1	very good	none	no
		2	very good	very poor	no
S13A,Y14A	residues at domain interface	1	good	none	no
		2	good	good	yes

Table S2. Table listing the GLR3.3 LBD mutants generated and tested by small-scale expression in *E. coli*. The corresponding results are reported. Condition 1: induction by 1 mM IPTG followed by shaking at 37 °C for 3 h. Condition 2: induction by 0.1 mM IPTG followed by shaking at 20 °C for 16 h.

Target protein	UniProtKB entry[a]	Araport identifier[b]	% id[c]	GMQE[d]	QMEAN Z-score[e]	MolProbity score[f]
AtGLR1.2 (LBD)	Q9LV72	AT5G48400	31.2	0.67	-2.32	2.14
AtGLR1.4 (LBD)	Q8LGN1	AT3G07520	32.1	0.65	-3.97	2.43
AtGLR3.1 (LBD)	Q7XJL2	AT2G17260	65.1	0.83	-0.42	1.56
AtGLR3.4 (LBD)	Q8GXJ4	AT1G05200	61.5	0.81	-0.95	1.82
AtGLR3.5 (LBD)	Q9SW97	AT2G32390	60.3	0.82	0.12	1.62

Table S3. Homology modelling statistics.
All models were generated using the online server SWISS-MODEL (swissmodel.expasy.org) (37). In all cases the areas affected by the lowest local reliability correspond to the exposed loop 1 (Fig. 3*A*), except for the *At*GLR1.4 LBD model, where all the exposed loops have a low quality score.
[a] UniProt: https://www.uniprot.org/
[b] Araport: https://www.araport.org/
[c] % sequence identity with GLR3.3 LBD
[d] Global Model Quality Estimation (number between 0 and 1) is a quality estimation which combines properties from the target–template alignment and coverage of the target.
[e] The QMEAN Z-score indicates how far the QMEAN score (44) of the model is from what one would expect from experimental structures of similar size. QMEAN Z-scores around zero indicate good agreement between the model structure and experimental structures of similar size. Scores of -4.0 or below indicate low quality of the model. The QMEAN score itself estimates global and local quality of geometry in one single model.
[f] Combines the clashscore, rotamer and Ramachandran evaluations giving one number that reflects the crystallographic resolution at which those values would be expected; from the server *MolProbity* (35).

SI REFERENCES

1. Clough SJ, Bent AF (1998) Floral dip: A simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J* 16(6):735–743.
2. Farquharson KL (2018) Small Talk: Protons Help Calcium Get the Message Across. *Plant Cell* 30(12):2885–2886.
3. Candeo A, Doccula FG, Valentini G, Bassi A, Costa A (2017) Light Sheet Fluorescence Microscopy Quantifies Calcium Oscillations in Root Hairs of Arabidopsis thaliana. *Plant Cell Physiol* 58(7):1161–1172.
4. Qi Z, Stephens NR, Spalding EP (2006) Calcium Entry Mediated by GLR3.3, an Arabidopsis Glutamate Receptor with a Broad Agonist Profile. *Plant Physiol* 142(3):963–971.
5. Krebs M, et al. (2012) FRET-based genetically encoded sensors allow high-resolution live cell imaging of Ca 2+ dynamics. *Plant J* 69(1):181–192.
6. Behera S, Kudla J (2013) High-Resolution Imaging of Cytoplasmic Ca2+ Dynamics in Arabidopsis Roots. *Cold Spring Harb Protoc* 8(7):670–674.
7. Romano Armada N, et al. (2019) Calcium Signalling Churchill.pdf. 1925:87–101.
8. Maizel A, Von Wangenheim D, Federici F, Haseloff J, Stelzer EHK (2011) High-resolution live imaging of plant growth in near physiological bright conditions using light sheet fluorescence microscopy. *Plant J* 68(2):377–385.
9. Cunningham KW, Fink GR (2015) Calcineurin inhibits VCX1-dependent H+/Ca2+ exchange and induces Ca2+ ATPases in Saccharomyces cerevisiae. *Mol Cell Biol* 16(5):2226–2237.
10. Wudick MM, et al. (2018) CORNICHON sorting and regulation of GLR channels underlie pollen tube Ca2+homeostasis. *Science (80-)* 360(6388):533–536.
11. Corso M, Doccula FG, de Melo JRF, Costa A, Verbruggen N (2018) Endoplasmic reticulum-localized CCX2 is required for osmotolerance by regulating ER and cytosolic Ca^{2+} dynamics in *Arabidopsis*. *Proc Natl Acad Sci* 115(15):3966–3971.
12. Bonza MC, Luoni L, De Michelis MI (2004) Functional expression in yeast of an N-deleted form of At-ACA8, a plasma membrane Ca2+-ATPase of Arabidopsis thaliana, and characterization of a hyperactive mutant. *Planta* 218(5):814–823.
13. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191.
14. Doublié S (1997) Preparation of Selenomethionyl Proteins for Phase Determination. *Methods Enzymol* 276:523–530.
15. Seidel SAI, et al. (2012) Label-free microscale thermophoresis discriminates sites and affinity of protein-ligand binding. *Angew Chemie - Int Ed* 51(42):10656–10659.
16. Guijarro M, et al. (2012) ID29: a high-intensity highly automated ESRF beamline for macromolecular crystallography experiments exploiting anomalous scattering. *J Synchrotron Radiat* 19(3):455–461.
17. Winn MD, et al. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr Sect D Biol Crystallogr* 67(4):235–242.

18. Kabsch W (2010) *XDS*. *Acta Crystallogr Sect D Biol Crystallogr* 66(2):125–132.
19. Evans PR, Murshudov GN (2013) How good are my data and what is the resolution? *Acta Crystallogr Sect D Biol Crystallogr* 69(7):1204–1214.
20. Karplus PA, Diederichs K (2012) Linking crystallographic model and data quality. *Science (80-)* 336(6084):1030–1033.
21. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
22. Rychlewski L, Li W, Jaroszewski L, Godzik A (2010) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* 9(2):232–241.
23. Schuermann JP, Tanner JJ (2003) MRSAD: Using anomalous dispersion from S atoms collected at Cu Ka wavelength in molecular-replacement structure determination. *Acta Crystallogr - Sect D Biol Crystallogr* 59(10):1731–1736.
24. Panjikar S, Parthasarathy V, Lamzin VS, Weiss MS, Tucker PA (2009) On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination. *Acta Crystallogr Sect D Biol Crystallogr* 65(10):1089–1097.
25. Potterton L, et al. (2018) CCP 4 i 2: The new graphical user interface to the CCP 4 program suite. *Acta Crystallogr Sect D Struct Biol* 74:68–84.
26. Skubák P, Pannu NS (2013) Automatic protein structure solution from weak X-ray data. *Nat Commun* 4:1–6.
27. Cowtan K (2006) The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr Sect D Biol Crystallogr* 62(9):1002–1011.
28. Smart OS, et al. (2012) Exploiting structure similarity in refinement: Automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr Sect D Biol Crystallogr* 68(4):368–380.
29. Blanc E, et al. (2004) Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr Sect D Biol Crystallogr* 60(12 I):2210–2221.
30. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr Sect D Biol Crystallogr* 66(2):213–221.
31. Vagin A, Teplyakov A (2010) Molecular replacement with MOLREP. *Acta Crystallogr Sect D Biol Crystallogr* 66(1):22–25.
32. Murshudov GN, et al. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr Sect D Biol Crystallogr* 67(4):355–367.
33. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot . *Acta Crystallogr Sect D Biol Crystallogr* 66(4):486–501.
34. Biswal HS, Wategaonkar S (2009) Sulfur, Not Too Far Behind O, N, and C: SH center dot center dot center dot pi Hydrogen Bond. *J Phys Chem a* 113(46):12774–12782.
35. Chen VB, et al. (2010) MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr Sect D Biol Crystallogr* 66(1):12–21.
36. Strong M, et al. (2006) Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *Proc Natl Acad Sci* 103(21):8060–8065.
37. Waterhouse A, et al. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 46(W1):W296–W303.
38. Cheng Y, Zhang X, Sun T, Tian Q, Zhang WH (2018) Glutamate Receptor Homolog3.4 is Involved in Regulation of Seed Germination under Salt Stress in Arabidopsis. *Plant Cell Physiol* 59(5):978–988.
39. Sievers F, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.
40. Robert X, Gouet P (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res* 42(W1):320–324.
41. De Bortoli S, Teardo E, Szabò I, Morosinotto T, Alboresi A (2016) Evolutionary insight into the ionotropic glutamate receptor superfamily of photosynthetic organisms. *Biophys Chem* 218:14–26.
42. Twomey EC, Yelshanskaya M V., Grassucci RA, Frank J, Sobolevsky AI (2016) Elucidation of AMPA receptor-stargazin complexes by cryo-electron microscopy. *Science (80-)* 353(6294):83–86.
43. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Crystallogr* 40(4):658–674.
44. Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* 27(3):343–350.

ABSTRACT

In the last decades of protein crystallography, the crystallization of contaminant proteins in place of the proteins of interest, or target proteins, has been reported several times despite the improvements in the expression and purification protocols, the availability of *ad hoc* software for contaminant check and the increasing awareness of crystallographers about this issue. In the vast majority of the cases, the contaminant protein comes from the expression organism (often *E. coli*) but the possibility of a contamination from other organisms exists and has been reported in rare and exceptional circumstances. Here, a case of contamination from a *Serratia* strain during attempts to crystallize a protein of interest is presented and discussed. The contamination led to the unintended crystallization of the cyanase hydratase from a bacterium of the *Serratia* genus in a novel crystal form. Oxalate, a natural inhibitor of the enzyme, was found in all the active sites. The origin of contamination, *Serratia*, is an opportunistic enterobacteria that can be found in a variety of habitats, including the laboratory environment where it grows in conditions similar to the ones of *E. coli*. This case shows that contamination from organisms other than the ones used for over-expression is not only possible but is likely to be more common and serious than expected. Furthermore, it suggests that a thorough check for contamination should become an essential and integral step in data analysis prior to any structure determination attempt and it encourages the deposition of known and unknown contaminant structures to further aid the identification of unintended proteins.

INTRODUCTION

It is often the case that, after the structure of a protein has been determined, it becomes interesting to study the structure of the same protein with one or more point mutations. Usually, the aim is to investigate the role of key residues present in the catalytic active site and thought to be involved in the protein activity and function. Structure determination of the mutant proteins is usually a straightforward task: this is because, in favorable cases, their structures are similar enough to the one of the wild-type to enable the use of Molecular Replacement with the wild-type protein as a search model. However, there are cases in which this task becomes more difficult and sometimes even impossible. Occasionally, even a single-residue mutation can induce a local or a global change in the structure, thus altering the conformation to the point that MR does no longer represent a viable phasing strategy. Rarely, it is also possible that a contaminant protein is crystallized in place of the protein of interest. In both situations, a considerable amount of time and efforts might be invested in unsuccessful MR phasing attempts before it becomes evident that either the conformation has significantly changed or that the nature of the crystal is not the one of the intended target (Niedzialkowska *et al.*, 2016).

The case of a contamination is probably worse than the situation when the crystallized protein is still the intended target, but with a significant conformational change. In fact, contamination is not always easy to spot, and the contamination hypothesis can appear unrealistic. This can be due to a number of factors: *i)* the search for structures with similar cell parameters might not return any hit, *ii)* the molecular weight of the target and of the contaminant proteins are similar enough to make very difficult to distinguish them from the gels and *iii)* only recently, powerful software have become available for the detection of contaminants (Ramraj *et al.*, 2012; Hungler *et al.*, 2016; Simpkin *et al.*, 2018).

Contamination is an unlikely event, and even more improbable is the contamination from an organism that is not the one used for during the over-expression. There is only a very limited number of such cases reported in literature (Musille & Ortlund, 2014; Butryn *et al.*, 2015).

Here, a case of contamination from a *Serratia* strain during attempts to crystallize another target protein is presented. The strategy that was devised to phase the unwanted protein is presented, as well as the steps that led to the discovery of the contamination and its source. Different hypothesis on the contamination are discussed, together with the numerous factors that complicated the identification of the contamination issue.

5. MATERIALS & METHODS

5.1. Protein purification and crystallization

Crystals of the *Serratia* cyanase hydratase were obtained during efforts to crystallize a mutant of 4B5C from *Burkholderia pseudomallei* (Gori *et al.*, 2013). The gene was cloned into the pET14b expression vector with thrombin cleavage site and transformed into BL21 (DE3) plyS competent cells. The expression was performed in SB culture media supplemented with Ampicillin and Cloroamphenicol (100mg/ml in water and 34 mg/ml in ethanol stock solutions, respectively) and the cells were induced with IPTG 0.5mM at 20°C overnight. The cells were harvested and lysed with cell disruptor, centrifuged for 1h at 18000g and the cell lysate was then loaded on a Profinia (BioRad®) column. The protein was eluted with 250 mM imidazole and desalted with a PD-10 column using a buffer with composition: 10mM Tris pH 8.0, 100mM NaCl and 10% glycerol. The final protein concentration was ~ 30mg/mL and it was screened for crystallization using the sitting-drop vapor diffusion method. 96-well plates (Molecular Dimensions®) using different protein concentrations (30, 50 and 70%) were set up, maintaining the final drop volume at 0.4 µL. After approximately two months, few crystals were obtained at 297 K from condition A6 of the JBScreen Classic 4 (15% w/v Polyethylene glycol 6000, 5% w/v Glycerol) from Jena Bioscience. These crystals were used for X-ray data collection.

5.2. Data collection and processing

Data collection statistics are reported in Table 1 (to be filled as soon as the refinement is completed). A single crystal grown in the optimized condition was soaked in a cryo-protectant solution (25% glycerol, 15% Polyethylene glycol 6000), flash frozen in liquid nitrogen and diffraction data was collected at 100 K on the ID29 beamline at the ESRF synchrotron (Sanctis *et al.*, 2012) using the Pilatus 6M-F pixel detector (Dectris). The native data set was collected at a wavelength of 1.000 Å and indexed in space group $P2_1$, truncating the resolution to 2.09 Å. The diffraction data set was processed using *XDS* (Kabsch, 2010) and scaled and merged with *AIMLESS* (Evans & Garib, 2013); the high-resolution data cut-off was based on the statistical indicators $CC_{1/2}$ and CC* (Karplus & Diederichs, 2015).

5.3. Analysis of the unit cell and solvent content

Solvent content calculation based on the Matthews coefficient (V_M) (Matthews, 1968, 1976) suggested the presence of a number of monomers in the ASU between 10 (V_M = 3.01 Å^3 Da^{-1} and ~ 59% solvent) and 15 (V_M = 2.01 Å^3 Da^{-1} and ~ 39% solvent), with maximum probability for 13 copies. The Self-Rotation Function was computed using the native data with *MolRep* (Vagin, A., Teplyakov, 1997) and *POLARFN* (https://services.mbi.ucla.edu/selfrot/), for different values of the integration radius and data resolution. In all cases, the SRF showed similar features (**Figure 5.1**), notably the five 2-fold peaks at k = 180° and the single 5-fold peak at k = 72°. These features suggested a complex with D5 symmetry, *i.e.* a double pentameric structure, with the two rings stacked onto each other. This hypothesis was supported by the analysis of the solvent content and was later used to plan and guide the phasing strategy.

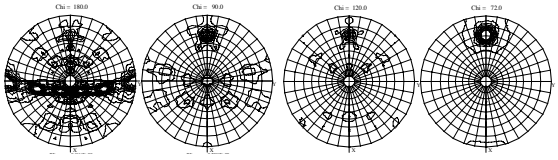


Figure 5.1: Self-Rotation function computed with MolRep for the native data set at different k sections (radius of integration = 15 Å, resolution = 2.32 Å)

5.4. Structure solution and refinement

All the attempts at structure solution using Molecular Replacement with a number of search models based on chain A of the 4B5C model were unsuccessful. As described below, a thorough check for contaminants eventually identified the crystallized protein as a cyanate hydratase. Molecular replacement with the contaminant model gave a clear solution. The MR-solution was subjected to *ARP/wARP* model building (Langer *et al.*, 2008), after which the origin of the contaminant protein was confirmed to be from a bacterium of the *Serratia* genus. Cycles of refinement with *REFMAC5* (Murshudov *et al.*, 2011) and *phenix.refine* (Adams *et al.*, 2010) (with the application of NCS-restraints) and *Coot* manual editing (Emsley & Lohkamp, 2010) completed and improved the model. The final structure also contains, in each of the enzyme active sites, one oxalate ion. Refinement statistics are reported in Table 1. The coordinates of the final model were deposited in the PDB with accession code WXYZ.

Table 1: Data collection and refinement statistics

Values in parenthesis are for the outer shell

Data collection	
Diffraction source	
Wavelength (Å)	
Detector	
Space group	
Cell parameters (*a*, *b*, *c*, Å)	
Cell parameters (α, β, γ, °)	
Mosaicity (°)	
Solvent content (%)	
Molecules per asymmetric unit	
Matthews coefficient ($Å^3$ Da^{-1})	
d_{max} - d_{min} (Å)	
Total No. of reflections	
No. of unique reflections	
Completeness (%)	
Multiplicity	
Mean *I/σ(I)*	
R_{meas}	
Overall *B*-factor from Wilson plot ($Å^2$)	
Refinement	
Resolution range	
No. of reflections	
Final R_{work}	
Final R_{free}	
Number of non-H atoms:	
Protein	
Ligand	
Ions	
Water	
R.m.s. deviations	
Bonds (Å)	
Angles (°)	
Average *B*-factors ($Å^2$):	
Protein	
Ligand	
Ion	
Water	
Ramachandran plot	
Favored regions (%)	
Additionally allowed (%)	
Outliers (%)	

6. RESULTS & DISCUSSION

6.1. Initial MR attempts

Initial attempts at solving the structure using Molecular Replacement with chain A of the 4B5C model were unsuccessful. At the beginning of the project, a contaminant search was performed by screening the entire PDB for structures having unit cell parameters similar to the ones of the collected data, but no hits were found. As a consequence, more sophisticated MR tests were performed which were based on the suggested oligomeric assembly deduced from the SRF and the solvent content analysis. A number of docking software (*SAM* (Ritchie & Grudinin, 2016), *HSYMDOCK* (Yan *et al.*, 2018), *ROSETTA SYMMETRY DOCKING* (Bradley *et al.*, 2007), *GalaxyWeb* (Ko *et al.*, 2012), *SYMMDOCK* (Schneidman-duhovny *et al.*, 2005) and *MZDOCK* (Pierce *et al.*, 2014)) were used to generate models with D5 symmetry starting from chain A of the 4B5C model: in a first step, the quality of these models (~ 2000) was tested with a self-written automated pipeline prepared in bash script with a Python wrapping. This pipeline employs *PHASER* (Mccoy *et al.*, 2007) and *MolRep* with default settings. The most promising solutions, as judged by the most important MR-indicators (TFZ equivalent, LLG and packing for *PHASER*; Score, contrast, TF/sigma and wRFac for *MolRep*) and the *R*-factors, were kept and used for the second step. Here, the models from the first step are tested in a different automated pipeline which uses *PHASER* and *MolRep* and varies some of the parameters known to be critical for the success of MR (data resolution and expected r.m.s.d. for *PHASER*; data resolution, similarity, completeness and number of rotation peaks for *MolRep*). The most promising models (selected with the same criteria used at the end of the first step and described above) were retained and subjected to *REFMAC5* refinement, *SHELXE*-expansion (Thorn & George, 2013) and/or NCS-averaging with DM. However, none of the MR solutions could be successfully refined, expanded or its density improved by any of the methods listed above. This suggested that many of the MR solutions, which were initially considered as promising, were in fact false-positives.

6.2. Contaminant search and identification of the source of contamination

At this point, a second, more thorough check for contaminants was carried out using the recently developed program *SIMBAD*. The program could not find a solution during the first step (*i.e.*: the screening of the PDB unit cell parameters), but quickly identified PDB ID 4Y42 (Butryn *et al.*, 2015) as the likely contaminant in the subsequent MR-search. In fact, a solution was initially found with the *AMoRE* rotation function (Z-score = 54.6), which was confirmed by full MR-search with *MolRep* (Score = 0.6798, TF-score = 19.16). After the *SIMBAD* result, the full MR-search with *MolRep* using 4Y42 as search model was repeated independently, followed by *REFMAC5* restrained refinement and one cycle of *ARP/wARP* model building, which confirmed the crystallized protein to be a cyanate hydratase, likely from *Serratia*. To confirm the contaminant origin, the following method was used. A main-chain only model was built with *ARP/wARP*, containing dummy atoms in place of the side-chains electron density. Then, using methods recently described (Chojnowski *et al.*, 2019), a Position-Specific Scoring Matrix (PSSM) of the contaminant sequence was generated. The PSSM was used to query a number of databases with *PSI-BLAST* and *HMMER* (Finn *et al.*, 2011) in order to find matching known sequences. The best matching sequence found thorough this first iteration (*E*-value of $3.8 \cdot 10^{-41}$, from a strain of *Serratia proteomaculans*) was used to build a full model with *ARP/wARP*. The quality of the model built in this second iteration allowed to unequivocally confirm the initial hypothesis of contaminant from *Serratia*. Further cycles of refinement with *REFMAC5* and *phenix.refine* (applying the NCS-restraints) and *Coot* manual editing led to the re-assignment of few residues owing to the better side chains electron density. Alignment with the sequence extracted from the final model shows that the protein comes from an organism of the *Serratia* genus, without showing the exact species.

6.3. Hypothesis about the contamination from *Serratia*

In order to better understand the origin of the contamination and, as a consequence, to reduce the possibility of this event to happen in the future, several hypotheses have been considered and analyzed. Among them, the contamination at the protein expression stage appears to be the most plausible one. In fact, several antibiotic-resistant *Serratia* strains are known for their ability to grow in the presence of ampicillin (Stock *et al.*, 2003). The fact that ampicillin is easily degraded, which would make easy for *Serratia* to grow in the culture media, supports the hypothesis of contamination during the expression step. Carbenicillin, normally used as a more stable alternative, was not employed as there have never been contamination problems using ampicillin in the past. In addition, *Serratia* grows in similar conditions to *E. coli* and can be found in the environment and even in the laboratory, particularly in the grooves of the floor. The combination of these factors (non-fresh or simply easily degradable ampicillin stocks,

combined with ubiquitous antibiotic-resistant *Serratia* strains and non-sterile laboratory environment) is likely to be responsible for the contamination during the expression stage. Purification of the His-tagged proteins using immobilized metal-ion affinity chromatography (IMAC) is often employed as a first step after protein expression. However, due to non-specific binding of non-target proteins to the nickel beads, IMAC is usually not sufficient to guarantee high protein purity (Joshua A. Bornhorst and Joseph J. Falke, 2000). Assuming that the contaminant protein is highly expressed by *Serratia*, the contaminant likely bound to the nickel resin and/or aggregated with the target protein and co-eluted from the IMAC column. The binding of the contaminant might have been facilitated by its large surface area, which is a consequence of its oligomerization state. For the contamination to happen, even small quantities of the contaminant are sufficient: indeed, a concentration as low as 5% in the protein preparation has been reported, which allowed the crystallization of the unintended target (Veesler & Cambillau, 2008). A contamination during the crystallization step seems implausible, too, but cannot be excluded.

6.4. Description of structure

All individual chains are virtually identical as they display an excellent structural match in their Cα traces, with a maximum Cα r.m.s.d. of 0.203 Å. The protein structure resembles very closely the ones of the cyaneses deposited in the PDB. It is composed of ten protomers, each of them consisting of two domains: the N-terminal domain forming a 5-helix bundle, and the C-terminal catalytic domain having a unique fold. Pairs of protomers are organized to form dimers through an intricate interaction of two C-terminal cyanase domains, and the dimers assemble into a decamer with D5 symmetry. The interface between dimers originates the five symmetrically disposed active sites of the enzyme, where the residues forming the catalytic triad (Arg96, Glu99, Ser122 and their NCS-related equivalents at the interface of two adjacent dimers) are responsible for the substrates binding. In our case, clear density for the substrate could be observed. In fact, after few refinement cycles, additional positive and symmetrical electron density started to appear from the $2F_o - F_c$ and $F_o - F_c$ electron-density maps in all the five active sites. Further refinement improved the density and allowed to unambiguously assign the ligand present in the active sites as oxalate ion. The presence of one oxalate ion in the five active sites was confirmed by calculating omit maps, and an example is shown in **Figure 6.1** (temporary figure to be replaced by a better one). Oxalate, together with other low-molecular weight dicarboxylic acids and mono-anions, is a known inhibitor of *E. Coli* CynS (Anderson *et al.*, 1987) and can easily be found in the culture media, possibly as a product of bacteria metabolism. Beside the oxalate ions, the structure also contains a number of glycerol molecules coming from the successful crystallization condition and from the cryo-solution.

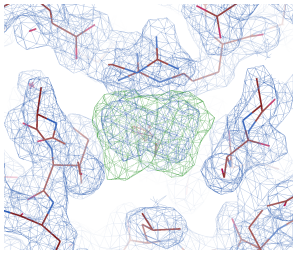


Figure 6.1: Verification of the presence of oxalate in the active sites of the enzyme. The 2.1 Å resolution $2mF_o - DF_c$ maximum-likelihood omit map contoured at 1σ (blue) and the corresponding $mF_o - DF_c$ map contoured at $+3\sigma$ (green) and -3σ (red) are shown. The omit maps were computed after omitting the ligand in question and subsequent 20 cycles of REFMAC5 restrained refinement.

→ For the computation of the omit maps, I followed two procedures which gave the same results (*i.e.*: they both confirmed the presence of the ligands):

1) **Removal of the ligand and subsequent *REFMAC5* restrained refinement** (as described above). The *REFMAC5* map obtained by this procedure contains two sets of map coefficients: FWT/PHWT (amplitude and phase for the weighted omit "$2F_o - F_c$" map, *i.e.*: $2mF_o - DF_c$) and DELFWT/PHDELWT (amplitude and phase for the weighted 'difference' map, *i.e.* $mF_o - DF_c$). I then decided to compute an omit map for the entire structure and not just for a single ligand:
2) **Calculation of a "full" composite omit map with *phenix.composite_omit_map***, which computes an omit map covering the full unit cell, with standard refinement and with simulated annealing. This omit map has coefficients $2mF_o - DF_c$.

7. CONCLUSIONS

In conclusion, we report here on the unintentional crystallization of a cyanate hydratase from a bacterium of the *Serratia* genus in complex with a natural inhibitor, in a novel crystal form and in a new crystallization condition. This result adds to two previous reports (Musille & Ortlund, 2014; Butryn *et al.*, 2015) of unexpected contamination from *Serratia*. These results are important under several aspects. First of all, they suggest that the contamination from organisms other than the ones used for the over-expression is possible and probably more likely to occur than one might expect. Secondly, it shows that not expected and pathogenic organisms can easily grow and proliferate in the laboratory environment and accumulate to the point that they can contaminate machineries, reagents etc…Thirdly, the results confirm that the contamination process is serendipitous by its nature and should be expected at any time. Problems with contamination never occurred in the case of 4B5C but manifested for the mutant, which confirms that even small changes in the sequence and/or in other experimental variables, can have a significant impact on the expression, purification and/or the crystallization steps. Some lessons can be learnt from this and previous reports: before anything, these cases highlight the importance of good laboratory practices. The proper and regular cleaning of the laboratory, including all the instrumentation, is of primary importance and is probably the most important way to reduce (and, hopefully, to avoid) any unintended contamination. This adds to the necessary checks (SDS-page, mass-spectrometry, chromatographic analysis in particular) that must be carried out during the expression and purification of the protein of interested. All these measures will probably not exclude the possibility of a contamination but will certainly reduce it more than any other practice. In the case presented here, several factors contributed together delaying the detection of the contaminant. First of all, the negative results obtained from the screening of the PDB for structures of similar cell parameters, explained by the absence of deposited structures with the same space group (this limitation has been recently overcome with the implementation of programs, as *SIMBAD*, for the rapid screening of large databases of structures by MR). Other factors are the false-positive MR results and the absence of any contamination problem with the wild-type protein. In addition to these factors, hypothesis made by the researchers can further delay the discovery of the contamination. In our case, a contaminant was initially ruled out because contaminants have, usually, a small size. Despite the fact that known contaminants have a wide range of molecular weights, from 10 to 140 kDa, the most of them weight between 20 and 40 kDa, which also corresponds to the weight range of 70% of all crystallized proteins reported in the PDB (Hungler *et al.*, 2016). This indeed confirms the rarity of this case, as the molecular weight of the *Serratia* CynS is approx. 170 kDa. Moreover, because of the time it took to the crystals to grow, partial proteolysis of the protein was suspected and assumed to be one of the factors complicating the MR search. As a consequence, it might be difficult to suspect and detect early on contamination problems. For all these reasons, a check for contaminants should become an essential part of the data analysis after data collection and processing and before any attempt at structure solution. The simplicity and the rapidity of a contamination check should convince crystallographers to include this step into the routine process of data analysis and it would most likely reduce cases similar to the one reported in this and other reports. In parallel, crystallographers should be encouraged to report similar cases and to deposit the contaminant structures, even when they are already known. Increasing the number of deposited contaminant structures will aid the identification of such contaminants by other crystallographers.

ACKNOWLEDGEMENTS

To be filled

REFERENCES

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221.

Anderson, P. M., Johnson, W. V., Endrizzi, J. A., Little, R. M. & Korte, J. J. (1987). *Biochemistry*. **26**, 3938–3943.

Bradley, P., Wang, C. & Baker, D. (2007). *PNAS.* **104**, 17656–17661.

Butryn, A., Stoehr, G. & Linke-winnebeck, C. (2015). *Acta Crystallogr. Sect. F*. **71**, 471–476.

Chojnowski, G., Pereira, J. & Lamzin, V. S. (2019). *Acta Crystallogr. Sect. D*. **75**, 753–763.

Emsley, P. & Lohkamp, B. (2010). *Acta Crystallogr. Sect. D*. 486–501.

Evans, P. R. & Garib, N. (2013). *Acta Crystallogr. Sect. D*. **69**, 1204–1214.

Finn, R. D., Clements, J. & Eddy, S. R. (2011). *Nucleic Acids Res.* 1–9.

Gori, A., Rinchai, D., Gourlay, L. J., Peri, C., Ferrer-navarro, M., Conchillo-sole, O., Thomas, R. J., Champion, O. L., Michell, S. L., Kewcharoenwong, C., Nithichanon, A., Lassaux, P., Perletti, L., Longhi, R., Lertmemongkolchai, G., Titball, R. W., Daura, X., Colombo, G. & Bolognesi, M. (2013). *Chem. Biol.* **20**, 1147–1156.

Hungler, A., Momin, A. & Arold, T. (2016). *J. Appl. Cryst.* **49**, 2252–2258.

Joshua A. Bornhorst and Joseph J. Falke (2000). *Methods Enzym.* **326**, 245–254.

Kabsch, W. (2010). *Acta Crystallogr. Sect. D*. **66**, 125–132.

Karplus, P. A. & Diederichs, K. (2015). *Curr. Opin. Struct. Biol.* **34**, 60–68.

Ko, J., Park, H., Heo, L. & Seok, C. (2012). *Nucleic Acids Res.* **40**, 294–297.

Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nat. Protoc.* **3**, 1171–1179.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Matthews, B. W. (1976). *Ann. Rev. Phys. Chem.* **27**, 493–523.

Mccoy, A. J., Grosse-kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 355–367.

Musille, P. & Ortlund, E. (2014). *Acta Crystallogr. Sect. F*. **70**, 166–172.

Niedzialkowska, E., Gasiorowska, O., Handing, K. B., Majorek, K. A., Porebski, P. J., Shabalin, I. G., Zasadzinska, E., Cymborowski, M. & Minor, W. (2016). *Protein Sci.* **25**, 720–733.

Pierce, B. G., Wiehe, K., Hwang, H., Kim, B., Vreven, T. & Weng, Z. (2014). *Struct. Bioinforma.* **30**, 1771–1773.

Ramraj, V., Evans, G., Diprose, J. M. & Esnouf, R. M. (2012). *Acta Cryst. D*. **68**, 1697–1700.

Ritchie, D. W. & Grudinin, S. (2016). *J. Appl. Cryst.* **49**, 158–167.

Sanctis, D. De, Beteva, A., Caserotto, H., Dobias, F., Giraud, T., Gobbo, A. & Guijarro, M. (2012). *J. Synchrotron Rad.* **19**, 455–461.

Schneidman-duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. (2005). *Nucleic Acids Res.* **33**, 363–367.

Simpkin, A. J., Simkovic, F., Thomas, J. M. H. & Savko, M. (2018). *Acta Crystallogr. Sect. D*. **74**, 595–605.

Stock, I., Burak, S., Sherwood, K. J., Grüger, T. & Wiedemann, B. (2003). *J. Antimicrob. Chemother.* 865–885.

Thorn, A. & George, M. (2013). *Acta Crystallogr. Sect. D*. **69**, 2251–2256.

Vagin, A., Teplyakov, A. (1997). *J. Appl. Cryst. (1997)*. 1022–1025.

Veesler, D. & Cambillau, C. (2008). *Acta Cryst. F*. **64**, 880–885.

Yan, Y., Tao, H. & Huang, S. (2018). *Nucleic Acids Res.* **46**, 423–431.