Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Estudos estruturais de hidrolases de glicosídeos em solução usando técnicas de espalhamento a baixo ângulo (SAS)

Vasilii Piiadov

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Bioenergia

Piracicaba
2019

Vasilii Piiadov
Físico

Estudos estruturais de hidrolases de glicosídeos em solução usando técnicas de espalhamento a baixo ângulo (SAS)

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **IGOR POLIKARPOV**

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Bioenergia

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Piiadov, Vasilii

Estudos estruturais de hidrolases de glicosídeos em solução usando técnicas de espalhamento a baixo ângulo (SAS) / Vasilii Piiadov. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2019.

83 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”. Universidade Estadual de Campinas. Universidade Estadual Paulista “ Julio de Mesquita Filho”

1. Bioinformática 2. Espalhamento a baixo ângulo 3. Hidrolases de glicosídeos I. Título

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Igor Polikarpov, pela oportunidade.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento do projeto 2014/00769-5.

CONTENT

RESUMO

Estudos estruturais de hidrolases de glicosídeos em solução usando técnicas de espalhamento a baixo ângulo (SAS)

As hidrolases de glicosídeos (GHs) exercem papéis fundamentais em vários processos biomédicos e aplicações industriais. A maioria destas enzimas possui vários domínios funcionais ligados entre si por peptídeos conhecidos como linkers. Informações sobre organização estrutural destas enzimas e sua mobilidade, posições e orientações mútuas de domínios individuais, bem como mudanças conformacionais introduzidas por ligantes ou por mudanças de condições bioquímicas (pH e T) podem ser muito informativas. Por esse motivo, é muito importante determinar a organização estrutural de GHs em termos de posição e orientação de seus domínios individuais e compreender a interação entre estes domínios em condições próximas às fisiológicas. Entretanto, atualmente, a conformação, dinâmica e função dos GHs com múltiplos domínios ainda não são totalmente compreendidas. Assim, o principal objetivo deste projeto foi conduzir estudos de hidrolases de glicosídeos em solução, usando SAS. Um grande número de GHs foi clonado e expresso em laboratório sob a direção do Prof. Dr. Igor Polikarpov (Grupo de Biotecnologia Molecular, IFSC / USP), seguindo protocolos já estabelecidos na literatura, para sua expressão e purificação. Experimentos SAXS foram realizados em colaboração com o Dr. Evandro Ares de Araújo (USP, São Carlos) e com o Prof. Dr. Mário de Oliveira Neto (UNESP, Botucatu). Para estudar as hidrolases de glicosídeos, foi utilizado o método de espalhamento a baixo ângulo, e em adição ao trabalho experimental, foi desenvolvido um novo pacote de software SAXSMoW2 para processar os dados do SAXS. Este pacote permite obter rapidamente os principais parâmetros estruturais de moléculas de proteínas, calcular o peso molecular e o estado oligomérico. Também foi aperfeiçoado e aplicado o método de acoplamento estatístico (statistical coupling analysis) , para complementar os dados estruturais experimentais, em especial para xiloses isomerases. Este método pode permitir uma melhor compreensão da relação entre as características estruturais evolutivas e sua funcionalidade biológica. Além disso, métodos de bioinformática foram desenvolvidos para complementar e compreender melhor as informações estruturais obtidas nos experimentos de SAXS. O primeiro foi um método para separar sequências de GH7 em duas categorias, exo e endogluconases. É útil analisar cada tipo de proteína dentro da família separadamente e estudar o papel dos loops funcionais - características estruturais que influenciam significativamente a atividade biológica. Outro método foi desenvolvido para encontrar o centro de atividade na nova enzima Xilose Isomerase obtida, usando uma estrutura relacionada, bem conhecida, da mesma família. Este método foi aplicado a enzimas cujas estruturas foram estudadas pela técnica de cristalografia em nosso laboratório no IFSC / USP. Inspirado pelo SCA, um método de detecção de comunidades difusas de aminoácidos em proteínas foi desenvolvido. Essa informação também pode complementar os resultados do SCA, indicando conjuntos fortemente correlacionados de aminoácidos na enzima. Outro novo método desenvolvido é uma estimativa de afinidade nas famílias de enzimas ativas em carboidratos utilizando similaridade dos modelos escondidos de Markov e bancos de dados open access de sequências de proteínas.

Palavras-chave: Bioinformática; Espalhamento a baixo ângulo; Hidrolases de glicosídeos

ABSTRACT

Structural studies of glycoside hydrolyses in solution using small-angle scattering (SAS) techniques

The Glycoside Hydrolases (GHs) play a key role in a number of biomedical processes and industrial applications. Most of these enzymes are multidomain proteins composed of different functional domains connected by linker peptides. Thus, it is very important to determine structural organization of glycoside hydrolases in terms of positions and orientations of their individual domains and comprehend the interplay between their multiple domains under close-to physiological conditions. To study the glycoside hydrolases, in this work a small-angle scattering method has been used. Currently, the conformation, dynamics and function of GHs with multiple domains are not fully understood. This is why the information on their structural organization and mobility; mutual position and orientation of the individual domains and conformational changes induced by interaction with the substrates or difference in biochemical conditions might be very informative. A large number of GHs have been cloned and expressed in the lab under direction of Prof. Dr. Igor Polikarpov (Molecular Biotechnology group, IFSC/USP) and we follow already established protocols for their expression and purification. SAXS experiments have been carried out in collaboration with Dr. Evandro Ares de Araujo (USP, São Carlos) and Prof. Dr. Mario de Oliveira Neto (UNESP, Botucatu). Additionally to experimental work, a new software package SAXSMoW2 for SAXS data processing has been developed. The software allows to obtain rapidly main structural parameters of the protein molecule, calculate molecular weight and oligomeric state. To supplement an structural data, the method of statistical coupling analysis (SCA) has been significantly improved and applied. The method allows a better understanding of interconnection between evolutionary caused structural features and their biological functionality. Also, various bioinformatic methods were developed to complete and understand better structural information obtained in SAXS experiments. The first one is a method for separating sequences from GH7 into the two bins of exo- and endogluconases. It is helpful to analyze each type of proteins inside the family separately and study the role of functional loops -- structural features that significantly influence on biological activity. Other developed method is for finding of activity center in the new obtained Xylose Isomerase enzyme using related well-known structure from the same family. This method was applied to the enzyme whose structure was studied using crystallography technique in our laboratory at IFSC/USP. Inspired by SCA, a method of aminoacid fuzzy communities detection in proteins has been developed as well. This information also can complete SCA results showing strong correlated sets of aminoacids in the enzyme. Another one new developed method is an estimation of carbohydrate-active family affiliation of unknown proteins using Markov hidden model similarities and open access databanks of protein sequences.

Keywords: Bioinformatics; Small-angle scattering; Glycoside hydrolases

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

SAS	Small-angle scattering
SAXS	Small-angle X-ray scattering
SANS	Small-angle neutron scattering
GH	Glycoside Hydrolase (Glycosidase)
GT	Glycosyltransferase
CAZyme	Carbohydrate-Active Enzyme
PL	Polysaccharides Lyase
CE	Carbohydrate Esterase
AA	Auxiliary Activities enzyme
SCA	Statistical Coupling Analysis
MSA	Multiple Sequence Alignmen
MD	Molecular Dynamics
LIC	Ligation Independent Cloning
DNA	Deoxyribonucleic acid
His-tag	Hexa Histidine-tag, trademarked name
CV	Column Volume
DAM	Dummy Atoms Model
PDB	Protein Data Bank
ICA	Independent Component Analysis
LNLS	Brazilian Synchrotron Light Laboratory (port. Laboratório Nacional de Luz Síncrotron)

1. INTRODUCTION

Plant biomass represents the most abundant and sustainable source of carbon which can be used as an alternative route to fossil fuels (Ragauskas et al. 2006; Gaurav et al. 2017) as well green chemicals productions (Alonso et al. 2017; Erythropel et al. 2018). Plant cell walls are a cross-linked matrix composed of polyphenolics, fibers, sugars, proteins, and polysaccharides (Keegstra 2010; Cosgrove 2014). Microbial and enzymatic deconstruction of lignocellulosic biomass has been a preferential biotechnological route for plant transformation to biomass-based products. Thus, it is important task to understand better the role of protein structural patterns in forming of their properties.

The synthesis and degradation of carbohydrates in the form of di-, oligo-, and polysaccharides, is crucial in all living domains. In fact, these processes have vital importance for providing sources of energy, cellular structure organization and intracellular communication. Carbohydrates are the main source of energy in heterotrophic organisms, in a form of polysaccharides, such as starch and glycogen, used as energy storage, which are subsequently hydrolyzed to monosaccharides capable for entering into metabolic cycle. Furthermore, carbohydrates in the form of glycoconjugates (glycoproteins and glycolipids) are responsible of important biological functions including cell-cell interactions, signal transduction, compartmentalization of proteins and antigenic response (Staudacher et al. 1999).

Two main classes of catalysts are involved in the modification of carbohydrates: glycoside hydrolases (GHs) and glycosyltransferases (GTs), which are responsible for the hydrolysis and synthesis of the carbohydrates, respectively. GTs catalyze the transfer of sugar moieties from activated donor molecules to specific acceptor, forming glycosidic bonds. Whereas GHs (also named Glycosidases) are a widespread group of enzymes that hydrolyze the glycosidic bond between two or more carbohydrates or between a carbohydrate and a non-carbohydrate moiety.

Online since 1998, CAZy http://www.cazy.org is a database dedicated to display and catalogue genomic, structural and biochemical information on Carbohydrate-Active Enzymes (CAZymes): glycoside hydrolases, glycosyltransferases, polysaccharides lyases (PLs), carbohydrate esterases (CEs) and auxiliary activities (AA) enzymes. These enzymes are classified in families on the base of their amino acid sequence and some families are further grouped in "clans" possessing conserved 3D-structures (Cantarel et al. 2009). The importance of this classification is based on the fact that the catalytic machinery, the stereospecificity and the nature of reaction mechanism (i.e. inverting or retaining) are conserved for all the GHs belonging to a certain family, therefore, CAZy proved to be extremely useful to attribute these characteristics to any new, not yet characterized, glycosidase. It is clear that the characterization of a novel GH and, eventually, the creation of a new family, always represent a crucial increment in the understanding of novel enzyme class for both basic and applied research. One important application of GHs in general and cellulases specifically is biomass hydrolysis for biofuel production. Cellulases are key enzymes used for saccharification of biomass in the process of liquid fuels production. Considerable efforts are currently directed toward reducing cellulases production costs using both structural approaches to improve the properties of individual cellulases and genomic approaches to identify new cellulases as well as other auxiliary proteins and enzymes capable of increasing the activity of cellulases in depolimerization of pretreated biomass materials.

Small-angle scattering (SAXS and SANS) methods are well-known techniques which are used to study biological samples such as macromolecules in solution (Lipfert and Doniach 2007; Jacques and Trewhella 2010). This is one of most effective physical technique to studying macromolecules on scale of 1-100 nm. The main advantage of this method is its generality: SAXS can be used for studying disordered objects and does not require special

preparation of samples. Using this method, it is possible to obtain detailed information, such as radius of gyration, volume, molecular surface and oligomerization state (Svergun and Stuhrmann 1991). The SAXS method allows for investigations of both, well-structured and disordered macromolecules in solution, without requiring crystallization procedures nor highly elaborate sample preparation. The experimental SAXS intensity associated to a set of proteins in dilute solution - after subtracting the parasitic scattering intensity produced by the buffer under the same experimental condition - is proportional to the scattering intensity produced by a single protein averaged over all possible orientations.

In modern laboratory setups and synchrotron radiation-based beamlines, SAXS data are recorded by two-dimensional (2D) detectors. For a dilute set of proteins with random orientations, the scattering intensity defined in the reciprocal space is isotropic. Thus, the angular-averaging of 2D detector patterns yields a one-dimensional (1D) scattering intensity as a function of the modulus of the scattering vector, $I(q)$. In order to derive structural information of proteins in solution from SAXS curves, several software packages for data analyses and evaluations, such as ATSAS (Franke et al. 2017) and SCATTER (Rambo 2015), are currently available.

Thus, SAXS is a powerful tool for structure validation and the quantitative analysis of flexible systems. At present, SAXS analysis methods have reached an advanced state, allowing for automated and rapid characterization of protein solutions in terms of low-resolution models, quaternary structure and oligomeric composition (Mertens and Svergun 2010). The small-angle scattering techniques are the methods of choice for determination of low resolution structures of the macromolecules in solution and can provide important information on the architecture and mobility of the multidomain proteins, revealing their compact or extended conformation and providing information of the flexibility of the macromolecules in close to physiological conditions. In contrast to diffraction, SAS methods do not require to crystallize the protein of interest.

It is known, for example, that cellulases have linkers of a different length, and that interaction with lignocellulose substrate and processivity of action are remarkably different for endoglucanases and exoglucanases, and even between two exoglucanases (CBH I and CBH II), but information on their molecular shape in full-length settings is not readily available for most of the enzymes and the influence of the linker flexibility and relative mobility of the correspondent CBMs on the interactions of the enzymes with their cognate substrates are not well understood as yet. In the recent publication (Lima et al. 2013) showed, based on the small-angle X-ray scattering and molecular dynamics (MD) simulations that the optimum distance between CCD and CBM is determined by the minimum energy conformation of the linker which is influenced by its amino acid sequence and degree of glycosylation. The linker length and conformation influence relative mobility of CBM with respect to CCD and thus interaction of the cellulase with the solid polymeric substrate (cellulose). Deglicosilation of CBHI leads to modifications of the catalytic activity of the enzyme, which might be related to the modifications in the flexibility and length of linker peptide.

In the work published in Science (Brunecky et al. 2013), Brunecky and collaborators demonstrated that *Caldicellulosiruptor bescii* multidomain cellulase CelA is able to hydrolyse none-pretreated biomass several times more efficiently than *T. reesei* CBHI which is the exoglucanase a most commonly used in modern commercial enzymatic prepares used for biomass enzymatic hydrolysis. *Caldicellulosiruptor bescii* cellulase CelA has a GH9 family catalytic domain appended to three type III cellulose-binding modules and a GH48 family catalytic domain. The authors demonstrated that such molecular organization is fundamental for novel molecular mechanism of cellulose hydrolysis, different for common fungal cellulase ablation of biomass (Brunecky et al. 2013). This discovery opens new avenues for boosting of enzymatic activities of enzymatic preparates developed for biomass hydrolysis and

depolimerization and further shows importance of studies of the domain organization and molecular shape of celluloses in solution, particularly in multidomain settings.

Thus, SAS studies which provide information on the shape and form of the enzyme, and as a consequence on the length of the linker and a distance between catalytic domains and CBMs, might provide very important information on the mobility and function of the cellulases that could be difficult, if not impossible, to obtain using other techniques, such as protein crystallography, for example. The use of SAS techniques can allow us to better understand the differences between molecular shapes and conformations of different exo- and endoglycosidases and GHs in general and to obtain important insights about their interactions with the polymeric substrates.

Basic theory, which connect scattering intensity with object structure, are defined by scattering ability of inhomogeneities only and their contrast relative to the solvant or solid matrix (Glatter and Kratky 1982). Thus, SAXS allows to study an objects in nano scale with different origin, particularly, biological macromolecules in diluted solution.

In the thesis, the author focused on SAXS intensity data corresponding to isotropic and dilute sets of monodisperse proteins hydrated by homogeneous buffers. Also, it is considered the electron densities of the proteins and the buffers are both spatially constant. Under this condition, the relevant parameter associated to SAXS intensity and related to electron densities is named as density contrast, which is defined as $\Delta\rho = \rho_{protein} - \rho_{buffer}$. The value of $\Delta\rho$ only affects the absolute value of the SAXS intensity but not the shape of the scattering intensity curve.

Robust determinations of molecular weight and oligomeric state of proteins in solution are fundamental for understanding their quaternary structure and function. On the other hand, a large number of proteins that are usually studied on SAXS beamlines at most of the existing synchrotron X-ray sources requires quick procedures for achieving quantitative information on molecular weights and oligomeric states of the proteins, during ongoing series of experiments. To achieve these goals, a molecular weight calculation method has been implemented as a online service SAXSMoW2 http://saxs.ifsc.usp.br.

Among other structural parameters, molecular weight and oligomerization state are leading in complete structural analysis. An estimate of molecular weight can be obtained from the SAXS curve in absolute scale (Orthaber et al. 2000) by:

$$M = \frac{I_0 N_a}{c(\Delta\rho\nu)^2},$$

where N_a - the Avogadro number, I_0 - zero-angle scattering intensity, c - concentration, ρ - electronic contrast and ν - partial specific volume of the protein. Also, it is possible to determine a molecular weight without using a data in absolute scale by comparison with data obtained from the reference protein, assuming that $N_a/(\Delta\rho\nu)^2$ is a constant.

The molecular weight of proteins in solution can be determined from experimental SAXS function from the value of the intensity in absolute scale extrapolated to zero angle, $I(0)$. Another method uses the extrapolated intensity $I(0)$ in relative scale and further comparison with intensity from a standard protein with known molecular weight (Orthaber et al. 2000). However, these methods exhibit several sources of errors that often introduce systematic bias in the assessment of a protein molecular weight (Trewhella et al. 2017).

Alternative method for determination of the molecular weight of protein in a dilute solution, that applies to SAXS data on a relative scale, that has been incorporated in the DatPorod program from ATSAS package (Franke et al. 2017). The method is based on the determination of the quotient between the extrapolated SAXS intensity and the Porod invariant:

$$Q = \int_0^\infty (I(q) - A)q^2\, dq,$$

where A, is a constant. The $I(0)$ value in relative units is determined by extrapolation down to $q = 0$ while the determination of the integral Q is more difficult because it requires the extrapolation of $I(q)$ up to $q = \infty$ that is performed up to the high q-range, where the scattering intensity is usually low and the relative statistical error is high. For high enough values of q, Kratky plot shows an asymptotic behavior making possible an extrapolation of the curve to infinite whose inclination gives the Porod invariant. Functionality of the software includes this extrapolation and, for accurate calculation of the Q-invariant, DatPorod automatically subtracts the contribution to SAXS intensity originating from the effects of protein flexibilities and from the fluctuations in density due to minor heterogeneities in the protein structures. This procedure follows by a further extrapolation of SAXS intensity up to $q = \infty$ by applying Porod equation $I(q) \propto q^{-4}$. After computing $I(0)$ and Q, the molecular volume of the protein is determined as

$$V = 2\pi^2 \frac{I(0)}{Q}$$

Finally, the product of the calculated protein volumes and the known mass density yields their molecular weight. This method gives the result with about 20% of discrepancy (Svergun et al. 2006).

Developed previously, SAXSMoW software (Fischer et al. 2010) is used to determine a molecular weight of monodisperse proteins in solution. The advantage of this package is in accuracy about 10-15% of protein weight calculation from single SAXS measurement in relative scale. This package uses GNOM, one of the utilities included in ATSAS package (Svergun et al. 2006), output file as input data that is not convenient feature. Also, SAXSMoW did not allow to choose any region of Kratky plot to calculate Q-invariant.

As mentioned above, the new package SAXSMoW2 has been developed. It is a utility for processing of SAXS data based on ideas of previous SAXSMoW package. This is a web application and can be used online without downloading and installation. Most significant difference from obsolete SAXSMoW is that the utility use "pure" SAXS data as input. It should be text file with widely used extension ".dat" which consists at least two columns: transferred momentum q, and intensity in arbitrary units. Obtaining ".dat" file, the software makes Guinier fitting, calculates radius of gyration, proposes intervals for Q-invariant calculation, makes Kratky and Porod plots and, finally, calculates molecular weight of protein.

Also, structural information obtained from experimental methods such as SAS, has been supplemented by bioinformatic methods to understand structural features better. To achieve this, statistical coupling analysis (SCA) method (Lockless and Ranganathan 1999; Socolich et al. 2005) has been applied directly to set of sequences and allowed to find a evolutionary conserved patterns in set of multiple sequence alignment (MSA) of protein family. Thus, this method was useful in understanding of correlations between structure given by SAS experiments and role

of the conserved patterns of the sites in amino-acid sequence. During the work, a few auxiliary bioinformatic methods were developed to support structural studies of the enzymes.

2. MATERIALS AND METHODS

2.1. Enzyme preparation

Many GHs have been cloned and expressed in the Lab (Molecular Biotechnology group, IFSC USP). To produce samples of enzymes we follow already established protocols for their expression and purification (Colussi et al. 2011, 2012; Vizoná Liberato et al. 2012; De Araújo et al. 2013; Dos Reis et al. 2013; Prates et al. 2013; Rosseto et al. 2013; Textor et al. 2013).

Briefly, cloning can be performed using ligation independente cloning (LIC) which utilizes site-specific recombination using 3'→ 5' exonuclease activity of T4 DNA polymerase. The proteins of interest should be inserted into plasmids for superexpression in bacteria which contains lac system and N-terminal His-tag with cleavage site recognized by specific proteases (TEV, thrombin), using culture media 2LB in this process. Although we mostly employ bacterial expression, in some cases we conduct expression in fungal systems such as *Aspergillus nidulans* and *Aspergillus niger* which are being actively used in the Molecular Biotechnology group of IFSC USP. Since proteins have been expressed with the His-tag, affinity chromatography should be used as a first choice. If further purification is necessary, we use second purification step using gel filtration (or size exclusion) column. Other purification steps, which explore ionic force of solution or protein hydrophobicity should be applied if needed. Cleavage of the His-tag can be accomplished with thrombin, TEV or ULP1 proteases, depending on the cloning vector. The protein samples must be prepared in an adequate buffer solution and concentrated for SAS studies.

Based on this outline, an initial culture of *E. coli Rosetta* (DE3) should be grown overnight in Luria-Bertani medium at $37\,℃$. After that, the culture must be grown in a shaker at $20\,℃$ and $150\,rpm$ until the OD_{600nm} reached a value of 0.6. The expression should be induced by $1\,mM$ of Isopropyl β-D-1-thiogalactopyranoside (IPTG) followed by incubation for $20\,h$ at $20°C$. The cells were harvested by centrifugation at $6000\,g$ and the pellets were suspended in $20\,mM$ Tris-HCl, $300\,mM$ NaCl and $pH\ 8.0$, lysed by sonication and centrifuged at $18000\,g$ during $20\,min$ at $4°C$ to remove debris.

The soluble lysate was applied in a Ni-NTA agarose columm (Qiagen), previously equilibrated in the same buffer, and incubated for $60\,min$. Afterwards, the column was washed with ten column volumes (CV) and subsequently eluted with 2 CV of $300\,mM$ of imidazole in the same buffer. The imidazole was removed by dialysis and the protein concentration was determined using a Nanodrop spectrophotometer (Thermo Scientific).

For all the samples of GH1 enzymes, His-tag should be removed with TEV protease at $8°C$ under mild stirring during $24\,h$. Then the proteins should be loaded to a Ni-NTA agarose columm (Qiagen) previously equilibrated with $20\,mM$ Tris-HCl pH 8.0 and $300\,mM$ NaCl and incubated for $60\,min$. The flowthrough should be collected and loaded onto the HiLoad Superdex 75 16/60 size-exclusion chromatography column (GE Healthcare) pre-equilibrated in $10\,mM$ of NaCl, $10\,mM$ of MES, pH 6.5, using ÄKTApurifier 10 system (GE Healthcare Life Sciences) to remove aggregates and impurities.

For GH3 enzyme samples, established protocols were used as well. It was carried out using Superdex 200 chromatographic column (16/60, GE Healthcare Life Sciences) equilibrated with buffer solution containing 50 mM HEPES (pH 7), 150 mM NaCl and 3% glicerol coupled to an HPLC system (ÄKTApurifier, GE Healthcare Life Sciences). For SAXS measurements, all the samples were centrifuged at $17000\,g$ for $10\,min$ at $4\,℃$. The final

protein concentration was determined measuring its absorbance at $280\,nm$ using a spectrophotometer (NanoDrop 2000, Thermo Scientific).

2.2. SAXS measurements

Small-angle scattering experiments were conducted in collaboration with Dr. Mario Oliveira Neto (UNESP, Botucatu) and Dr. Evandro Araujo (IFSC/USP, Sao Carlos) at D02A-SAXS2 beamline at National Synchrotron Light Laboratory https://www.lnls.cnpem.br/ and at B21 beamline at Diamond Light Source Synchrotron http://www.diamond.ac.uk.

Small-angle experiments at D02A-SAXS2 beamline are conducted using a two-dimensional detector that allows to measure with a transferred momentum of a range of $0.01\text{Å}^{-1} < q < 0.35\text{Å}^{-1}$. The wavelength of the incoming monochromatic X-ray beam is $\lambda = 0.154\,nm$. Bi-dimensional patterns were integrated using the FIT2D program (Hammersley et al. 1996), which is included in the standard software package on the beam-line.

At B21 beamline, the chromatography-coupled small-angle X-ray scattering data are collected using a Pilatus2M detector (Dectris) with a transferred momentum covering a range of $0.0040\ \text{Å}^{-1} < q < 0.4075\ \text{Å}^{-1}$. The protein sample was loaded onto a 2.4 mL Superdex 200 (GE Healthcare) equilibrated with a buffer solution coupled to an Agilent 1200 HPLC system. The data frames were collected at three seconds exposure intervals and the buffer frames have been used to substract the background. Primary data processing involves averaging on azimuth angles in the detector plane and scaling using ScÅtter software (Rambo 2015).

The data has been collected at National Synchrotron Light Laboratory similarly: the scattering patterns are primarly treated using the FIT2D program (Hammersley et al. 1996). Independently to the software, it uses definition of transferred momentum $q = 4\pi\frac{\sin\theta}{\lambda}$, where 2θ is a scattering angle observed between the incident and scattered beam, λ - a wavelength of the incident beam.

Further treatment includes a calculation of radius of gyration R_g that estimated by Guinier equation $I(q) = I(0)\exp(-q^2R_g^2/3)$ applying to an interval named Guinier region $0 < q < q_{max}$, where q_{max} is limited by $1/3R_g$ value (Mortensen and Posselt 1998; Perry and Tainer 2013; Konarev and Svergun 2015). Molecular weight was obtained by our SAXSMoW2 package (Piiadov et al. 2018). This method allows to obtain protein weight information, since the system is monodispersed and diluted. The software has $> 90\%$ accuracy to calculate molecular weight without the necessity to obtain SAXS curves in absolute scale.

Starting from the DAMSTART model \citep{atsaspub}, the final model was improved with the DAMMIN software (Svergun 1999). To select most typical model, the minimal discrepancy between experimental and calculated SAXS curves was quantified with the minimized χ^2-parameter. Using each DAM-model for each protein sample, at least 15 models were averaged using DAMAVER program (Volkov and Svergun 2003). Using obtained *ab-initio* model, 3D shape of the molecule can be represented with PyMOL software (DeLano 2002).

Analysis of the $p(r)$ using GNOM program (Svergun et al. 2006) provide us with structural parameters such as maximum diameter D_{max}, radius of gyration Rg, shape, etc. *Ab initio* modelling of "dummy atoms models" (DAM) is performed using annealing that is used in packages DAMMIF (Franke and Svergun 2009) and the average model was calculated by DAMAVER program (Volkov and Svergun 2003).

The program CRYSOL (Svergun et al. 1995) was used to simulate the scattering patterns from the structures resolved in "Protein Data Bank" (PDB). By comparing the curves which are obtained from the

experimental data $I(q)$ with simulated curve using PDB structure, we can determine an oligomeric state of the enzyme in solution and relative position of their individual domains with great precision. Modeling with available high-resolution structures treating them as rigid bodies and by allowing for conformational adjustments by computationally changing angles between the subunits or domains of a protein may be performed. To this, the packages MASSHA (Konarev et al. 2001), SASREF and BUNCH (Petoukhov and Svergun 2005) can be used. This approach has been successfully applied in our lab to several proteins before, leading in some cases to new and unexpected results (Rojas et al. 2005; Grimm et al. 2006; Santos et al. 2012; Lima et al. 2013).

2.3. Needleman-Wunsch pairwise alignment

We have used pairwise alignment method for statistical characterization of a set of sequences used in statistical coupling analysis. An alignment, as an editorial distance calculation allows us to calculate identity and similarity of sequences inside a set. To this, Needleman-Wunsch pairwise alignment algorithm was used (Needleman and Wunsch 1970). The algorithm uses dynamic programming approach and includes following steps:

- Initialization

$$F_{0,0} = 0; F_{0,j} = d \cdot j; F_{i,0} = d \cdot i$$

- Recursion

$$F_{i,j} = \max\{ F_{i-1,j-1} + G\big(s_1(i), s_2(j)\big); F_{i-1,j} - d; F_{i,j-1} - d$$

- Finalization

  Last element (bottom right element) $F_{n,m}$ has a maximum score. Optimal alignment will be obtained by passage through the matrix from the last element.

Here, $F_{i,j}$ is a matrix containing weights of alignments of subsequences having length i and j, $G\big(s_1(i), s_2(j)\big)$ - elements of substitution matrix chosen for aminoacids at positions i, j from sequences s_1, s_2, d - gap penalty.

A substitution matrix $G(a,b)$ (a and b are aminoacids) describes an evolutionary model and it is defined as:

$$G(a,b) = \log\frac{p_a m_{a,b}}{q_a q_b}$$

where p - frequency of amino acid, $m_{a,b}$ - probability of mutation for a into b aminoacid.

In bioinformatics, substitution matrix BLOSUM (Henikoff and Henikoff 1992) and PAM (Dayhoff et al. 1978) are most often used. Therefore, we have chosen the matrix BLOSUM to use.

For our goals, we have used MATLAB implementation of Needleman-Wunsch algorithm compiled as a standalone library.

2.4. Multiple sequence alignment (MSA)

Multiple sequence alignment is a generalization of pairwise alignment described previously and used directly in statistical coupling analysis and in our carbohydrate-active enzyme family detecting. In comparison to pairwise algorithm, MSA is much more difficult for calculations: the time growth exponentially with number of

sequences. Thus, there is some approaches to accelerate computation: a few euristic algorithms such as progressive methods, genetic algorithms, branch & bound, etc. Progressive methods include well known algorithms such as CLUSTALW (Thompson et al. 2002), T-Coffee (Notredame et al. 2000), MUSCLE (Robert C 2004). In our work we have used a progressive method implemented in Bioinformatic Toolbox of MATLAB package because of higher performance in comparison to others.

Progressive method of MSA involves the following steps:

1. Build all possible pairwise alignments in a set;
2. Pairwise alignment of the most closely related sequences is used in the method as fixed;
3. Create a distance matrix with alements in accordance with evolutionary distances defined by pairwise alignments;
4. Build phylogenetic tree based on distance matrix;
5. Add closest sequence from obtained tree to alignment and preserve gaps;
6. Repeat the algorithm over all the sequences.

The method is fastest in comparison to others MSA techniques, but, however, a quality is highly dependent to the first pairwise alignment.

2.5. Protein distance calculation

For protein distance calculation which was used in our implementation of statistical coupling analysis we have used the algorithm implemented in package EMBOSS (Rice 2000) as **fprotdist** prtogram. The method is based on maximum likelihood estimates and it can use five models of aminoacid substitution. In our work we were used Dayhoff point accepted mutation matrix (PAM) (Dayhoff et al. 1978, 1979) implemented in the program. The PAM is an empirical matrix containing scaled probabilities of aminoacid mutations accepted by natural selection. To build the matrix, 1572 changes in 71 families were analized. Dayhoff et al. have used ungapped alignments of well-conserved parts from evolutionary close proteins which have at least 85% of identity. Thus, probabilities of mutations represented in PAM matrix are based on empirical frequencies and give evolutionary distance between sequences of aminoacids.

2.6. Protein homologues searching

As an option in our implementation of statistical coupling analysis, a set of sequences can be built automatically using search of homologous for the reference sequence using BLAST method (Altschul et al. 1997). Homologous searching is based on search of similar sequences to the reference: BLAST do it over comprehensive database of more than ten million proteins. However, homologous does not always mean significant similarity, but statistically significant similarity signs homology.

BLAST does local alignment because of different proteins can have similar patterns or domains. Search procedure implemented in BLAST starts with 3-letter words taken from the sequence of interest and does searching in the database of non-redundant protein sequences. In case of match, the algorithm increases a length of matched word without using of gaps and with it. After the highest possible increasing of word length, BLAST selects an alignment with highest number of matches for each pair formed by the sequence of interest and another one from a database. A list of found sequences are suggested as a result.

Next, a resultant list of sequences must be ranged by similarity with the reference. For similarity calculation, BLAST uses a substitution matrix BLOSUM62 (BLOcks SUbstitution Matrix of 62% of identity). To characterize degree of significance of homology numerically, BLAST calculates E-values (expected value) for each pair of reference sequence and one from searching database. This parameter is defined as:

$$E = m \cdot n \cdot 2^{-B}$$

In the equation, m, n are lengths of reference sequence and B a sequence from searching database, respectively. Parameter B (bit score) reflects similarity of sequences and it is calculated as shown below:

$$B = (P \cdot S - \ln K)/\ln 2$$

where K - a statistical parameter as a natural scale for search space size, S - row score of similarity and P - probabilty of occurrence at least one high-scoring segment pair (HSP) with row score greater than S. Thus, resultant list is ordered by E-value, where more homologous sequences are in top.

Implementation of the method we have using is provided at https://blast.ncbi.nlm.nih.gov as **blastp** application.

2.7. Construct profile hidden Markov models (HMM)

Profile hidden Markov model describes a consensus of a multiple sequence alignment. Profiling on HMM is based on score system to detect an information about position specific conservations in multiple alignment of aminoacid sequences. As it can be understood from name of the technique, the method has using Markov process and it is stochastic model of this kind of process with unknown parameters which can be detected by observing of results obtained by HMM. In application to biology, HMM allows to transform a multiple sequence alignment in position specific scoring system for searching of distant homologous sequences (Eddy 1998). HMM profiling is based on linear algorithm of match state blocks which are corresponded with positions in studied MSA. A "match state" term means given aminoacid can be aligned to current column of alignment in accordance with a distribution of allowed residues in the column. The linear Markov process chain used in the method is shown in Figure 1. As shown in Figure 1, besides match states, it has two other blocks: inserts and deletes with emission probabilities distributions over all possible amino acids. Thus, sequence aligning with HMM profile is search of most probable path through insert and deletion blocks. In most simple case of a tested sequence equivalent to the consensus of the original alignment, the path in the scheme on Figure 1 will be linear way from match state to match state till the finish without deletions or inserts.

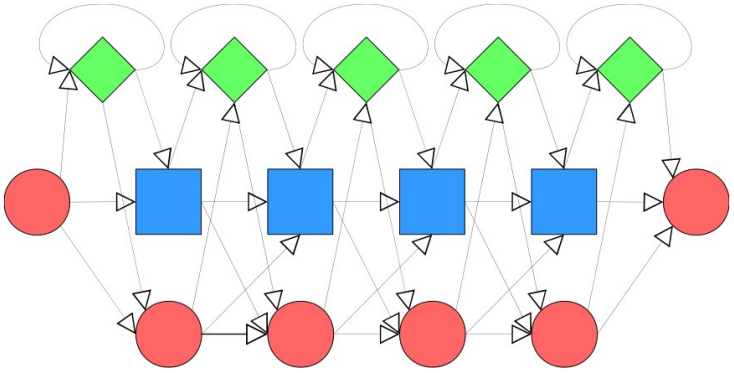

Figure 1. HMM algorithm. Blue squares - match states, red circles - non-emmiting states such as delete, start and finish, green diamonds - insert states

In our work we have used an implementation of the method known as HMMER package of version described in (Eddy 2011). The package contains **hmmbuild** and **hmmsearch** utilities which were used to build HMM profiles for carbohydrate-active enzyme families and to compare a studied sequence with new created profiles.

2.8. Statistical coupling analysis

Statistical coupling analysis (SCA) is used to characterize the pattern of evolutionary constraints between amino acid positions in a protein family. The method analyzes representative multiple sequence alignment (MSA) of the family and implements the quantitatively measurements of the overall functional constraint at each sequence position (positional correlations) and analyzing the coupled functional constraint on all pairs of sequence positions (pairwise correlations) (Lockless and Ranganathan 1999; Rivoire et al. 2016).

Coupling of site pairs have been introduced statistical coupling analysis. This such of coupling defines degree of statistical correlation between frequencies of pairs of aminoacid sites. SCA operates with multiple sequence alignment of proteins. Set of proteins for analysis from the same family normally is the best choice because they represent similar biological functions which can be captured by statistical method such SCA. First steps of analysis are computing of aminoacid conservations over all the positions in MSA. In SCA, the conservation measured by Kullback-Leibler relative entropy (Halabi et al. 2009; Reynolds et al. 2013):

$$D_i^a = f_i^a \ln \frac{f_i^a}{q^a} + (1 - f_i^a) \ln \frac{1 - f_i^a}{1 - q^a}$$

where f_i^a - frequency of aminoacid a in position i, q^a - probability of aminoacid estimated from protein database. Kullback-Leibler entropy indicates an estimation of expectation of frequencies relatively to expected q^a in general.

Based on this, second-order statistics can be obtained: conservation-weighted correlation matrix $\tilde{C}_{ij}$ that represents the coevolution of each pair of position in MSA (Reynolds et al. 2013). It can be calculated using weighted correlation tensor $\tilde{C}_{ij}^{(ab)}$.

$$\tilde{C}_{ij}^{(ab)} = \phi_i^a \phi_j^b C_{ij}^{(ab)}$$

where $C_{ij}^{(ab)} = f_{ij}^{(ab)} - f_i^a f_j^b$ - raw correlation of aminoacids, ϕ - conservation-based weighting function as a gradient of relative entropy: $\phi_i^a = |\partial D_i^a / \partial f_i^a|$. Resulting tensor contains correlations of all the possible aminoacid pairs. Frobenius norm can be used to reduce dimensions of the tensor:

$$\tilde{C}_{ij} = \sqrt{\sum_{a,b} \left(\tilde{C}_{ij}^{ab}\right)^2}$$

This matrix contains all the correlations for each pair of aminoacids. In SCA method has further steps to analyze the correlations by spectral analysis.

Thus, the method involves general sequence of the computational steps broadly described below:

1. Multiple sequence alignment and frequencies calculation. MSA can be calculated by progressive method such as used in MATLAB Bioinformatica Toolbox or MUSCLE software. Frequencies of amino acids at defined position are the number of sequences in the alignment having given aminoacid at the position, divided by total number of sequences.
2. Definition of the position specific conservation. Conservation of the aminoacid at given position, calculated independently of other positions, can be measured by the statistical quantity named Kullback-Leibler relative entropy.
3. Correlated conservation. This step involves a few general considerations:
    i. Calculation of covariance matrix which represents pair-wise correlations between amino acids at given positions. Fundamental principle of SCA method is computing of the correlations not for raw alignments, but for a weighted.
    ii. Choosing of weights. The approach of SCA is to consider the effect on the conservation of each position upon removing each sequence. The idea is that this "perturbation" will provide an estimate of the significance of each amino acid at each position in the alignment by its impact on the measure of conservation used (relative entropy). Finally, we have obtained SCA-tensor describing positional and pair-wise correlations which is 4-dimensional array of L positions × L positions × 20 amino acids × 20 amino acids.
    iii. Reduction to positional correlations. In this step we should reduce SCA tensor to 2D matrix of positional correlations. This can be achieved by method of singular value decomposition. Thus, set of each pairs of positions can be reduced to the 20×20 matrix of amino acid correlations.
    iv. Spectral decomposition (independent component analysis). From this point, identifying sectors from SCA positional correlation matrix starts. The presence of significant correlations between positions in the matrix indicates that treating the amino acid positions as the basic functional units of proteins is not the most informative representation. Eigenvalue (spectral) decomposition produces a reparameterization of the protein. The first modes of the spectra represent a collective aminoacid groups which co-evolve per

the positional correlation matrix. Thus, we can find the co-evolved groups of amino acids named sectors.

3. RESULTS AND DISCUSSION

3.1. SAXS measurements

3.1.1. GH1 from *Saccharophagus degradans*

For the experiments, GH1 from *Saccharophagus degradans* (SdBgl1B) were prepared. The samples were obtained at the concentration of $3\ mg/ml$ in Tris-HCl $50\ mM$ buffer solution at pH 7.5. This work is described in our article in "Molecular Biotechnology" (Brognaro et al. 2016). The SAXS scattering curve and its primary analysis are given in Figure 2.

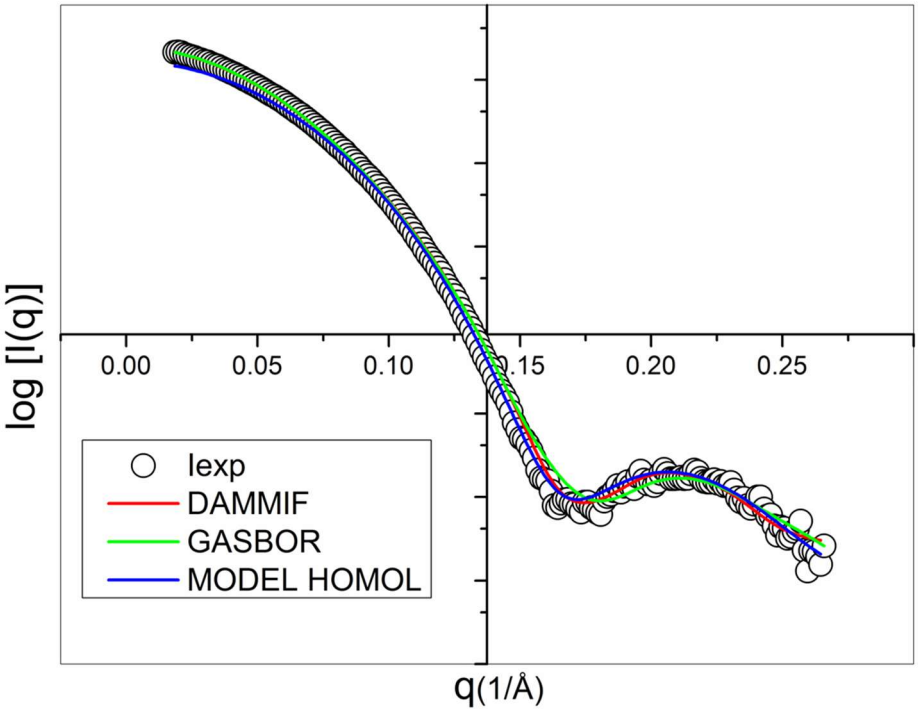

Figure 2. Experimental and modeled intensities from GH1 (SdBgl1B)

Radius of gyration R_g of 23.90 ± 0.30 Å was computed from the inclination of the Guinier area defined by criteria $q_{max} \cdot R_g < 1.3$ (Feigin, Svergun,1987). Pair-distance distribution function $p(r)$, obtained by indirect Fourier transformation of the scattering curve shows a symmetric profile typical for compact/globular macromolecules (Figure 3a). The maximum protein size, D_{max}, obtained from this curve is 67 Å and radius of gyration, which had been calculated from distance distribution function ($R_g = 23.65$Å), is in agreement with the value obtained from Guinier analysis (23.90 ± 0.30 Å).

We further computed Kratky plot, $I(q)q^2\ vs.\ q$, which can be used to access globularity and flexibility of macromolecules in solution Figure 3b. In the case of SdBgl1B, the Kratky plot exhibits a bell-shaped peak at low q and converges to the q-axis at high values. This indicates that the enzyme is a compact and well-folded protein.

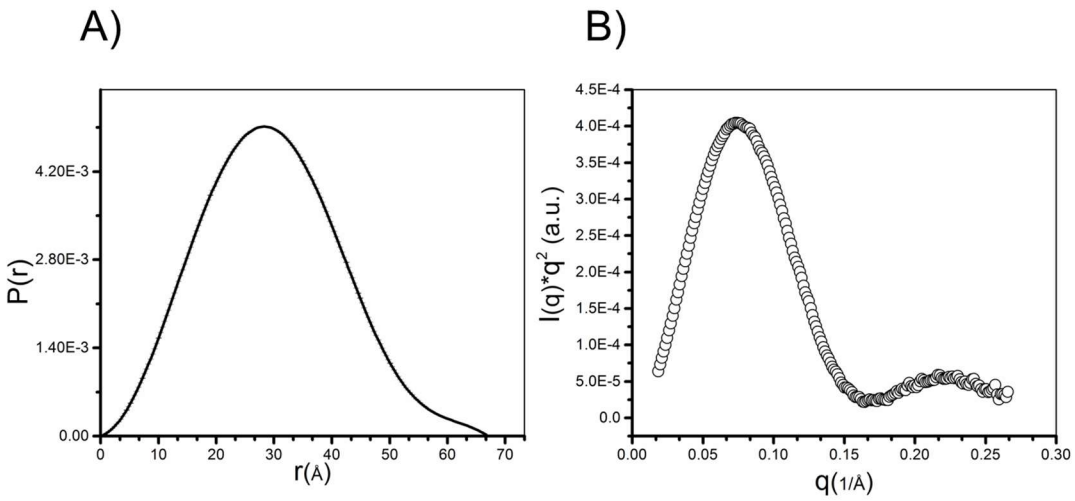

Figure 3. Analysis of the SAXS data from GH1 (SdBgl1B). A) Pair-wise distribution function, B) Kratky plot

Next, we reconstructed the 3D low-resolution shape of the protein using $ab-initio$ modeling performed with DAMMIF package (Franke and Svergun 2009). The package performs several independent computations (usually, ten) from different initial randomly chosen models to reveal the most persistent molecular shapes in the averaged DAMs solution. The scattering intensity from the final DAM was compared directly with the experimental scattering curve. Also, Figure 2 shows intensity reconstructed from ab-initio simulated DAMs as red line fitted with experimental SAXS-curve (open-circle in black) and the GASBOR-derived scattering curve (green line).

As one can see, the reconstructed scattering curve is in very good agreement with experiment ($\chi = 1.69$), as well as the simulated scattering curve computed from the homology model ($\chi = 1.76$). Obtained finally DAM model clearly shows that SdBgl1B has an approximately globular molecular shape in solution (see Figure 4 and Table 1).

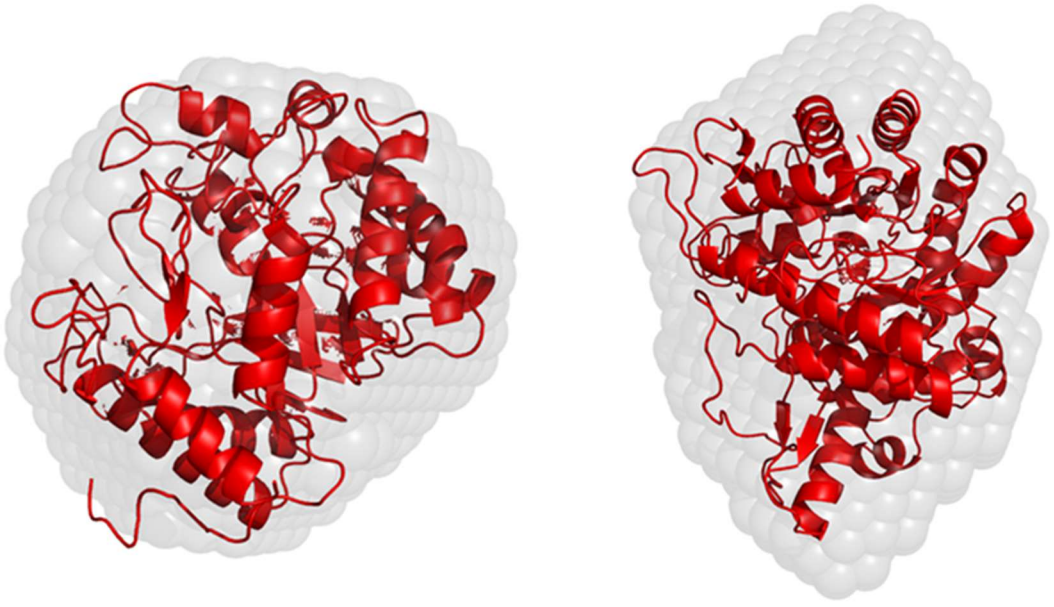

Figure 4. Front and side view of SdBgl1B. Monomeric ab initio shape in grey obtained from SAXS measurements with superimposed the SdBgl1B model (in red) obtained by I-tasser using the template crystal structure with PDB ID 3WH6 as an initial model

Table 1. SAXS data collection and experimental parameters

Data collection parameters	
Instrument	SAXS2 (LNLS)
Wavelength ($Å^{-1}$)	0.015418
q-range ($Å^{-1}$)	0.015 to 0.2658
Exposure time (s)	300
Concentration (mg/ml)	2
Temperature (℃)	20
Structural parameters	
R_g (Å) from Guinier analysis	23.90 ± 0.30
R_g (Å), estimated from $p(r)$	22.47
D_{max} (Å)	67
χ, Experiment/DAMMIF	1.69
χ, Experiment/GASBOR	1.90
χ, Experiment/HOMOL	1.76
Molecular weight determination	
Sequence molecular weight (kDa)	49.9
Molecular weight (kDa) from SAXSMoW2	50.93
Discrepancy (kDa)	2.08
Software for data treatment	
Primary data reduction and processing	Fit2D and ATSAS suite
Ab initio analysis	DAMMIF and GASBOR
Align obtained *ab initio* models	DAMAVER
3D structure superimposes	SUPCOMB
Oligomer state and molecular weight	SAXSMoW2

To build the homology model of SdBgl1B using I-Tasser (Petersen et al. 2011), we separated 3D-structures from the PDB database with the highest sequence identities (within homogenity range of 35 to 43%) and which sequence cover more than 90% of SdBgl1B aminoacid sequence. This comparison shows that SdBgl1B exhibits the globular form similar to the observed for structures of the β-glucosidase from *Paenibacillus polymyxa* (PDB ID 2O9P, (Neuhoff et al. 1985), β-glycosidase from *Thermus nonproteolyticus* HG102 (PDB ID 1NP2, (Zhang and Lynd 2004), β-glucosidase from soil metagenome (PDB ID 3CMJ, (Miller 1959) and β-glucosidase from *Humicola insolens* (PDB ID 4MDO, (Leatherbarrow and Enzfitter 1987). We estimated SdBgl1B molecular mass from the experimental SAXS curve using SAXSMoW2 package. The molecular mass of the protein in solution had been computed to be 50.93 kDa, which differs from theoretical molecular weight calculated based on the aminoacid sequence (49.75 kDa) only by 2%. This further confirms that SdBgl1B forms monomers in solution. This is different from some other GH1 β-glucosidases that form large molecular assemblies, such as dimers or tetramers in solution. For example, Zanphorlin and collegues (Provencher 1982) showed GH1 β-glucosidase from *Exiguobacterium*

antarcticum B7 (EaBglA) forms tetramers in solution. This quaternary structure stabilizes native conformation of the enzyme in solution and augments its activity 10 times as compared to the enzyme monomeric form. Homology model of SdBgl1B snugly fits low-resolution SAXS-derived DAM leaving no doubts about approximately spherical molecular shape and a monomeric form of the enzyme in solution (Figure 4).

3.1.2. GH3 from Bifidobacterium adolescentis

Two GH3 enzymes from *Bifidobacterium adolescentis* were purified in our laboratory at IFSC/USP and new SAXS experiments were carried out. These enzymes were prepared and SEC-SAXS measureaments performed at B21 beamline at Diamond Light Source Synchrotron. Experimental data are given in Figure 5 and Figure 6.

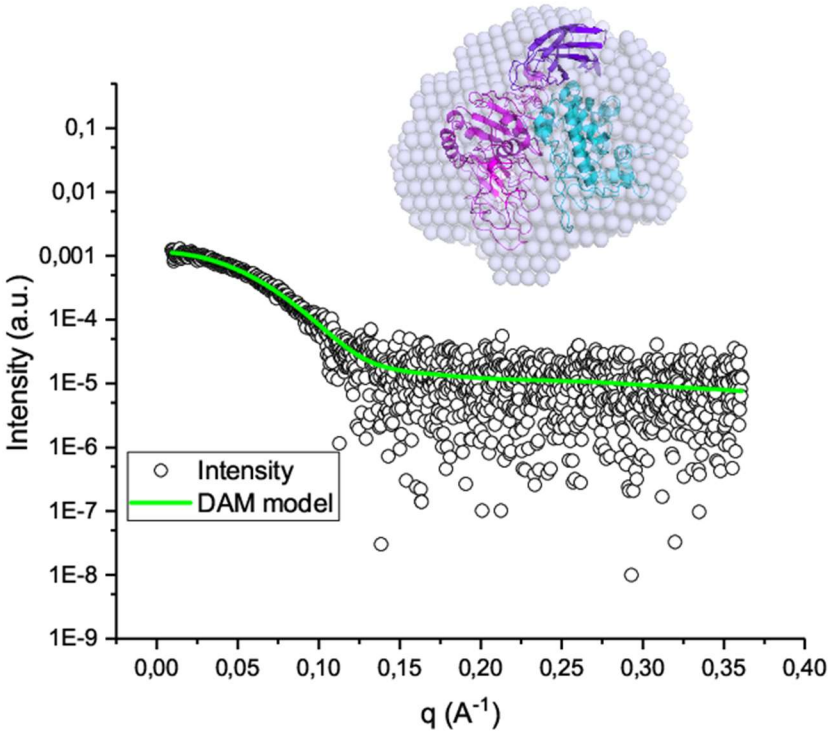

Figure 5. Experimental SAXS data on GH3 from *Bacillus adolescentis*, theoretical curve restored from DAM model. In insert: Low-resolution shape combined with homology model

GH3 from *Bacillus adolescentis* presented in Figure 5 is a monomer with three domains and sequence molecular weight of $100\ kDa$. Three-dimensional shape of low-resolution restored from an experimental data was combined with homology model. For homology modeling of GH3 from *Bacillus adolescentis*, we constructed the atomic models using structural homology-modelling approach (Webb and Sali 2014) combining the crystal structures such as β-glucosidase 3b from *Thermotoga Neapolitana* with PDB id 2X40, β-glucosidase from *Kluyveromyces marxianus* with PDB id 3AC0 and β-glucosidase from *Streptomyces Venezuelae* with PDB id 3ABZ determined by X-ray crystallography.

Based on known crystallographic structures as well as the homology model, theoretical scattering profile was computed using the FOXS software (Schneidman-Duhovny et al. 2016). Discrepancy between experimental and theoretical SAXS curves was quantified minimization of χ-parameter (Schneidman-Duhovny et al. 2016). In order to

evaluate the molecular weight of the protein in solution, the software SAXSMoW2 (Piiadov et al. 2018) has been used.

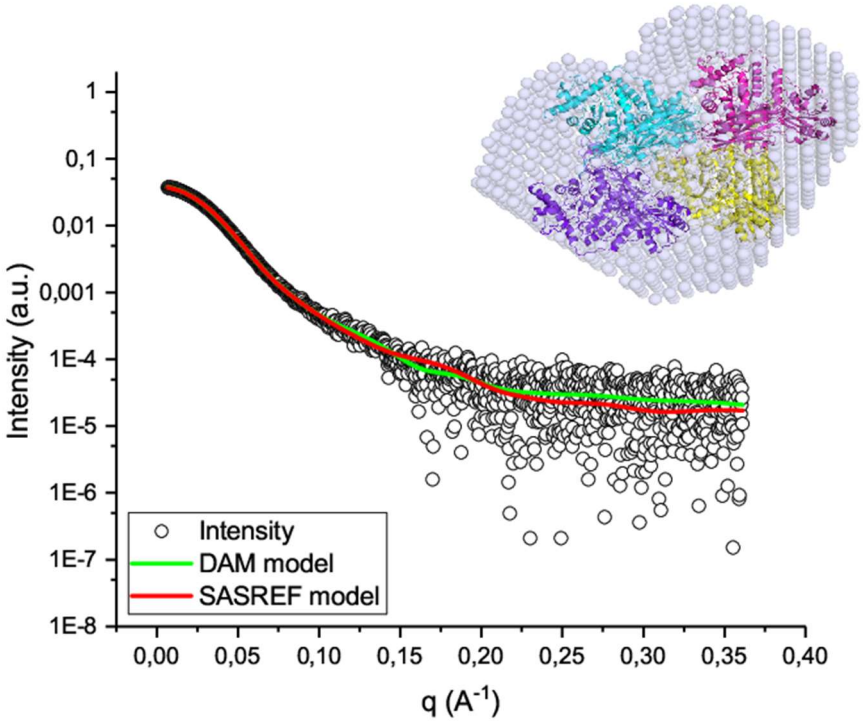

Figure 6. Experimental SAXS data on GH3 from *Bacillus adolescentis*, theoretical curves restored from DAM model and from SASREF quaternary structure modeling. In insert: Low-resolution shape combined with crystallography structure

Another studied enzyme was also GH3 from *Bacillus adolescentis*. The molecule of this protein represents a tetramer with three domains in each chain. For the protein, crystallography structure was obtained in our laboratory at IFSC/USP and it is shown in the insert of Figure 6. It has tetrameric crystallography structure and it is the same structure as in solution. Since this GH3 from *Bacillus adolescentis* forms oligomers, modelling of quaternary structure was performed using SASREF package (Petoukhov and Svergun 2005). Sequence molecular weight of the protein is 80 kDa. Molecular weight based on the experimental profile was calculated using our package SAXSMoW2 giving an estimated value of 307.2 kDa. Thus, discrepancy from theoretical molecular weight of tetramer was about 4%.

This, the both proteins shown in both Figure 5 and Figure 6 have perfect agreement between theoretical and experimental data that that confirms a quality of the calculated three-dimensional shapes.

3.2. SAXSMoW2 software for data processing

The previous version of SAXSMoW (Fischer et al. 2010) was developed for determining the molecular weight of proteins in dilute solution starting from a single SAXS curve measured at a relative scale. Instead of calculating the "true invariant", Q, as DatPorod does, this program determines a truncated or "apparent" Porod invariant, Q'. SAXSMoW is a simple to use, fast and relatively precise method for determinations of the molecular weight of proteins in dilute solution (Guttman et al. 2013). The program has been widely used during the last decade

to determine the molecular weights of many proteins with different shapes and forms, including globular, elongated, flexible and glycosylated proteins, and also protein complexes (Ferreira et al. 2011; Zheng et al. 2012; Carter et al. 2015; Gruszka et al. 2015; Chang et al. 2018; de Araújo et al. 2018; Glauninger et al. 2018; Kadowaki et al. 2018).

SAXSMoW is also currently used as an accurate tool for quick diagnosing of protein molecular weights, such as reported for UltraScan-SOMO SAXS pipeline (Brookes et al. 2016) at SOLEIL synchrotron SWING beamline, for BioXTAS RAW SAXS pipeline (Hopkins et al. 2017) at the Cornell High Energy Synchrotron Source bioSAXS CHESS beamline, for BL16B1 SAXS beamline at the Shanghai Synchrotron Radiation Facility (Zeng et al. 2017) and summarized in the workflow for determinations of molecular weights and quaternary structure of proteins in solution (Korasick and Tanner 2018).

In this section a new version SAXSMoW 2.0 is described and the user's workflow is presented in Figure 7 using an example of Bovine Serum Albumin (PDB 3V03) in solution. It is a web-based utility for processing SAXS data which is available at http://saxs.ifsc.usp.br/. This web application can be used online without the need of downloading. The input is a one-dimensional SAXS intensity text file (".dat" file) containing at least two columns: The modulus of the scattering vector q and the scattering intensity, $I(q)$, in arbitrary or relative units. All other columns in the uploaded file, if any, are discarded.

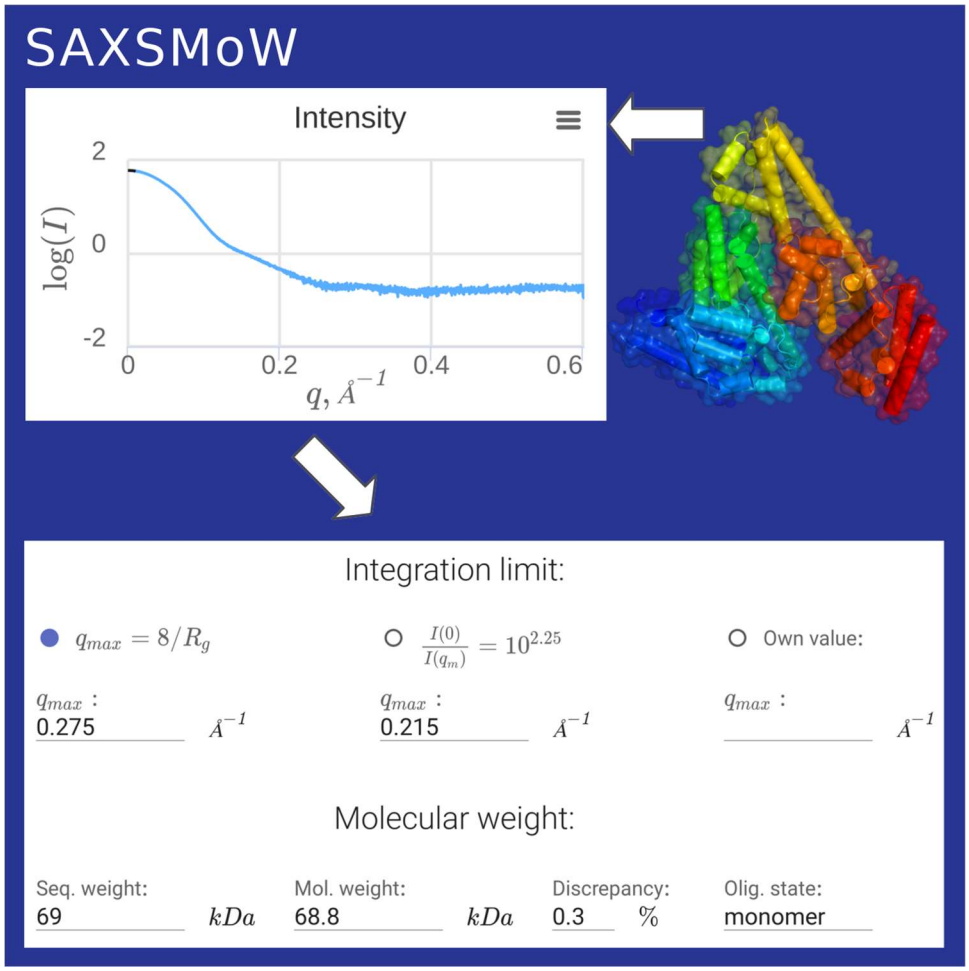

Figure 7. SAXSMoW2 workflow diagram: from SAXS to molecular weight

The program automatically performs Guinier fitting, computes the molecular radius of gyration R_g, generates Kratky $I(q)q^2$ and Porod $I(q)q^4$ plots, suggests q-intervals for the calculation of the Q'-invariant, and determines the molecular weight of protein from the SAXS data recorded in experiments.

The input of the previous version of the SAXSMoW program (Fischer et al. 2010) is the regularized intensity function obtained by Indirect Fourier Transform (IFT) of the scattering intensity. This transformation is performed using the GNOM program of ATSAS package (Petoukhov et al. 2012). The newly developed SAXSMoW 2.0 program directly applies the Guinier extrapolation to raw experimental SAXS data (after subtracting the scattering intensity from the buffer), without a need to use ATSAS package.

3.2.1. Computation of the molecular radius of gyration and extrapolation of SAXS intensity to $q = 0$

The radius of gyration of homogeneous particles with a constant electron density defined as $R_g = [(1/V)\int r^2 dv]^{1/2}$ characterizes their size and compactness. This parameter can be determined from SAXS data in several different ways. SAXSMoW 2.0 utilizes a method based on Guinier approximation for the scattering intensity, which applies to SAXS curves at low q (Guinier 1939). Guinier approximation can be written as

$$\ln I\,(q) = \ln I\,(0) - q^2 R_g^2/3 \tag{3.1}$$

Thus, R_g is determined from the slope of a straight line that asymptotically (at low q) fits the experimental Guinier ($\log I(q)$ versus q^2) plot.

The accuracy of R_g determined from Guinier analysis depends on several factors. First, interference effects on SAXS curves due to spatial correlation of protein positions may strongly affect the low-q region of the scattering curve. In order to eliminate correlations or interference effects in the SAXS curves, the proteins in solution should be studied in dilute conditions. To accomplish this, SAXS measurements are carried out for several protein concentrations and the results are extrapolated to zero concentration. The experience gained by frequent users of SAXS beamlines associated to synchrotron X-ray sources, usually allows them a priori estimations of the required protein concentrations for achieving dilute condition, without the time-consuming extrapolation procedure described above. Protein concentrations in typical dilute solutions are of the order of a few mg/ml. Also, concentrated solutions may cause inter-particle repulsion depended of molecular charge and pH of solution. In the case, this effect can affect on scattering curve and make dufficult to determine parameters of scattering as well as aggregation which is another source of systematic errors in data analysis. It is the eventual partial formation of aggregates which leads to a significant increase in the scattering intensity at very small angles. The scattering intensity from aggregates overlaps the signal from the remaining non-aggregated molecules and changes the shape of the scattering curve.

SAXSMoW 2.0, by default, performs Guinier fitting automatically, but the program also offers an option for manually defining values for q-range of fitting interval. The implemented strategy of Guinier fit search is based on testing of all possible Guinier fits and selecting the one with the best fit, within the range $R_g \cdot q_{max} < 1.3$ by combined criteria of Pearson correlation coefficient and a length of the fitting q^2-range. The area of search is limited by q_{max} value of 0.15 Å^{-1}. Depending on the resolution of the experimental curve, a minimal q^2-range length of

acceptable fit was defined as $3 \cdot 10^{-3}\text{Å}^{-1}$ or q-range corresponding to the first five experimental points of the dataset if it corresponds to a higher resolution. Moreover, imported data having more than 10^6 points is limited to this number.

For globular particles, analysis of the accuracy of R_g calculations obtained from experimental SAXS data in the interval $q \cdot R_g < 1.3$ results in a systematic error lower than a few percents. Thus, small errors in this interval allow for a reliable determination of R_g. Over the $q \cdot R_g < 1.5$ interval, the deviation can reach 20-30%, while over the $q \cdot R_g > 2$ region, this approximation becomes highly imprecise (Guinier 1939; Feigin and Svergun 1987). SAXSMoW 2.0 checks $q \cdot R_g$ values for each tested fitting line and discards the ones for which the mentioned relation is not satisfied. A final set of possible fits is first selected by maximizing the Pearson coefficient and, second, by maximizing the length of the fit range, in such a way that considering two fits with similar Pearson coefficients, the fit with a larger linearity range is selected.

Theoretically, the best Guinier approximation should be the one that exhibits the highest Pearson coefficient. However, this allows for cases situations in which, for example, the fitting line with Pearson coefficient 0.995 and fit interval length of 5 experimental points would be preferred to another fitting straight line with a Pearson coefficient equal to 0.994 and a total length of 100 points. Obviously, choosing the first option would be a wrong decision because Pearson coefficients 0.995 and 0.994 indicate similar (equal) quality of the approximation in the statistical sense. On the other hand, fittings over higher numbers of experimental points provide a more robust result, which is less sensitive to eventual local artifacts of data set. Thus, in such cases, the second option is selected. To implement this, fitting lines with differences in Pearson coefficients smaller than 0.001 are assumed to be with same fitting quality and thus, in these cases, the line with a larger fitting interval is selected. After application of such decision filter, the SAXSMoW 2.0 algorithm selects a Guinier fitting with a higher Pearson coefficient and assumes it as the best fitting line to be used for $I(0)$ calculation.

Thus, regardless of the protein internal structure, the SAXSMoW 2.0 analysis of scattering curves at low q yields R_g and $I(0)$, which depend on the size and compactness of the particle, and on the amount of scattering matter, respectively.

3.2.2. Determinations of the protein volume and molecular weight

For the determination of the molecular volume and molecular weight, SAXSMoW 2.0 starts from the calculation of the apparent protein volume, V', derived from the following equation:

$$V' = 2\pi^2 \frac{I(0)}{Q'} \tag{3.2}$$

where $I(0)$ is the SAXS intensity extrapolated to $q = 0$ which is derived from the linear fitting procedure described in the previous sub-section and Q' is named apparent Porod invariant, which is the truncated integral of the Kratky $I(q)q^2$ function from $q = 0$ up to a selected q_{max}:

$$Q' = \int_0^{q_{max}} I(q)q^2 \, dq \tag{3.3}$$

Notice that V' and Q' are named as apparent volume and apparent Porod invariant, respectively, because the determination of their true values requires integration of the Kratky function from $q = 0$ up to $q = \infty$.

The first choice for upper limit of integration used by SAXSMoW 2.0 is given by

$$q_{max} \sim 8/R_g \tag{3.4}$$

which corresponds to the estimated maximum value of q which contains relevant information associated with perfectly homogeneous particles. This q_{max} value is often used as, for example, in ATSAS software (Petoukhov et al. 2012).

Another option for determining of q_{max} is suggested in (Kayushina et al. 1974; Feigin and Svergun 1987):

$$\log\frac{I(0)}{I(q)} \sim 2 \dots 2.5 \tag{3.5}$$

Thus, the equation $\log[I(0)/I(q)] = 2.25$ was implemented as a second option of q_{max} in SAXSMoW 2.0. Value of 2.25 was chosen as an average value for the interval from Eq. 3.5.

The next step of SAXSMoW 2.0 is to establish the relations between the true protein volume, V, and the apparent protein volumes associated to different q_{max} values, $V'(q_{max})$. For this purpose, CRYSOL program (Franke et al. 2017) is used for determinations of the SAXS functions of a large number (1148) of proteins with known 3D high-resolution structures deposited in the PDB. The integrals of Kratky functions truncated at different q_{max} values and the values of $I(0)$ are determined for all selected proteins, which allow for the calculation of their apparent volumes $V'(q_{max})$. Moreover, the true volumes, V, of all selected proteins are easily computed from their known high-resolution structures. Thus, as described with more detail in a previous work (Fischer et al. 2010), the true protein volumes V was found to exhibits dependences on the apparent volume V' for all selected q_{max}, given by

$$V = A + B \cdot V' \tag{3.6}$$

The A and B coefficients were determined in the first version of SAXSMoW for several q_{max} values corresponding to the $V(V')$ function built up by starting from the known high-resolution structures of 1148 selected proteins downloaded from the PDB (Fischer et al. 2010).

The linear equation Eq. 3.6 is used for determining true volumes of proteins from their apparent volume computed from experimental SAXS curves truncated at one of the q_{max} values for which the coefficient A and B were reported in (Fischer et al. 2010). In SAXSMoW 2.0, coefficients A and B are interpolated over the whole q-range, from $q = 0.1\text{Å}^{-1}$ up to 0.5Å^{-1}, by the following polynomials:

$$A[\text{Å}^3] = -2.114 \cdot 10^6 q_{max}^4 + 2.920 \cdot 10^6 q_{max}^3 - 1.472 \cdot 10^6 q_{max}^2 + 3.349 \cdot 10^5 q_{max} - 3.577 \cdot 10^4$$
$$B = 12.09 q_{max}^3 - 9.39 q_{max}^2 + 3.03 q_{max} + 0.29 \tag{3.7}$$

in which $[q_{max}] = \text{Å}^{-1}$. A and B values for different q and their polynomial approximation given by Eq. 3.7 are shown in Figure 8 and Figure 9.

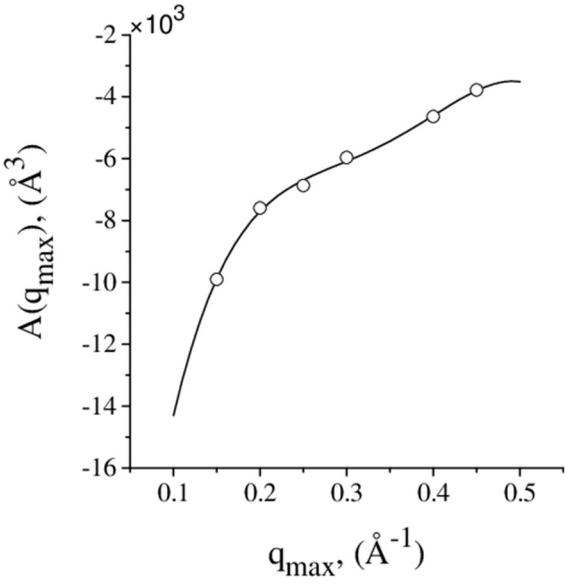

Figure 8. Plot of the polynomial which define $\boldsymbol{A(q_{max})}$ for all $\boldsymbol{q_{max}}$ values from $\mathbf{0.1\text{Å}^{-1}}$ to $\mathbf{0.5\text{Å}^{-1}}$

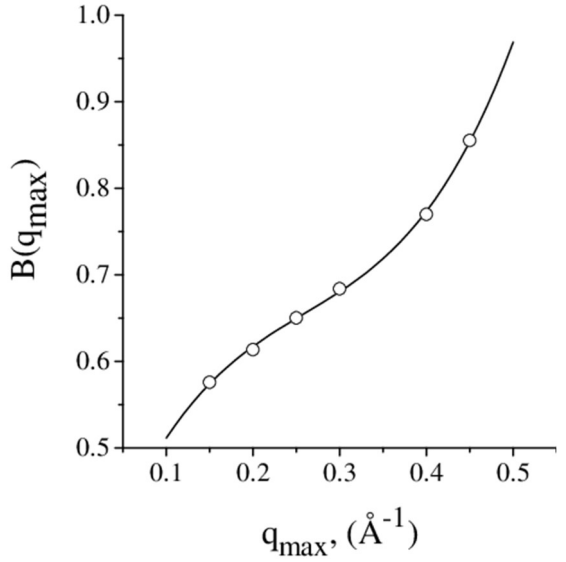

Figure 9. Plot of the polynomial which define $\boldsymbol{B(q_{max})}$ for all $\boldsymbol{q_{max}}$ values from $\mathbf{0.1\text{Å}^{-1}}$ to $\mathbf{0.5\text{Å}^{-1}}$

Finally, the molecular weight of proteins is calculated from their volume V by

$$MW[kDa] = \frac{\rho_m[g/cm^3]V[cm^3]}{1.662 \cdot 10^{-21}[g/kDa]} \tag{3.8}$$

where $\rho_m = 1.37 g/cm^3$ is the mass density of proteins. This value is assumed to be constant because density is the reciprocal of the partial specific volume which is not vary greatly from one globular protein to another (Gekko and Noguchi 1979; Squire and Himmel 1979).

3.2.3. Testing SAXSMoW 2.0 on experimental datasets

In order to evaluate the accuracy in the results yielded by SAXSMoW 2.0, we have applied it to a number of SAXS curves downloaded from SASBDB (Valentini et al. 2014) and BIOISIS http://www.bioisis.net databases. These databases contain open experimental SAXS data of many proteins, and protein complexes such as protein-protein, protein-DNA and protein-RNA.

We have excluded in the selection of SAXS datasets those corresponding to (i) aggregated or not purified proteins, (ii) protein-DNA/RNA complexes, (iii) partially folded/unfolded or very disordered proteins, and (iv) metalloproteins. Furthermore, 175 datasets containing SAXS intensity curves corresponding to well-folded proteins were selected for analysis. Notice that the expected molecular weights of all selected proteins are known.

Furthermore, in order to better understand the influence of non-globularity on the quality of the SAXSMoW 2.0 calculations, we have selected 18 experimental SAXS curves from proteins having aspect ratios ranging between 1.2 and 18. The aspect ratios were computed by dividing the lengths of two main axes in a spheroid approximation for the protein shape, based on its 3D model. Both datasets were processed by SAXSMoW 2.0 in automatic mode using both suggested options for q_{max}, as described in the previous subsection.

3.2.4. User interface and usage

The user interface of SAXSMoW 2.0 is shown in Figure 10. To get started, users must upload the selected ".dat"-file containing the experimental data to be analyzed. A data file is uploaded and processed on the server side automatically and then results are displayed on the same page. "Guinier fitting" section shows the relevant parameters derived from a linear fitting in a Guinier plot, namely the fitting interval in Å^{-1} units, the extrapolated intensity $I(0)$, the radius of gyration R_g and the $q \cdot R_g$ relation associated to the fit.

The second part of the interface regards the choice of the upper limit of integration (q_{max}) for the calculation of the apparent Porod invariant Q'. As mentioned above, there are three available options (i) $q_{max} = 8/R_g$, which is the default option, (ii) q_{max} satisfying the condition $\log\frac{I(0)}{I(q)} = 2.25$ and (iii) q_{max} manually selected by the user. For options 1 and 2, the values of q_{max} are automatically calculated by SAXSMoW 2.0.

The third part of the web interface displays the calculated molecular weight. If the expected molecular weight is known (which can be, for example, computed based on a known aminoacid sequence) and specified, the program also displays the oligomeric state and the discrepancy between calculated molecular weight and the expected value. If necessary, the calculation of molecular weight can be repeated, for example, after manually updating Guinier fitting or varying the upper integration limit.

Figure 10 display the results of calculations shown in the screen using experimental SAXS data from Bovine Serum Albumin (BSA) https://www.sasbdb.org/data/SASDA32/, taken from SASBDB database (Valentini et al. 2014). Four plots are displayed in Figure 10 showing: (i) experimental SAXS intensity, (ii) Guinier plot $\log I\,(q)$ vs q^2, (iii) Kratky function $I(q)q^2$ vs q and Porod function $I(q)q^4$ vs q. Finally, a file with all the results can be downloaded.

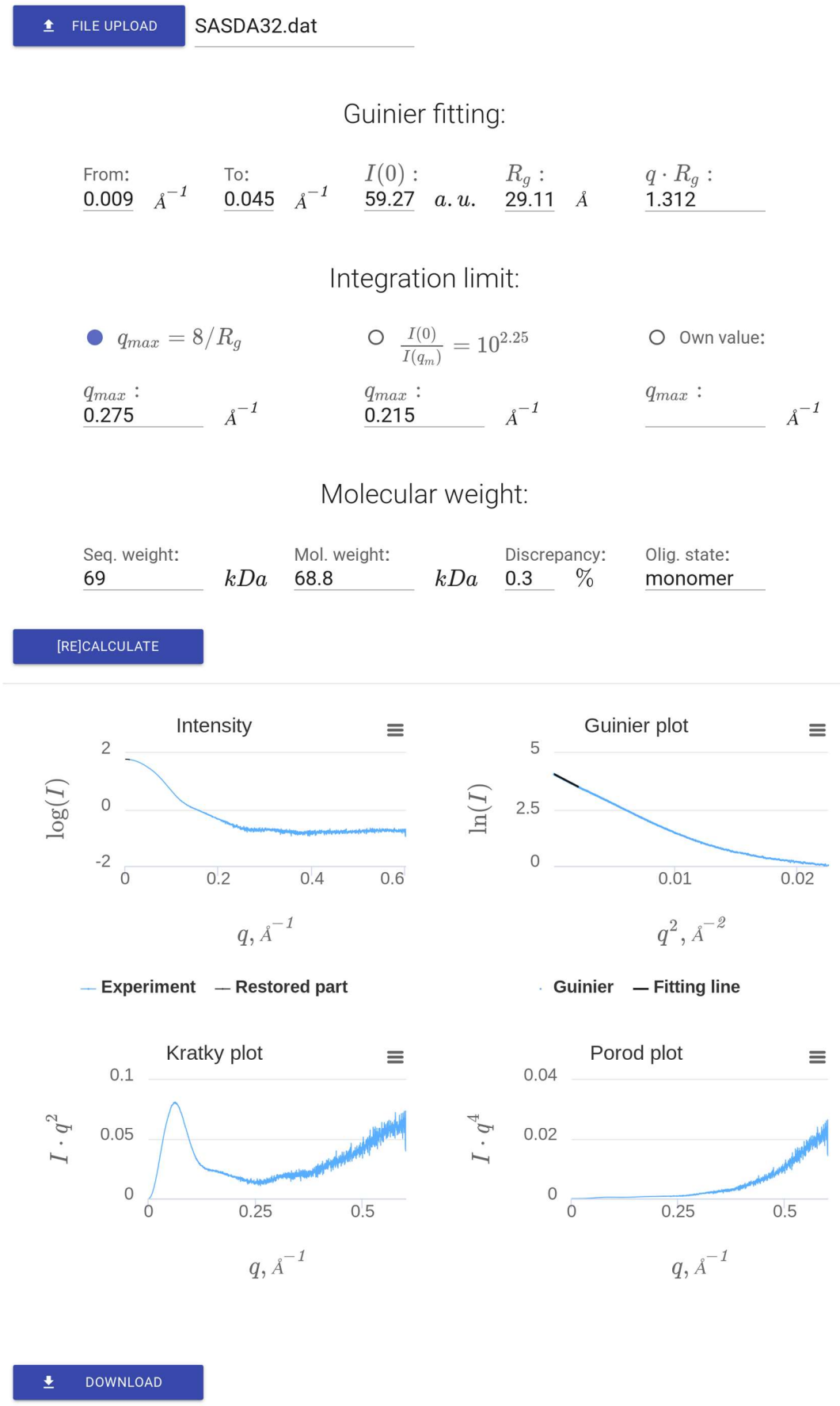

Figure 10. SAXSMoW 2.0 interface displaying results associated to the SASDA32 dataset from SASBDB

3.2.5. Precision of molecular weight determinations: Globular proteins

The magnitude of the relative error in the molecular weight determined by SAXSMoW 2.0 can be considered as a relative discrepancy of the output value from the program with respect to a nominal molecular weight, i. e.

$$D = |\frac{m}{m_0} - 1| \cdot 100\% \tag{3.9}$$

where m and m_0 are calculated molecular weight and that expected, respectively. The distribution of relative errors D for test SAXS datasets measured from globular proteins is displayed in Figure 11 and Figure 12A-B for both suggested options for q_{max} selection. Notice that the figures do not represents datasets for which automated search of q_{max} is not available.

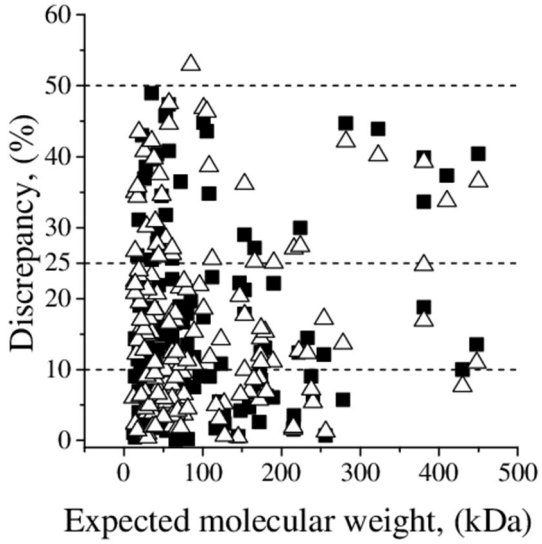

Figure 11. Discrepancy distributions for different expected molecular weights. (■) - Option 1: $q_{max} = 8/R_g$; (△) - Option 2: derived from equation $I(0)/I(q_{max}) = 10^{2.25}$

Figure 11 displays the distribution of calculated molecular weights and their respective discrepancies with respect to their expected values. The molecular weights for most of the evaluated SAXS datasets exhibit discrepancies lower than 10%. This statistic leads us to the conclusion that well-folded and compact proteins, SAXSMoW 2.0 in most cases allows for determining molecular weights with good accuracy. However, the observed distribution also includes a few outliers with discrepancies larger than 25%.

Figure 12 show that a number of proteins with a given discrepancy in their molecular weights monotonically decreases for increasing discrepancy values. One can see that the use of the first (default) option, $q_{max} = 8/R_g$, leads to determinations of the molecular weights with somewhat lower discrepancies, resulting in a more compact discrepancy distribution. This is evidenced by comparing Figure 12A with Figure 12B. Distribution of discrepancy is not symmetric, but monotonically decreases because of modulus in Eq. 3.9. Without the modulus brackets when the distribution is becoming symmetry, we obtain false estimation of mean error of the method because of mutual compensation of errors from different proteins. For example, positive discrepancy 50% will erase

an impact of negative discrepancy of -50% in mean estimation of the method precision. To remove this effect, we introduce a modulus brackets in Eq. 3.9. Also, since we have tested datasets from different proteins with a lot of various parameters such a shape, weight (note, an error of weight is normalized differently for different proteins in definition of discrepancy), measure conditions, etc., we cannot fit the obtained distributions by Gaussian even without the modulus in Eq. 3.9. Also, an error of the method obviously is different for different proteins because of, at least, limits of theoretical assumptions behind. Thus, a median was chosen as robust statistical measure of central tendency for obtained monotonically decreased error (discrepancy) distribution. The median, as 50%-quantile can be as well used as a distribution width estimation relatively to $D = 0$.

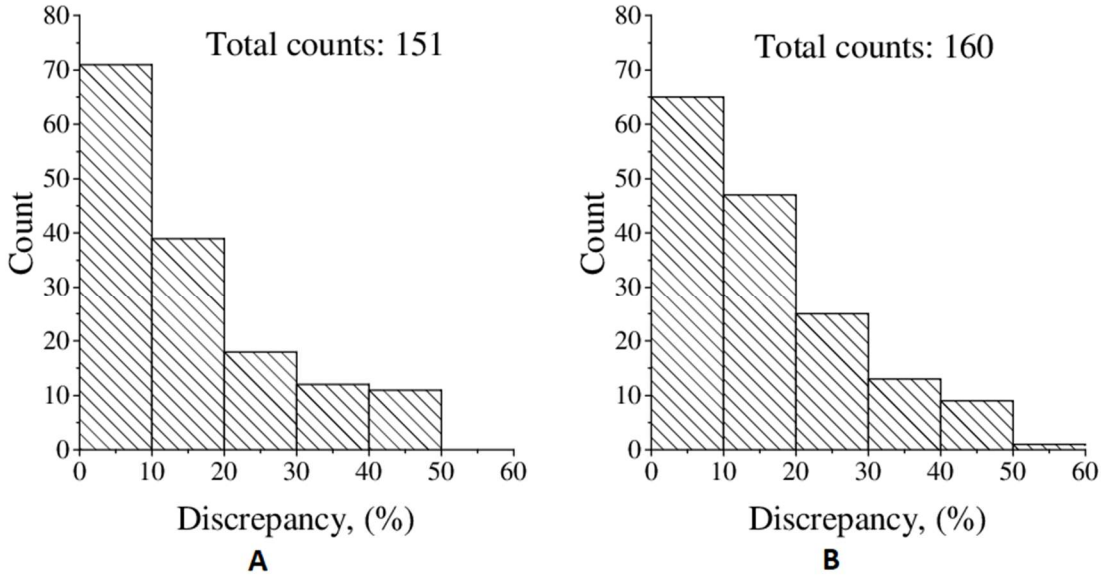

Figure 12. Discrepancy distributions in molecular weights computed for a set of globular proteins using different q_{max} values: (A) $q_{max} = 8/R_g$ and (B) q_{max} from equation $\log\frac{I(0)}{I(q_{max})} = 2.25$

Table 2 reports statistical features of distributions of discrepancies in molecular weights associated to each suggested option for automatic determination of q_{max} available in SAXSMoW 2.0. Table 2 shows for 50% of the dataset, SAXSMoW 2.0 calculates molecular weights with an error smaller than 11.01% when q_{max} is defined as $8/R_g$ (option 1) and smaller than 12.25% when q_{max} is defined by the equation $\log\frac{I(0)}{I(q_{max})} = 2.25$ (option 2). In both cases, the median discrepancy is lower than 12.5%.

Table 2. Statistics on distributions of discrepancy D for globular proteins set

Option 1: $q_m = 8/R_g$		
Minimum	Median	Maximum
0.08	11.01	48.95
Option 2: q_{max} by eq. $I(0)/I(q_{max}) = 10^{2.25}$		
Minimum	Median	Maximum
0.33	12.25	52.90

3.2.6. Precision of molecular weights determinations: Elongated proteins

Eighteen datasets were selected to establish the influence of non-globularity on the accuracy of the results yielded by SAXSMoW 2.0, for elongated molecules with estimated aspect ratio ranging from 1.2 to 18. As shown in Figure 13, relative discrepancies between computed molecular weights and those a priori expected, clearly follow an increasing trend as the protein shapes become more elongated. Nevertheless, SAXSMoW 2.0 is quite successful in calculations of the molecular weights of proteins with aspect ratio lower than 8.0 (with discrepancies of about 9.4%) and with aspect ratio of 10 to 18 (with discrepancies of about 21%). Interestingly, q_{max} computed by equation $I(0)/I(q_{max}) = 10^{2.25}$ (option 2 in Figure 13) generally achieves better results for molecules having high aspect ratio.

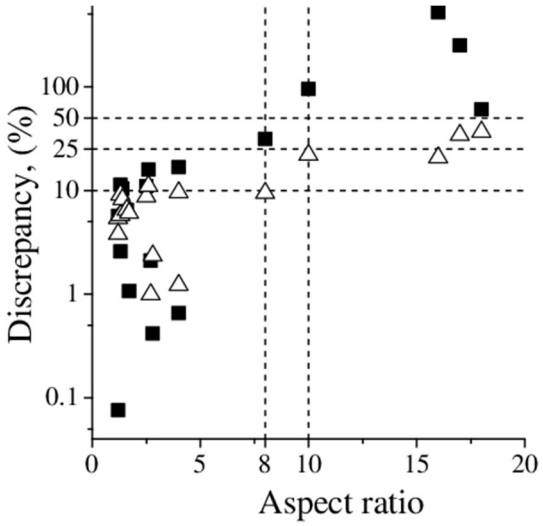

Figure 13. Discrepancies in molecular weights associated to elongated proteins with different aspect ratios for both suggested options for q_{max} values. First option (■) - $q_{max} = 8/R_g$. Second option (△) - derived from equation $I(0)/I(q_{max}) = 10^{2.25}$

The reported analyses indicate that SAXSMoW 2.0 can determine the molecular weights of a number of proteins with different aspect ratios. However, for SAXS data corresponding to elongated proteins with aspect ratio higher than 10, calculations of the molecular weights become progressively less precise.

Furthermore, cases for which molecular weights are difficult to assess, such as for highly elongated and/or very flexible proteins, Kratky plots do not exhibit a well-defined maximum peak and only show a flat plateau at high q. In these more challenging cases, SAXSMoW 2.0 could still be useful for evaluating the oligomeric states of proteins in solution, even though the discrepancy is about 20%.

3.2.7. Comments on determinations of molecular weights

As described above, the molecular weights of folded and compact proteins can be determined by using the automatic mode and default option for the upper integration limit, yielding values with average discrepancy of about 10%. Furthermore, in more complex cases, such as those referring to elongated or flexible proteins, an advanced user can select the q_{max} value manually as it was also implemented in the previous version of SAXSMoW program (Fischer et al. 2010).

As an example of the program application to well-behaved globular proteins, the molecular weights were calculated for different SAXS datasets using available data for a BSA monomer, BSA dimer and xylose Isomerase from *Streptomyces rubiginosus*, which are deposited respectively as entries SASDBJ3, SASDBK3 (Jeffries et al. 2016) and SASDAB6 (Franke et al. 2015) in SASBDB repository. Using default mode for integration limit determining, obtained molecular weights for monomeric and dimeric BSA were $64.7\ kDa$ and $123.2\ kDa$ with discrepancies of 2% and 7.4%, respectively.

The molecular weight of tetrameric *Streptomyces rubiginosus* was estimated as $157.1\ kDa$ with a 9.2% discrepancy. This value of the molecular weight agrees with those previously determined by (Jeffries et al. 2016) and SASDAB6 (Franke et al. 2015).

However, in cases for which the aspect ratio is approximately 10 or higher, discrepancies in molecular weight estimates are above 10% (see Figure 13) as it is observed for myelin-associated glycoprotein (entry SASDB56) (Pronker et al. 2016) and surface G protein (entry SASDA37) (Gruszka et al. 2015) having aspect ratio of 10 and 18, respectively.

When the upper limit q_{max} is manually selected, SAXSMoW 2.0 consistently leads to a discrepancy in molecular weight up to 10% for globular proteins and larger than 10% for proteins with aspect ratios of 10:1 to 18:1 as shown in Figure 13.

As expected, large discrepancies in molecular weight determinations were observed for unfolded/disordered, metal-depend, and aggregated proteins. This is illustrated, for example, by scattering behavior of human persulfide dioxygenase ETHE1, which is a metal-dependent protein and for its metal-free forms (entries SASDAH7, SASDAJ7, SASDAK7, SASDAL7, SASDAM7, and SASDAN7; https://www.sasbdb.org/project/76/) which are highly elongated. In these cases, linear behaviors of $\log I(q)$ versus q^2 at low q are not apparent. Thus, Guinier fitting with acceptable accuracy and extrapolation of $I(q)$ down to $q = 0$ cannot be done. However, for a metal-bound form of the enzyme (entry SASDAF7), the molecular weight determined by SAXSMoW 2.0 has a discrepancy of 14.7% with respect to the expected molecular weight. Similarly, in the case of the protein ORF 2047.1 from *Pyrococcus furiosus* (Hura et al. 2009), which is an unfolded macromolecule; the program has been unsuccessful in determination of its molecular weight.

3.2.8. Criteria for selecting the upper limit for integration of Kratky function

After the input of the dataset containing raw SAXS curve, SAXSMoW 2.0 users need to decide which upper integration limit, q_{max}, to select for the calculation of the truncated integral of the Kratky function (Eq. 3.3).

The suggestion is to select the default option ($q_{max} = 8/R_g$), which is widely used in several packages for analyses of SAXS results. It is noteworthy that the SAXS intensity $I(q)$ up to this upper q-limit contains most of the relevant structural information associated to strictly homogeneous particles. This implies that the comparatively weak effects from molecular flexibility and density fluctuations are expected to strongly affect the Kratky function mainly above $\sim 8/R_g$. A strong contribution to the integral of the $I(q)q^2$ function above $q = 8/R_g$ can clearly be seen in the example of Kratky plot corresponding to the SASDA32 data set shown in Figure 10. The same behavior is apparent in Porod plots of SAXS curves of many other proteins.

In some cases, the first option for q_{max} may lie outside the available $q_{max} = 0$ range ($0.1\text{Å}^{-1} < q_{max} < 0.5\text{Å}^{-1}$) over which the A and B parameters of the linear functions $V(V')$ are defined. For these SAXS curves, the

second option for q_{max} can be tested. If selecting the second option yields $q_{max} > 0.5\text{Å}^{-1}$, the suggested choice is to use $q_{max} = 0.5\text{Å}^{-1}$.

For many proteins with the molecular weights below $20\ kDa$, the upper limit q_{max} for the integration of the Kratky function is higher than $0.5\ {Å}^{-1}$. For the analysis of these small proteins, the use of the maximum value $q_{max} = 0.5\text{Å}^{-1}$ is advisable.

Alternatively, users may opt for a manual mode of q_{max} selection. SAXSMoW 2.0 allows for choosing any q_{max} value between 0.1Å^{-1} and 0.5Å^{-1}. In this case, users should avoid selecting too low q_{max} to avoid strong truncation of the $I(q)q^2$ function and keep q_{max} below the high q-range over which the contribution to the integral Q' from density fluctuations is high.

SAXSMoW 2.0 is a user-friendly program for robust and quick online determinations of the molecular weight of proteins in dilute solution from experimental SAXS intensity data collected on a relative scale. This program builds up on its previous version (Fischer et al. 2010), which was widely applied during the last decade. The SAXSMoW 2.0 exhibits new features with respect to the previous version (Fischer et al. 2010), namely:

- Input of background-subtracted SAXS intensity curves without the need to use auxiliary packages;
- display of experimental data as $I(q), I(q)q^2$ and $I(q)q^4$ for visual examination;
- automatic Guinier fitting of the experimental SAXS intensity at low q, calculation of the molecular radius of gyration, R_g, and determination of $I(0)$ by extrapolation of SAXS curve down to $q = 0$;
- suggestions of two options of upper integration limits for calculation of the truncated integrals of Kratky functions which allow for quick and automatic determinations of molecular weights;
- possibility for calculations of molecular weight by selecting any value for the upper integration limit q_{max} within the $0.1\text{Å}^{-1} < q_{max} < 0.5\text{Å}^{-1}$ range to compute the apparent molecular volume.

The test analyses of many openly available SAXS datasets indicate that SAXSMoW 2.0 allows for determining molecular weights with a median discrepancy lower than 12% for globular (i.e. not very elongated) and homogeneous proteins. For elongated proteins having aspect ratios up to 18 and highly flexible proteins, the discrepancies are much higher.

The program SAXSMoW 2.0 is fully implemented online and freely available at http://saxs.ifsc.usp.br webpage. Also, the work on SAXSMoW 2.0 was described in the paper (Piiadov et al. 2018).

3.3. Modifications of statistical coupling analysis method

The method of SCA was significantly modified to improve analysis of highly correlated sets of coevolving aminoacids. Cloud implementation of statistical coupling analysis (SCA) with methodical improvements was developed as a web application. New methodical features include ability of analysis of sequence in FASTA format without known structure, automated search of set of homologous using BLAST method and its automated multiple sequence alighnment in case of user does not have an own alignment for analysis.

The modified method was applied to several families of enzymes such as GH7 (both exo- and endogluconases), GH74, GH3, GH1, GH48 and Xylose Isomerases.

3.3.1. Postprocessing improvements of the analysis

SCA method is used to characterize the pattern of evolutionary constraints between amino acid positions in a protein family. The method is applicable to multiple sequence alignment (MSA) of the family and allows measurements of the overall functional constraint at each sequence position (positional correlations) and analysis of the coupled functional constraint on all pairs of sequence positions (pairwise correlations) (Lockless and Ranganathan 1999; Socolich et al. 2005). MSA alignment we have obtained using MUSCLE algorithm, one of the bests alignment algorithms (Robert C 2004; Dereeper et al. 2008), and MATLAB software.

As a key parameter, SCA uses a frequency of amino acids at defined position that is a lot of sequences in the alignment having given amino acid at given position, divided by the total number of sequences. On this parameter, definition of the position specific conservation is based. Conservation of an amino acid at given position can be measured by the statistical parameter named Kullback-Leibler relative entropy (Kullback and Leibler 1951) that is used in SCA.

During usage of SCA method, it was observed that a quality of the results is highly dependent from a size of the sequence alignment. We have developed a several features to post-processing procedures in SCA method.

After SCA calculated independent components (IC), our new developed software makes a normalization of the SCA-correlation matrix and plots each pair correlation inside each IC. To separate stronger correlations, we apply a filter having 5 bins for normalized correlation values: "perfect" - $0.9 \div 1.0$; "good" - $0.7 \div 0.9$; "average" - $0.5 \div 0.7$; "poor" - $0.1 \div 0.5$ and "garbage" - $0.0 \div 0.1$. We considered to work with correlations from "perfect", "good" and "average" bins. These bins characterize the quality of the pair correlations, so we can study subsets with stronger correlations and weaker correlations separately. We introduce a parameter characterizing filtered IC named degree of coherence. A coherence of a set of positions implies a strong correlation between each pair in given filtered IC. Filtered set of pairs we can represent as a graph $G(E,V)$ where edges E are correlations between positions which represented as vertices V. In terms of graph theory, degree of coherence can be defined as a density of graph ρ:

$$\rho = \frac{E}{N}$$

where E - number of edges, N - number of vertices.

As it can be seen from the equation, high density means more edges between vertices, or, in other words, more position pairs have strong correlations between them. Obviously, in extremal cases, density can by calculated as:

$$\rho_{max} = \frac{1}{N}\binom{N}{2}$$
$$\rho_{min} = 0$$

Here, $\binom{N}{2}$ is maximal possible number of edges in graph which is simple binomial coefficient.

For more convenience we can use normalized ρ:

$$\rho_n = \frac{\rho}{\rho_{max}}$$

where ρ_{max} - constant defined above. Therefore, after simple substitution and calculation of binomial coefficient we have:

$$\rho_n = \frac{2E}{N(N-1)}$$

Thus, $\rho \simeq 1$ means a case when each position inside the sector strongly correlated with each or approximately each other. And, contrarily, $\rho \simeq 0$ means that the sector has only pair correlations.

Further we automated procedure of searching if interference between ICs. To remind, ICs having high interference between them form only one sector whereas ICs without interference form separate sectors. To evaluate degree of interference betwen ICs, we have written special procedure which calculate average quality of interefering set of correlations from two ICs and place the pair of ICs in one of three bins: "independent" - calculated quality of interference less than 0.3; "pseudo-independent" - case of $0.3 \div 0.7$ and "joint" - case of $0.7 \div 1.0$. Thus, if pair of ICs highlighted as "joint", they should be merged in the same sector.

In addition, we introduced the term named “coherent core”. This is the maximal subset of positions and correlations (vertices and edges in graph) that has $\rho_n = 1$. In graph theory, a subset of vertices having this property named maximal clique. Finding maximal cliques is well-known mathematical NP-complete problem. An effective algorithm to resolve this problem was offered by Joep Kerbosch and Coenraad Bron (Bron and Kerbosch 1973) which we use in our work.

Thus, coherent core (or cores) inside a sector represents collective coevolution of the amino acids in given sequential positions whereas others positions of the sector represents only pair correlations.

3.3.2. Methodical improvements and cloud implementation of SCA

A cloud implementation of statistical coupling analysis (SCA) was developed. Generally, the method of SCA and our modifications have been described in previous section and here the cloud implementation with minor methodical modifications is described. The term "cloud" means type of programming architecture which has been used. The diagram showing the architecture of the application is shown in Figure 14.

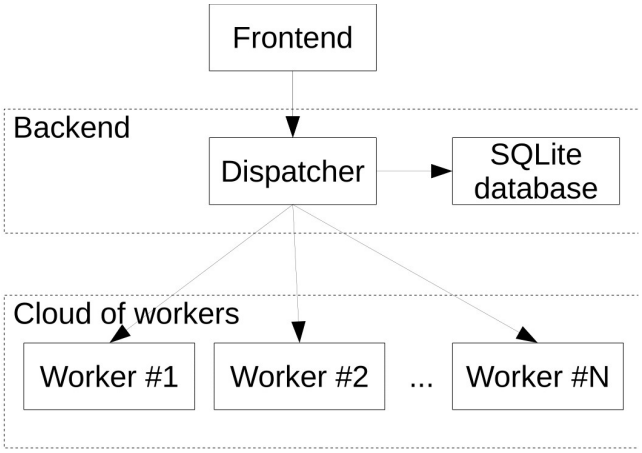

Figure 14. CloudSCA architecture diagram

In Figure 14 has three layers: **frontend** - JavaScript reactive application launched in a user browser, **backend** - server application launched at IFSC/USP to dispatch user requests between underlying layer of **workers** - instances doing all the calculations of SCA in parallel mode for different users in the same time. Also, the application uses SQLite database to mantain user sessions and to store a job metadata. As can be seen from Figure 14, the application can be horizontally scaled to any number of workers that hardware resources allow.

User interface of the frontend part is shown in Figure 15. The figure shows input data for analysis of structure of the anzyme from GH48 family as an example.

Figure 15. CloudSCA interface with a data prepared for analysis

In the new implementation we introduce features allow an analysis without prepared set of sequences. The application can find out a set of homologous for an input sequence using BLAST method (Altschul et al. 1997). Also, it allows FASTA format for reference sequence and does not require a structure. It could be useful for an analysis of new proteins with unknown structure to detect coevolution patterns that must be reflected in the structure.

Result should be obtained, normally, in a few hours and a page with results of SCA calculations is represented in Figure 16.

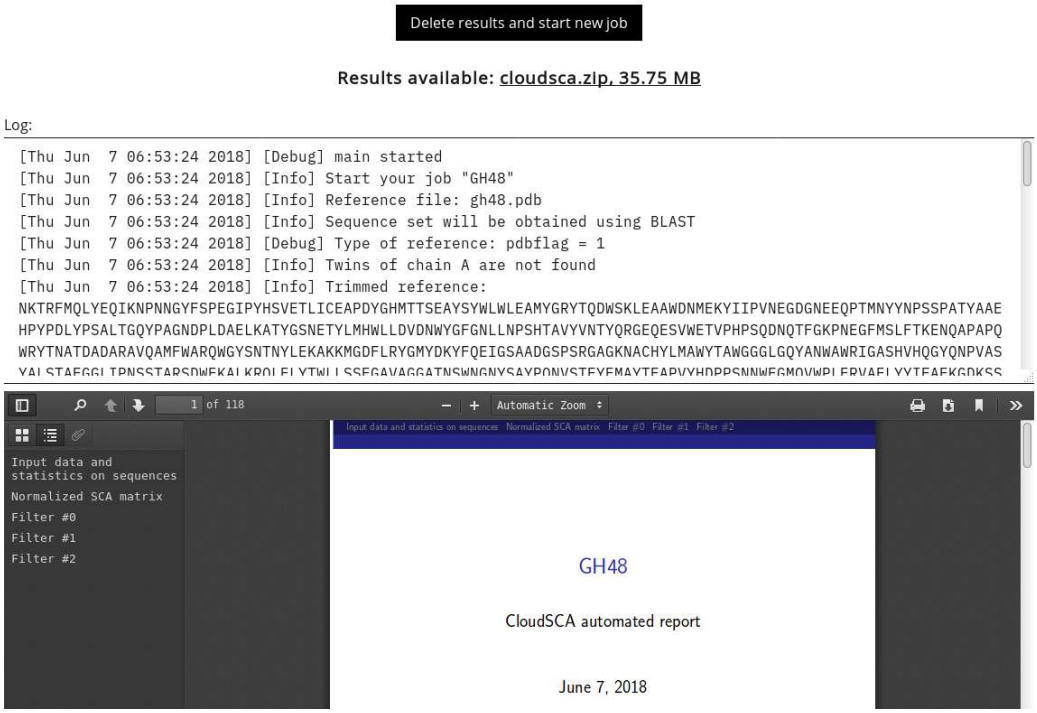

Figure 16. CloudSCA results after calculations

A page with results is available after calculations and contains a log with technical information and PDF report compiled automatically using $LaTeX$, engine. All the images using to create PDF report are also available to download as high-quality images and as a raw data for reusing. Also, if a structure of the reference was provided for analysis, the application produces "pml"-files containing all the found coevolution patterns of aminoacids projected in the structure. These files should be used with PyMol application (DeLano 2002) to examine structural and functional features of found coevolution patterns.

On example of GH48 analysis, automated report shown in Figure 16 contains statistical characteristics of identity and similarity for analyzed set of sequences as histogram (Figure 17), "SCA-matrix" representing all the statistical correlations between aminoacids (Figure 18), general statistic information on each pattern such as size of sector, coherent cores, density, etc. The resulting report on SCA calculations contain images of structure of the reference sequence with projected sector (Figure 19) and graphs representing a network of correlations between aminoacid positions in the reference sequence (Figure 20).

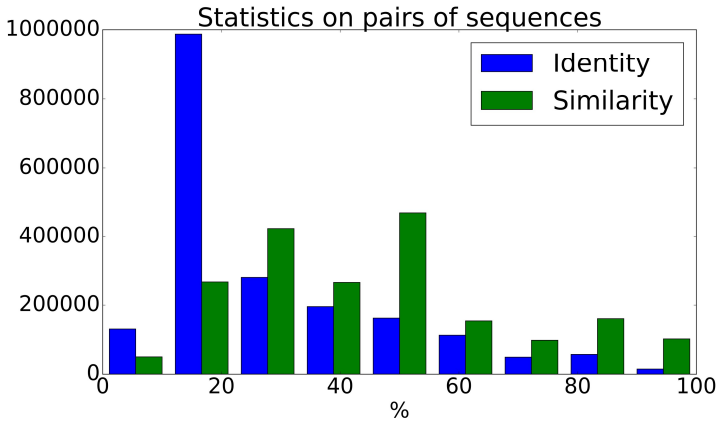

Figure 17. Statistical check of quality of studied set of sequences: identity and similarity for analized bunch of sequences

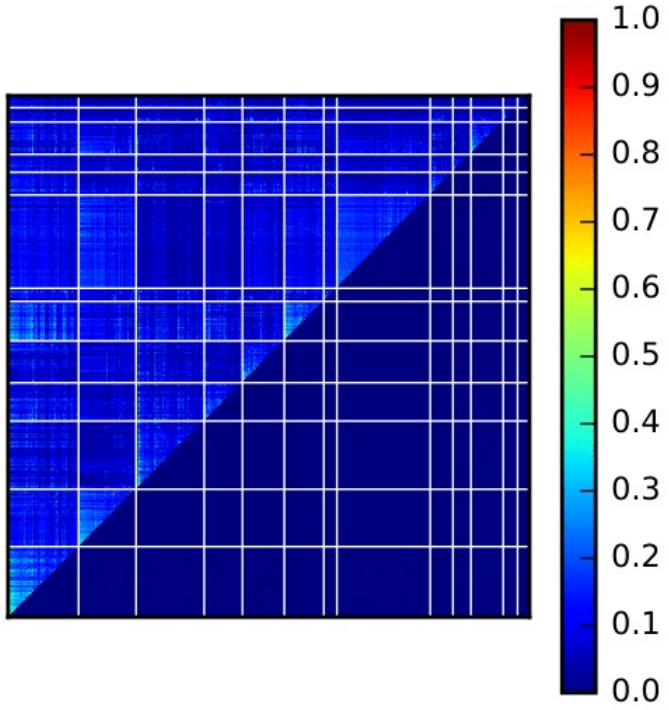

Figure 18. SCA matrix representing aminoacid pair correlations

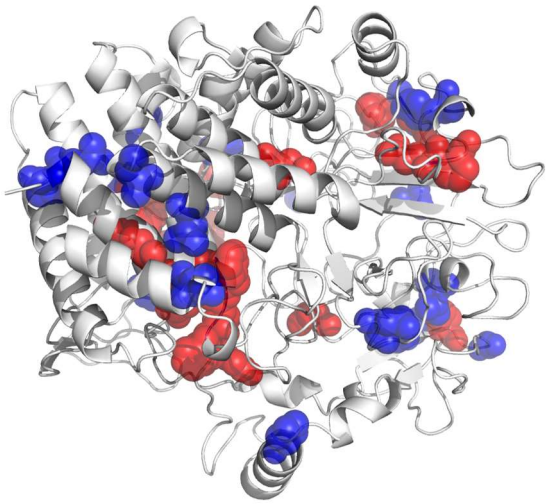

Figure 19. Reference GH48 structure with projected pattern formed by red and blue coloured aminoacids: "red" subset has density equal to 1

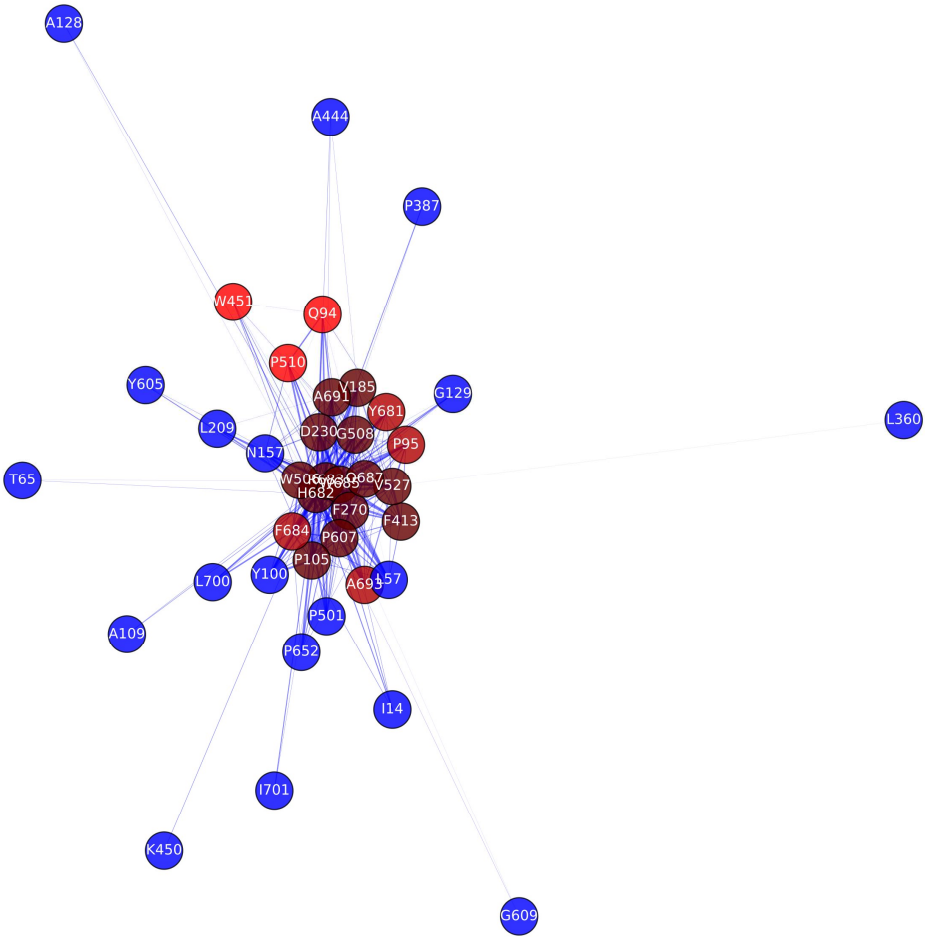

Figure 20. The graph showing pair correlations inside a pattern: two tones of red show two sets of collective strong correlations, blue represents pair SCA correlations inside the sector

3.4. Statistical coupling analysis applying

3.4.1. GH7 enzyme family

Cellobiohydrolases (CBHs, 3.2.1.91) from GH7 family are enzymes that hydrolyse β-1,4 linked glycoside bonds that connect glucose moieties into a linear polymer known as cellulose. Cellobiohydrolases however are mostly able to attack cellulose polymer chains at their termini, processively clipping off dimers of glucose (cellobiose molecules) (Henrissat et al. 1995). Endoglucanases (EC 3.2.1.4) on the other hand choose a random internal, mostly amorphous, cellulose chain as the hydrolysis sites thus generating new termini and liberating stalled CBHs, which in turn result in synergism when combined with cellobiohydrolases (Väljamäe et al. 1999, 2003). Because the substrate binding site involves four, five, six or more glucose molecules and the catalytic center only hydrolyses one glycosidic bond per enzymatic cycle, the breakdown product typically is two or three sugar units long. β-glucosidases, on the other hand, are enzymes with a similar catalytic mode of action but specialized in hydrolysing glucose oligomers producing single glucose units (Chauve et al. 2010).

The substrate of cellulases, cellulose, is a linear polymer consisting of D-glucose units tethered solely to each other by glycoside β-1,4 bonds supporting long linear chains with lengths of 2,000 to 25,000 glucose residues. Individual chains cooperate through an intricate hydrogen bonding pattern which alternates with the idiosyncratic arrangement of the pyranose ring and conformational changes of hydroxymethyl groups, forming microfibrils which exhibit various degrees of crystallinity \(Marchessault et al. 1983). Accessing crystalline hydrophobic cellulosic microfibrils by enzymes that contain a hydrolytic catalytic active center is problematic (Liu et al. 2011) – they require water and enough porosity to access the substrate.

Most fungal genomes encode multiple forms of two types of cellobiohydrolases, the GH7 reducing and GH6 non-reducing end cleaving enzymes, while typically present in multiple isoforms, at least one in each family contains an extra linker and a cellulose binding (CBM1) domain (Van Tilbeurgh et al. 1986). GH7 cellobiohydrolase with a CBM1 is the major enzyme secreted by most fungi (often half of all secreted protein mass) in biomass hydrolysis conditions (Nummi et al. 1983; Momeni et al. 2013) that acts processively from the reducing terminus from a single cellulose chain with retention of substrate configuration (Davies and Henrissat 1995; Boisset et al. 2000; Schülein 2000). Because of the processive nature along with activity on single cellulose chains makes these enzymes notoriously inefficient (low turnover rates with high affinity constants) thus justifying their massive overproduction. Nevertheless, this overproduction also illustrates that the nature of cellobiohydrolase cleavage of glycosidic bonds is important in cellulose degradation making worthwhile the invested energy in synthesizing and secreting such sizable amounts of a single protein.

In SCA analysis we have used the structure of GH7 enzyme MtGH7 obtained in our laboratory at IFSC/USP.

In Figure 21 and Figure 22 normalized pair correlations for endoglucaneses and exoglucanases are shown. On diagonal triangles we can observed considered independent components of the SCA correlation matrix. We found two sectors #1 and #3 for the set of endoglucanases. Sector #1 contains 23 positions with degree of coherence of 0.19. Coherent core has size of 7 aminoacids and highlighted in red in Figure 23.

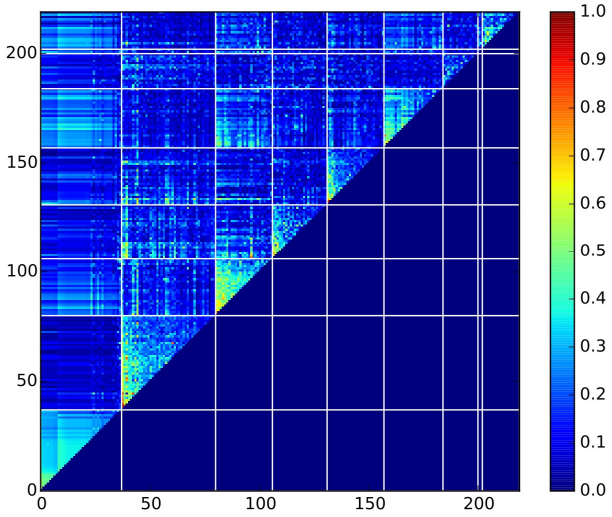

Figure 21. SCA pair correlations inside each IC for GH7 endoglucanases

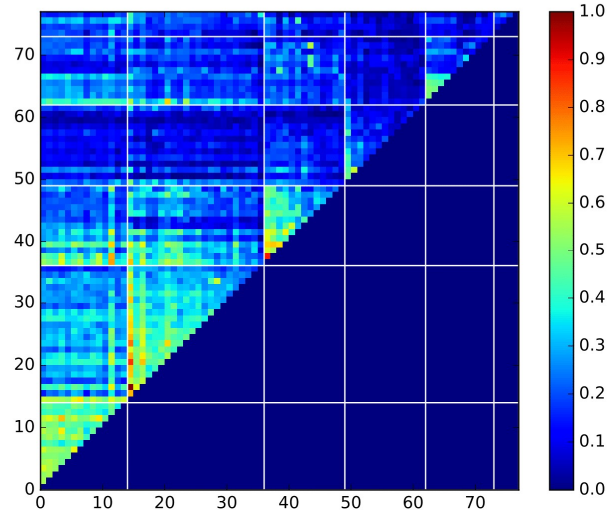

Figure 22. SCA pair correlations inside each IC for GH7 exoglucanases

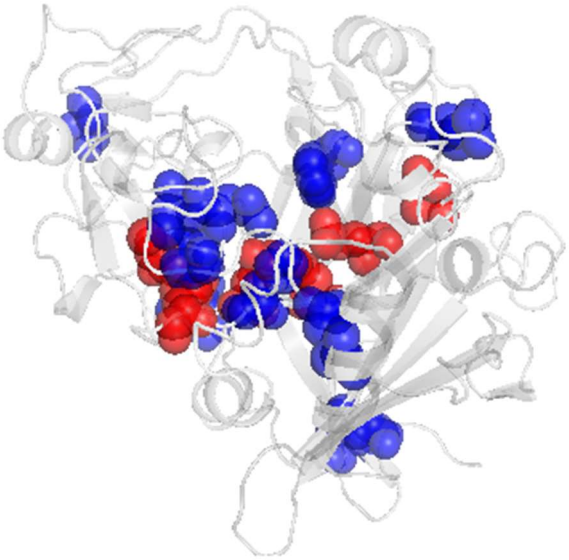

Figure 23. MtGH7 structure with highlighted first sector (endoglucanase)

We can observe that coherent core has compact continous shape and placed on the open side of molecule where is considered active site for endoglucanases. Thus, this sector may have functional importance of the correlated positions. Third sector shown in Figure 24 has 25 positions with degree of coherence of 0.47 that shows stronger internal correlations in comparison with the first sector. As expected from sector with high degree of coherence, this sector has bigger coherent core with size of 10 positions. These positions are placed in compactly on the one side of molecule and, probably, are responsible for the structure factor.

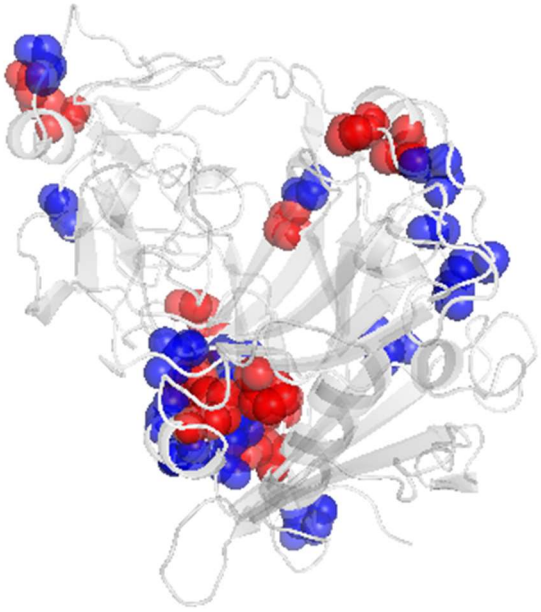

Figure 24. MtGH7 structure with highlighted third sector (endoglucanase)

In case of exogluconase sectors which are represented in Figure 25 and Figure 26, we can observe two continuous sectors. First sector shown in Figure 25. It contains 14 positions (reds on the figure) and has very high degree of coherence - 0.9. This strongly correlated sector is placed inside the molecule near supposed active site and may have responsibility for exoglucanase functioanality.

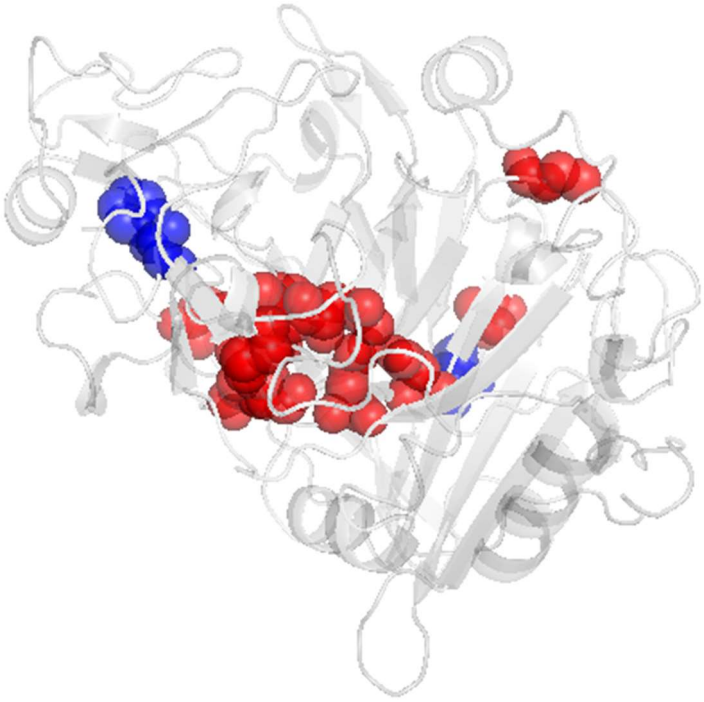

Figure 25. MtGH7 structure with highlighted first sector (from exoglucanase)

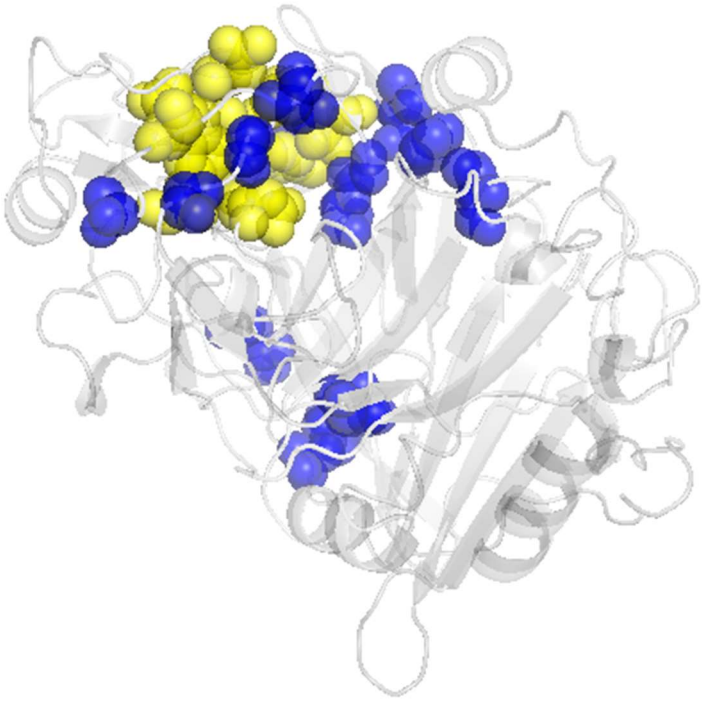

Figure 26. MtGH7 structure with highlighted second sector (from exoglucanase)

Second sector is represented in Figure 26 and has interesting feature: in the figure we can observed compact set of positions highlighted in yellow, it is coherence core with very high correlations ≥0.7. It points to high importance of this sector in the structure of exoglucanase. This sector contains 22 positions with degree of coherence of 0.4. Coherent core has size of 8 amino acids.

3.4.2. GH74 enzyme family

The xyloglucan-specific β-(1→4)-glucanases belong to GH74 Family. They are relatively poorly understood CAZyme family with only 5.7% enzymes characterized biochemically and only 1.9% structures determined to date. Based on structural point of view, some of those enzymes present a restrict cleavage pattern where the glucose is required at -1 subsite. For *Xanthomonas* species, a sequence analysis revealed that the enzymes have a distinct GH74 endo-xyloglucanase substrate. The molecular basis of such mode of action remains a conundrum highlighting a need for deeper understanding of structure-function relationships of the enzymes from GH74 family. We have applyed modified SCA method to set of 1000 sequences separated from CAZy database http://www.cazy.org/GH74_all.html using XcGH74 as a reference.

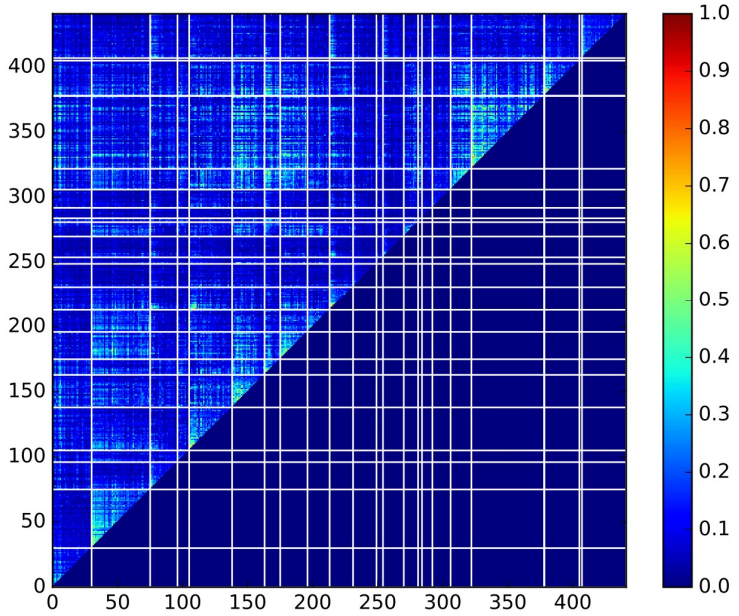

Figure 27. Obtained SCA pair correlations inside each IC for GH74

In Figure 27 shows SCA correlations inside each IC. Each IC highlighted by white rectangle. SCA analysis found 23 independent components forming the sectors. Our additional analysis shows that all the ICs are independent. Applying of the filter significantly reduced number of correlated pairs in 2 times, approximately. Statistically, we have obtained 23 independent components that form the sectors. Most of them is to small to consider them as significant. More continuous sectors shown in Figure 28 and Figure 29. In these figures, coherent cores are highlighted in red. We can see that in Figure 28 coherent core of 8 amino-acids located near activity site and center of the sector. It may point to functional role of this sector in GH74 family. Oppositely, in Figure 29 we observed coherent core of 5 amino-acids on the side that may point to conserved structural feature in GH74 family.

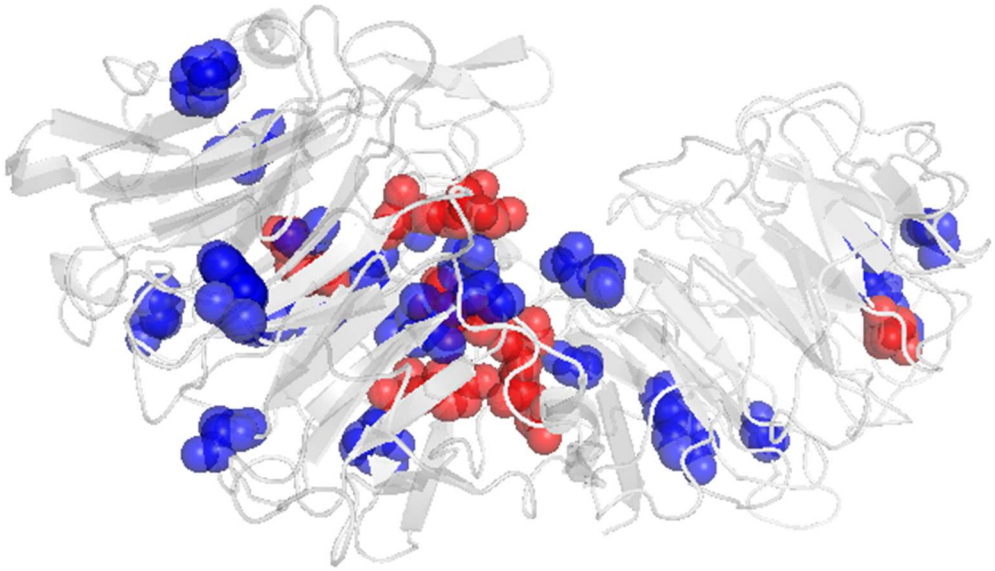

Figure 28. XcGH74 structure with highlighted first sector

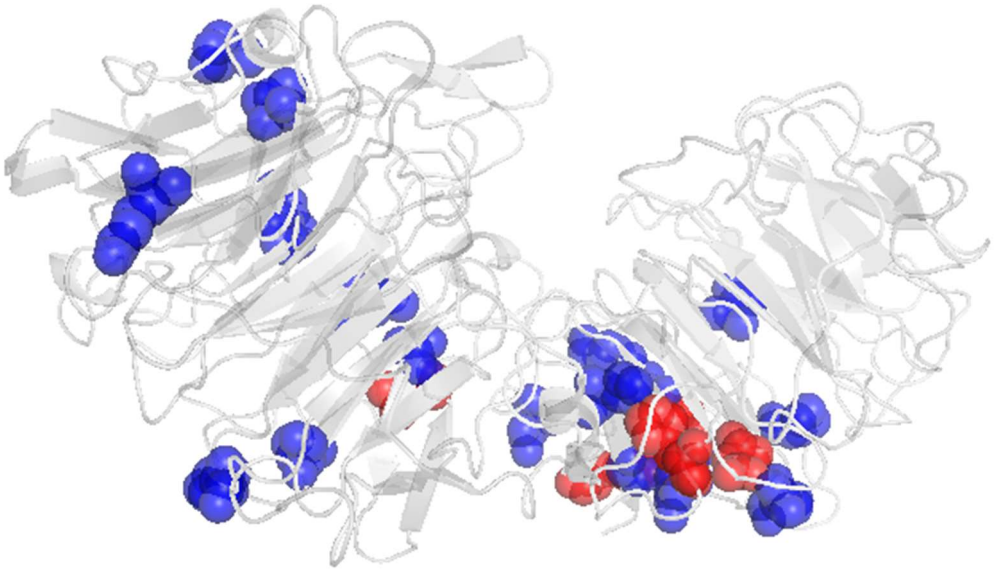

Figure 29. XcGH74 structure with highlighted fourth sector

3.4.3. Xylose Isomerase

Xyloses Isomerases (XI), E.C. 5.1.3.5 have widely-known industrial application in bioenergy. They are actively studied in our Lab at IFSC USP and a novel structure of the XI from *Saccharophagus degradans* (SdXI) was obtained. In collaboration with Lorenzo Briganti (IFSC/USP), we have studied a coupling of site pairs in obtained structure using statistical coupling analysis. This such of coupling defines degree of statistical correlation between frequencies of pairs of aminoacid sites in the sequence that allow to understand better a coevolution of structural and functional features.

For coupling analysis, we have used 2000 homologous of the novel structure. These sequences were obtained using BLAST software and sequences with identity to our reference structure more than 80% were

removed. Then, the resulting set was aligned using MUSCLE software (Robert C 2004). Calculation of correlations between aminoacid sites and clustering into orthogonal clusters was performed using PySCA implementation of the method (Rivoire et al. 2016). For postprocessing, correlation values corresponding all the pairs of aminoacid sites have been normalized and, then, to understand better an impact of individual sites on its cluster, we sequentially removed pairs with normalized correlations less than 0.2, 0.3 and 0.4. This procedure allows to separate strongest correlations which should have more influence. Resulting pair correlations were represented as a graph and post-processed using "quality bins" which are thresholds to indicate strongest correlations inside coevolving sectors of aminoacids. Eventually, the sectors were mapped into mentioned novel structure and analyzed using PyMol software as well.

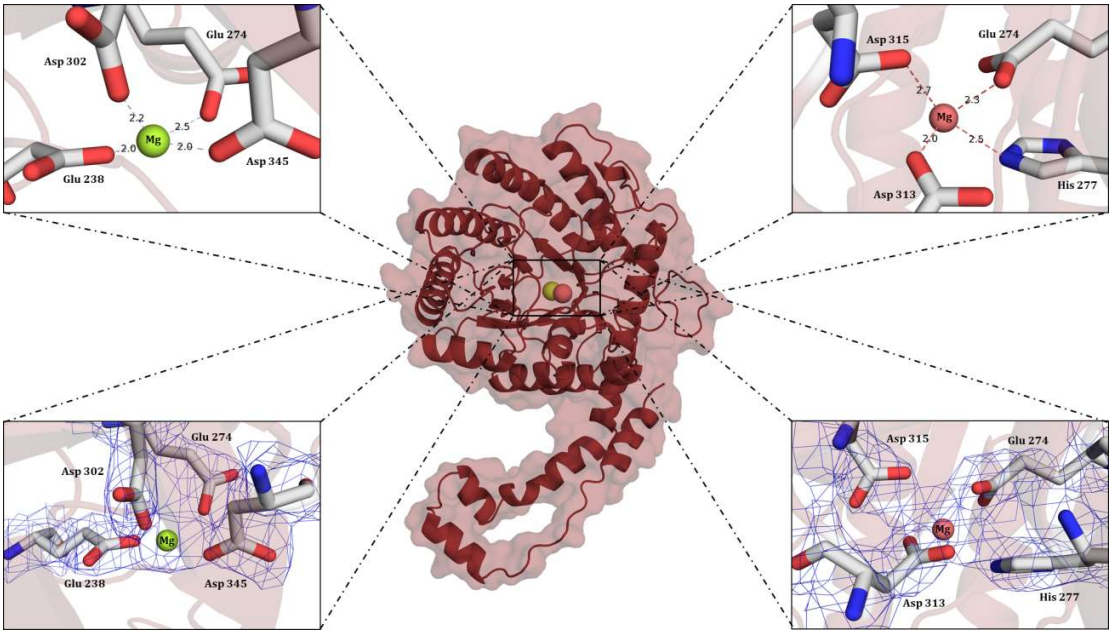

Figure 30. The structure of SdXI monomer with two metal sites. Upper panels indicate atomic distance and the residues involved in metal coordination. Bottom panels show electron density in blue mesh

The result of SCA calculations is shown in Figure 31 as SCA matrix. It represents pair coevolution in MSA and diagonal clusters of pairs show groups of aminoacids which are top eigenvectors obtained by independent component analisys (ICs) of the raw pair correlations (Rivoire et al. 2016).

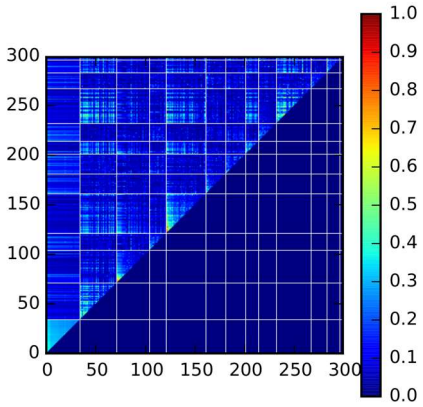

Figure 31. Normalized SCA matrix: diagonal squares indicate found ICs

During analysis, we have cut off a level of correlation by 0.4 of normalized values that helps to separate more strongest pair coevolutions which should help to understand better roles of aminoacid groups in the structure of Xylose Isomerase. Considering the SCA matrix generated and functional insights, besides interface residues, apparently IC 1 and 2 perform the same function, being the same sector (Sector 1). This sector is responsible for the great majority of interface residues and in a deeper analysis, also responsible for residues within the catalytic core (Figure 32).

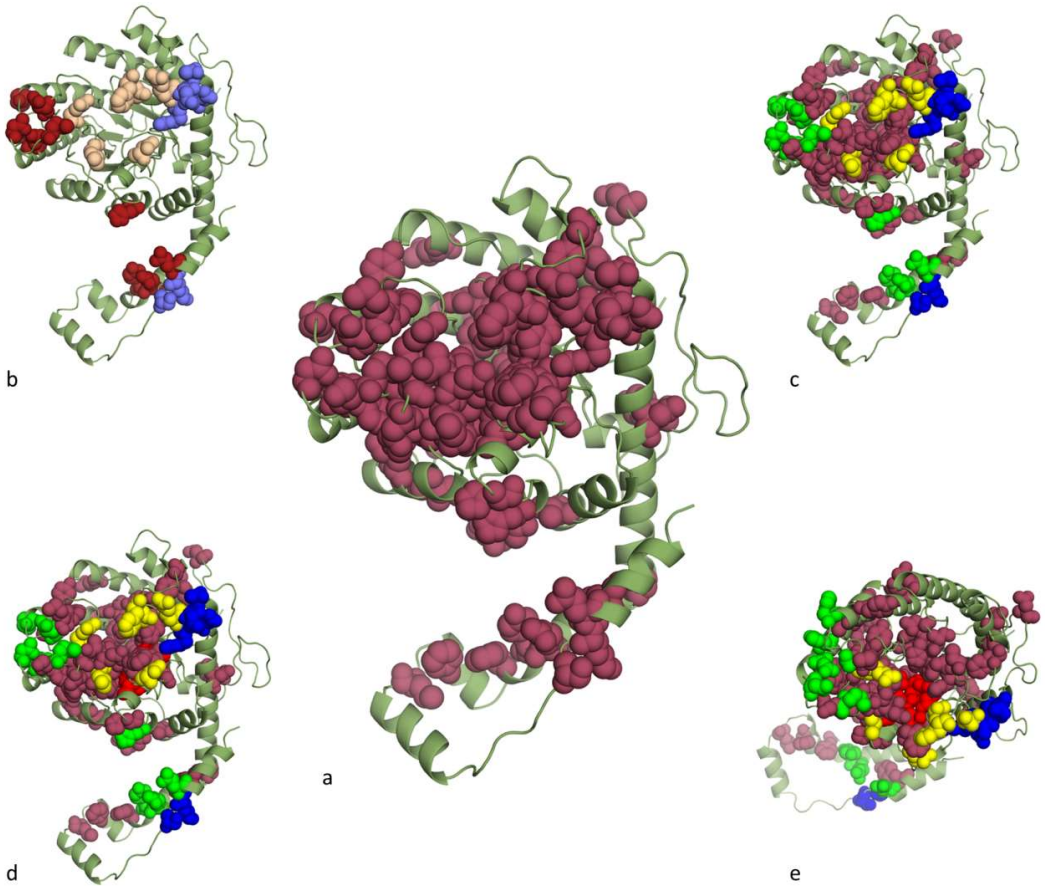

Figure 32. SCA analysis for SdXI. In the central figure (a), is represented the residues within Sector 1. (b): Interface residues located in Sector 1 – red for dimer interface, blue for tetramer interface and salmon for cross-link interface. (c): Sector 1 and interface residues highlighted (green for dimer interface, blue for tetramer and yellow for cross-link). (d): same as (c), but with the catalytic core in red. (e): upper vision of (d)

Similarly, when function is considered, ICs 3, 4 and 5 should be joined within a single sector. Also, we have observed some aminoacids inside IC2 and inside IC3 form a functional pattern in the structure. Therefore, significance of interference between IC2 from sector 1 and next IC3 was observed. Thus, for continuity purposes, IC 2 should be part of the sector 2 as well. This comparison can be observed in Figure 32. SCA analysis revealed a strong correlation between oligomeric structure and function, indicated by grouping residues within the same IC. This suggests that coevolution in XI classes occurred simultaneously for structure (tetramer formation) and function (sugar isomerization).

3.4.4. GH48 enzyme family

As the enzyme to study we have taken BlCel48 from *Bacillus licheniformis* obtained in ourt lab at IFSC/USP and applied the method of SCA to the family using BlCel48 structure as the reference. We have applied statistical coupling analysis to the set of BlCel48 homologous with similarity of 20% - 90%. As the result, we obtained three coevolving patterns of aminoacids, see Figure 33.

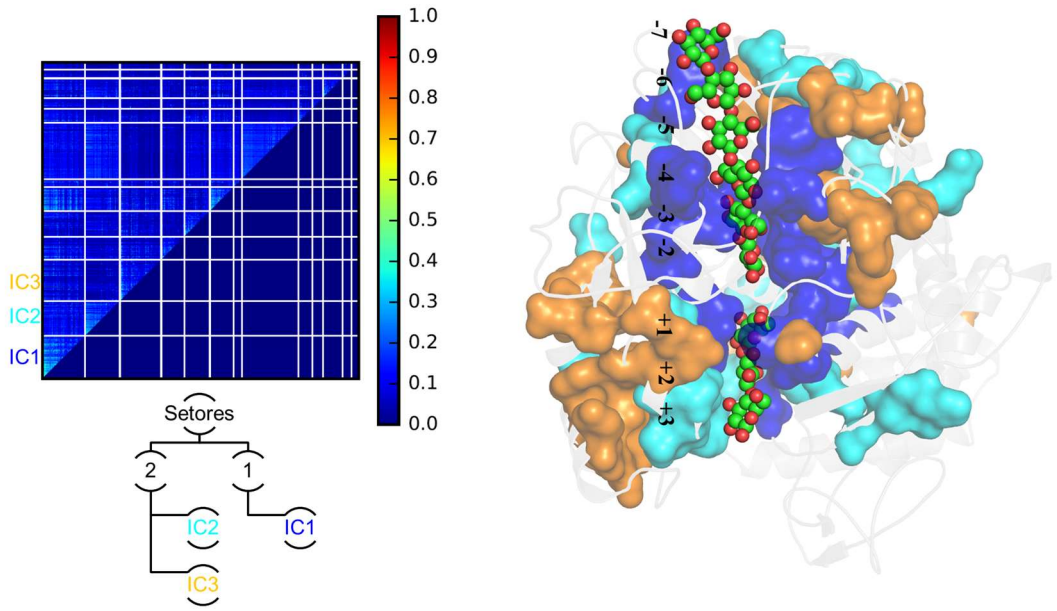

Figure 33. SCA results obtained on BlCel48 from *Bacillus licheniformis*: SAC matrix is on left side, sector projection on the structure are in the right side

We associated first three principal components obtained by spectral analysis of SCA correlations with structural patterns of aminoacid sets having functional load. Also, we have estimated inter-pattern interference as a ration between correlations inside thhe pattern and inter-pattern pairs and found them insignificant in case of first pattern (blue in the Figure 33) while second (yellow) and third (cyan) patterns are not. Thus, we have concluded the blue pattern are independent from others and having different role in enzyme functionality whereas yellow and cyan patterns have significant interference and should should be jointed in a single functional unit.

In Figure 33 shows all the patterns mapped ion the structure of BlCel48. The blue pattern is located manly beside to the substrate binding site. The blue sector reveals a single cluster spanning the tunnel binding sites (-7 to -1 subsites) to tunnel exit part of products subsites (+1 to +2) while the yellow sector surrounds the open cleft surface located at +3 subsite, which includes a few amino-acids residues in the loop Trp665-Glu673. This indicates the straightforward relationship between the co-evolving amino acid positions in blue and yellow sectors and the functional units of GH48 family.

The aminoacids, included in blue sector, are placed along the binding sites (+1 to +2 and -7 to -2 subsites) which allows to conclude the pattern plays significant role in enzymatic mechanism. The co-evolving residues in blue pattern are critical for the ligand binding while the aminoacids in yellow sector should be related to unproductive ligands binding.

Analyzing the blue sector, we found the residues involved in ligands interaction stand out as top-ranked. The top-ranked co-evolving residues on the catalytic tunnel part of BlCel48 include the Trp130 (-7 subsite), Gln220

(-6 subsite), Tyr269 (-5 subsite), Lys268/Tyr297/Gln178 (-4), Thr224/Gln178/Asn175 (-3 subsite), and Trp296/Tyr323 (-2). Similarly, on the open cleft subsites, the Trp418/Glu38 (+1 subsite) as well as Trp412/Asp535 (+2 subsite), are highly correlated. Interestingly, the residue Trp130 (-7 subsite), as a part of blue pattern, is essential for the substrate accessing on the catalytic site and processivity steps.

Furthermore, the residues from yellow co-evolving pattern are mainly localized around the open side at the tunnel exit. Thus, we may consider the role of yellow pattern is involved on product recognition at the positive subsites. The top-ranked co-evolving residues Asp595, His593 and Glu591 are located at subsite +3. The equivalent residues of Cel48F (PDB id is IFBW) are His593 (Arg544) and Glu591 (Glu542) are involved on hydrogen-bonding network with the cellotriose. These hydrogen networks are important to modulate the conformation of the residues at product expulsion subsites and may account for decreased activity of many GH48 enzymes.

3.5. Active site of Xylose Isomerase G5

In the previous section, statistical coupling analysis on Xilose Isomerase (XI) has been described. Here is continue of this work on XIs, aiming to map an active site of the experimental structure designmated as G5 and resolved in our laboratory at IFSC/USP using crystallographic method.

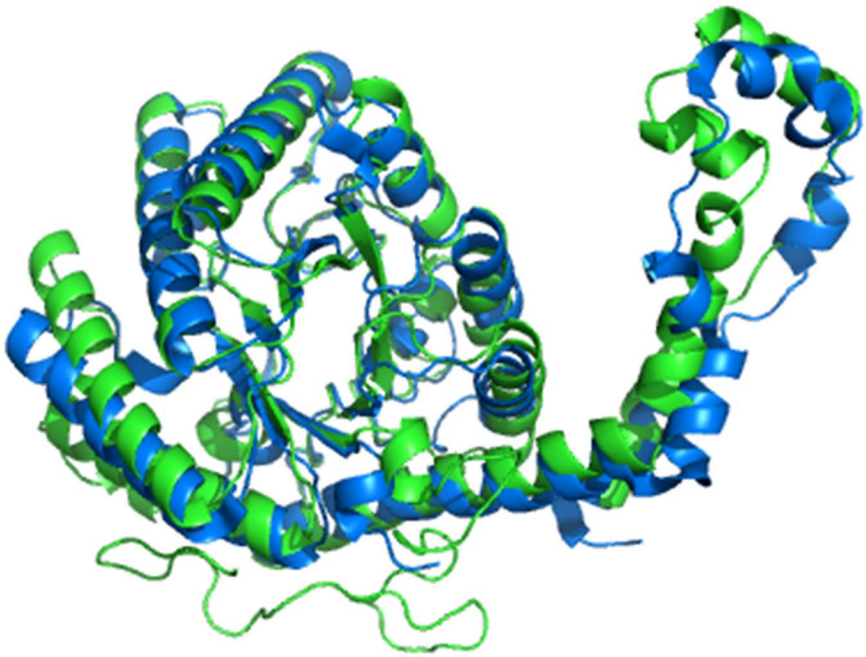

Figure 34. Chains A of the structures 1XYG (blue) and G5 (green)

To do this, we have developed a method based on finding of an active site in one of the homologous structures with a known ligand. Figure 34 shows two structures: experimentally determined structure and a homologue with PDB id 1XYG. Using Needleman-Wunsch algorithm (Needleman and Wunsch 1970) we aligned both sequences. After we know all the amino acids interacting with the ligand in a homologous structure and their pairwise alignment, we may project these positions in the sequence of homologous to the G5 sequence and obtain positions of considered active site in G5 structure obtained from experiments.

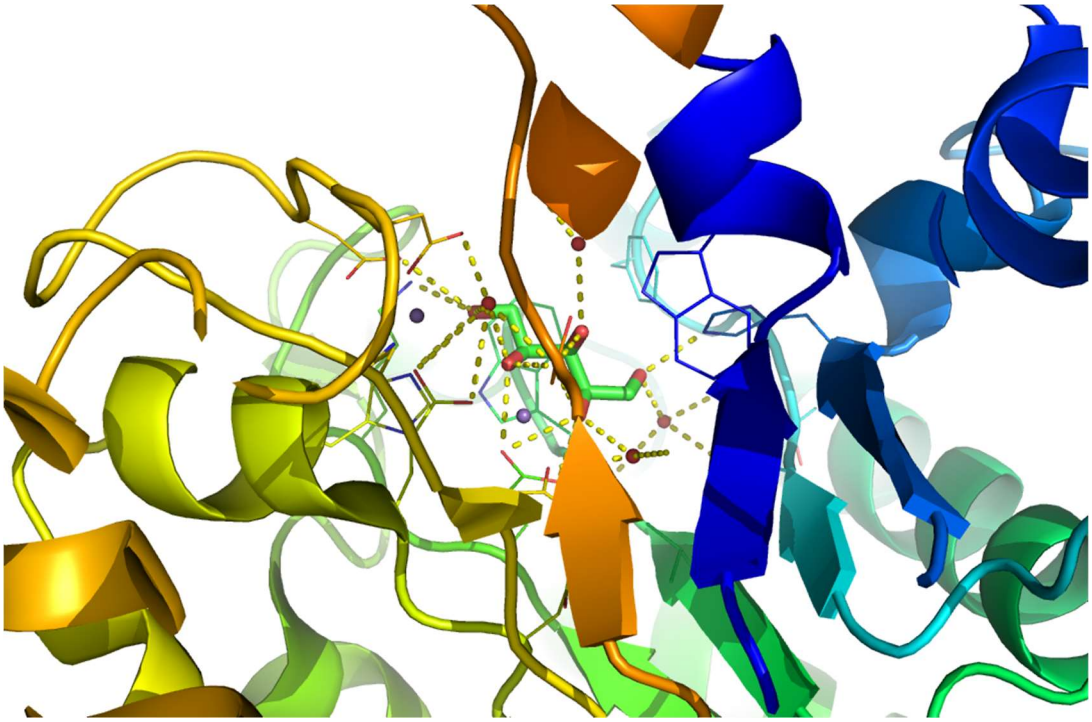

Figure 35. Active site of 1XYG with the ligand (dash lines)

```
Identities = 119/452 (26%), Positives = 224/452 (50%)
001  MIVLGDKEYYKGIGQIKFEGKESDNPLAFKYYNPDQIVAGKTMREHFKFAIAYWHTFCGQGSDP
001  ---------Y----Q----------P------TPED-------R--FTFGL--W-TVGWQGRDP

065  FGPGTQQFAWDASSDPYQAAKDKADAAFEFISKMGFDYFCFHDYDLIAEGATFAESEKRLAFIT
024  FGDATRR-A--L--DPVESVQ-R--LA-E-LGAHG---VTFHDDDLIPFG-S-SDSE-REEHVK

129  DYLKQKKAESGIKLLWGTSNCFSNPRFMNGAATNPDFNVVARAGGQVKLALDATIALGGENYVF
072  RF-RQALDDTGMKVPMATTNLFTHPVFKDGGFTANDRDVRRYALRKTIRNIDLAVELGAETYVA

193  WGGREGYMSLLNTDMGRELDHMAQFLAMSRDYARAQGFKGTFFIEPKPMEPSKHQYDFDS-ATA
135  WGGREGAESGGAKDVRDALDRMKEAFDLLGEYVTSQGYDIRFAIEPKPNEP-RGDILLPTVGHA

256  IGFLKNYGLDKDFKINIEVNHATLAQHTFQHELEVAAKAGMLGSIDANRGDYQNGWDTDQ-FPN
198  LAFIERLERPELYGVNPEVGEQMAGLNFPHGIAQALWAGKLFHIDLN-G--QNGIKYDQDLRF

319  NIQETTEAM-LV-FLKAGGLQGGGVNFDAKIRRNSTDLEDVFLAHIGGADTFARALLTADKIIT
259  GAGDLRAAFWLVDLLESAG-YSGPRHFDFKPPR-TEDFDGVWASAAGCMRNYL-ILKERAAAFR

381  SSP--YEKLRKERYSSF--DS-GKG-KDFADGKLSLKDL-YTIAHENGELNLQSGKQELFENII
320  ADPEVQEALRASRLDELARPTAADGLQALLDDRSAFEEFDVDAAAARG-MAFERLDQLAMDHLL

438  NQYI
383  GARG
```

Figure 36. Pairwise alignment of the sequences of 1XYG and experimentally determined structure of G5. Selection shows translation of active site positions from 1XYG to G5

Using PyMOL software (DeLano 2002), 12 residues have been manually selected as an active site in the structure of 1XYG: 52, 88, 89, 135, 179, 181, 215, 218, 243, 253, 255, 285. In Figure 35, active site with the bound ligand is shown.

As a next step we translated selected positions of 1XYG to the experimental structure using pair alignment of the sequences. Obtained pair alignment is shown in Figure 36 where selected positions are marked by rectangles. Using this translation, we calculated position numbers for considered active site in our experimental structure and visualized it using PyMOL software.

As a result, obtained active site of Xylose Isomerase G5 is shown in Figure 37.

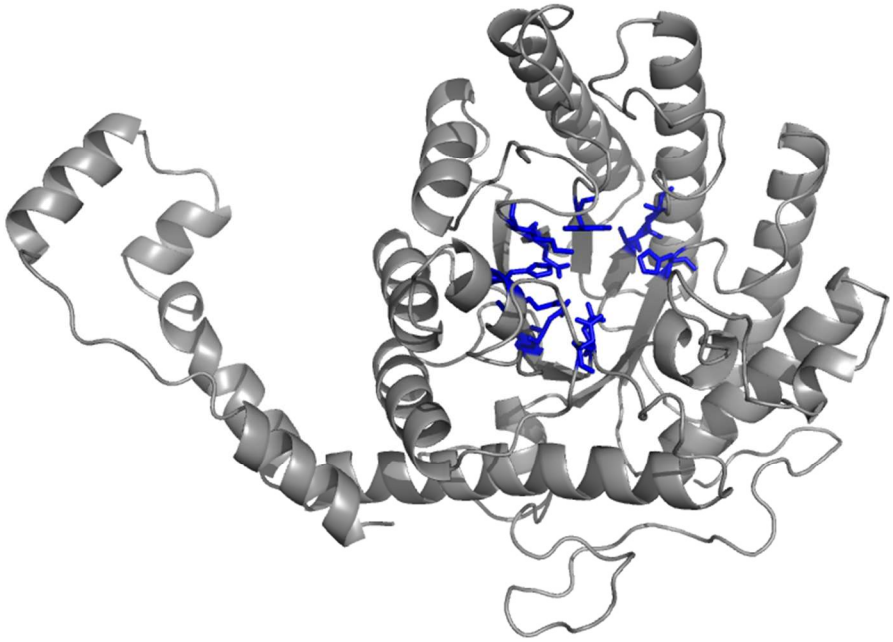

Figure 37. Xilose Isomerase G5 chain structure with found active site highlited in blue

3.6. Aminoacid fuzzy communities in proteins

In this section we describe the newly developed method of coupling communities of aminoacids which analyzes SCA pair correlations using fuzzy communities techniques (Newman 2006; Zhang et al. 2007). Thus, the new method can indicate collective coevolving groups of aminoacids which can be important for functional and structure features of proteins.

This analisys has been inspired by method of uncovering fuzzy communities (Zhang et al. 2007), and we have applied it to correlation matrix $\tilde{C}_{ij}$ obtained using SCA. For this, we define a graph $G(E,V)$ from the matrix:

- Weighted edges formed by elements of $\tilde{C}_{ij}$;

- Vertices represent positions of aminoacids and noted number-letter form, such a "35V", where number is the position of specific aminoacid "V".

The algortithm is based on non-negative factorization (NMF) of feature matrix which can be obtained by normalizing the kernel of the laplacian of the graph. The autors of (Zhang et al. 2007) define the kernel as:

$$K \equiv \exp \beta L = \lim_{n\to\infty} \left(1 + \frac{\beta L}{n}\right)^n$$

where β is coefficient of diffusion of the communities, L - laplacian of the graph with elements

$$L_{ij} = 1 : i \frown j, -d_i : i = j, 0 : otherwise$$

Symbol " $\frown$" represents a connection between vertices i and j.

Thus, feature data from the graph is repesented in matrix B:

$$B_{ij} = \frac{K_{ij}}{\sqrt{K_{ii} K_{jj}}}$$

As it can be seen from definition, B is a symmetric matrix where each column (or row) corresponds to attributes of the network. Using approximate factorization $B \simeq W \cdot H$ we can obtain matrix W (or H as equivalent) of rank $k \times n$ defines clustering partition. In our work we used NMF implementation from package Nimfa for Python 2.7.

Thus, matrix W cluster n vertices of the graph into k clusters. Number of clusters can be defined by maximum of modularity function (Newman 2006). To find the value, we repeated modularity calculations for diffirent k-values: k-value corresponding to maximum of modularity was chosen as a count of clusters. As postprocessing, resulting clusters can be projected on molecular structure using Pymol software (DeLano 2002).

3.7. Aminoacid fuzzy communities in GH48 from B. licheniformis

Several fungal and bacterial organisms secrete a synergistic multicomponent of enzymes system which including endoglucanases (EG), cellobiohydrolases (CBHs), β-D-glucosidases, and the lytic polysaccharide monooxygenases (Vaaje-Kolstad et al. 2010; Paine et al. 2015). Among cellulases, extensive studies have been conducted focusing on the structure-function relationship of processive cellulases (Horn et al. 2012) which can convert cellulose directly into a cellobiose molecule during hydrolytic pathways (Paine et al. 2015), specially CBHs belonging to glycoside hydrolase families 6, 7, and 48 (T. Teeri 1997). Most of Glycoside Hydrolase Family 48 (GH48) enzymes are produced by bacterial, fungal, and insect organisms and show their hydrolytic activity mainly on cellulose and chitin substrates (Sukharnikov et al. 2012; Nguyen et al. 2018) and have played an important role in the breakdown of plant cell walls (Liao et al. 2011). The 3D X-ray structure of GH48 enzymes reveal an (αα)6-helix barrel topology with a tunnel-shape and cleft regions covering the productive and unproductive binding sites (Parsiegla et al. 1998). Recently, the biochemical characterization of Blcel48 from *B. licheniformis* was reported (de Araújo et al. 2018). The structure of the enzyme was obtained in our lab at IFSC/USP and it is not published yet.

However, we have chosen the structure to apply our method for the analysis. Figure 38 shows experimentally obtained structure of Blcel48 from *B. licheniformis*.

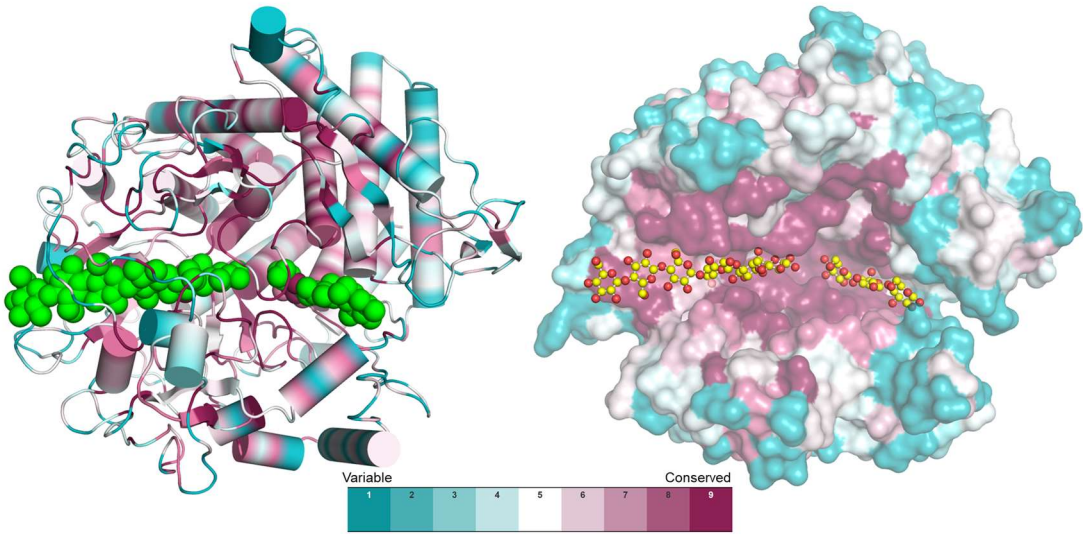

Figure 38. 3D structure of GH48 enzyme Blcel48 from *B. licheniformis* with substrat molecule (green). The color scale show dummy conservation degree of aminoacids

To make analysis, we have downloaded set of sequences for family GH48 from Pfam database (Finn et al. 2016). The set contents 184 full sequences, we have added our Blcel48 from *B. licheniformis* and made alignment using MATLAB software. We characterized the set by identity and similarity to make sure that aminoacid in the positions are not over conserved or very variable, see Figure 39.

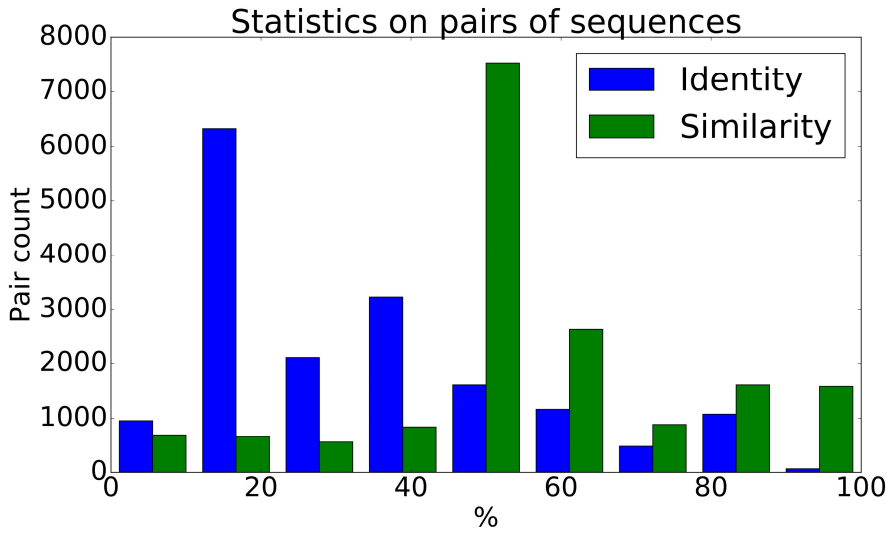

Figure 39. GH48 set of sequnces characterization by identity and similarity

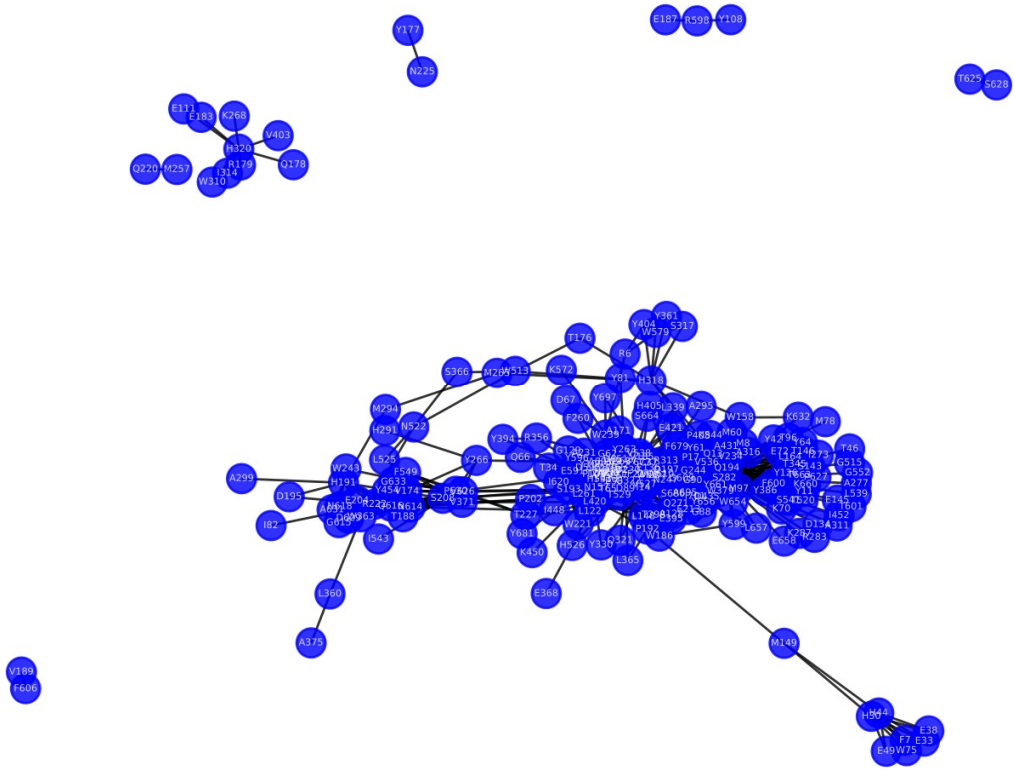

Figure 40. The graph representing correlations above 30% of normalized values of correlation SCA

From Figure 39 we can see that the set does not have big amount of very identical pairs of sequences (over 90%). It is a good signal meaning a conservation will not interfere with coevolving patterns.

After building the graph from normalized correlation matrix, we have made a cutoff on 30% value to separate strongest correlation and reduce noise impact. Resulting graph of 217 aminoacids is shown on Figure 40 and Figure 41 shows histogram of degrees of nodes. As it can be seen, the graph is sufficiently sparse: it has many nodes of low degree.

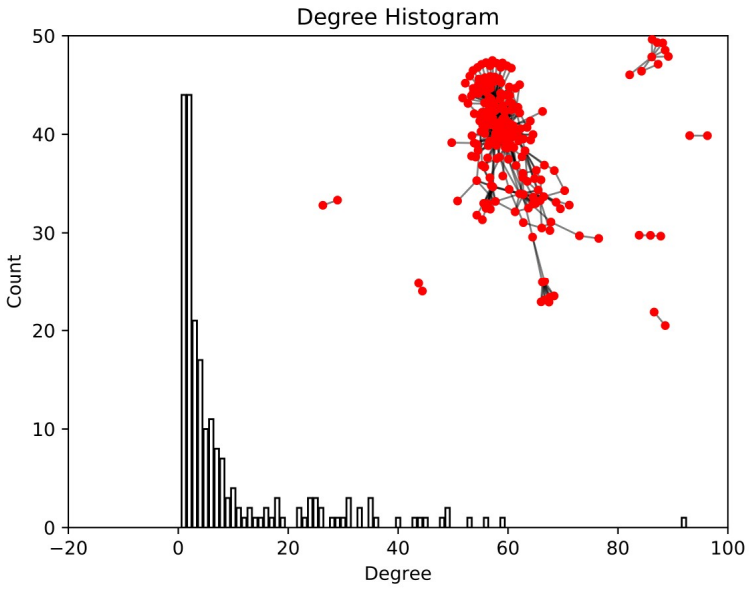

Figure 41. Histogram degree for the graph of GH48 family

Since we must find coevolving groups of aminoacids, we have no interest in nodes with very low degree, thus we eliminate the graph from nodes with $d \leq 3$. In result we obtained a graph of 99 nodes shown in Fig. \ref{fig:rm}. Also, degree histogram reflects the changes in Figure 42.

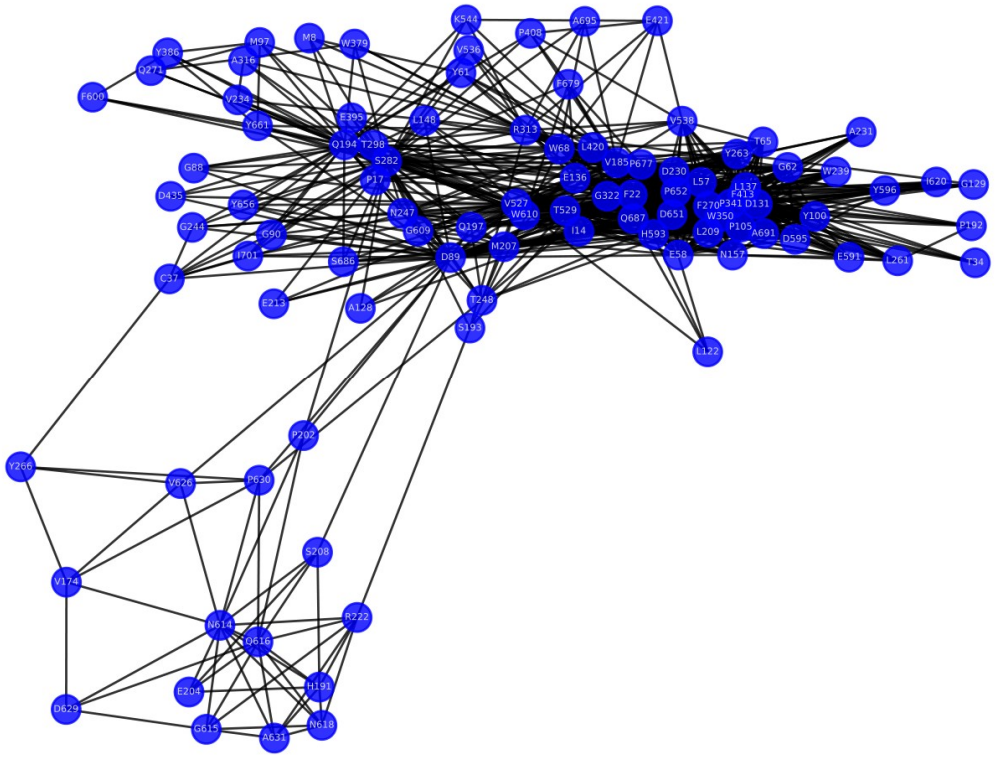

Figure 42. The graph of GH48 family after removing low degree nodes

Using a strategy of cluster number searching described in previous section, we calculated modularity for $k = 1 \ldots 20$ and selected k value corresponding with maximal modularity. Figure 43 shows modularity values for different number of clusters.

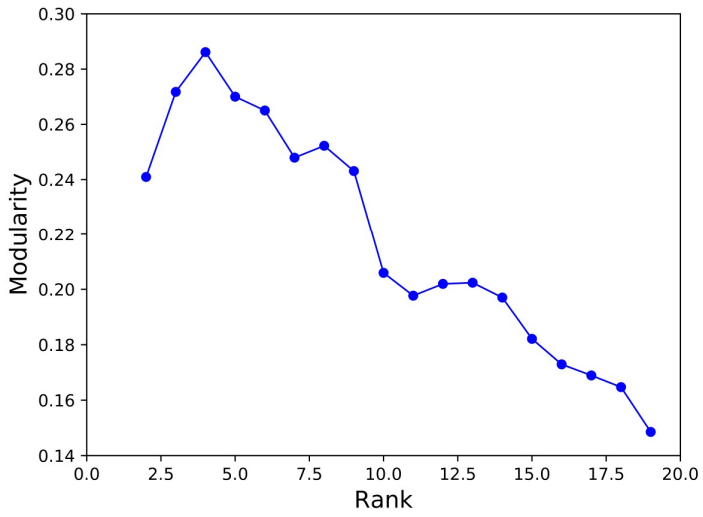

Figure 43. Modularity values depending to rank of W matrix (number of communities)

Thus, we have found four fuzzy communities in our network representing coevolving correlations for GH48 family of proteins. These communities (clusters) contain 40, 24, 20 and 15 aminoacids. Note, many nodes do not have strong membership for their cluster, thus the discovered communities are very diffuse. This result is shown in Figure 44.

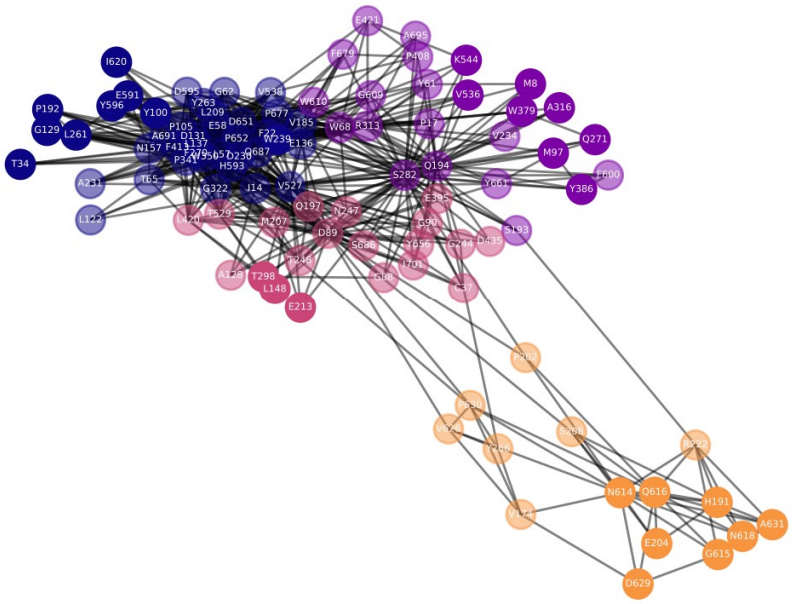

Figure 44. Fuzzy communities of the network. Clusters are selected by colors; more transparent nodes represent more diffusing aminoacids

Figure 45 shows four clusters projected on the structure of Blcel48 from *B. licheniformis*. Interesting that all the clusters are cover the region near active center. It confirms that the method has ability to discover important motifs inside a structure. Besides that, the clusters show importance of backside of molecule: it indicates non-obvious importance of the region in forming of active center.

Thus, the results of our method applied on GH48 family allowed to find fuzzy communities which can be confirmed by our knowledge about the family and, also, indicate an importance of backside part which is not obvious. The results can be completed by statistical coupling analysis of found clusters to understand role of each cluster and discover an importance of backside region for biological function and structure forming.

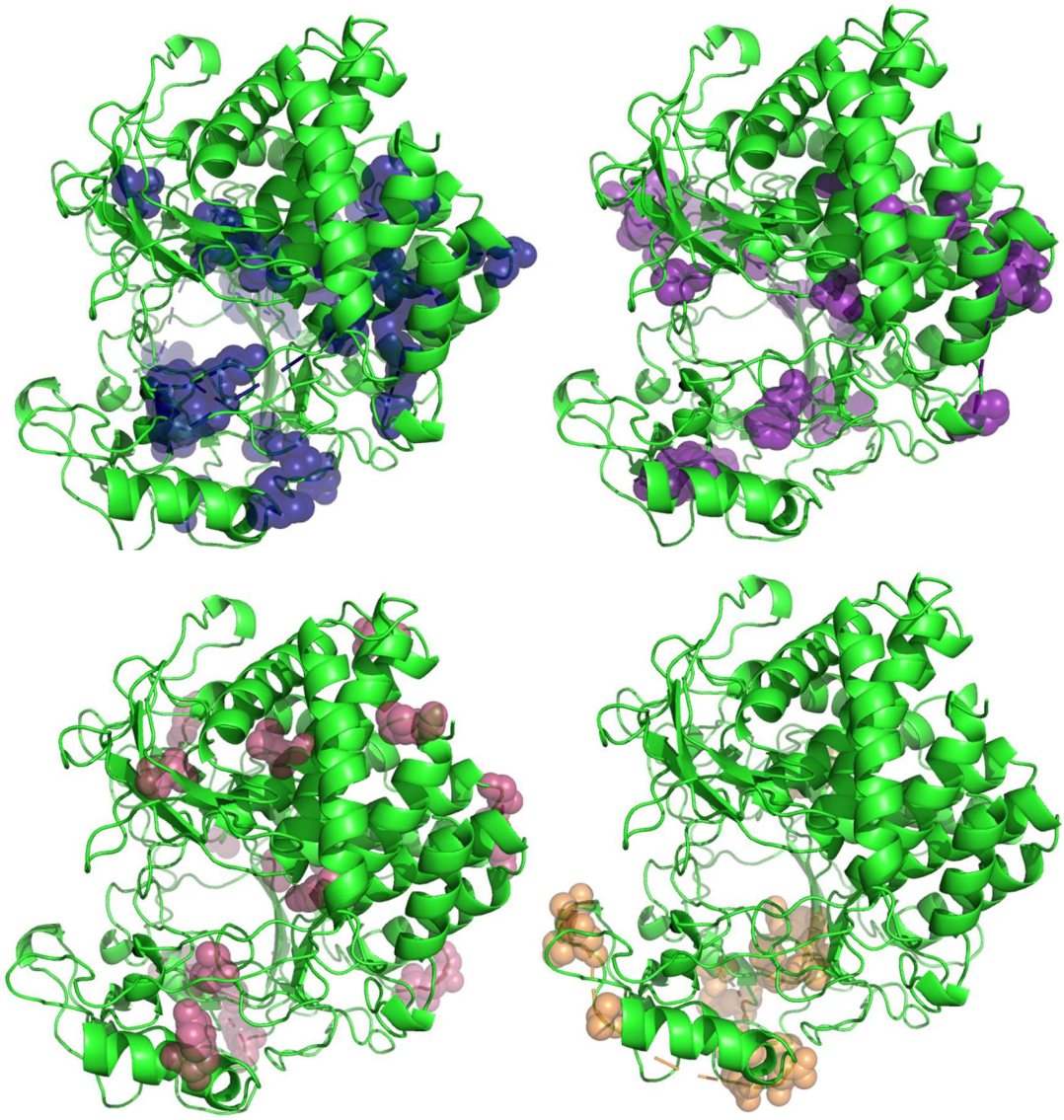

Figure 45. Four coevolving communities projected on the structure of BlCel48 from *B. licheniformis*

3.8. Bioinformatic separation of GH7 endo- and exogluconases

Glycoside hydrolase family GH7 contains proteins with two types of activity: endoglucanase (EC 3.2.1.4) and exoglucanases (cellobiohydrolases, EC 3.2.1.91). All the GH7 entries from PFAM using UNIPROT databank (http://pfam.xfam.org/family/PF00840) have been separated. This data set was intended to be used for SCA analysis, but the results can be distorted by the diferences between endo- and exogluconases if they will be processed in the same bunch. Therefore, it was necessary to separate the sequence alighnment in two sets, one with exoglucanases and another one with endoglucanases only. To do this, a method based on structural features detection was developed.

Firstly, 6413 mixed sequences of the same GH7 family was downloaded from PFAM website. Since the difference between two types of enzymes in the set is in a sequence patterns that represent functional loops in the structure (exoglucaneses) or their absence (endoglucanases), we can make pairwise alignments of preliminary known sequence of exoglucanase and the studied one. If we will observe characteristic continuous pattern of gapes in the

studied sequence, we can consider that this is absence of functional loops in the structure and, thus, the studied sequence has a structure of endoglucanase. Based on this principle, we can formalize the constrains for the method of separation:

- Endo- has sector(s) of gapes in pair alignment with exo-. It represents absence of functional loops in the structure.
- The sector should be internal.
- The sector should not be too small (\textgreater 10 positions) but not too large (\textless 25 positions). These criterion values are empirical by analysis of 3D structure justifyed exoglucanase. Thiese values should be justifyed (or improved) by statistics on manual analysis as large as possible set of justifyed exoglucanases. For now, we have used from 10 to 25.
- Thus, if investigated sequence have at least 1, but better 2 or 3, gap sectors in alignment with justifyed exoglucanases, we can consider it as an endoglucanase.
- To check the decision, we should map the gap sectors to subsequences of amino acids in reference exoglucanase 3D structure in PyMOL \citep{delano2002pymol}. If it is correct, we should see selected funcional loops on the structure.

To implement the method, we have used the sequence with PDB 1CEL as a confirmed exoglucanase available at http://www.rcsb.org/pdb/explore.do?structureId=1CEL and a list of studied sequences from GH7 as input data.

The developed program can be running in two modes:

a) to study unknown sequences and separate endo-s of the list;

b) to characterize well-known sequences. The option must be used to justify the list contains only endoglucanases.

Thus, the algorithm includes following steps:

1. Pair alignment: taking the input list, the program aligns each sequence from the list to the reference. After, it saves output aligned pairs "reference - being studied" where is easy to see the features of the alignment;
2. Pair analysis and filtering: in analysis of aligned sequence, firstly, we take away ending gaps (not internal) because these gapes are features relating to a specific alignment. After, the program separates positions of continous internal gap sectors but only those that match continuos subsequences of amino acids in the reference. Thus, each continuos gap sector can match a functional loop in the reference with high probability.
3. Save subset of endoglucanases (mode 1) or make statistics on sectors over the set (mode 2). In the first option, the program writes a file with suggested endo- sequences. It can filter them by number of found gap sectors. For example, the program can save only sequences with number of sectors more than two: it helps to make a set cleaner from artifacts of alignment method. Second option is to characterize a quality of the studied set. If the set consists only endodoglucanases, the program confirms it clearly or, if the set has exoglucanases in significant quantities, it also demonstrates this. Statistical quality parameters of sequence set are:
    i. Number of gap sectors for each sequence;
    ii. Average sector length for each sequence;

iii. Sum of positions over all the sectors for each sequence;
iv. Sector weight, it shows degree of conservation of the positions inside a sector over all the set. High conservation of subsequence inside reference sequence above the gap sector in pair alignment significantly increases probability that the program found a functional loop in the structure correctly and, accordingly, a chance for investigated sequence to be chosen as endoglucanase. This meaning is based on obvious consideration of absence a functional "exo-"-loop in structure of endoglucanse.

It should be best practice to combine of both options in analysis: to separate list of endoglucanases and after that to verify obtained set.

We have used the following methodology of using of developed package, tests on the known sets, and verification of the result:

1. As a reference we have used reference exogluconase (PDB id: 1CEL);
2. We have manually separated already classified full endo-glucanases and exo-glucanases sequences of GH7 family from CAZy database. There are 19 classified endoglucanases and 45 classified exoglucanases. These sets should be used for verification of the written software.
3. For verification, we have run the program with following different initial conditions:
    a. "exo-endo": 1CEL as a reference vs. 19 classified endoglucanases;
    b. "exo-exo": 1CEL as a reference vs. 45 classified exoglucanases;
    c. "endo-exo": In this case, we need an endoglucanase as a reference. Considered endoglucanase is the sequence with pdb id 1EG1 (http://www.rcsb.org/pdb/explore.do?structureId=1EG1). Thus, the program was running with 1EG1 vs. 45 classified exoglucanases;
    d. "endo-endo": 1EG1 as a reference vs. 19 classified endoglucanases.

    Results of each combination of options lets us to see how well the program can identify the extreme cases on well-known sets.
4. Next run is 1CEL vs. full set of the sequnces of our interest separated from PFAM. In this run we separate a several sets of supposed endoglucanases based on filtered numbers of obtained sectors in each sequence. For example, the program produced a file with a list of sequences in FASTA format which contain at least two gap sectors.
5. Obtained sets should be checked by processing of the set through the program using **mode 2** described above.

The test "exo-endo", shows that most of the sequences have more than 1 heavy-weighted sectors, and at least one sector with weight is very close to 1.0 (Figure 46).

To verify manually, we can copy sectors from generated automatically pml-script to PyMol package with opened pdb file contgaining the reference 1CEL, as it shown in Figure 47. In the figure, we can see 2 found sectors from sequence #0 (first line in section "sector weights" of output file shown in Figure 46). These sectors demonstrate loops in 1CEL which are conserved well through all the set with weights 0.61 and 0.91, as can be seen in Figure 46.

It is interesting to compare results from list of known endoglucanases with list of exoglucanases. Results of this test is shown in Figure 48. In the figure, we can see no sectors, or, theoretically, we could see sectors with very low (less than 0.3 ... 0.4) weight. It is what is expected for searching endoglucanases in a list of exoglucanases.

pairs3.pl | stat-xe.txt

```
sec num for each seq:
2 2 4 2 2 4 2 4 2 4 2 3 4 4 6 6 6 6 5

avg sector len for each seq:
13.00 13.00 13.25 13.00 13.00 13.25 13.00 13.25 13.00 13.25 13.00 14

sum of sectors:
26 26 53 26 26 53 26 53 26 53 26 44 53 53 88 92 87 92 76

sector weights:
0.61,0.91,
0.57,0.91,
0.63,0.54,0.40,0.76,
0.61,0.93,
0.61,0.93,
0.63,0.54,0.40,0.93,
0.61,0.93,
0.63,0.54,0.40,0.93,
0.61,0.91,
0.63,0.54,0.40,0.93,
0.57,0.93,
0.63,0.53,0.93,
0.63,0.54,0.40,0.76,
0.63,0.54,0.40,0.76,
0.14,0.09,0.58,0.52,0.05,0.91,
0.21,0.18,0.58,0.48,0.32,0.91,
0.21,0.18,0.58,0.35,0.32,0.91,
0.21,0.18,0.58,0.48,0.32,0.91,
0.05,0.22,0.56,0.56,0.93,
```

Figure 46. Result for "exo-endo" test; weights of gap sectors are highlighted

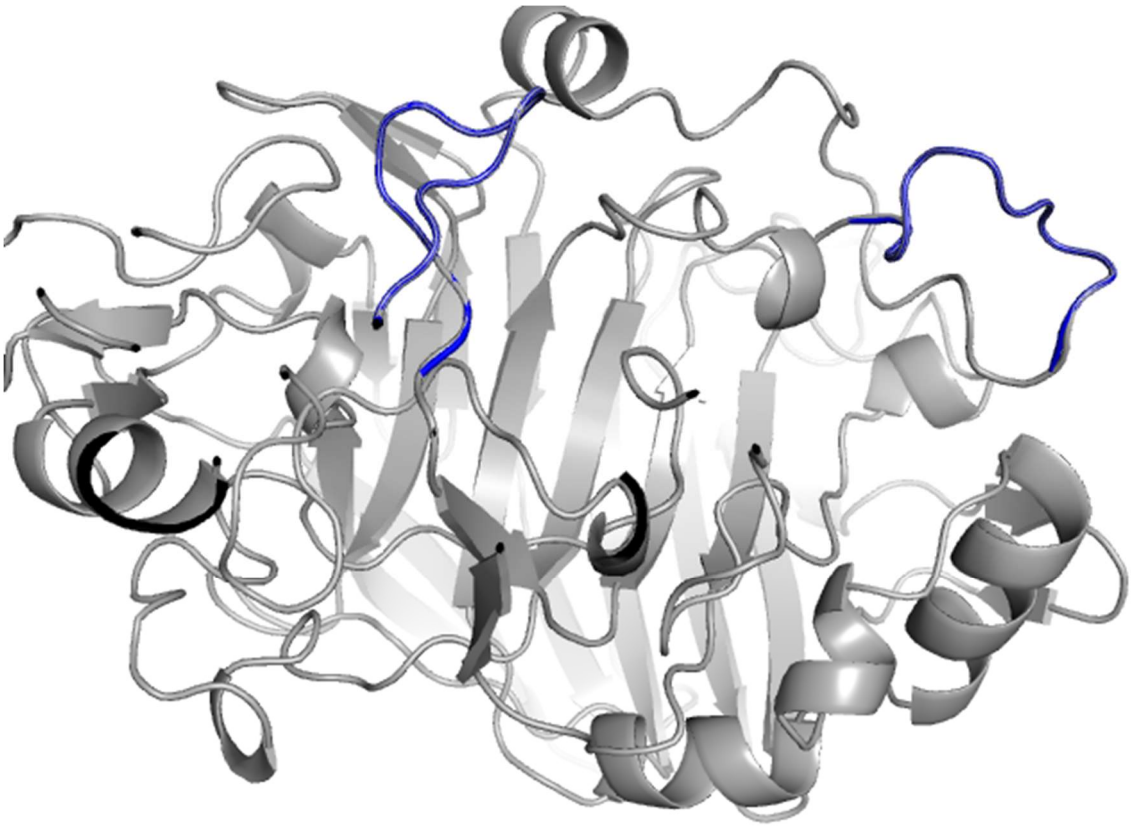

Figure 47. Structure of 1CEL with automatically selected regions of gaps in sequence #0

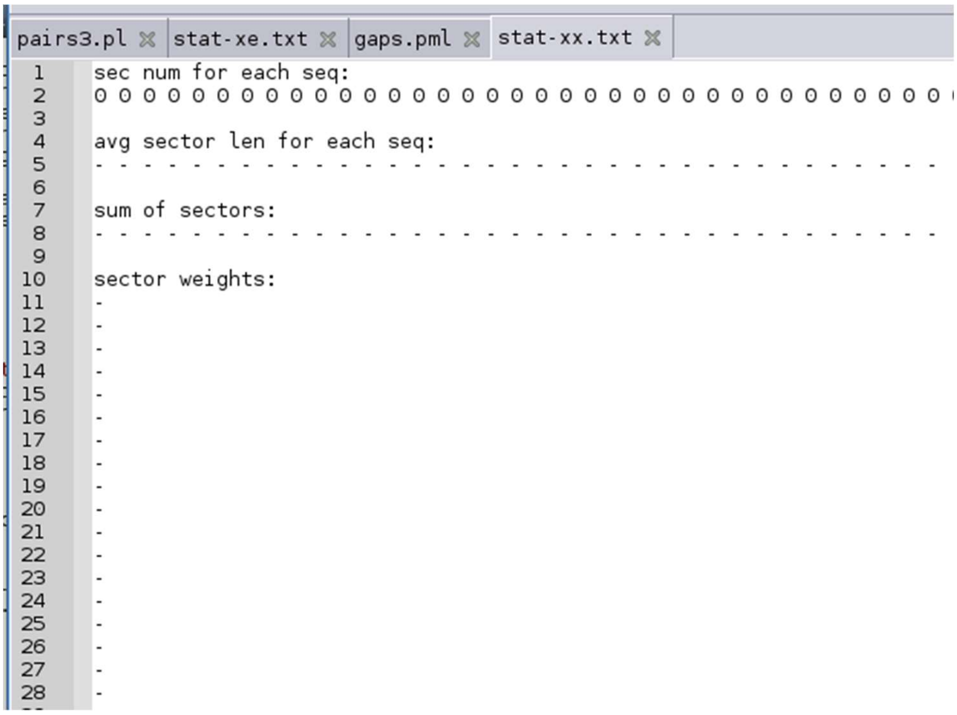

Figure 48. Result for "exo-exo" test; no gap sectors observed

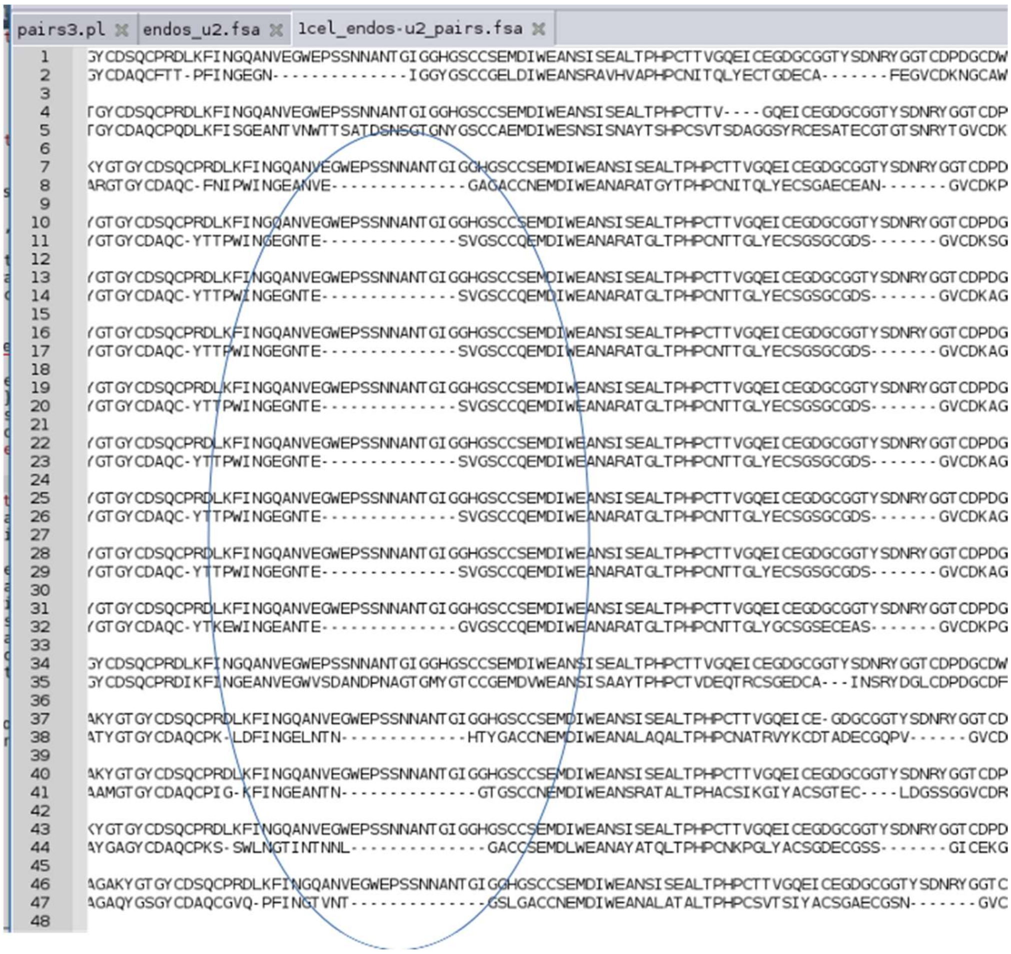

Figure 49. Pair alignments of each sequence from resulting list of endoglucanases with the reference 1CEL

Figure 46 and Figure 48 demonstrate that developed method works well on test sets of known sequences and can separate 100% of endogluconases.

After verification, we have processed the set of interest which contains 6413 sequences. Thus, we obtained three output sets with different filter on quantity of sectors inside a sequence. These three options were: to separate the sequences containing >1, >2 or >3 sectors.

Obtained sets were processed by the program as unknown lists to confirm (or refute) high content of endogluconases statistically. It is suitable to look on pair alignments of every sequence of the set with the reference for each case. We found gap intervals in studied sequences, as it shown in Figure 49. In the figure, an example of well conservated (that means with high weight) sectors of gaps are selected from the set containing >2 sectors.

We considered the set containing >2 sectors (293 sequences) is more believable than the set containing >1 sectors (511 sequences) because of high concentration of sectors having greater weight. The set containing >3 sectors is even better in terms of weights but contains very small amount of sequences (89).

Thus, we suppose the set containing >2 sectors is more realistic from statistical point of view. Developed method and its implementation show perfect results in test on well-known sets of sequences and allowed us to divide a large set of GH7 sequences into two bins: exo- and endoglucanases.

3.9. Carbohydrate-active enzymes detection

The metagenomics was obtained in our biomolecular Lab at IFSC/USP. This data represents hundreds of thousands of sequences recorded in a database. Thus, it was necessary to develop a method of detecting of carbohydrate-active enzymes. At first, a new database "cazybank" was created. An idea of the database is maintenance of all known carbohydrate-active sequences classified by their families. To do this, we have used free available data from database CAZY https://www.cazy.org: all the genes from all known carbohydrate-active families was downloaded. Further, all the data from GenBank database was downloaded as well. Since the GenBank data contain information about sequences for each known gene, it was possible to combine data from these two databases to obtain tables of sequences tied to Cazy families.

As the next step, a set of sequences for each Cazy family was selected and aligned using progressive method described in the Methods section. For each obtained MSA, we have made hidden Markov (HMM) profilies. Since HMM describes a consolidation of set of sequences, thus HMM built from MSA of sequences from carbohydrate-active family must describe their family. To check the quality of obtained HMMs, the software **hmmsearch** (Eddy 2011) was used. Each sequence tied to known family was compared with each HMM and resulting scores as a metric of similarity were fixed. As an expected result, score of a sequence from a family must be much greater value than score of comparison with another family. Thus, this checking allows to rate the quality of HMM and all the built HMM passed the test with 100% accuracy. Second test was held on known sequences contained in metagenomic data obtained in our lab at IFSC/USP. There are 956 sequences which was compared with all built HMMs and results were compared with already known inormation for their families:75% of sequences were identified correctly. Thus, we can evaluate this value as a probability that unknown sequence belongs to found family. In further work this accuracy of the method can be increased by manual improvement of MSA quality for each family. The developed method was implemented as web application available at http://saxs.ifsc.usp.br:7000. Figure 50 shows interface of the application with loaded example sequence and next, in Figure 51, the result is

shown. The result is a table with ranking of suggested families in descending order of probability that the sequence belongs to the family.

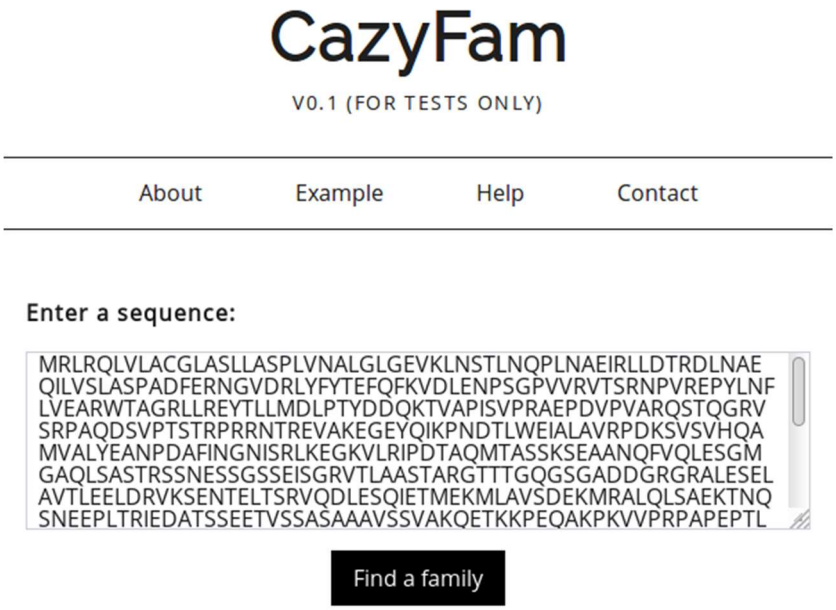

Figure 50. CazyFam interface with an example sequence

Rank	Suggested families
1	CBM50
2	AA11

Figure 51. CazyFam result representing most probable families for specified sequence

REFERENCES

Alonso DM, Hakim SH, Zhou S, et al (2017) Increasing the revenue from lignocellulosic biomass: Maximizing feedstock utilization. Sci Adv 3:e1603301. doi: 10.1126/sciadv.1603301

Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Boisset C, Fraschini C, Schülein M, et al (2000) Imaging the enzymatic digestion of bacterial cellulose ribbons reveals the endo character of the cellobiohydrolase Cel6A from Humicola insolens and its mode of synergy with cellobiohydrolase Cel7A. Appl Environ Microbiol 66:1444–1452. doi: 10.1128/AEM.66.4.1444-1452.2000

Brognaro H, Almeida VM, de Araujo EA, et al (2016) Biochemical Characterization and Low-Resolution SAXS Molecular Envelope of GH1 β-Glycosidase from Saccharophagus degradans. Mol Biotechnol 58:777–788. doi: 10.1007/s12033-016-9977-3

Bron C, Kerbosch J (1973) Algorithm 457: finding all cliques of an undirected graph. Commun ACM 16:575–577. doi: 10.1145/362342.362367

Brookes E, Vachette P, Rocco M, Pérez J (2016) US-SOMO HPLC-SAXS module: Dealing with capillary fouling and extraction of pure component patterns from poorly resolved SEC-SAXS data. J Appl Crystallogr 49:1827–1841. doi: 10.1107/S1600576716011201

Brunecky R, Alahuhta M, Xu Q, et al (2013) Revealing nature's cellulase diversity: The digestion mechanism of Caldicellulosiruptor bescii CelA. Science (80-) 342:1513–1516. doi: 10.1126/science.1244273

Cantarel BL, Coutinho PM, Rancurel C, et al (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37:D233--D238

Carter L, Kim SJ, Schneidman-Duhovny D, et al (2015) Prion Protein—Antibody Complexes Characterized by Chromatography-Coupled Small-Angle X-Ray Scattering. Biophys J 109:793–805

Chang A, Abderemane-Ali F, Hura GL, et al (2018) A Calmodulin C-Lobe Ca2+-Dependent Switch Governs Kv7 Channel Function. Neuron 97:836–852.e6. doi: 10.1016/j.neuron.2018.01.035

Chauve M, Mathis H, Huc D, et al (2010) Comparative kinetic analysis of two fungal β-glucosidases. Biotechnol Biofuels 3:3. doi: 10.1186/1754-6834-3-3

Colussi F, Garcia W, Rosseto FR, et al (2012) Effect of pH and temperature on the global compactness, structure, and activity of cellobiohydrolase Cel7A from Trichoderma harzianum. Eur Biophys J 41:89–98. doi: 10.1007/s00249-011-0762-8

Colussi F, Serpa V, da Silva Delabona P, et al (2011) Purification, and biochemical and biophysical characterization of cellobiohydrolase I from trichoderma harzianum IOC 3844. J Microbiol Biotechnol 21:808–817. doi: 10.4014/jmb.1010.10037

Cosgrove DJ (2014) Re-constructing our models of cellulose and primary cell wall assembly. Curr Opin Plant Biol 22:122–131. doi: 10.1016/j.pbi.2014.11.001

Davies G, Henrissat B (1995) ScienceDirect.com - Structure - Structures and mechanisms of glycosyl hydrolases. Structure 3:853–859

Dayhoff M, Schwartz R, Orcutt BC (1979) Matrices for detecting distant relationship. Atlas Protein Seq 353–358

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A Model of Evolutionary Change in Proteins. In: In Atlas of protein sequence and structure. National Biomedical Research Foundation Silver Spring, MD, pp 345–352

de Araújo EA, Manzine LR, Piiadov V, et al (2018) Biochemical characterization, low-resolution SAXS structure and an enzymatic cleavage pattern of BlCel48 from Bacillus licheniformis. Int J Biol Macromol 111:302–310. doi: 10.1016/j.ijbiomac.2017.12.138

De Araújo EA, Tomazini A, Kadowaki MAS, et al (2013) Crystallization and preliminary X-ray diffraction analysis of a new xyloglucanase from Xanthomonas campestris pv. campestris. Acta Crystallogr Sect F Struct Biol Cryst Commun 69:676–678. doi: 10.1107/S174430911301275X

DeLano WL (2002) Pymol: An open-source molecular graphics tool. CCP4 Newsl Protein Crystallogr 40:82–92

Dereeper A, Guignon V, Blanc G, et al (2008) Phylogeny . fr : robust phylogenetic analysis for the. Access 36:465–469. doi: 10.1093/nar/gkn180

Dos Reis CV, Bernardes A, Polikarpov I (2013) Expression, purification, crystallization and preliminary X-ray diffraction analysis of Bifidobacterium adolescentis xylose isomerase. Acta Crystallogr Sect F Struct Biol Cryst Commun 69:588–591. doi: 10.1107/S174430911301110X

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763

Erythropel HC, Zimmerman JB, De Winter TM, et al (2018) The Green ChemisTREE: 20 years after taking root with the 12 principles. Green Chem 20:1929–1961. doi: 10.1039/c8gc00482j

Feigin LA, Svergun DI (1987) Structure Analysis by Small-Angle X-Ray and Neutron Scattering. Springer Science+Business Media, LLC

Ferreira FM, Oliveira LC, Germino GG, et al (2011) Macromolecular assembly of polycystin-2 intracytosolic C-terminal domain. Proc Natl Acad Sci 108:9833–9838. doi: 10.1073/pnas.1106766108

Finn RD, Coggill, Penelope Eberhardt RY, Eddy SR, et al (2016) The Pfam Protein Families Database: Towards A More Sustainable Future. Nucleic Acids Res 44:D279–D285

Fischer H, de Oliveira Neto M, Napolitano HB, et al (2010) The molecular weight of proteins in solution can be determined from a single SAXS measurement on a relative scale. J Appl Crystallogr 43:101–109

Franke D, Jeffries CM, Svergun DI (2015) Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. Nat Methods 12:419–422. doi: 10.1038/nmeth.3358

Franke D, Petoukhov M V., Konarev P V., et al (2017) ATSAS 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions. J Appl Crystallogr 50:1212–1225. doi: 10.1107/S1600576717007786

Franke D, Svergun DI (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Crystallogr 42:342–346. doi: 10.1107/S0021889809000338

Gaurav N, Sivasankari S, Kiran GS, et al (2017) Utilization of bioresources for sustainable biofuels: A Review. Renew Sustain Energy Rev 73:205–214. doi: 10.1016/j.rser.2017.01.070

Gekko K, Noguchi H (1979) Compressibility of globular proteins in water at 25.degree.C. J Phys Chem 83:2706–2714. doi: 10.1021/j100484a006

Glatter O, Kratky O (1982) Small angle scattering. Academic Press

Glauninger H, Zhang Y, Higgins KA, et al (2018) Metal-dependent allosteric activation and inhibition on the

same molecular scaffold: the copper sensor CopY from Streptococcus pneumoniae. Chem Sci 9:105–118

Grimm ED, Grimm ED, Portugal R V, et al (2006) Structural analysis of an Echinococcus granulosus actin-fragmenting protein by small-angle x-ray scattering studies and molecular modeling. Biophys J 90:3216–23

Gruszka DT, Whelan F, Farrance OE, et al (2015) Domains Drives Assembly of a Strong Elongated Protein. Nat Commun 6:1–9. doi: 10.1038/ncomms8271

Guinier A (1939) La diffraction des rayons X aux très petits angles : application à l'étude de phénomènes ultramicroscopiques. Ann Phys (Paris) 11:161–237. doi: 10.1051/anphys/193911120161

Guttman M, Weinkam P, Sali A, Lee KK (2013) All-atom ensemble modeling to analyze small-angle X-ray scattering of glycosylated proteins. Structure 21:321–331. doi: 10.1016/j.str.2013.02.004

Halabi N, Rivoire O, Leibler S, et al (2009) Supplemental Data Theory Protein Sectors: Evolutionary Units of Three-Dimensional Structure. Cell 138:774–786

Hammersley AP, Svensson SO, Hanfland M, et al (1996) Two-dimensional detector software: From real detector to idealised image or two-theta scan. High Press Res 14:235–248. doi: 10.1080/08957959608201408

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89:10915–10919. doi: 10.1073/pnas.89.22.10915

Henrissat B, Callebaut I, Fabrega S, et al (1995) Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. Proc Natl Acad Sci 92:7090–7094. doi: 10.1073/pnas.92.15.7090

Hopkins JB, Gillilan RE, Skou S (2017) BioXTAS RAW: Improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. J Appl Crystallogr 50:1545–1553. doi: 10.1107/S1600576717011438

Horn SJ, Sorlie M, Vårum KM, et al (2012) Measuring processivity. In: Methods in Enzymology. Elsevier, pp 69–95

Hura GL, Menon AL, Hammel M, et al (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods 6:606

I Svergun D (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys J 76:2879–2886

Jacques DA, Trewhella J (2010) Small-angle scattering for structural biology - Expanding the frontier while avoiding the pitfalls. Protein Sci 19:642–657. doi: 10.1002/pro.351

Jeffries CM, Graewert MA, Blanchet CE, et al (2016) Preparing monodisperse macromolecular samples for successful biological small-Angle X-ray and neutron-scattering experiments. Nat Protoc 11:2122–2153. doi: 10.1038/nprot.2016.113

Kadowaki MAS, Higasi P, de Godoy MO, et al (2018) Biochemical and structural insights into a thermostable cellobiohydrolase from Myceliophthora thermophila. FEBS J 285:559–579. doi: 10.1111/febs.14356

Kayushina RL, Rolbin YA, Feigin LA (1974) Determination of the volume of biological macromolecule by mean of small-angle x-ray scattering. Sov Phys Crystallogr 19:1161–1165

Keegstra K (2010) Plant cell walls. Plant Physiol 154:483–486

Konarev P V., Petoukhov M V., Svergun DI (2001) MASSHA - A graphics system for rigid-body modelling of macromolecular complexes against solution scattering data. J Appl Crystallogr 34:527–532. doi: 10.1107/S0021889801006100

Konarev P V., Svergun DI (2015) A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. IUCrJ 2:352–360. doi: 10.1107/S2052252515005163

Korasick DA, Tanner JJ (2018) Determination of protein oligomeric structure from small-angle X-ray scattering. Protein Sci 27:814–824. doi: 10.1002/pro.3376

Kullback S, Leibler RA (1951) On Information and Sufficiency. Ann Math Stat 22:79–86. doi: 10.1214/aoms/1177729694

Leatherbarrow RJ, Enzfitter (1987) a non-linear regression data analysis program for the IBM PC. Biosoft

Liao H, Zhang XZ, Rollin JA, Zhang YHP (2011) A minimal set of bacterial cellulases for consolidated bioprocessing of lignocellulose. Biotechnol J 6:1409–1418. doi: 10.1002/biot.201100157

Lima LHF, Serpa VI, Rosseto FR, et al (2013) Small-angle X-ray scattering and structural modeling of full-length: Cellobiohydrolase I from Trichoderma harzianum. Cellulose 20:1573–1585. doi: 10.1007/s10570-013-9933-3

Lipfert J, Doniach S (2007) Small-Angle X-Ray Scattering from RNA, Proteins, and Protein Complexes. Annu Rev Biophys Biomol Struct 36:307–327. doi: 10.1146/annurev.biophys.36.040306.132655

Liu Y-S, Baker JO, Zeng Y, et al (2011) Cellobiohydrolase hydrolyzes crystalline cellulose on hydrophobic faces. J Biol Chem 286:11195–11201

Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science (80-) 286:295–299

Marchessault RH, Sundararajan PR, others (1983) Cellulose. The polysaccharides 2:11–95

Mertens HDT, Svergun DI (2010) Structural characterization of proteins and complexes using small-angle X-ray solution scattering. J Struct Biol 172:128–141. doi: 10.1016/j.jsb.2010.06.012

Miller GL (1959) Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal Chem 31:426–428

Momeni MH, Payne CM, Engström Å, et al (2013) Enzymology : Structural , biochemical , and computational characterization of the glycoside hydrolase family 7 cellobiohydrolase of the tree-killing fungus Heterobasidion irregulare Henrik Hansson , Nils Egil Mikkelsen , Jesper Gregg T . Beckham and Jerry. J Biol Chem 288:5861–5872. doi: 10.1074/jbc.M112.440891

Mortensen K, Posselt D (1998) Small-angle scattering of x-rays and neutrons. John Wiley & sons, Inc.

Needleman S, Wunsch C (1970) a General Method Applicable To Search for Similarities in Amino Acid Sequence of 2 Proteins. J Mol Biol 48:443–453. doi: 10.1016/0022-2836(70)90057-4

Neuhoff V, Stamm R, Eibl H (1985) Clear background and highly sensitive protein staining with Coomassie Blue dyes in polyacrylamide gels: A systematic analysis. Electrophoresis 6:427–448. doi: 10.1002/elps.1150060905

Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103:8577–8582. doi: 10.1073/pnas.0601602103

Nguyen STC, Freund HL, Kasanjian J, Berlemont R (2018) Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy. Appl Microbiol Biotechnol 102:1629–1637. doi: 10.1007/s00253-018-8778-y

Notredame C, Higgins DG, Herringa J (2000) Elephant Shark Genome Project\rT-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217

Nummi M, Niku-Paavola ML, Lappalainen A, et al (1983) Cellobiohydrolase from Trichoderma reesei. Biochem J 215:677–683

Orthaber D, Bergmann A, Glatter O (2000) SAXS experiments on absolute scale with Kratky systems using water as a secondary standard. J Appl Crystallogr 33:218–225. doi: 10.1107/S0021889899015216

Paine, Payne CM, Knott BC, et al (2015) Fungal Cellulases. Chem Rev 115:1308–1448. doi: 10.1021/cr500351c

Parsiegla G, Juy M, Reverbel-Leroy C, et al (1998) The crystal structure of the processive endocellulase CelF of Clostridium cellulolyticum in complex with a thiooligosaccharide inhibitor at 2.0 Å resolution. EMBO J 17:5551–5562. doi: 10.1093/emboj/17.19.5551

Perry JJP, Tainer JA (2013) Developing advanced X-ray scattering methods combined with crystallography and computation. Methods 59:363–371. doi: 10.1016/j.ymeth.2013.01.005

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786. doi: 10.1038/nmeth.1701

Petoukhov M V., Franke D, Konarev P V., et al (2012) New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr 50:342–350. doi: 10.1107/S1600576717007786

Petoukhov M V, Svergun DI (2005) Global Rigid Body Modeling of Macromolecular Complexes against Small-Angle Scattering Data. Biophys J 89:1237–1250. doi: 10.1529/biophysj.105.064154

Piiadov V, de Araújo EA, Oliveira Neto M, et al (2018) SAXSMoW 2.0: Online calculator of the molecular weight of proteins in dilute solution from experimental SAXS data measured on a relative scale. Protein Sci. doi: 10.1002/pro.3528

Prates ÉT, Stankovic I, Silveira RL, et al (2013) X-ray Structure and Molecular Dynamics Simulations of Endoglucanase 3 from Trichoderma harzianum: Structural Organization and Substrate Recognition by Endoglucanases That Lack Cellulose Binding Module. PLoS One 8:e59069. doi: 10.1371/journal.pone.0059069

Pronker MF, Lemstra S, Snijder J, et al (2016) Structural basis of myelin-associated glycoprotein adhesion and signalling. Nat Commun 7:13584. doi: 10.1038/ncomms13584

Provencher SW (1982) CONTIN: A general purpose constrained regularization program for inverting noisy linear algebraic and integral equations. Comput Phys Commun 27:229–242. doi: 10.1016/0010-4655(82)90174-6

Ragauskas AJ, Williams CK, Davison BH, et al (2006) The path forward for biofuels and biomaterials. Science (80-) 311:484–489

Rambo RP (2015) Scatter, a java-based program for bioSAXS

Reynolds KA, Russ WP, Socolich M, Ranganathan R (2013) Chapter Ten – Evolution-Based Design of Proteins. In: Methods in Enzymology. Elsevier, pp 213–235

Rice P (2000) (DSAP) EMBOSS : The European Molecular Biology Open Software Suite. Science (80-.). 16:2–3

Rivoire O, Reynolds KA, Ranganathan R (2016) Evolution-based functional decomposition of proteins. PLoS Comput Biol 12:e1004817

Robert C E (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. doi: 10.1186/1471-2105-5-113

Rojas AL, Fischer H, Eneiskaya E V., et al (2005) Structural insights into the β-xylosidase from Trichoderma reesei obtained by synchrotron small-angle x-ray scattering and circular dichroism spectroscopy. Biochemistry 44:15578–15584. doi: 10.1021/bi050826j

Rosseto FR, Puhl AC, Andrade MO, Polikarpov I (2013) Crystallization and preliminary diffraction analysis of the catalytic domain of major extracellular endoglucanase from Xanthomonas campestris pv. campestris. Acta Crystallogr Sect F Struct Biol Cryst Commun 69:137–140. doi: 10.1107/S1744309112051408

Santos CR, Paiva JH, Sforça ML, et al (2012) Dissecting structure–function–stability relationships of a thermostable GH5-CBM3 cellulase from *Bacillus subtilis* 168. Biochem J 441:95–104. doi: 10.1042/BJ20110869

Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2016) FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. Nucleic Acids Res 44:W424–W429. doi: 10.1093/nar/gkw389

Schülein M (2000) Protein engineering of cellulases. Biochim Biophys Acta (BBA)-Protein Struct Mol Enzymol 1543:239–252

Socolich M, Lockless SW, Russ WP, et al (2005) Evolutionary information for specifying a protein fold. Nature 437:512–518

Squire PG, Himmel ME (1979) Hydrodynamics and protein hydration. Arch Biochem Biophys 196:165–177. doi: 10.1016/0003-9861(79)90563-0

Staudacher E, Altmann F, Wilson IBH, März L (1999) Fucose in N-glycans: From plant to man. Biochim Biophys Acta - Gen Subj 1473:216–236. doi: 10.1016/S0304-4165(99)00181-6

Sukharnikov LO, Alahuhta M, Brunecky R, et al (2012) Sequence, structure, and evolution of cellulases in glycoside hydrolase family 48. J Biol Chem 287:41068–41077. doi: 10.1074/jbc.M112.405720

Svergun D, Barberato C, Koch MH (1995) CRYSOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 28:768–773. doi: 10.1107/S0021889895007047

Svergun DI, Konarev P V, Petoukhov M V, Volkov V V (2006) ATSAS 2.1, a program package for small-angle scattering data analysis. J Appl Crystallogr 39:277–286. doi: 10.1107/S0021889806004699

Svergun DI, Stuhrmann HB (1991) New development in direct shape determination from small-angle scattering. 1. Theory and model calculations. Acta Crystallogr A47:736–744. doi: Doi 10.1107/S0108767391006414

T. Teeri T (1997) Crystalline cellulose degradation: New insight into the function of cellobiohydrolases. Trends Biotechnol 15:160–167

Textor LC, Colussi F, Silveira RL, et al (2013) Joint X-ray crystallographic and molecular dynamics study of cellobiohydrolase i from Trichoderma harzianum: Deciphering the structural features of cellobiohydrolase catalytic activity. FEBS J 280:56–69. doi: 10.1111/febs.12049

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple Sequence Alignment Using ClustalW and ClustalX. Curr Protoc Bioinforma 2–3. doi: 10.1002/0471250953.bi0203s00

Trewhella J, Duff AP, Durand D, et al (2017) 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: An update. Acta Crystallogr Sect D Struct Biol 73:710–728. doi: 10.1107/S2059798317011597

Vaaje-Kolstad G, Westereng B, Horn SJ, et al (2010) An Oxidative Enzyme Boosting the Enzymatic Conversion of Recalcitrant Polysaccharides. Science (80-) 330:219–222. doi: 10.1126/science.1192231

Valentini E, Kikhney AG, Previtali G, et al (2014) SASBDB, a repository for biological small-angle scattering data. Nucleic Acids Res 43:D357--D363

Väljamäe P, Kipper K, Pettersson G, Johansson G (2003) Synergistic cellulose hydrolysis can be described in terms of fractal-like kinetics. Biotechnol Bioeng 84:254–257. doi: 10.1002/bit.10775

Väljamäe P, Sild V, Nutt A, et al (1999) Acid hydrolysis of bacterial cellulose reveals different modes of synergistic action between cellobiohydrolase I and endoglucanase I. Eur J Biochem 266:327–334. doi: 10.1046/j.1432-1327.1999.00853.x

Van Tilbeurgh H, Tomme P, Claeyssens M, et al (1986) Limited proteolysis of the cellobiohydrolase I from Trichoderma reesei. FEBS Lett 204:223–227. doi: 10.1016/0014-5793(86)80816-X

Vizoná Liberato M, Cardoso Generoso W, Malagó W, et al (2012) Crystallization and preliminary X-ray diffraction analysis of endoglucanase III from Trichoderma harzianum. Acta Crystallogr Sect F Struct Biol Cryst Commun 68:306–309. doi: 10.1107/S1744309112000838

Volkov V V., Svergun DI (2003) Uniqueness of ab initio shape determination in small-angle scattering. J Appl Crystallogr 36:860–864. doi: 10.1107/S0021889803000268

Webb B, Sali A (2014) Protein structure modeling with MODELLER. Protein Struct Predict 1–15

Zeng J, Bian F, Wang J, et al (2017) Performance on absolute scattering intensity calibration and protein molecular weight determination at BL16B1, a dedicated SAXS beamline at SSRF. J Synchrotron Radiat 24:509–520. doi: 10.1107/S1600577516019135

Zhang S, Wang RS, Zhang XS (2007) Uncovering fuzzy community structure in complex networks. Phys Rev E - Stat Nonlinear, Soft Matter Phys 76:46103. doi: 10.1103/PhysRevE.76.046103

Zhang Y, Lynd L (2004) Toward an aggregated undestanding of enzymatic hydrolysis of cellulose. Biotechnol Bioeng 88:797–824

Zheng J, Gay DC, Demeler B, et al (2012) Divergence of multimodular polyketide synthases revealed by a didomain structure. Nat Chem Biol 8:615–621. doi: 10.1038/nchembio.964

APPENDICES

APÊNDICE A. Brognaro, H., Almeida, V. M., de Araujo, E. A., Piyadov, V., Santos, M. A. M., Marana, S. R., & Polikarpov, I. (2016). Biochemical Characterization and Low-Resolution SAXS Molecular Envelope of GH1 β-Glycosidase from Saccharophagus degradans. *Molecular biotechnology*, *58*(12), 777-788.

Mol Biotechnol
DOI 10.1007/s12033-016-9977-3

ORIGINAL PAPER

Biochemical Characterization and Low-Resolution SAXS Molecular Envelope of GH1 β-Glycosidase from *Saccharophagus degradans*

Hevila Brognaro[1] · Vitor Medeiros Almeida[2] · Evandro Ares de Araujo[1] · Vasily Piyadov[1] · Maria Auxiliadora Morim Santos[1] · Sandro Roberto Marana[2] · Igor Polikarpov[1]

Abstract The marine bacteria *Saccharophagus degradans* (also known as *Microbulbifer degradans*), are rod-shaped and gram-negative motile γ-proteobacteria, capable of both degrading a variety of complex polysaccharides and fermenting monosaccharides into ethanol. In order to obtain insights into structure–function relationships of the enzymes, involved in these biochemical processes, we characterized a *S. degradans* β-glycosidase from glycoside hydrolase family 1 (SdBgl1B). SdBgl1B has the optimum pH of 6.0 and a melting temperature T_m of approximately 50 °C. The enzyme has high specificity toward short D-glucose saccharides with β-linkages with the following preferences β-1,3 > β-1,4 ≫ β-1,6. The enzyme kinetic parameters, obtained using artificial substrates *p*-β-NPGlu and *p*-β-NPFuc and also the disaccharides cellobiose, gentiobiose and laminaribiose, revealed SdBgl1B preference for *p*-β-NPGlu and laminaribiose, which indicates its affinity for glucose and also preference for β-1,3 linkages. To better understand structural basis of the enzyme activity its 3D model was built and analysed. The 3D model fits well into the experimentally retrieved low-resolution SAXS-based envelope of the enzyme, confirming monomeric state of SdBgl1B in solution.

Keywords Glycosidase hydrolase family 1 · *Saccharophagus degradans* · Enzymatic characterization · Modelling · SAXS structure

✉ Igor Polikarpov
ipolikarpov@ifsc.usp.br

Hevila Brognaro
hbrognaro@usp.br

Vitor Medeiros Almeida
vitorma3@gmail.com

Evandro Ares de Araujo
evandro.ufpa.fisica@gmail.com

Vasily Piyadov
piyadov@usp.br

Maria Auxiliadora Morim Santos
santosma@ifsc.usp.br

Sandro Roberto Marana
srmarana@iq.usp.br

[1] Instituto de Física de São Carlos, Universidade de São Paulo, Avenida Trabalhador São Carlense 400, São Carlos, SP 13566-590, Brazil

[2] Instituto de Química, Universidade de São Paulo, Avenida Prof. Lineu Prestes, 748, Bloco 10, Sala 1054, São Paulo, SP 05508-900, Brazil

Introduction

Deconstruction of lignocellulosic biomass is one of the most important natural recycling processes of plant materials on Earth. Large amounts of biomass available for bioconversion promoted ever-growing interest in the development of second generation biofuel technologies, green chemistry and sustainable materials technologies [1]. To achieve this, distinct biomass and/or agricultural wastes have to be transformed and hydrolyzed enzymatically to produce platform molecules (such as glucose, for example), which can be further converted into liquid biofuels and sustainable chemical products [2]. However identification, characterization and optimizations of the enzymes useful in such transformations still remain a bottleneck for their utilization in the bioprocesses [3, 4]. Therefore, during the last years, a number of suitable microorganisms and enzymes useful for this purpose have been extensively studied.

Published online: 26 September 2016

APÊNDICE B. Araújo, E. A., Manzine, L. R., Piiadov, V., Kadowaki, M. A. S., & Polikarpov, I. (2018). Biochemical characterization, low-resolution SAXS structure and an enzymatic cleavage pattern of BlCel48 from Bacillus licheniformis. *International journal of biological macromolecules*, *111*, 302-310.

International Journal of Biological Macromolecules 111 (2018) 302–310

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: http://www.elsevier.com/locate/ijbiomac

Biochemical characterization, low-resolution SAXS structure and an enzymatic cleavage pattern of *Bl*Cel48 from *Bacillus licheniformis*

Evandro Ares de Araújo, Lívia Regina Manzine, Vassili Piiadov, Marco Antonio Seiki Kadowaki, Igor Polikarpov *

Instituto de Física de São Carlos, Universidade de São Paulo, Av. Trabalhador São-carlense, 400, São Carlos, SP CEP 13560-970, Brazil

ARTICLE INFO

Article history:
Received 19 November 2017
Received in revised form 17 December 2017
Accepted 25 December 2017
Available online 30 December 2017

Keywords:
Bacillus licheniformis
Cellulase
GH48 family
SAXS

ABSTRACT

Economic sustainability of modern biochemical technologies for plant cell wall transformations in renewable fuels, green chemicals, and sustainable materials is considerably impacted by the elevated cost of enzymes. Therefore, there is a significant drive toward discovery and characterization of novel carbohydrate-active enzymes. Here, the *Bl*Cel48 cellulase from *Bacillus licheniformis*, a glycoside hydrolase family 48 member (GH48), was functionally and biochemically characterized. The enzyme is catalytically stable in a broad range of temperatures and pH conditions with its enzymatic activity at pH 5.0 and 60 °C. *Bl*Cel48 exhibits high hydrolytic activity against phosphoric acid swollen cellulose (PASC) and bacterial cellulose (BC) and significantly lower activity against carboxymethylcellulose (CMC). *Bl*Cel48 releases predominantly cellobiose, and also small amounts of cellotriose and cellotetraose as products from PASC hydrolysis. Small-angle X-ray scattering (SAXS) data analysis revealed a globular molecular shape and monomeric state of the enzyme in solution. Its molecular mass estimated based on SAXS data is ~77.2 kDa. *Bl*Cel48 has an $(\alpha\alpha)_6$-helix barrel-fold, characteristic of GH48 members. Comparative analyses of homologous sequences and structures reveal the existence of two distinct loops in *Bl*Cel48 that were not present in other structurally characterized GH48 enzymes which could have importance for the enzyme activity and specificity.

1. Introduction

The plant cell walls are constituted mainly by a complex cross-linked matrix composed of polyphenolics, fibers, sugars, and some proteins and polysaccharides [1–3]. Cellulose, a major component of plant cell walls polysaccharides, is an abundant natural, renewable and biodegradable linear homopolymer formed by poly-β(1,4)-D-glucopyranose [3–6]. Extensive hydrogen bonding, high crystallinity, and recalcitrance of cellulose make this polymer insoluble in most solvents hindering the processing of plant biomass [5,7].

At present, biochemical conversion of lignocellulosic biomass using enzymatic technology became a preferential biotechnological route for plant biomass depolymerization [8–14]. Enzymatic hydrolysis is carried out by a diversity of enzymes classified in CAZy (Carbohydrate-Active EnZymes) database, which can differ structurally and mechanistically [15,16]. To overcome recalcitrance of lignocellulosic substrates, synergic action of several enzymes is required for its efficient hydrolysis [7,10, 17–20]. Bacterial biomass degrading systems play an important role in the breakdown of plant cell walls, including cellulases from glycoside hydrolase family 48 (GH48) [21]. GH48 contains mostly bacterial enzymes, with CelS from *Clostridium thermocellum* [22], CelF from *Clostridium cellulolyticum* [23], cellulase from *Caldocellum saccharolytic* [24] and cellobiohydrolase B from *Cellulomonas fimi* [25] being its founding members. GH48 includes mainly, but not always, reducing end-acting cellobiohydrolases, which make part of complex bacterial cellulose degrading systems.

These enzymes are capable of synergistically enhancing activities of GH9 processive *endo*-cellulases in the process of crystalline cellulose hydrolysis, as has been described for *Tf*Cel48A and *Tf*Cel9A from *Thermobifida fusca* [26], *Cb*Cel9A and *Cb*Cel48A from *Caldicellulosiruptor bescii* [27], and *Cc*Cel9B and *Cc*Cel48A from *Clostridium cellulosi* [20]. Unlike GH9 enzymes, GH48 genes are frequently present as single copies in the genomes of cellulolytic organisms, such as *Bacillus* species, for example [28].

A gram-positive bacterium, *Bacillus licheniformis* is a facultative anaerobe, which is widely distributed in the environment [29]. *B. licheniformis* has a number of important biotechnological, agricultural and industrial applications [30,31]. It is used for production of antibiotics, chemicals [30] and some of the industrially important hydrolytic enzymes including penicillinase, pentosanases, proteases, α-amylases, glucoamylase, pectinases and several cellulase enzymes [32–34]. In this work, we report the heterologous expression, purification and biochemical characterization of *Bl*Cel48 from *Bacillus licheniformis* (ATCC 14580). We also built and analyzed its low resolution SAXS molecular envelope and its 3D homologous model.

* Corresponding author at: Universidade de São Paulo, Departamento de Física e Ciência Interdisciplinar, Instituto de Física de São Carlos, Av. Trabalhador São-carlense, 400, São Carlos, SP CEP 13560-970, Brazil.
E-mail address: ipolikarpov@ifsc.usp.br (I. Polikarpov).

https://doi.org/10.1016/j.ijbiomac.2017.12.138
0141-8130/

APÊNDICE C. Piiadov, V., Ares de Araújo, E., Oliveira Neto, M., Craievich, A. F., & Polikarpov, I. (2019). SAXSMoW 2.0: Online calculator of the molecular weight of proteins in dilute solution from experimental SAXS data measured on a relative scale. *Protein Science*, *28*(2), 454-463.

THE PROTEIN SOCIETY

SAXSMoW 2.0: Online calculator of the molecular weight of proteins in dilute solution from experimental SAXS data measured on a relative scale

Vassili Piiadov,[1] Evandro Ares de Araújo,[1] Mario Oliveira Neto,[2] Aldo Felix Craievich,[3] and Igor Polikarpov[1]*

[1]Institute of Physics of São Carlos, University of São Paulo, São Carlos, São Paulo, Brazil
[2]Institute of Biosciences, University Estadual Paulista, Botucatu, São Paulo, Brazil
[3]Institute of Physics, University of São Paulo, São Paulo, São Paulo, Brazil

Received 2 August 2018; Accepted 8 October 2018
DOI: 10.1002/pro.3528
Published online 13 December 2018 proteinscience.org

Abstract: Knowledge of molecular weight, oligomeric states, and quaternary arrangements of proteins in solution is fundamental for understanding their molecular functions and activities. We describe here a program SAXSMoW 2.0 for robust and quick determination of molecular weight and oligomeric state of proteins in dilute solution, starting from a single experimental small-angle scattering intensity curve, *I*(*q*), measured on a relative scale. The first version of this calculator has been widely used during the last decade and applied to analyze experimental SAXS data of many proteins and protein complexes. SAXSMoW 2.0 exhibits new features which allow for the direct input of experimental intensity curves and also automatic modes for quick determinations of the radius of gyration, volume, and molecular weight. The new program was extensively tested by applying it to many experimental SAXS curves downloaded from the open databases, corresponding to proteins with different shapes and molecular weights ranging from ~10 kDa up to about ~500 kDa and different shapes from globular to elongated. These tests reveal that the use of SAXSMoW 2.0 allows for determinations of molecular weights of proteins in dilute solution with a median discrepancy of about 12% for globular proteins. In case of elongated molecules, discrepancy value can be significantly higher. Our tests show discrepancies of approximately 21% for the proteins with molecular shape aspect ratios up to 18.

Keywords: SAXS; proteins; molecular weight; on-line calculator

Introduction

Small-angle X-ray scattering (SAXS) is an experimental technique frequently applied to low-resolution structural studies of macromolecules embedded in a homogeneous liquid medium, over a molecular size scale within the 1–100 nm range. The SAXS method allows for investigations of both, well-structured and disordered macromolecules in solution, neither requiring crystallization procedures nor highly elaborate sample preparations. The experimental SAXS intensity associated to a set of proteins of the same

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Conselho Nacional de Desenvolvimento Científico e Tecnológico 140667/2015-6158752/2015-5303988/2016-9405191/2015-4440977/2016-9; Grant sponsor: Fundação de Amparo à Pesquisa do Estado de São Paulo 2014/00769-52015/13684-0.

*Correspondence to: I. Polikarpov, Institute of Physics of São Carlos, University of São Paulo São Carlos São Paulo, Brazil. Email: ipolikarpov@ifsc.usp.br

V. Piiadov and E. Ares de Araújo contributed equally to this work.