Computational and Structural Biotechnology Journal 18 (2020) 1864–1876

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A convenient protein library for spectroscopic calibrations

Joëlle De Meutter [a], Erik Goormaghtigh [b,*]

[a] Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, Campus Plaine CP206/02, Brussels, Belgium
[b] Université Libre de Bruxelles CP206/2, B1050 Brussels, Belgium

ARTICLE INFO

Article history:
Received 21 May 2020
Received in revised form 30 June 2020
Accepted 2 July 2020
Available online 10 July 2020

Keywords:
Protein selection
FTIR spectroscopy
Secondary structure
Protein spectroscopy

ABSTRACT

While several Raman, CD or FTIR spectral libraries are available for well-characterized proteins of known structure, proteins themselves are usually very difficult to acquire, preventing a convenient calibration of new instruments and new recording methods. The problem is particularly critical in the field of FTIR spectroscopy where numerous new methods are becoming available on the market.

The present papers reports the construction of a protein library (cSP92) including commercially available products, that are well characterized experimentally for their purity and solubility in conditions compatible with the recording of FTIR spectra and whose high-resolution structure is available.

Overall, 92 proteins were selected. These proteins cover well the CATH space at the level of classes and architectures. In terms of secondary structure content, an analysis of their high-resolution structure by DSSP shows that the mean content in the different secondary structures present in cSP92 is very similar to the mean content found in the PDB.

The 92-protein set is analyzed in details for the distribution of helix length, number of strands in β-sheets, length of β-strands and amino acid content, all features that may be important for the interpretation of FTIR spectra.

© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Correct protein 3D structure is responsible for enzymatic activity as well as for functional recognition between protein partners. Whether we want to understand how a protein is working or we want to produce therapeutic proteins, which are now the first domain of innovation and investment in the Pharma, we need to assess protein structure as well as minute protein structure changes. X-ray diffraction, nuclear magnetic resonance (NMR) and cryo-electron microscopy (Cryo-EM) are the gold standards to obtain high-resolution protein structures. While they certainly provide the highest resolution structures, they suffer from a number of drawbacks, which make them difficult to use on a routine basis. The price of high resolution is that methods are cumbersome, slow, expensive and do not work for all proteins. Furthermore, observing the effect of environmental changes on structure remains difficult and limited to a small number of cases.

On the other hand, spectroscopies such as circular dichroism (CD), Raman and infrared (FTIR) are fast, cheap and workable in various environments, e.g. [1]. The drawback is that no high-resolution structure can be obtained. Yet, when investigating the conformational behavior of an enzyme, we are rather interested in differences that may exist between different states. Similarly, when producing a new batch of a therapeutic antibody or a biosimilar, the key question is the presence or absence of difference with respect to the reference molecule. Interpretation of spectra strongly depends on reference techniques, essentially X-ray diffraction, and on the quality of the protein library used to create mathematical models that relate spectral shape to structural features. The quality of the protein library depends on both adequate coverage of the conformational space and quality of the samples considering potential unfolding or aggregation. Several protein spectral libraries have been established in the past. In 1981 Hennessey and Johnson [2] presented a protein set of 16 proteins used for CD spectrum analysis and in 1992 Prestrelski et al. [3] built a set of proteins used for FTIR analysis. Twenty one reference proteins were similarly used for combining both FTIR and CD spectra by Pribic et al. [4]. In 2003 we presented a set of 50 proteins used to build a predictive model in Attenuated infrared reflection mode (ATR) on Germanium surfaces [5,6]. Yet, all proteins were not easily accessible and some proteins present in the previous reference sets are no longer commercially available. Currently, the most complete spectral library is the SMP180 CD spectral library built

* Corresponding author.
E-mail address: egoor@ulb.ac.be (E. Goormaghtigh).

https://doi.org/10.1016/j.csbj.2020.07.001
2001-0370/© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Nomenclature

Aa	amino acid residue
ATR	Attenuated Total Reflection
CATH	Class(C), Architecture(A), Topology(T) and Homologous superfamily (H) structure classification
CD	circular dichroism
cSP92	convenient Soluble Protein set containing 92 proteins
DSSP	Dictionnary of Secondary Structure of Proteins algorithm
FTIR	Fourier transform infrared
PDB	Protein Data Bank
SDS PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
SRCD	synchrotron radiation CD

over the years by the group of B.A. Wallace [7]. This dataset includes CD spectra for 30 membrane proteins and 98 soluble proteins.

While the protein spectral libraries described above are excellent, the protein themselves are not available to anyone. First, spectra have been accumulated over the years and proteins are not all available anymore, second, many of the proteins included are gift from numerous laboratories and are not easily accessible.

In the present paper we designed a protein library that can be easily rebuilt by any researcher who needs a well-characterized protein library to test a new structure prediction method. This is important, in particular for FTIR spectroscopy. While CD, including SRCD, reference spectra remain valid over the years if adequately recorded, FTIR spectroscopy has so many modes of recording that may impact the exact spectral shape e.g. transmission, ATR (with various internal reflection elements, various incidence angles), microscopy, imaging, etc. that a protein set that can be easily acquired is most needed to calibrate the different approaches. While the present work is essentially guided by the need of an easily accessible protein set for FTIR spectroscopy purposes, it can be used for any other aim.

The protein set proposed here is made out of 92 soluble proteins. We show how the structures present in the library span the secondary structure space and the fold space as described by CATH [8]. We also paid great attention to the match between the sequence of the acquired protein and the sequence of the reference protein structure found in the PDB [9], about the quality of the PDB structure and finally about the actual solubility of the commercially available protein sample and the purity.

2. Experimental procedures

2.1. Protein selection

The purpose of this selection is to identified commercially available proteins with a published high resolution structure.

a) Protein pre-selection

The first step of the protein selection process was the search of commercially available proteins. Sigma Aldrich, Worthington Enzyme and Biochemical and ENZO Life sciences catalogs were consulted for this purpose. More than 1200 references were identified. All proteins included in the Sigma Aldrich and Worthington catalogs were examined one by one. In the first round of selection, to be retained as potential candidates, suitable proteins had to be:

- available in 1 or 2 mg quantities (or less in some exceptional cases)
- claim at least 85% purity as determined by SDS-PAGE. If the purification process of the protein included crystallization or affinity chromatography, it was considered sufficiently pure at this stage of the selection.
- present in the PDB, either from the same source or represented by a similar structure. In the latter case, the Universal Protein knowledgebase (UniProt Consortium [10] was consulted to identify another biological species for the protein and the corresponding PDB entry. To compare the sequence of the acquired protein with the sequence of other PDB entries corresponding to other biological species, EMBOSS pairwise Needle (a part of the EMBL European Bioinformatics Institute, https://www.ebi.ac.uk/Tools/psa/emboss_needle/, [11]) program was used. If the score was below 85% for the sequence identity or if no candidate matched, the protein was rejected.

At the end of this first step, 104 proteins from Sigma Aldrich catalogs, matching the criteria were preselected, Similarly 23 candidates from the Worthington Enzyme and Biochemical Catalog and 9 proteins from Enzo Life Sciences were preselected.

b) Selection refinement based on sequence similarity and quality of reference structure

The second step consisted in examining, among all the structures present in the PDB for a particular protein, the one which best represents the protein of interest considering structure quality and protein sequence. For this purpose, PDBSum [12,13] was used as it offers an ordered classification of PDB identities by decreasing order of similarity to the sequence reported in UniProt [14] including modifications mutations/mismatches, and decreasing resolution of the X-ray or NMR structures. The best % of similarity as well as the highest quality structures (better resolution) were selected. Structures determined by X-ray crystallography were preferred over NMR or Cryo EM structures when the choice was possible.

At the end of this process, 127 candidates were selected and acquired: 99 proteins from Sigma Aldrich, 14 from Worthington Enzyme and Biochemical, 8 from Enzo life Sciences. ZneA, ZneB, SilB-C, Sil β-NM [15–17], β-Lactamase [18] and ApoE3 [19] were obtained from Robotein (Belgium). Robotein is one of the European Instruct Centers. Table S1 shows all the proteins preselected for the establishment of a protein library. The proteins discarded after selection refinement, as discussed later, are highlighted in grey.

2.2. Experimental evaluation

Protein solubilization was carried out to take into account constraints of spectroscopies, in particular FTIR spectroscopy which requires removal of many buffer or additive molecules such as EDTA, acetate etc. Proteins acquired as dry materials were solubilized at a final concentration of 10–20 mg/ml. To avoid contributions of the original buffers, salts and/or additives present in the commercial sample, proteins were de-salted and buffer exchanged against 4 mM Hepes, 85 mM NaCl, pH usually between 7.4 and 7.6 as precised in Table S2. Low NaCl concentration (85 mM) was added to the buffer as, at low concentration, NaCl increases

solubility by suppressing electrostatic protein–protein interaction (salting in effect). Attention was paid to the isoelectric point of the proteins, which is reported in Table S2. When solubility issues were detected, the pH was modified to keep away from the isoelectric point. The actual experimental pH is reported in Table S2 for each protein. Buffer exchange was achieved through 5 cycles of filtration (Amicon Ultra-0.5 ml Centrifugal Filters 3 K) of 100–200 μl protein solution (i.e. 1–2 mg protein) or 2 passes through size exclusion centrifuge mini column (Bio-Rad Micro Bio-Spin 3kD), equilibrated with the final buffer.

Purity and integrity of the acquired proteins were then controlled by SDS Page (4–20% Mini-PROTEAN Precast Protein Gels, Bio-Rad). Four to five μg of protein were deposited in each well. Protein bands were finally revealed by Coomassie Blue staining. Gels were then scanned (Bio-Rad GS-80 Calibrated Densitometer) and the Bio-Rad Quantity One program was used to process the scanned images and obtain the density profile of each lane. After background subtraction, the lane profiles were exported to Kinetics, a homemade program running under Matlab (The MathWorks Inc.). Purity was estimated in Kinetics by integrating each band from the profile. The area under the protein band identified by its molecular weight was divided by the sum of all integrated bands for each lane. If the estimated purity was below 85%, the protein was discarded. The purity values for each protein can be found in Table 1 for the selected proteins.

Solubility was in fact the main issue but purity check also resulted in the rejection of several proteins. Whenever possible, another similar protein was acquired using the tools described above in Protein Selection. Finally, out of the 127 selected proteins, only 92 were judged to be suitable for spectroscopy.

2.3. Extraction of secondary structure from PDB files

The secondary structure values from the PDB files were obtained according to definitions provided by DSSP, a method originally described by Kabsch and Sander [20] and improved in DSSPcont [21–23]. DSSPcont identifies eight secondary structures states: 3_{10}-helix (G), α-helix (H), pi-helix (I), helix-turn (T), extended β-sheet (E), β-bridge (B), bend (S) and other/loop (L). Only α-helix (H) and β-sheet (E) are reported in Table 1.

Secondary structures features such as α-helix length, β-sheet length, number of strands per β-sheet, proportion of parallel and antiparallel β-sheet were extracted from the DSSP files by a module of the home made Kinetics software running under Matlab. They are reported in Table S4. As previously suggested by Kalnin et al. in 1990 [24] for FTIR spectroscopy and by Sreerama et al. in 1999 [25] for CD, we also computed the fraction of "ordered" and "disordered" helices. Disordered helix content was obtained by considering the two amino acid residues at the end of each α-helix as "disordered" helix while the core of the helix was assigned to "ordered" helix. Similarly, the fractional content of each amino acid per protein was obtain from the protein sequence taken from the DSSP file (Table S3). The Kinetics software tabulated the data, provided the histograms, and required statistics.

In a second step, the sequence of the full protein present in the test tube was compared with the sequence of the crystallized protein described in the PDB. For a certain number of proteins, a short sequence was missing in the crystallized species. When that was the case, the secondary structure content was recalculated supposing the missing sequence had no ordered structure. The α-helix and β-sheet content could change by 2–3% in general. In two case (ApoE3 and SBTI), the crystallized species was longer than the acquired protein and secondary structure was adjusted considering in the PDB structure only the fraction corresponding to the actual protein available. Similarly, amino acids that were not resolved in the high resolution structure were considered as having unordered structures.

3. Results

At the end of the selection process described in Methods, only 92 proteins met the purity, sequence identity, structure quality, crystallographic and solubility criteria. This set is referred below to as cSP92, standing for convenient Soluble Protein set containing 92 candidates. Table 1 presents the 92 proteins retained. They constitute the reference set. The 35 candidates discarded on the basis of their behavior in our experimental conditions designed for FTIR spectroscopy are highlighted in grey in Table S1.

Table 1 presents the PDB identities for each one, the resolution expressed in Å, a comparison between the sequence of the acquired protein and of the crystallized protein. This comparison includes the % of gaps, i.e. the % of missing residues observed after alignment of sequences of the acquired protein and of the crystallized one. The % of identity between the sequence of the crystallized protein and the sequence of the commercially available protein is also indicated, as well as the % of similarity (as defined by EMBOSS Needle [11]) along with the estimated purity of each protein, the % α -helix and β -sheet structure as assigned by the DSSPcont algorithm. When there was a difference in the number of residues between the crystallized and commercially available protein, the secondary structure value was recalculated taking into account this difference, assuming that the missing residues adopt a random structure.

Table 1 List of the proteins included in cSP92. The first columns provide the common name, the PDB ID and the resolution of the structure when obtained from X-ray diffraction. When the PDB entry is followed by "*", it means that the PDB entry refers to another species (see Table S1). The "Gap" column reports in % a difference between the sequence of the acquired protein and of the crystallized proteins. Such differences mostly occur when a few amino acid residues from one end of the protein are missing in the crystallized protein or when an additional His tag is present on the acquired protein but not on the crystallized protein (or not considered in the DSSP analysis). The "Identity" and "Similarity" of the sequences have been obtained from EMBOSS Needle [11]. Purity of the acquired proteins was evaluated from SDS-PAGE as described in Methods. The "%H" and "%E" columns report the α -helix and β -sheet content respectively as defined by DSSP. The "Chain ID" column reports the chain that has been used for the DSSP analysis. For chains of identical sequences, the chain with the best resolution and with the smallest number of unresolved amino acid residues was selected. The label "all" indicates that the mean of all chain was computed, as required when the acquired protein was constituted of several chains of different sequences. The results of the DSSP analysis were corrected for taking into account any mismatch between the sequence of the acquired protein and of the crystallized protein. Any of these sequences was assigned to "Other" structure and the H and E content recomputed accordingly

3.1. Characteristics of the protein library

3.1.1. Coverage of structural classes and architectures

The development of bioinformatics has led to hierarchical and systematic classification systems of secondary structure units at different levels. The CATH database developed by Orengo et al [8] is such a classification system. It is considered that two proteins that have the same secondary structure elements in the same orientation and connectivity will possess identical fold. Some motives are reoccurring among the protein population because of the limited number of spatial organizations allowed. CATH is a

Table 1
cSP92 proteins and their main characteristics.

Protein name	PDB ID	Resolution (Å)	Gaps (%)	Identity (%)	Similarity (%)	Purity (%)	%H DSSP	%E DSSP	Chain ID	CATH superfamily			
Aldolase A	1zah	1.8	0	100	100	100	41.87	13.77	A	3.20.20.70			
Alpha-2-Macroglobulin	4acq	4.3	0	100	100	90.8	12.62	30.37	C	rejected			
Alpha-2-MRAP	2p03	NMR	0	100	100	95.8	63.15	0	A	1.20.81.10			
Alpha-Amylase	1vjs	1.7	0	99.8	99.8	100	23.39	19.87	A	3.20.20.80	2.40.30.140	2.60.40.1180	
Alpha-Crystallin B chain	2ygd*	EM, 9.4	0	97.7	97.7	93.8	9.71	28	all	rejected			
Amidase	2uxy	1.25	1.4	98.6	98.6	100	30	21.39	A	3.60.110.10			
Amino acid oxidase, D-	1ve9	2.5	0	100	100	91.6	25.36	26.08	A	3.40.50.720	3.30.9.10		
Apolipoprotein E3	1h7i	1.9	9.8	89.6	89.6	98.9	65.02	0	A	1.20.120.20			
Apo-Transferrin	4h0w	2.4	0	100	100	99	29.89	18.11	A	3.40.190.10			
Aprotinin (Trypsin inhibitor Kunitz) type)	4y0y chain I	1.25	0	100	100	100	13.79	24.13	I	2.40.10.10			
Avidin	1vyo	1.48	0	100	100	97.9	0	46.48	B	2.40.128.30			
Beta-Amylase	1fa2	2.3	0	99.8	99.8	91.2	31.12	11.44	A	3.20.20.80			
Beta-Galactosidase	5a1a	2.2	0.1	99.9	99.9	100	10.46	37.67	A	2.60.120.260	2.60.40.10	3.20.20.80	2.70.98.10
Beta-Glucunoridase	3lpf	2.26	0.3	99.7	99.7	96.99	16.74	27.44	A	2.60.120.260	2.60.40.10	3.20.20.80	
Beta-Lactamase TEM	1xpb	1.9	0	100	100	100	39.92	17.11	A	3.40.710.10			
Beta-Lactoglobulin	3npo	2.2	0	100	100	100	9.87	40.74	A	2.40.128.20			
Bowman-Birk proteinase inhibitor	5j4q	2.3	0	100	100	100	0	29.68	B	2.10.69.10			
Calmodulin	1prw	1.7	13.5H-tag	99.3	99.3	100	50.53	2.35	A	1.10.238.10			
Carbonic anhydrase 1	1hcb	1.6	0	100	100	99.3	8.46	28.84	A	3.10.200.10			
Carbonic anhydrase 2	1v9e	1.95	0	100	100	100	7.91	28.95	A	3.10.200.10			
Carboxyl esterase	1k4y	2.5	2.4	97.6	97.6	92.3	30.89	13.52	A	3.40.50.1820			
Carboxypeptidase A1	2ctb	1.5	0.6	99.4	99.4	99.3	36.8	16.28	A	3.40.630.10			
Carboxypeptidase Y	1ysc	2.8	0	100	10	100	35.39	14.25	A	3.40.50.1820	1.10.287.410		
Catalase	3rgp	1.88	5.1	94.9	94.9	100	26.28	15.68	A	1.10.8.1230	2.40.180.20	1.20.1370.60	
Cathepsin G	1cgh	1.9	4.7	95.3	95.3	95*	6.65	29.55	A	2.40.10.10			
Ceruloplsamin	4enz	2.6	1.8	98.2	98.2	93.5	9.34	36.41	A	2.60.40.120			
Choline oxidase	4mjw	1.95	0	100	100	100	21.61	20.67	A	3.50.50.60	4.10.450.10	1.10.1220.10	3.30.410.40
Chymotripsinogen A	2cga	1.8	0	100	100	99.3	7.34	32.04	A	2.40.10.10			
Citrate synthase	3enj	1.78	0	100	100	92.3	58.81	3.43	A	1.10.580.10	1.10.230.10		
Concanavalin A (Lectin)	1i3h	1.2	0	100	100	89.6	0	43.45	A	2.60.120.200			
Creatine (phospho)kinase	1u6r	1.65	0.3	99.5	99.7	89	33.68	14.21	A	3.30.590.10	1.10.135.10		
Cyclophyllin A	3k0n	1.39	12.3H-tag	100	100	100	10.88	27.82	A	2.40.100.10			
Cytochrome *c*	1hrc	1.9	0.1	99	99	100	40.95	0	A	1.10.760.10			
Deoxyribonuclease-1	3dni	2	0	100	100	99.6	25.38	26.15	A	3.60.10.10			
DT-diaphorase	1d4a	1.7	0.4	99.6	99.6	100	29.02	11.35	A	3.40.50.360			
Elafin	1fle	1.9	0	100	100	90	0	21.27	I	4.10.75.10			
Elastase	1qnj	1.1	0	100	100	95.95	5.83	30.41	A	2.40.10.10			
Endo-1,4-beta-xylanase	2jic*	1.5	6.9	80.9	86.3	100	4.9	56.86	A	2.60.120.180			
Enolase	1ebh	1.9	0	99.8	100	100	38.64	16.97	A	3.30.390.10	3.20.20.120		
Galactose oxidase	2eie	1.8	0	100	100	100	0.62	39.59	A	2.60.120.260	2.130.10.80	2.60.40.10	
Gelonin	3ktz	1.6	0	100	100	100	34.66	21.11	A	3.40.420.10	4.10.470.10.		
Glucagon	1nau	NMR	10	86.7	90	100	58.65	0	A	rejected			
Glucose Oxidase	1cf3	1.9	0	100	100	96.2	26.92	19.21	A	3.50.50.60	4.10.450.10	3.30.560.10	
Glutamate oxaloacetate transaminase 1	5toq	1.2	0.5	98.8	99.5	85.63	44.17	13.34	A	3.90.1150.10*	3.40.640.10*		
Glutathione Reductase	3djg	1.8	0.2	99.8	99.8	100	27.88	23.48	X	3.50.50.60	3.30.390.30		
Glyceraldehyde-3-phosphate dehydrogenase	1j0x	2.4	0.3	99.4	99.4	100	24.84	25.52	O	3.40.50.720	3.30.360.10		
Glycogen phosphorylase-b	1axr	2.3	0	100	100	100	44.65	14.25	A	3.40.50.2000			
Hemoglobin	2qsp	1.85	0	100	100	97.9	67.3	0	all	1.10.490.10			

(*continued on next page*)

Table 1 (*continued*)

Protein name	PDB ID	Resolution (Å)	Gaps (%)	Identity (%)	Similarity (%)	Purity (%)	%H DSSP	%E DSSP	Chain ID	CATH superfamily				
Hexokinase	1ig8	2.2	0.2	99.8	99.8	97	38.47	16.04	A	3.30.420.40	3.40.367.20	1.10.287.-		
Immunoglobuling G	1hzh	2.7	–	–	–	100	3.19	42.41	all	2.60.40.10				
Insulin (neat)	3w7y	0.92	0	100	100	100	42.15	5.88	all	rejected				
Lactate Dehydrogenase	3h3f	2.38	0	100	100	97.5	39.5	19.75	A	3.40.50.720	3.90.110.10			
Lactoferrin (Lactotransferrin) human	1cb6	2	0	100	100	100	29.95	18.81	A	3.40.190.10				
Lactoferrin bovin	1blf	2.8	0	100	100	95.9	29.6	17.56	A	3.40.190.10				
Lactoperoxidase	6a4y	1.92	2.8	97.2	97.2	100	32.67	5.55	A	1.10.640.10*				
Lectin	1len	1.8	0.4	69.6	97.9	100	1.71	48.06	all	2.60.120.200				
Leptin	1ax8	2.4	0	99.3	99.3	92.5	56.16	0	A	1.20.1250.10				
Lipoxidase	1f8n	1.4	0	100	100	89.1	32.53	13.11	A	2.60.60.20	4.10.375.10	4.10.372.10	3.10.450.60	1.20.245.10
Lysostaphin	4lxc	3.5	3.5H-tag	96.5	96.5	99.1	2.43	43.9	A	2.70.70.10	2.30.30.140			
Lysozyme	4lzt	0.95	0	100	100	100	31	6.2	A	1.10.530.10				
Metallothionein-2A	4mt2*	2	3.2	82.5	85.7	94.1	0	0	A	4.10.10.10				
Micrococcal Nuclease	1ey0	1.6	0	100	100	100	22.14	26.84	A	2.40.50.90				
Myoglobin	1wla	1.7	0	100	100	100	73.85	0	A	1.10.490.10				
Myokinase (Adenylate kinase 1)	2c95	1.71	1	98	98	94.5	56.52	12.78	A	3.40.50.300				
Ovalbumin	1ova	1.95	0.3	99.7	99.7	100	27.39	28.88	A	2.30.39.10	3.30.497.10			
Ovotransferrin (Conalbumin)	1ovt	2.4	0	100	100	100	27.55	17.63	A	3.40.190.10				
Pepsin A	4pep	1.8	0	99.7	99.9	100	11.04	43.25	A	2.40.70.10				
Pepsinogen	2psg	1.8	0	99.5	100	95.2	7.02	36.21	A	2.40.70.10				
Peroxidase	1hch	1.57	0.6	99.4	99.4	100	44.44	1.96	A	1.10.520.10	1.10.420.10			
Phosphatase, Alkaline	1y6v	1.6	0.2	99.8	99.8	100	27.83	18.15	A	3.10.130.10				
Phosphoglucomutase 1	5epc	1.85	3.9H-tag	96.1	96.1	100	32.97	24.24	A	3.40.120.10	3.30.310.50			
Phosphoglycerate kinase	1qpg	2.4	0	99.8	100	100	35.42	16.62	A	3.40.50.1260				
Phospholipase A2	2osh*	2.2	0.8	78.2	83.2	100	42.01	6.72	A	1.20.90.10				
Protein disulfide isomerase	4el1*	2.88	9.8H-tag	86.5	88.5	100	25.96	18.05	A	3.40.30.10				
Pyrophosphatase inorganic	1i40	1.8	0.3	99.7	99.7	100	17.4	32.57	A	3.90.80.10				
Pyruvate Kinase	1a49	2.1	0	100	100	100	35.4	18.86	A	3.40.1380.20	3.20.20.60	2.40.33.10		
Ribonuclease A	1kf5	1.15	0	100	100	100	17.74	33.06	A	3.10.130.10				
Ribonuclease T1	1rls	1.9	0	100	100	100	16.34	27.88	A	3.10.450.30				
Serum Albumin	1n5u	1.9	0	100	100	95.9	68.88	0	A	1.10.246.10				
SilB-C	2 l55	NMR	0	100	100	100	0	48.78	A	2.40.50.320				
Silb-NM2	5a4g	NMR	0	100	100	100	2.22	25	A					
Subtilisin Calsberg	3unx	1.26	0	98.5	99.6	100	29.56	17.88	A	3.40.50.200				
Superoxide Dismutase (Fe)	1isa	1.8	0	100	100	95.1	47.39	10.93	A	holding pen				
Superoxide Dismutase (CU Zn)	1q0e	1.15	0.7	99.3	99.3	100	2.63	38.81	A	2.60.40.200				
Thaumatin	3aok	1.27	0	100	100	100	10.62	35.74	A	2.60.110.10				
Transketolase	2r8o	1.47	0.9H-tag	99.6	99.6	100	42.82	13.6	A	3.40.50.970	3.40.50.920			
Transthyretin (Préalbumin)	1tta	1.7	0	100	100	93.4	4.72	48.03	A	2.60.40.180				
Triose Phosphate Isomerase	1ypi	1.9	0	100	100	97.4	37.65	16.19	A	3.20.20.70				
Trypsin inhibitor A Kunitz type SBTI	1ba7	2.5	0	100	100	100	0	33.42	A	2.80.10.50				
Ubc9	2pe6 chain A	2.4	2.5	97.5	97.5	100	33.75	18.17	A	3.10.110.10				
Ubiquitin	2wwz	1.4	0	100	100	100	15.78	31.57	A	3.10.20.90				
ZneB	3lnn chainA	2.8	0	100	100	89.5	17.3	33.72	A	2.40.420.20	2.40.30.170	2.40.50.100	1.10.287.470	

hierarchical classification. Briefly, the first level refers to the nature of the secondary structure content and is subdivided in 4 classes: 1) Mainly Alpha; 2) Mainly Beta, 3) Alpha Beta and 4) Few secondary structure. The second level in the classification is Architecture. It refers to the general arrangement of the secondary structures ignoring the connection between them. The third one, the Topology level, or 'fold' level, takes into account the spatial arrangement of secondary structures units in the chain. The last one, Homologous Superfamily concerns the relations of domains to a potential common ancestor. We will focus here only on the first second and third level of classification.

All four classes are represented in the selected protein set, totalizing 17 architectural designs: 2 out of 5 in class 1; 7 out of 21 in class 2; 7 out of 14 in class 3 and 1 out of 1 in class 4 (Table 1). In class 1, that is mainly Alpha; 1.10 (Orthogonal bundle) and 1.20 (Up–down bundle) are the most populated architectures in term of folds since they gather 291 and 104 folds respectively, totalizing 395 out of 405 Folds. These 2 major architectures are represented in cSP92. The remaining 3 architectures total only 10 Folds altogether. Class 2, mainly β-sheet proteins, presents a higher diversity of architectures than class 1; since it gathers 21 architectures instead of 7. The 7 architectures represented in the set count for 179 folds out of 244 folds. The third class, Alpha Beta, groups 14 architectures. Once again, the 7 architectures represented in cSP92 are the most populated in term of fold since they totalize 618 folds out of 634. Class 4; irregular, counts only 1 architecture and cSP92 totalizes 6 folds out of 108.

Table 2 shows the number of unique folds represented in each architecture type. The number of unique folds both in the all-alpha is 20 (represented by 23 proteins in cSP92) and in the all-beta class is also 20, represented by 40 proteins in cSP92. In the alpha–beta class 3, 30 unique folds and represented by 58 proteins.

The largest number of topologies (folds) present cSP92 set is found in class 1, the orthogonal buddle (architecture 1.10) with 14 different folds and in class 2 the β barrel (architecture 2.40), 9 folds. In class 3, α/ β, we find the 2-layer sandwich, 9 folds, and 3-layer (aba) sandwich architectures, 10 folds. Class 4, Irregular architecture, is represented too by 6 folds. In this particular architecture, Metallothionein (4.10.10) and Elafin (4.10.75) are small size proteins (6 kDa) both with one unique domain. Metallothionein is characterized by a high abundance of cysteines that serve as metal ion coordination sites. It is an intrinsically disordered protein with no α-helix, no β-sheet, and 100% of "other" structure. Elafin displays a compact structure maintained by four conserved disulfide bridges and no α-helix, around 21% β -sheet and 79% of "other" structure. The other irregular folds (4.10.-) appear in proteins that possess more than one domain (multidomain) (see Table 1). The discovery of intrinsically disordered proteins has changed the idea that protein function depends only on its three-dimensional folding. [26–28]. Flexibility may be essential for such disordered proteins or for some segments of well-structured ones to be fully functional. Yet, it is difficult to obtain a high-resolution 3D structure of a disordered protein and these proteins are therefore underrepresented with respect to their expected natural abundance [27,28].

Another structure that is difficult to sample in significant quantity is the parallel β-sheet. In class 2, all beta proteins, many proteins contain mixture of both parallel and anti-parallel β sheets. However, strict parallel β sheet is found in Tim Barrel fold (3.20.20), a specific fold of the αβ barrel architecture. Both Triose phosphate isomerase, β -Amylase and Aldolase contain a single Tim barrel fold. Some multidomain proteins also have Tim barrels, e.g. Enolase, β-Galactosidase, Pyruvate kinase, β -Glucuronidase and α -Amylase for example. A fair number of parallel beta sheets is also found in the α β 3-Layer (aba) sandwich architecture and the Rossman fold (3.40.50) present in Myokinase, Subtilisin Carlsberg, DT-Diaphorase, Phosphoglycate kinase, Amino acid oxidase, Carboxylesterase, Transketolase and Glycogen phosphorylase B for example. Altogether, 56 proteins contain parallel β-sheet but the content does not go beyond 16% (Table S4).

Antiparallel β-sheet is much more abundant and found in 75 proteins in cSP92. It is found for instance in the β -sandwich architecture, Jelly Rolls fold (2.60.120) e.g. Concanavalin A and Lectin.

3.1.2. Protein secondary structure distribution in cSP92:

As for any analytical calibration, the protein library needs to span the full range of structure fraction content. The number of representative individuals is also important because the classical secondary structures such as α- α-helix and β-sheet are not classes of homogenous conformations of the polypeptide chain as many present distortions compared with ideal canonical models. These distortions are reflected, for instance, in the infrared spectrum of the proteins e.g. [3]. It must also be stressed that the length of the helices is variable, helices can be distorted or present kinks, βsheets are either parallel or antiparallel, of different lengths and with different numbers of strands. The parallel β sheets give an infrared absorbance spectrum which is different from the antiparallel one, allowing for instance to distinguish oligomers from fibers in aggregate amyloid proteins [29,30]. Furthermore, β sheets are not completely flats but generally twisted. A recent report presents the large effect of the twist on CD spectra [31]. All these variations and others are reflected in the infrared spectra by shift of the absorbance bands and band shape variations. A large protein set such as cSP92 allows to better sample these variations.

3.1.2.1. Protein secondary structure distribution in cSP92. Protein structure information is available in the PDB or PDBe websites. The values of the secondary structures elements were assigned by algorithms such as DSSP cont [21–23]. Eight secondary structure states are identified in DSSPcont. The following discussion concerns α-helices (H) and β-sheets (E) only.

Fig. 1 reports the α -helix (H) and the β -sheet (E) distribution in cSP92. It can be observed that cSP92 set has a good coverage of H and E content. Alpha helix (H) content is covered between 0 and 78%, β sheet content (E) between 0 and 60%. The distribution of structure between H and E is well balanced, both in the extremes values and in combinations. There is a lack of representative proteins with few secondary structure (few α-helix and β-sheet content, class 4, irregular architecture). As already mentioned, such examples are difficult to find as there are few commercially avail-

Table 2
Range of architectures according to CATH classification in cSP92. Column 1. CATH Architecture code, column 2. Architecture name (A), column 3. Number of particular folds represented in A.

CATH A code	A: Architecture name	T varieties in A
1.10	α Orthogonal Bundle	14
1.20	α Up–down Bundle	6
2.10	β Ribbon	1
2.130	β 7 Propeller	1
2.30	β Roll	2
2.40	β Barrel	9
2.60	β Sandwich	4
2.70	β Distorted Sandwich	2
2.80	β Trefoil	1
3.10	αβ Roll	5
3.20	αβ Alpha-Beta Barrel	1
3.30	αβ 2-Layer Sandwich	9
3.40	αβ 3-Layer(aba) Sandwich	10
3.50	αβ 3-Layer(bba) Sandwich	1
3.60	αβ 4-Layer Sandwich	2
3.90	αβ Complex	2
4.10	Irregular	6

able proteins for which a high resolution structure exist. However the set includes two representatives of class 4: Metallothionein which is 100% disordered and Elafin which contain 79% disordered and 21% β sheet.

Fig. 1 also emphasizes some inverse proportionality relation that exists between α-helix and β-sheet content. ln order to visualize the distribution of the different secondary structures, cSP92 proteins were sorted successively in ascending order for H, E and O structure content where "O" represents the rest of the structures not described by H or E.

Fig. 2, left column, reports the α-helix, β -sheet and Other structure content in the 92 proteins of cSP92. Each bar represent one protein. It must be noticed that 28 of the cSP92 proteins have none or less than 10% α-helix content. On the other hand, four proteins have more than 60% and one more than 70% helix content. In between there is a rather continuous increase in helix content, suggesting cSP92 spans adequately the helix content range present in soluble proteins. The distribution reported in the right column indicates a somewhat less good representation between 15 and 25% helix content. Regarding β-sheet structure, 8 proteins have no β-sheet and 14 contain less than 10% β-sheet. The group which contains few beta structure (<10%) has important helix content (between 31 and 73%). The 10 protein group that contains high level of β-sheet (>40%) has fewer than 6% of α- helices, except Pepsin (11%) andβ-Lactoglobulin (10%). The β -sheet content distribution indicates a good coverage up to 40%. Fig. 2 also shows that 74% of the proteins contain between 40 and 60% of "O" with 44% of the proteins between 45 and 55% "O". The corresponding histogram illustrates that there is much less variance in the O structure than in H and E structures.

In conclusion, the cSP92 library has a good coverage of the H E space. Yet, the content of the "O" structure is characterized by a smaller variance.

3.2. Specific features of protein secondary structure distribution in cSP92

As mentioned earlier, the infrared spectra of proteins depends mostly on the secondary structure content, but also depends very significantly on some specific structural features of α-helices and β-sheet. For instance the length and distortion of the α-helical structure result in significant shifts of the amide I band [32,33] and the length and number of strands in β-sheet also correlated with definite spectral change in FTIR spectroscopy [33,34]. It was therefore important to establish the coverage of cSP92 in terms of α-helix length, β-sheet length and number of strands in β-sheets.

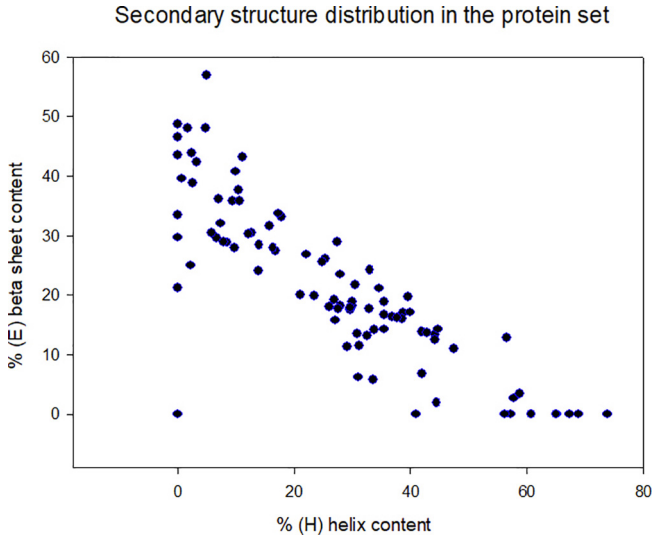

Fig. 1. Secondary structure distribution in cSP92. For each protein, it β-sheet content is reported as a function of its α-helix content. Each dot represents a protein from cSP92.

The length of each strand as well as the number of strand in each β-sheet was obtained from the DSSP files. The distribution of α-helix lengths reported in Fig. 3A shows that out of the 1,063 helices identified, cSP92 contains a large number of short helices (<8 aa). This number decreases as length increases. Though this is interesting in its own right, spectroscopies such as FTIR are sensitive to the amount of amino acid residues involved in a structure rather than to the number of structure elements and obviously longer helices contain more amino acid residues than shorter ones. Fig. 3B reports the fraction (in %) of amino acid found for each helix length category with respect to the total number of amino acids present in cSP92. This distribution shows that there is a good coverage up to about 20 amino acid long helices. The contribution of helices longer than 25 amino acids is reduced. As the FTIR spectrum dependency on helix length disappears for lengths above 14 amino acid [32,33], cSP92 adequately represents the variability of this structure. The proteins with the longest helices (more than 18 amino acids) are (% helix longer than 18 amino acid / total helix content): Apolipoprotein E3 (46.2/63.5%), Alpha-2-MRAP (44.0/63.2%), Leptin (42.4/56.1%), Serum Albumin (23.7/68.6%), Citrate synthase (21.1/58.8%), Myoglobin (16.3/71.2%) and Phospholipase A2 (16.0/42.0%) The fraction of the amino acids present in β- sheet structure is presented in Fig. 3C as a function of the β-strand length and in Fig. 3D as a function of the number of strands in the β-sheet, two factors well-known to affect the shape of FTIR spectra [34]. The longest β-sheets (more than 10 amino acid long) are found in (%β-sheet longer than 10 amino acid residues / total β-sheet content): Avidin (24.3/47.9%), Alpha-Crystallin B chain (7.4/28.9%), Ubc9 (7.0/17.7%) and Alpha-2-Macroglobulin (6.5/30.0%) and the sheets with the largest number of strands (% of β-sheet with more than 8 strands / total β-sheet content) are Superoxide Dismutase (6.6/39.1%), Lysostaphin (6.5/42.1%) and Carbonic anhydrase (4.6/29.0%). Lysozyme (4LZT) is remarkable by a 16% content of very short sheets (1 to 3 residues) which could obviously result in unclassical spectral contribution.

3.3. Amino side chain distribution in cSP92

Some amino acid side chains contribute to the infrared absorbance spectrum in the Amide I and II region used to predict the secondary structure [33,35–38] with sometimes dramatic effects on the band shape in the amide I and II region of the spectrum [39]. Their contribution needs therefore to be carefully analyzed. Major contributions of absorbance are due to arginine, asparagine, glutamine, lysine, aspartic and glutamic acid and to a less extend tyrosine, histidine and phenylalanine. The distribution of amino acid in the proteins of cSP92 is reported in Fig. 4. The majority of proteins contain less than 10% of any of these amino acids, but some particular proteins contain much more and will require a special attention for interpretation of their FTIR spectra. In the infrared spectrum, arginine side chains have a major contribution overlapping the amide I protein band. The most intense contribution is found at 1673 cm^{-1}, a second significant one is found at 1633 cm^{-1}, exactly overlapping turns and β -sheet contributions (reviewed in [33,37,38]). Not considering arginine content could drive to misinterpretation of the spectra. While the median content is about 3.8%, some proteins have more than 8% (Alpha-Crystallin, Alpha-2-MRAP, Lysozyme, Apolipoprotein E3, Aprotinin) up to 15% (Cathepsin G) which is going to bring a very significant contribution to the amide I region of the spectrum. At the opposite, other

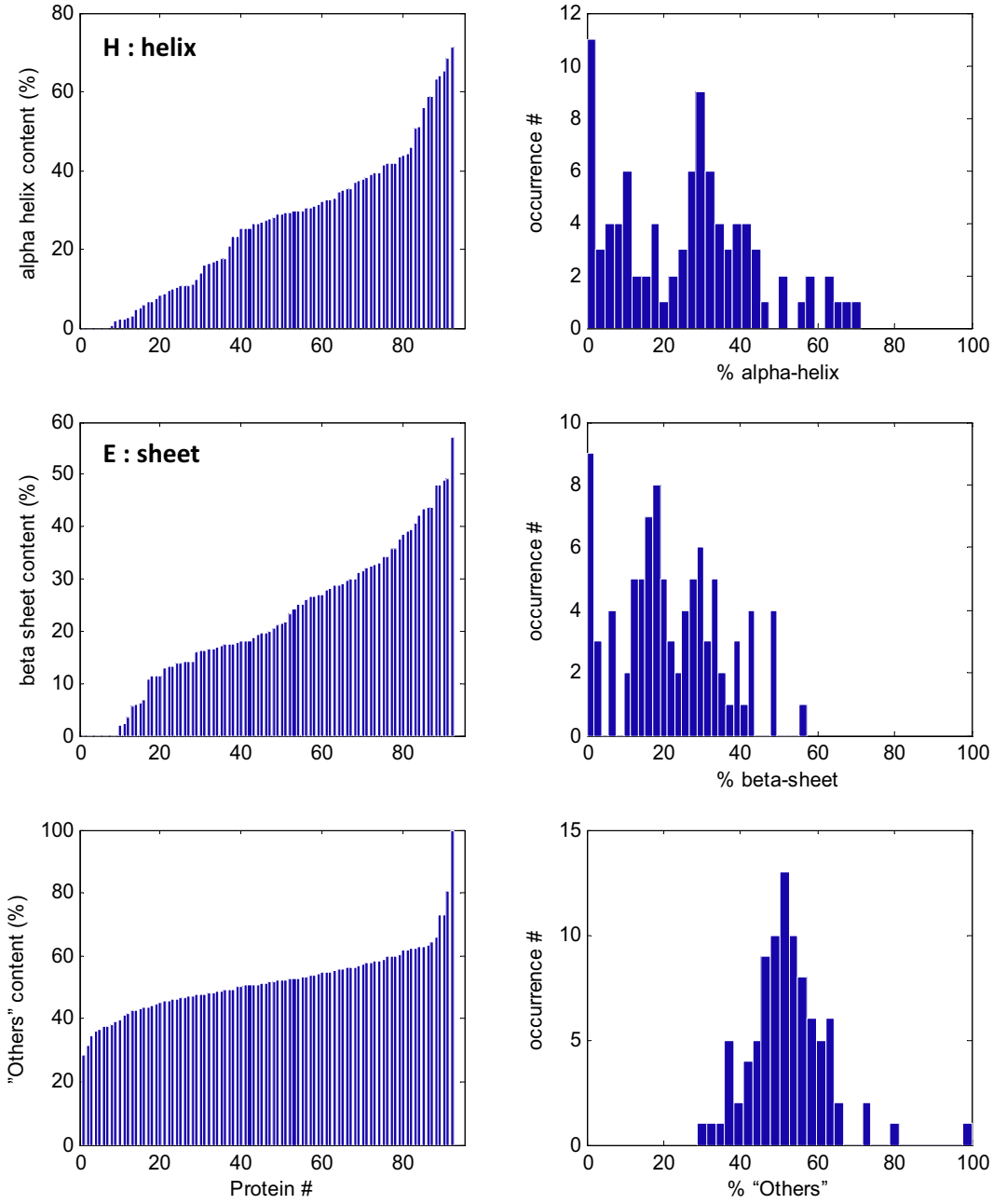

Fig. 2. Secondary structure content in cSP92 for α-helix (H), β-sheet (E) and all the other structures (O). For each structure, the proteins have been sorted in order of increasing content. The left column reports the % of structure in each protein, the right column reports the histogram describing the distribution of the structures with respect to the number of proteins with a particular secondary structure content.

proteins (Metallothionein-2A, Pepsin A) do not have any or insignificant amount, which might also contribute to poor secondary structure prediction from the infrared spectra as mathematical models include the average contribution of amino acid side chains. The contribution of glutamine and asparagine side chains with a major band near 1675 cm^{-1} and a significant one near 1620 cm^{-1} are additional contributions that complicate the interpretation of the amide I band. Together, they make up from about 2% (Transthyretin) to 14% (Endo-1,4-beta-xylanase) of the side chains. Similarly, the distribution of lysine side chains, which have a significant contribution near 1629 and 1526 cm^{-1}, is broad, from almost 0 (Pepsin A, Ribonuclease T1) to 15% in Micrococcal Nuclease and 18% in Cytochrome *c*, which is known to bring considerable contribution in the amide I region of the infrared spectrum. The distribution of the combined occurrence of the 4 amino acid side chains which most interfere with the analysis of the infrared amide I band (i.e. Arg, Lys, Asn, Gln) is presented in Figure S1. These combined amino acids represent more than 25% of the total amino acid content for 7 proteins: Ribonuclease A, Deoxyribonuclease-1, Alpha-2-Macroglobulin, Pyruvate Kinase, Cytochrome *c* and Avidin. The two carboxylic acid containing amino acids, aspartate and glutamate contribute more in the amide II region near 1580–1560 cm^{-1} and near 1720–1710 cm^{-1} when protonated. The sum of both varies between 4% (Endo-1,4-beta-xylanase) and 19% in Alpha-2-MRAP and 22% in Calmodulin. As amide II is sometimes the spectral region best correlated with some secondary structures [40], this effect needs to be carefully considered too. Because they are less intense and with narrow,

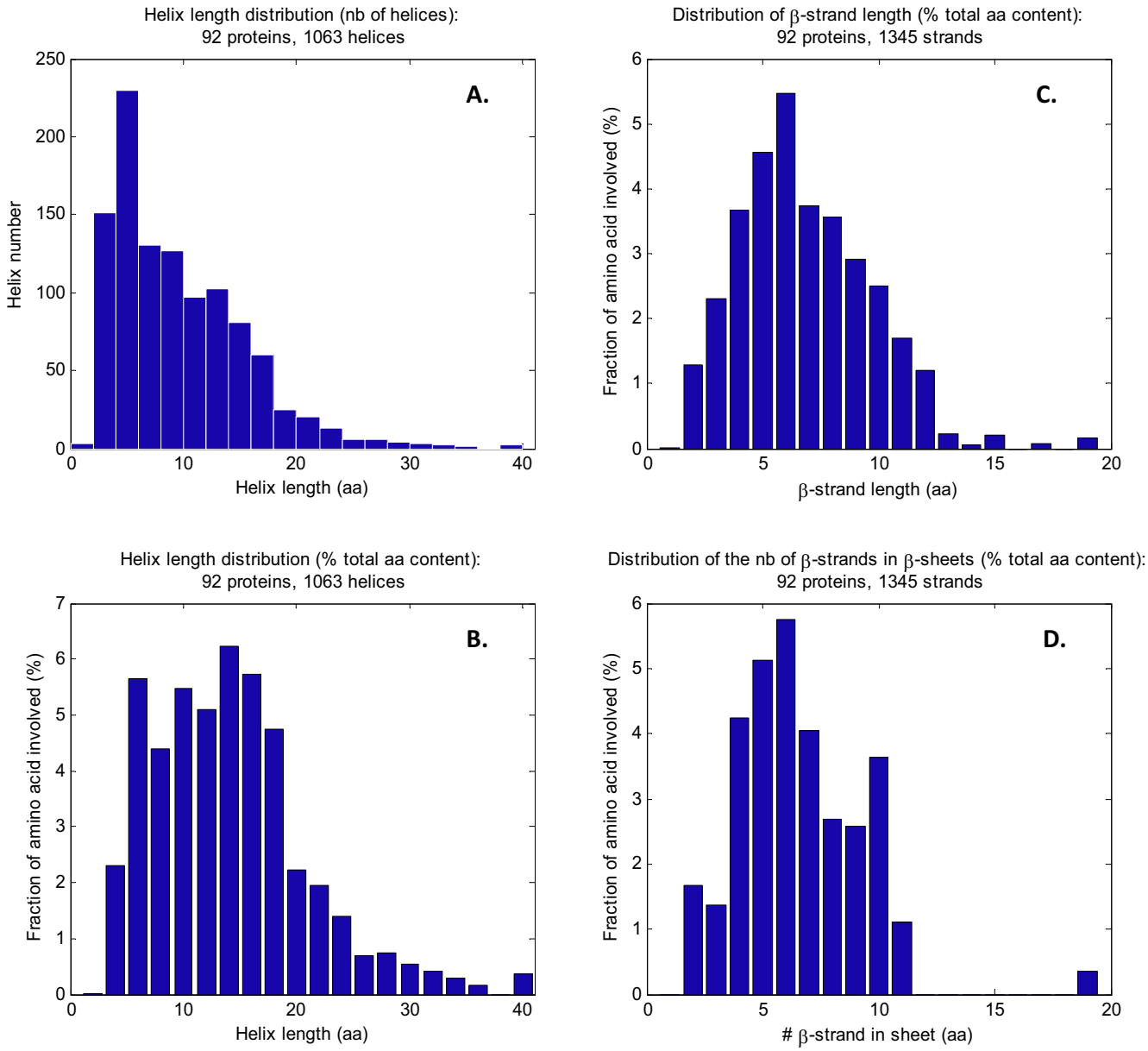

Fig. 3. Distribution of secondary structure features in cSP92 proteins. A. distribution of helices with a particular helix length (in amino acid residues), B. distribution of amino acid residues (expressed in % of total aa in cSP92) involved in helices of different length (in amino acid residues), C. distribution of amino acid residues (expressed in %) involved in β-sheet structures as a function of β-strand length, D. distribution of amino acid residues (expressed in %) involved in β-sheet structures as a function of the number of strands in the sheet.

well-localized contribution, histidine, phenylalanine and tyrosine are less of a problem. Nevertheless, recognizing their spectral contribution can be important, for instance the tyrosine narrow band near 1517 cm^{-1} can be used as in internal standard to scale spectra in the course of H/D exchange experiments [41].

Only the amino acid residues contributing significantly in the amide I – amide II region of the FTIR spectrum are reported here. The distribution for the 20 amino acids can be found in Table S3.

4. Discussion

Open Source is becoming the rule in scientific publication and access to raw data is generally granted through specific databases. Yet, chemical compounds are much less accessible, preventing the researchers to re-use the compounds utilized in published work. Protein sets used to calibrate Raman, CD or FTIR spectra for analytical determination of structural features are no exception. Yet, new instruments, new recording methods appear at a fast pace and constantly require new calibrations. In the absence of easily available and well-characterized proteins, such calibrations cannot take place and comparison with previous work cannot be achieved. In the field of FTIR spectroscopy only, transmission cells for aqueous solution [42], microfluidic modulation FTIR [43], vibrational circular dichroism [44], ATR with various incidence angles and internal reflection elements of different refractive indices [45,46], microscopy or imaging of 2D arrays of proteins [47,48] or human tissue sections [49–51] and new techniques as AFMIR [52] etc. produce spectra that, though similar, display specific features that prevent a single spectral database to be used for all approaches.

The present papers reports the construction of a protein library made out of commercially available products, they are well

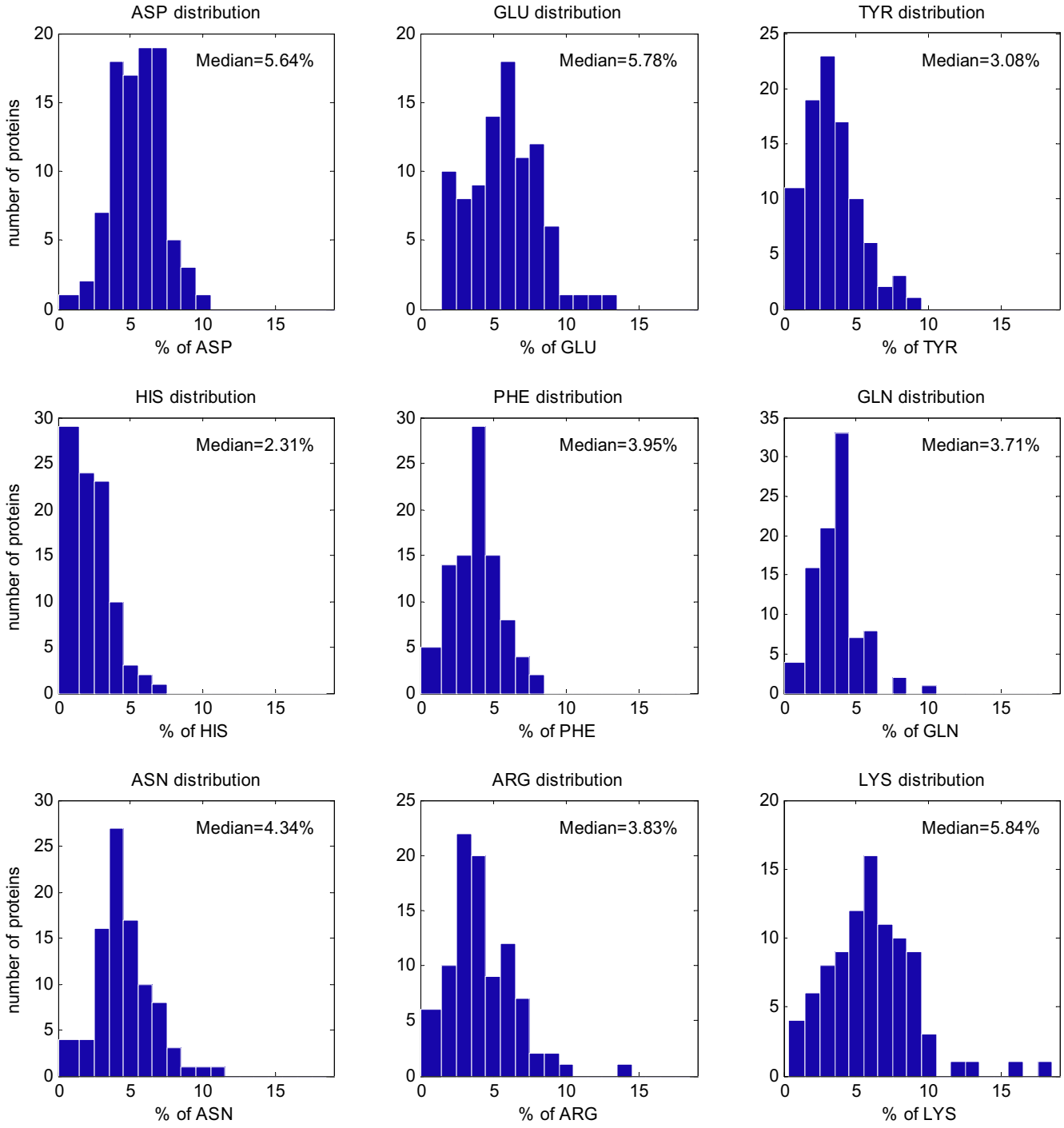

Fig. 4. Amino acid distribution in Csp92. The median value is indicated for each residue.

Table 3
Comparison of the mean secondary structure content in cSP92 and in the PDB according to Andersen and Rost [53].

	PDB	cSP92
α-helix (H)	31.3	26.0
β-sheet (E)	20.4	22.0
Anti// β-sheet	15.7	17.1
// β-sheet	5.7	4.1
Other (G, I, B, T, S)	48.3	52.0

characterized experimentally for their purity and solubility in conditions compatible with the recording of FTIR spectra and whose high-resolution structure is available. The most tedious part of the work was to cross the commercial catalogs, protein by protein, with the PDB and make sure the sequence of the crystallized protein matches the sequence of the commercial protein. Acquiring the proteins and rejecting them because of low solubility and/or poor purity also resulted in repeated searches for replacement. Overall, 92 proteins could be selected. These proteins cover well the CATH space at the level of classes and architectures. In terms of secondary structure content (Table 3), an analysis of the PDB

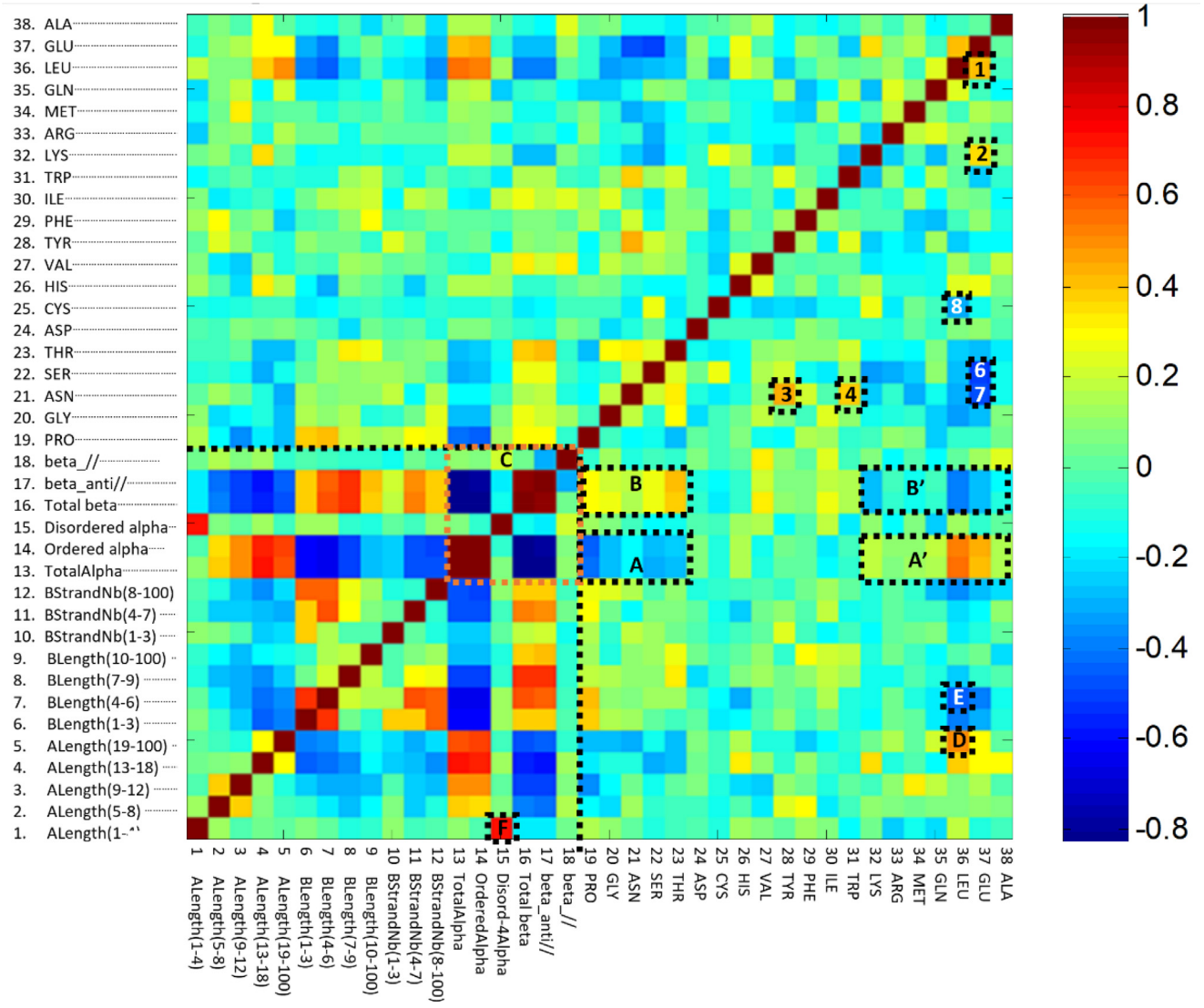

Fig. 5. Correlation between the relative abundance of secondary structure features and amino acids. The abbreviations used are beta_// and beta_anti// for parallel and antiparallel β-sheet respectively, Alength(i-j) for helix lengths comprised between i and j amino acid, Blength(i-j) for sheet lengths comprised between i and j amino acid, BStrandNb(i-j) for sheets containing between i and j strands.

by DSSP [53] shows that the mean content in the different secondary structures present in cSP92 is very similar to the mean content found in the PDB. A more recent PDB analysis reported by Micsonai et al. [31] is very similar.

The relatively small set of the proteins present in cSP92 with respect to the PDB shows other features similar to those found in the entire PDB, for instance in the correlations among structural features. Highlighting these correlations is also important as highly correlated features will be difficult to resolve independently from spectral data. Fig. 5 reports an auto-correlation analysis of the different structural features and amino acid content. Below the diagonal, the lower left part of the figure delineated by black dotted lines concerns correlations between secondary structure features and the upper right corner reports correlation among amino acid abundance in the protein set. Finally, the lower right part of the figure reports correlations between secondary structure features and amino acid abundance.

Negative correlations between α-helix content and β-sheet content is quite strong as already expected from Fig. 1. Frame **C**, orange, highlights the negative correlation between the two main secondary structures: on the one hand α-helix and ordered α-helix, on the other hand total β-sheet content or total antiparallel β-sheet content. It must be stressed that the amount of parallel β-sheet is not strongly anti-correlated with α-helix content. This can be expected as parallel β-sheet containing folds often also contain α-helices as in β-barrels for instance.

Interestingly, α-helix is anti-correlated with Pro, Gly, Asn, Ser, Thr (Fig. 5, Frame **A**) and positively correlated with the relative abundance of Lys, Arg, Met, Gln, Leu, Glu, Ala (Frame **A'**), a well-known property generally found in the PDB [54,55], while the β-sheet structure displays the opposite trend (**B** and **B'**) suggesting that the selected protein structures have amino acid composition representative of what is generally found in the PDB. Among amino acid residue relative abundance, we find positive correlation for Glu-Leu (**1**), Glu-Lys (**2**), Asn-Tyr (**3**), Asn-Trp (**4**) while there is a marked anti-correlation between Glu-Ser (**6**), Glu-Asn (**7**) and Leu-Cys (**8**). Finally, there is an interesting significant correlation (**D**) between Leu and long helices, i.e. the cluster of helices longer than 18 amino acid residues, and an anticorrelation with most β -sheets (**E**). It is also interesting to note that short helices are, as expected by definition, well correlated with disordered helices but do not show any specific correlation with other helix lengths or β structures.

The observations reported in Fig. 5 are significant in two aspects. First they demonstrate that the particular propensity of certain amino acids to belong to specific secondary structure known in the general case is similar in Csp92 proteins. Second, the extraction of structural information from spectroscopies can only be obtained if there is no major correlation between the structures investigated. Fig. 5 already indicates it is irrelevant to attempt to separate the total β -sheet content from the antiparallel β -sheet content or the total α -helix content from the ordered α -helix content, using the proteins of Csp92.

Finally, it must be stressed that the structure of the protein described here is the structure determined in a specific condition, essentially in protein crystals. Whether the structure in solution (transmission FTIR) or on dry films (ATR-FTIR) will be identical to the structure in the crystal is a matter of concern for the user. The validity of dried films has been mostly confirmed for ATR-FTIR spectroscopy [45,56], including the effect of pH on carboxylic acid ionization [57]. Yet, in particular for disordered proteins, the experimental condition could have a significant impact on structure. Whether they remain fully disordered in measurement conditions can be addressed by monitoring hydrogen/deuterium exchange kinetics, which can be easily achieved using FTIR spectroscopy [41,58,59]. When monitoring hydrogen/deuterium exchange kinetics by FTIR spectroscopy, a fully disordered protein will fully exchange very rapidly, providing a control for full exchange and simultaneously confirming the reality of the fully disordered conformation.

In conclusion, we report here a selection of well-characterized proteins that can be easily obtained from commercial sources. The distribution of their structural features has been extensively characterized and cover a wide range of structural content. cSP92 should be very useful for the calibration of spectroscopic methods.

5. Role of funding source

Study design; collection, analysis and interpretation of data; writing the publication and decision to submit the article for publication were all completely independent from the funding sources.

CRediT authorship contribution statement

J. De Meutter was responsible for conceptualization, data curation, and formal analysis. E. Goormaghtigh obtained funding and was responsible for project administration. Writing the original draft and review and editing was equally shared by both authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Fonds de la Recherche Scientifique - FNRS under Grant n°O001518F (EOS-convention # 30467715). We thank the Walloon Region (SPW, DGO6, Belgium) for supporting the ROBOTEIN project within the frame of the EQUIP2013 program convention 1318159). E.G. is Research Director with the National Fund for Scientific Research (Belgium).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.07.001.

References

[1] Wang Y, Boysen RI, Wood BR, Kansiz M, McNaughton D, Hearn MTW. Determination of the secondary structure of proteins in different environments by FTIR-ATR spectroscopy and PLS regression. Biopolymers 2008;89:895–905. https://doi.org/10.1002/bip.21022.
[2] Hennessey Jr., J P, Johnson Jr. W. Information content in the circular dichroism of proteins. Biochem JID - 0370623 1981;20:1085–94.
[3] Prestrelski SJ, Byler DM, Liebman MN. Generation of a substructure library for the description and classification of protein secondary structure. II. Application to spectra-structure correlations in fourier transform infrared spectroscopy. Proteins Struct Funct Genet 1992;14:440–50. https://doi.org/10.1002/prot.340140405.
[4] Pribic R, van Stokkum IH, Chapman D, Haris PI, Bloemendal M. Protein secondary structure from Fourier transform infrared and/or circular dichroism spectra. AnalBiochem 1993;214:366–78.
[5] Oberg KA, Ruysschaert JM, Goormaghtigh E. The Optimization of Protein Secondary Structure Determination with Infrared and CD Spectra. EurJBiochem 2004;271:2937–48.
[6] Oberg KA, Ruysschaert JM, Goormaghtigh E. Rationally selected basis proteins: A new approach to selecting proteins for spectroscopic secondary structure analysis. ProtSci 2003;12:2015–31.
[7] Abdul-Gader A, Miles AJ, Wallace BA. A reference dataset for the analyses of membrane protein secondary structures and transmembrane residues using circular dichroism spectroscopy. Bioinformatics 2011;27:1630–6. https://doi.org/10.1093/bioinformatics/btr234.
[8] Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH - a hierarchic classification of protein domain structures. Structure 1997;5:1093–108.
[9] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–42. https://doi.org/10.1016/s0022-2836(77)80200-3.
[10] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004;32:115D–9D. https://doi.org/10.1093/nar/gkh131.
[11] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–53. https://doi.org/10.1016/0022-2836(70)90057-4.
[12] Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM. PDBsum: Structural summaries of PDB entries. Protein Sci 2018;27:129–34. https://doi.org/10.1002/pro.3289.
[13] Laskowski RA, Hutchinson EG, Michie AD, Wallace AC, Jones ML, Thornton JM. PDBsum: a web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 1997;22:488–90. https://doi.org/10.1016/S0968-0004(97)01140-7.
[14] Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
[15] Bersch B, Derfoufi K-M, De Angelis F, Auquier V, Ngonlong Ekendé E, Mergeay M, et al. Structural and Metal Binding Characterization of the C-Terminal Metallochaperone Domain of Membrane Fusion Protein SilB from *Cupriavidus metallidurans* CH34. Biochemistry 2011;50:2194–204. https://doi.org/10.1021/bi200005k.
[16] De Angelis F, Lee JK, O'Connell JD, Miercke LJ, Verschueren KH, Srinivasan V, et al. Metal-induced Conformational Changes in ZneB Suggest an Active Role of Membrane Fusion Proteins in Efflux Resistance Systems. Proc Natl Acad Sci U S A 2010;107. https://doi.org/10.1073/PNAS.1003908107.
[17] Pak JE, Ekendé EN, Kifle EG, O'Connell JD, Angelis De, Fabien T, et al. Structures of Intermediate Transport States of ZneA, a Zn(II)/proton Antiporter. Proc Natl Acad Sci U S A 2013;110. https://doi.org/10.1073/PNAS.1318705110.
[18] Fonzé, E., Charlier, P, To'th, Y, Vermeire, M, Raquet, X, Dubus, A, Frère J. TEM1 Beta-Lactamase Structure Solved by Molecular Replacement and Refined Structure of the S235A Mutant. Acta Crystallogr D Biol Crystallogr 1995;51. https://doi.org/10.1107/S0907444994014496
[19] Raussens V, Fisher CA, Goormaghtigh E, Ryan RO, Ruysschaert J. The Low Density Lipoprotein Receptor Active Conformation of Apolipoprotein E. Helix Organization in N-Terminal Domain-Phospholipid Disc Particles. J Biol Chem 1998;273. https://doi.org/10.1074/JBC.273.40.25825.
[20] Kabsch W, Sander S. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.
[21] Andersen CAF, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure 2002;10:175–84. https://doi.org/10.1016/s0969-2126(02)00700-1.
[22] Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks

and profiles. Proteins Struct Funct Genet 2002;47:228–35. https://doi.org/10.1002/prot.10082.
[23] Carter P. DSSPcont: continuous secondary structure assignments for proteins. Nucleic Acids Res 2003;31:3293–5. https://doi.org/10.1093/nar/gkg626.
[24] Kalnin NN, Baikalov IA, Venyaminov SY. Quantitative IR spectrophotometry of peptides compounds in water (H2O) solutions. III. Estimation of the protein secondary structure. Biopolymers 1990;30:1273–80.
[25] Sreerama N, Venyaminov SY, Woody R. Estimation of the Number of Alpha-Helical and Beta-Strand Segments in Proteins Using Circular Dichroism Spectroscopy. Protein Sci 1999;8. https://doi.org/10.1110/PS.8.2.370.
[26] Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. DisProt: the Database of Disordered Proteins. Nucleic Acids Res 2007;35:D786–93. https://doi.org/10.1093/nar/gkl893.
[27] Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. J Mol Biol 2004;337:635–45. https://doi.org/10.1016/j.jmb.2004.02.002.
[28] Weathers EA, Paulaitis ME, Woolf TB, Hoh JH. Insights into protein structure and function from disorder-complexity space. Proteins Struct Funct Bioinforma 2006;66:16–28. https://doi.org/10.1002/prot.21055.
[29] Sarroukh R, Cerf E, Derclaye S, Dufrêne YF, Goormaghtigh E, Ruysschaert J-M, et al. Transformation of amyloid β(1–40) oligomers into fibrils is characterized by a major change in secondary structure. Cell Mol Life Sci 2011;68:1429–38. https://doi.org/10.1007/s00018-010-0529-x.
[30] Cerf E, Sarroukh R, Tamamizu-Kato S, Breydo L, Derclaye S, Dufrene YF, et al. Antiparallel beta-sheet: a signature structure of the oligomeric amyloid beta-peptide. Biochem J 2009;421:415–23.
[31] Micsonai A, Wien F, Kernya L, Lee Y-H, Goto Y, Réfrégiers M, et al. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. Proc Natl Acad Sci 2015;112:E3095–103. https://doi.org/10.1073/pnas.1500851112.
[32] Nevskaya NA, Chirgadze YN. Infrared spectra and resonance interactions of amide-I and II vibrations of alpha-helix. Biopolymers 1976;15:637–48. https://doi.org/10.1002/bip.1976.360150404.
[33] Goormaghtigh E, Cabiaux V, Ruysschaert JM. Determination of soluble and membrane protein structure by Fourier transform infrared spectroscopy. I. Assignments and model compounds. SubcellBiochem 1994;23:329–62.
[34] Chirgadze YN, Nevskaya NA. Infrared spectra and resonance interaction of amide-I vibration of the paraellel-chain pleated sheets. Biopolymers 1976;15:627–36.
[35] Chirgadze YN, Fedorov OV, Trushina NP. Estimation of amino acid residue side-chain absorption in the infrared spectra of protein solutions in heavy water. Biopolymers 1975;14:679–94.
[36] Venyaminov SYY, Kalnin NN. Quantitative IR spectrophotometry of peptides compounds in water (H2O) solutions. I. Spectral parameters of amino acid residue absorption band. Biopolymers 1991;30:1243–57.
[37] Barth A. The infrared absorption of amino acid side chains. Prog Biophys Mol Biol 2000;74:141–73.
[38] Barth A. Infrared spectroscopy of proteins. Biochim Biophys Acta 2007;1767:1073–101.
[39] Cabiaux V, Agerberth B, Johansson J, Homble F, Goormaghtigh E, Ruysschaert JM. Secondary structure and membrane interaction of PR-39, a Pro+Arg-rich antibacterial peptide. EurJBiochem 1994;224:1019–27.
[40] Goormaghtigh E, Ruysschaert JM, Raussens V. Evaluation of the information content in infrared spectra for protein secondary structure determination. BiophysJ 2006;90:2946–57.
[41] Raussens V, Ruysschaert JM, Goormaghtigh E. Analysis of H-1/H-2 exchange kinetics using model infrared spectra. ApplSpectrosc 2004;58:68–82.
[42] Fabian H, Lasch P, Naumann D. Analysis of biofluids in aqueous environment based on mid-infrared spectroscopy. J Biomed Opt 2005;10:. https://doi.org/10.1117/1.1917844031103.
[43] Zonderman LWIAPSR-BJ. Native Measurement of a Biotherapeutic without Interference from Excipients Using Microfluidic Modulation Spectroscopy n.d.
[44] Keiderling TA. Structure of Condensed Phase Peptides: Insights from Vibrational Circular Dichroism and Raman Optical Activity Techniques. Chem Rev 2020:acs.chemrev.9b00636. https://doi.org/10.1021/acs.chemrev.9b00636
[45] Goormaghtigh E, Raussens V, Ruysschaert JM. Attenuated total reflection infrared spectroscopy of proteins and lipids in biological membranes. BiochimBiophysActa 1999;1422:105–85.
[46] Srour B, Bruechert S, Andrade SLA, Hellwig P. Secondary Structure Determination by Means of ATR-FTIR Spectroscopy. Methods Mol Biol 2017;1635:195–203. https://doi.org/10.1007/978-1-4939-7151-0_10.
[47] De Meutter J, Vandenameele J, Matagne A, Goormaghtigh E. Infrared imaging of high density protein arrays. Analyst 2017;142:1371–80. https://doi.org/10.1039/c6an02048h.
[48] De Meutter J, Derfoufi MK, Goormaghtigh E. Analysis of protein microarrays by FTIR imaging. Biomed Spectrosc Imaging 2016;5:145–54.
[49] Bonnier F, Rubin S, Debelle L, Venteo L, Pluot M, Baehrel B, et al. FTIR protein secondary structure analysis of human ascending aortic tissues. J Biophotonics 2008;1:204–14.
[50] Kuepper C, Großerueschkamp F, Kallenbach-Thieltges A, Mosig A, Tannapfel A, Gerwert K. Label-free classification of colon cancer grading using infrared spectral histopathology. Faraday Discuss 2016;187:105–18. https://doi.org/10.1039/c5fd00157a.
[51] Ami D, Mereghetti P, Leri M, Giorgetti S, Natalello A, Doglia SM, et al. A FTIR microspectroscopy study of the structural and biochemical perturbations induced by natively folded and aggregated transthyretin in HL-1 cardiomyocytes. Sci Rep 2018;8:12508. https://doi.org/10.1038/s41598-018-30995-5.
[52] Dazzi A, Glotin F, Carminati R. Theory of infrared nanospectroscopy by photothermal induced resonance. J Appl Phys 2010;107:. https://doi.org/10.1063/1.3429214124519.
[53] Andersen CAF, Rost B. Secondary Structure Assignment. Struct Bioinforma 2005:339–63. https://doi.org/10.1002/0471721204.ch17.
[54] Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:222–45. https://doi.org/10.1021/bi00699a002.
[55] Fujiwara K, Toda H, Ikeguchi M. Dependence of alpha-helical and beta-sheet amino acid propensities on the overall protein fold type. BMC Struct Biol 2012;12:18. https://doi.org/10.1186/1472-6807-12-18.
[56] Goormaghtigh E, Gasper R, Benard A, Goldsztein A, Raussens V, Bénard A. Protein secondary structure content in solution, films and tissues: Redundancy and complementarity of the information content in circular dichroism, transmission and ATR FTIR spectra. Biochim BiophysActa-Proteins Proteomics 2009;1794:1332–43. https://doi.org/10.1016/j.bbapap.2009.06.007.
[57] Goormaghtigh E, de-Jongh HH, Ruysschaert JM. Relevance of protein thin films prepared for attenuated total reflection Fourier transform infrared spectroscopy: significance of the pH. ApplSpectrosc 1996;50:1519–27
[58] Vigano C, Smeyers M, Raussens V, Scheirlinckx F, Ruysschaert JM, Goormaghtigh E. Hydrogen-deuterium exchange in membrane proteins monitored by IR spectroscopy: A new tool to resolve protein structure and dynamics. Biopolymers 2004;74:19–26.
[59] de Jongh HH, Goormaghtigh E. Ruysschaert JM, de-Jongh HH. Amide-proton exchange of water-soluble proteins of different structural classes studied at the submolecular level by infrared spectroscopy. Biochemistry 1997;36:13603–10.