Turkish Computational and Theoretical Chemistry

Turkish Comp Theo Chem (TC&TC)

Volume(Issue): 7(2) – Year: 2023 – Pages: 70-83

e-ISSN: 2602-3237

https://doi.org/10.33435/tcandtc.1195150

Received: 22022 ***Accepted:*** 05.06.2022 *Research Article*

Web Server-based structure prediction as a supplementary tool for basic and acidic FGF secondary structure analysis using FTIR spectroscopy and a case study comparing curve-fit with the model-based structure inspection of the FTIR data

Filiz Korkmaz[a], Ayca Mollaoglu[b], Yekbun Adiguzel[c, 1]

[a]Physics Group, Faculty of Engineering, Atilim University, Ankara, Turkey.
[b]Department of Physiology, School of Medicine, Altinbas University, Istanbul, Turkey.
[c]Department of Medical Biology, School of Medicine, Atilim University, Ankara, Turkey.

Abstract: Fourier Transform Infrared (FTIR) spectroscopy can provide relative proportion of secondary structure elements in a protein. However, extracting this information from the Amide I band area of an FTIR spectrum is difficult. In addition to experimental methods, several protein secondary structure prediction algorithms serving on the Web can be used as supplementary tools requiring only protein amino acid sequences as inputs. In addition, web-server based docking tools can provide structure information when proteins are mixed and potentially interacting. Accordingly, we aimed to utilize web-server based structure predictors in fibroblast growth factor (FGF) protein structure determination through the FTIR data. Seven such predictors were selected and tested on basic FGF (bFGF) protein, to predict FGF secondary structure. Results were compared to available structure-files deposited in the Protein Data Bank (PDB). Then, FTIR spectra of bFGF and the acidic form of the protein with 50 folds more bovine serum albumin as carrier protein (1FGFA/50BSA) were collected. Optimized Amide I curve-fit parameters of bFGF with low (<5) root mean square deviation (RMSD) in the PDB data and the predictions were obtained. Those parameters were applied in curve-fitting of 1FGFA/50BSA data. Secondary structure was inspected also through applying models derived from the previously established methods. Results of model-based secondary structure estimation from FTIR data were compared with secondary structure calculated as 1 part contribution from 1FGFA/1BSA complex and 49 parts contribution from BSA. Complex structure was obtained through docking. RMSD in the PDB data and the predictions were respectively 3.05 and 2.39 with the optimized parameters. Those parameters did not work well for the 1FGFA/50BSA data. Models are better in this case, wherein one model (Model-1') with the lowest average RMSD has 8.38 RMSD in the bFGF and 4.78 RMSD in the 1FGFA/50BSA structures. Model-based secondary structure predictions are better for determining bFGF and 1FGFA/50BSA secondary structures through the curve-fit approach that we followed, under non-optimal conditions like protein/BSA mixtures. Web servers can assist experimental studies investigating structures with unknown structures. Any web-based structure prediction supporting the experimental results would be enforcing the findings, but the unsupported results would not necessarily falsify the experimental data.

Keywords: Protein structure, secondary structure, structure prediction, ATR FTIR, FGF, protein docking

1. Introduction

Secondary structure determination and detecting changes in them are of profound importance since the function of a protein is directly related to its secondary and tertiary structure [1–5]. Thus, experimental techniques giving information on both are valuable. Fourier Transform Infrared Spectroscopy (FTIR) is among those and preferred

[1] Corresponding Authors

e-mail: yekbun.adiguzel@atilim.edu.tr

due to its ease in use and speed of data acquisition. On the other hand, secondary structure analysis requires mathematical processing of spectra. There are several mathematical methods reported in literature that are in-house developed and are not commercially available such as the utilization of neural network [6,7], multivariate regression methods and other algorithms [8,9], method of co-fitting [10,11], etc. There is also curve-fitting option in certain brands of commercially available software that are primarily present for the spectrometer control. However, the success of this method depends heavily on the user experience and on the initial parameter set. Therefore, any improvement in structure prediction with FTIR is greatly appreciated.

In this study, we evaluated seven Web Server-based predictors and compared their performances for the secondary structure analysis of the basic fibroblast growth factor (bFGF) protein [12], to combine with spectroscopy, during curve-fitting of protein FTIR data. The use of Web server-based predictors is highly attractive as supplements to experimental techniques like NMR, X-ray crystallography and FTIR because only the primary structure of the protein is required as input. Using computational approaches in combination with FTIR spectroscopy [7,13,14] has much to offer in protein secondary structure determination and structure-function relation, and it was known that studies involving FGF structure can benefit from FTIR spectroscopy [15,16]. FGF is a protein involved in distinct biological activities [12,17–22]. Following curve-fit analysis approach, we present a better performing model-based approach in this occasion. However, secondary structure predictions through amino acid sequences do not predict secondary structure changes upon interactions. Another type of web-based resource can serve with this purpose, enabled by the improvements in docking algorithms and computer hardware [23], to generate protein-interaction model, as a new knowledge [24]. Accordingly, predictions of the protein complex structures are possible e.g., with Hex [25] and server-based GRAMM [23]. Both examples perform protein docking to estimate interactions, thorough which any change in protein secondary structure percentages in the interacting protein complexes can be obtained. Other than enabling the used to download softwares, such services can also automate docking, which was first performed in 1978 [26]. Present docking approaches employ different search and scoring strategies. The exemplified Hex utilizes grid-free spherical polar Fourier approach [27] with rotational rather than translational correlations to be calculated with one-dimensional fast Fourier transforms, i.e., FFTs [28]. It first generates candidate solutions which are then rescored using shape only or if required, shape and electrostatic, correlations [25]. The other mentioned-tool, GRAMM-X (discontinued, the new GRAMM Docking Web Server can be found at http://gramm.compbio.ku.edu/), is also FFT-based.

2. Computational Method

2.1. PDB data, prediction tools, and prediction performance calculations

The bFGF protein structure deposited in the Protein Data Bank (PDB) with the ID 1BFF was selected as an arbitrary reference to compare with the other structures (PDB ID: 2BFH, 1BLD, 1IIL:A, 1FGA, 1II4:A, 1EV2:A, 1BFB, 1FQ9, 1BAS). Secondary structure prediction tools that were used in this study were SOPM, Yaspin, 1D-PSPS, PredictProtein, DPM, DSC, and SSpro8. We performed calculation of the root mean square deviation (RMSD) of alpha (α)-helices, beta (β)-sheets, turns, and (random) coils' secondary structure percentages of PDB data and prediction results from the experimental data, and the generation of confusion matrices for the predictor-performances in estimating the structure of bFGF (PDB ID: 1BFF). Details of these calculations and the tools are provided in the supplementary file.

2.2. Curve-fit vs model-based secondary structure, based on RMSD to 1:49 (docked complex:BSA)

We obtained attenuated total reflectance (ATR) FTIR spectra of bFGF and acidic FGF (FGFA) with 50 folds more bovine serum albumin as carrier (1FGFA/50BSA) and inspected their secondary structure percentages through curve-fitting and model-based approaches, as detailed in the supplementary file. Curve-fit was based on the approach developed earlier [4,29,30] while models were derived through the models published by Goormaghtigh et al. [9,31]. RMSD values were calculated with the 1:49 docked complex and BSA, as the reference. For docking, A chains of PDBs with IDs 2AFG (for FGFA) and 3V03 (for BSA) were used.

Hex [25] was used for docking. We downloaded Hex software, which also has a server (http://hexserver.loria.fr). Hex docking program has steric scan and final scan stages, wherein the former is the fast, low-resolution phase while the latter is the higher resolution, smaller step-lengths, of search in the top scoring orientations, within a user-defined "distance range", which was 40 Angstrom, by default. Substeps of 2 was used for the latter. Angular ranges of the receptor and ligand are also user-defined, along with the angular sizes of the steps that cover those angular ranges. Those were 180 and 7.5 degrees correspondingly, for the ranges and the step sizes, by default. Before docking, chains B-D and its solvent molecules and related connections were removed from the PDB (2AFG) file of FGFA. Similarly, chain B and its solvent molecules and related connections were removed from the PDB (3V03) file of BSA. Docking was performed after loading those files as ligand and receptor, correspondingly, and following a macro docking, due to the large sizes of the molecules. After docking, Stride was utilized as the program-based secondary structure recognition from the atomic coordinates [32], to obtain the secondary structure of the complexed proteins in the best docking. It is mentioned above, but in detail, the best result was utilized through estimating its contribution as 1/50 of the measured sample, assuming one-to-one interaction. Accordingly, 1/50 of the sample structure was the docked structure, and 49/50 of the sample structure was the secondary structure of A chain in the PDB (3V03) file of BSA, to comply with the BSA structure used for docking. Secondary structure information obtained through using Stride, as the post-docking analyses. To understand the secondary structure of the interacting proteins in the model, Stride was utilized as the program-based secondary structure recognition from the atomic coordinates [32]. RMSD of the FTIR data-based secondary structures to those obtained as described, was used to evaluate the performance of the applied-models to calculate secondary structures from the FTIR data. Those applied models were additionally modified to diminish RMSD.

GRAMM docking server was also used for protein–protein docking, which generated a docked complex [23]. Ten docked complex models were generated by GRAMM for the same processed PDB files of the two proteins and secondary structure information obtained through using Stride, as before, for the best complex (model_1). BSA and FGFA complex images were generated by Swiss-PdbViewer v4.1.0 at http://www.expasy.org/spdbv. We compared Hex with GRAMM for the secondary structure percentages of the top results.

3. Results and discussion

Protein structure prediction algorithms are already tested with a set of proteins so that their performances are proven before being released for public use. However, usually one structure per protein is assumed and used to construct such sets, although generally several structures are deposited for one protein in the PDB. This is partly because the data resolution is different, yielding slightly different structures and/or partly because of the varying experimental conditions. Yet, it is also true that the protein structures are sensitive to changes in the environment and different PDB files may well be representing this flexibility. Using structure predictors is quite tempting since only the primary sequence of the protein is required as input data. After that, the predicted structure could be used for the starting point of a curve-fitting processes.

The secondary structure prediction tools that were used in this work are SOPM [33], Yaspin [34], 1D-PSPS [35], PredictProtein [36], DPM [37], DSC [38], and SSpro8 [39]. An important issue to note here is that the predictors Yaspin, 1D-PSPS, PredictProtein, and DSC do not predict the turns. Yet, they were preferred to be included in the study since such type of analysis still gives a good approach for the protein secondary structure analysis and prediction studies, as well as the FTIR based secondary structure analysis. So, the prediction of helices, sheets, and coils without the turns content in a reliable manner is still beneficial. Here, the predictors were used to predict the secondary structure of bFGF, of which the X-ray crystallography structure data is available at 2.0 Å resolution (PDB ID: 1BFF). This structure (Figure 1) was compared with the prediction results to test the performance of these predictors. There are also different experimentally obtained structure data of the same protein at different resolutions. Accordingly, by using ten different structure data deposited in PDB, RMSD among experimental results were determined by comparing the secondary structure percentages of the 1BFF structure to that of the other structures. Furthermore, determination of the deviation among several PDB data of the bFGF protein was used to judge the performance of Web-based predictors. Then, a similar RMSD calculation was done with the prediction results.

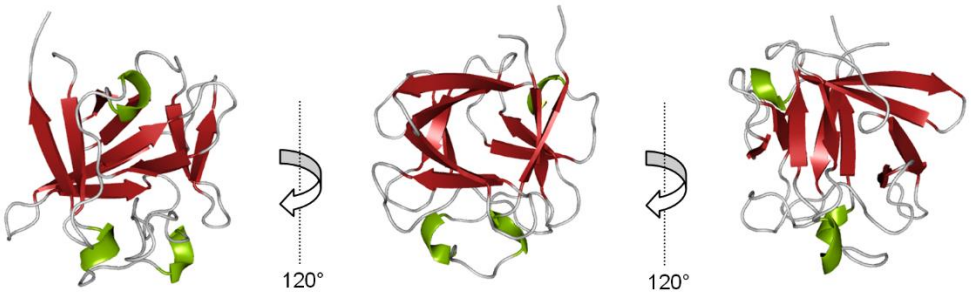

Figure 1. The basic fibroblast growth factor (bFGF) structure with the PDB ID 1BFF from three different angles. It is mainly a β-sheet protein (53%) shown in red with only 9% α-helix, shown in green.

Residue-based comparison of the secondary structure of bFGF among ten different PDB data is presented in Alignment S1, supplementary file. The primary structure of the protein is obtained from the Uniprot file, coded P09038. Alignment S2 shows residue-based comparison of the predicted secondary structures of the same sequence by the predictors, together with the reference PDB data on top.

3.1. Deviations in the predicted secondary structure percentages: Comparison of different PDB data of bFGF

The secondary structure of bFGF is categorized under four main secondary structure types as helices, sheets, turns and coils. Secondary structure percentages by PDB data are given in Table 1, where the first row shows the structure given by the 1BFF data. Deviations of the secondary structure percentages of the other PDB files' data from that of 1BFF are represented with the RMSD1 values. We have then calculated the RMSD of each structure from each of the other structures separately and took their averages (RMSD2, Table 1). There, we see deviations of secondary structure percentages of all the other PDB-based bFGF structures are highest on average from the protein with PDB ID: 1FGA (RMSD2: 5.13), and lowest on average from the protein with PDB ID: 1BFB (RMSD2: 2.33) and 1BFF (RMSD2: 2.42).

According to the RMSD1 values in Table 1, the deviation is, in general, proportional to the resolution of the X-ray crystallography data. It may be expected that the RMSD1 value increases as the data resolution deviates more and more from that of 1BFF, which is 2.0 Å. However, this is not the case. For instance, RMSD1 of 1EV2 is 3.47 although the resolution of the 1EV2 data is 2.2 Å, which is close to that of 1BFF. The structure data of 1FGA (resolution 2.2 Å) deviates the most from 1BFF with an RMSD1 value of 4.14. The difference is mainly in the fraction of sheets and coils. The closest data with 0.55 RMSD1 belongs to the 1BFB (resolution 1.9 Å), which yields ~1% difference in turns and coils, while the fractions of sheets and helices match precisely with those of 1BFF. In general, structural data that is obtained by the same experimental method and at similar resolutions is not the same. Results deviate from the 1BFF secondary structure percentages by an RMSD1 of minimum 0.55 and maximum 4.14.

Table 1. Secondary structure percentages of 10 experimental data of bFGF. RMSD1 values represent deviation of the other experimental secondary structure percentages from that of the 1BFF. RMSD2 values are the averages of the remaining experimental structures' RMSDs to the respective experimental structure. The average of the RMSD1 values is 2.45 while that of the RMSD2 values is 3.21.

PDB ID	Sheets%	Helices%	Turns%	Coils%	RMSD1	RMSD2
1BFF	**41.09**	**6.98**	**19.38**	**32.56**		2.42

2BFH	41.09	9.30	18.60	31.01	1.45	2.93
1BLD	40.31	2.33	19.38	37.98	3.59	4.28
1IIL:A	42.64	6.98	20.93	29.46	1.90	2.79
1BAS	41.09	9.30	18.60	31.01	1.45	2.93
1BFB	41.09	6.98	20.16	31.78	0.55	2.33
1FGA	35.66	6.98	18.60	38.76	4.14	5.13
1FQ9	42.64	2.33	23.26	31.78	3.15	3.20
1II4:A	42.64	4.65	22.48	30.23	2.39	2.78
1EV2:A	42.64	2.33	24.03	31.01	3.47	3.42

Table 2. Secondary structure percentages and RMSD values representing the deviations of prediction results from the experimental (exp.) data of 1BFF in the PDB. 1BFF structure content and best prediction results are written in bold**.**

Predictor or 1BFF	Sheets%	Helices%	Turns%	Coils%	RMSD
1BFF (exp.)	**41.09**	**6.98**	**19.38**	**32.56**	
SOPM	**41.09**	**6.20**	**17.83**	**34.88**	**1.45**
Yaspin	47.29	0.00	N/A	52.71	5.41*
1D-PSPS	51.16	3.10	N/A	45.74	7.19*
Predict-Protein	51.16	0.00	N/A	48.84	7.30*
DPM	9.30	37.21	11.63	41.86	22.75
DSC	49.61	0.00	N/A	50.39	6.42*
SSpro8	**41.09**	**4.65**	**23.26**	**31.01**	**2.39**

* RMSD of the predictors that do not provide the turns structure in their outputs are calculated based on 3 secondary structure types, as explained in the Extended Materials and Methods at the supplementary file.

Table 3. RMSD values representing the deviations of the predictor results from all the inspected experimental data for bFGF, separately. Average of the RMSD values (RMSDav.) for each predictor is at the last column. Predictor results with the lowest RMSD to the structure with the given PDB ID as the respective column title are written in bold.

PDB ID of bFGF:		1BFF	2BFH	1BLD	1IIL:A	1BAS	1BFB	1FGA	1FQ9	1II4:A	1EV2:A	$RMSD_{av.}$
Predicted by	**SOPM**	**1.45**	**2.51**	**2.63**	3.24	**2.51**	**1.98**	**3.38**	3.76	3.47	4.21	2.91
	Yaspin*	5.41	6.70	5.02	5.02	6.70	5.41	8.28	3.29	3.80	3.29	5.29
	1D-PSPS*	7.19	7.19	9.19	6.04	7.19	7.19	11.41	7.30	6.42	7.30	7.64
	PredictProtein*	7.30	7.93	8.08	6.42	7.93	7.30	10.98	6.23	6.04	6.23	7.45
	DPM	22.75	22.11	23.74	23.80	22.11	22.90	20.41	25.32	24.62	25.49	23.33
	DSC*	6.42	7.30	6.85	5.70	7.30	6.42	9.87	5.02	5.02	5.02	6.49
	SSpro8	2.39	3.29	4.17	**1.98**	3.29	**1.98**	5.40	**1.45**	**0.95**	**1.45**	**2.63**

* RMSD values of the predictors that do not provide the turns structure in their outputs are calculated based on 3 secondary structure types, as explained in the Extended Materials and Methods at the supplementary file.

Table 4. Confusion matrices of secondary structure predictions for bFGF and 1BFF as the actual structure. True predictions are written in bold. Total structure percentages in columns of each matrix are written in the last row. RMSD of each predictor-predictions from the actual values is written in the last column.

			Actual (1BFF)				Predicted Total%	RMSD of predicted to total%
			Sheets%	Helices%	Turns%	Coils%		
Predictor rows of 7 separate confusion matrices	**SOPM**	Sheets%	**30.23**	2.33	1.55	6.98	41.09	1.45
		Helices%	3.88	**0.00**	1.55	0.78	6.20	
		Turns%	3.88	0.78	**12.40**	0.78	17.83	
		Coils%	3.10	3.88	3.88	**24.03**	34.88	
	Yaspin*	Sheets%	**35.66**	2.33	N/A	9.30	47.29	5.41*
		Helices%	0.00	**0.00**	N/A	0.00	0.00	
		Turns%	N/A	N/A	**N/A**	N/A	N/A	

		Coils%	5.43	4.65	N/A	**42.64**	52.71	
	1D-PSPS*	Sheets%	**35.66**	3.10	N/A	12.40	51.16	7.19*
		Helices%	3.10	**0.00**	N/A	0.00	3.10	
		Turns%	N/A	N/A	**N/A**	N/A	N/A	
		Coils%	2.33	3.88	N/A	**39.53**	45.74	
	Predict-Protein*	Sheets%	**34.11**	1.55	N/A	15.50	51.16	7.30*
		Helices%	0.00	**0.00**	N/A	0.00	0.00	
		Turns%	N/A	N/A	**N/A**	N/A	N/A	
		Coils%	6.98	5.43	N/A	**36.43**	48.84	
	DPM	Sheets%	**6.20**	0.00	0.78	2.33	9.30	22.75
		Helices%	21.71	**1.55**	3.88	10.08	37.21	
		Turns%	0.78	2.33	**5.43**	3.10	11.63	
		Coils%	12.40	3.10	9.30	**17.05**	41.86	
	DSC*	Sheets%	**35.66**	2.33	N/A	11.63	49.61	6.42*
		Helices%	0.00	**0.00**	N/A	0.00	0.00	
		Turns%	N/A	N/A	**N/A**	N/A	N/A	
		Coils%	5.43	4.65	N/A	**40.31**	50.39	
	SSpro8	Sheets%	**38.76**	0.00	0.00	2.33	41.09	2.39
		Helices%	0.00	**2.33**	1.55	0.78	4.65	
		Turns%	0.00	3.10	**17.05**	3.10	23.26	
		Coils%	2.33	1.55	0.78	**26.36**	31.01	
Actual Total%			41,09	6,98	19.38	32.56		

* RMSD of the predictors that do not provide the turns structure in their outputs are calculated based on 3 secondary structure types, as explained in the Extended Materials and Methods at the supplementary file. Also, N/A is written in the Turns% row of those predictors' prediction results and the turns of the reference that are predicted as the other structure types are included in the respective false predictions of coils.

Table 5. Comparison of true prediction rates, precisions, and accuracies of confusion matrices of secondary structure predictions for bFGF (1BFF) as the actual structure. The highest true prediction and precision in each structure type, and the highest overall accuracy are observed in case of SSpro8 prediction results, except for the coils.

Structure type:		**Sheets%**		**Helices%**		**Turns%**		**Coils%**		**Accuracy**
		True Prediction Rate	**Precision**	**True Prediction Rate**	**Precision**	**True Prediction Rate**	**Precision**	**True Prediction Rate**	**Precision**	
Predicted by	**SOPM**	0.74	0.74	0.00	0.00	0.64	0.70	0.74	0.69	0.67
	Yaspin*	0.87	0.75	0.00	No data	N/A	N/A	0.82	0.81	0.78
	1D-PSPS*	0.87	0.70	0.00	0.00	N/A	N/A	0.76	0.86	0.75
	PredictProtein*	0.83	0.67	0.00	No data	N/A	N/A	0.70	0.75	0.71
	DPM	0.15	0.67	0.22	0.04	0.28	0.47	0.52	0.41	0.30
	DSC*	0.87	0.72	0.00	No data	N/A	N/A	0.78	0.80	0.76
	SSpro8	**0.94**	**0.94**	**0.33**	**0.50**	**0.88**	**0.73**	0.81	0.85	**0.84**

*In case of predictors that do not provide the turns as output, Turn% row of the predictors' results was eliminated and the turns of the reference that were predicted as the other structure types were included in the respective, false coils predictions. Accordingly, respective eliminated values with N/A in the matrices were eliminated also from the calculations of the true secondary structure rates, precision, and accuracy values.

Table 6. Comparison of true prediction rates, precisions, and accuracies of confusion matrices of predictions without the turns, for bFGF. The 1BFF is still accounted as the actual structure, though. The highest true prediction and precision in each structure type, and the highest overall accuracy are observed in case of SSpro8 prediction results.

Structure type:	**Sheets%**		**Helices%**		**Turns%**		**Coils%**		**Accuracy**
	True Prediction Rate	**Precision**	**True Prediction Rate**	**Precision**	**True Prediction Rate**	**Precision**	**True Prediction Rate**	**Precision**	

Predicted by	**SOPM**	0.74	0.74	0.00	0.00	N/A	N/A	0.79	0.78	0.71
	Yaspin	0.87	0.75	0.00	No data	N/A	N/A	0.82	0.81	0.78
	1D-PSPS	0.87	0.70	0.00	0.00	N/A	N/A	0.76	0.86	0.75
	PredictProtein	0.83	0.67	0.00	No data	N/A	N/A	0.70	0.75	0.71
	DPM	0.15	0.67	0.22	0.04	N/A	N/A	0.67	0.65	0.43
	DSC	0.87	0.72	0.00	No data	N/A	N/A	0.78	0.80	0.76
	SSpro8	**0.94**	**0.94**	**0.33**	**0.50**	N/A	N/A	**0.91**	**0.87**	**0.88**

* Turn% row of the predictor-results was eliminated in all the predictors confusion matrices and the turns of the reference that were predicted as the other structure types were included in the respective false predictions of coils, for comparability with the results of predictors without turns as an output. Accordingly, respective eliminated values with N/A in the matrices were excluded also from the calculations of the true secondary structure rates, precisions, and accuracies.

Average of RMSD1 values in Table 1 is 2.45 while that of the RMSD2 values is 3.21. This difference indicates that the structure of 1BFF is deviating from the other bFGF structures to a lesser extent in general, in comparison to the deviation of the other bFGF structures from the remaining. According to the RMSD2 values, 1BFB is the least and 1FGA is the most deviating one from the other bFGF structures. Of note, all the other PDB structures except 1FGA and 2BFH have serine residues instead of cysteine residues at the 211th and the 229th positions. However, substituting serine instead of cysteine does not seem to induce a consistent secondary structure change.

Another factor that contributed to the variation among the experimental data might be the slight changes in the total lengths of the analysed proteins, measurement conditions of the proteins and the utilisation of techniques other than the X-ray crystallography, in obtaining those structures. All such sources of structural alterations can be considered as experimental variations. As an example, it is observed in Table 1 that the structure that deviates the most from the 1BFF, thus having the highest RMSD1 and RMSD2 values, belongs to the 1FGA, wherein the heparin binding sites of the protein was determined by replacing the ammonium sulphate in the crystallisation medium with the ammonium selenite, for the electron-denser selenite ion [40]. Also, 1BLD structure was obtained with NMR; 1EV2:A had ligand binding domains in the analysed sample; and 1FQ9 is the crystal structure of a ternary FGF-FGFR-heparin complex.

Comparison of the prediction results with the 1BFF structure

According to the secondary structure information provided in 1BFF, there are 6.98% helices, 41.09% sheets, 19.38% turns, and 32.56% coils (Table 2, 1st row). Secondary structure prediction of the same sequence by SOPM (Table 2) yields the best results (RMSD 1.45) compared to 1BFF. Maximum deviation is ~2% in turns and coils. RMSD values of both SSpro8 and SOPM prediction results are less than 2.5 and comparable to the range of deviations in the PDB data (Table 1). On the other hand, algorithms like Yaspin, Predict-protein and DSC predict bFGF as an all-beta sheet protein with RMSD values within 5.41–7.30. Among the results, DPM predicts bFGF as an overwhelmingly helical protein with few residues in sheets and turns that deviate to the most from the PDB data with an RMSD of 22.75.

Comparison of the prediction results with all the inspected PDB data of the bFGF

The best outcomes are obtained by the predictors SOPM and SSpro8 when all the predictors are tested for all the inspected PDB data of bFGF (Table 3). Yaspin, 1D-PSPS, PredictProtein, and DSC do not predict turns. Therefore, their average RMSD were calculated by omitting the turns. SOPM and SSpro8 still seems to be the best performing predictors for all the inspected PDB data of bFGF even if turns and coils are counted together in case of all predictors (data not shown).

Confusion matrices

When confusion matrices for the predictor-performance compared to 1BFF are generated, it reveals that the faulty assignments are quite high (Table 4, Table 5). Evidently, source of previously presented good predictions with low RMSD values are the compensating changes in the secondary structure elements since some incorrect predictions in one direction are compensated by those in the other direction (e.g., sheet predicted as helices compensated by helices, coils, or turns, predicted as sheets). The highest true prediction and precision in

each structure type, and the highest overall accuracy are given by SSpro8, except for the coils (Table 5). The reason that coils was not among those best predictions is the elimination of the turns from the calculations of the predictors without the turns in their outputs. Turns were included in the coils content in those. Therefore, wrong predictions of turns were automatically eliminated from their predictions. The same calculation for all predictors leads to the highest true prediction and precision in each structure type and the highest overall accuracy in SSpro8 (Table 6).

3.2. Curve-fit:

Protein spectra were obtained by measuring the powder form of the protein to eliminate the need of water or buffer subtraction from the spectrum or the need to dry the sample on the surface. Inspection of the FGF secondary structure experimental data through curve-fit was performed by obtaining a parameter set (Table S4) that resulted in the optimized parameters (Table S5). Curve-fit result of the spectral data of bFGF is presented in Figure 2 and Table 7.

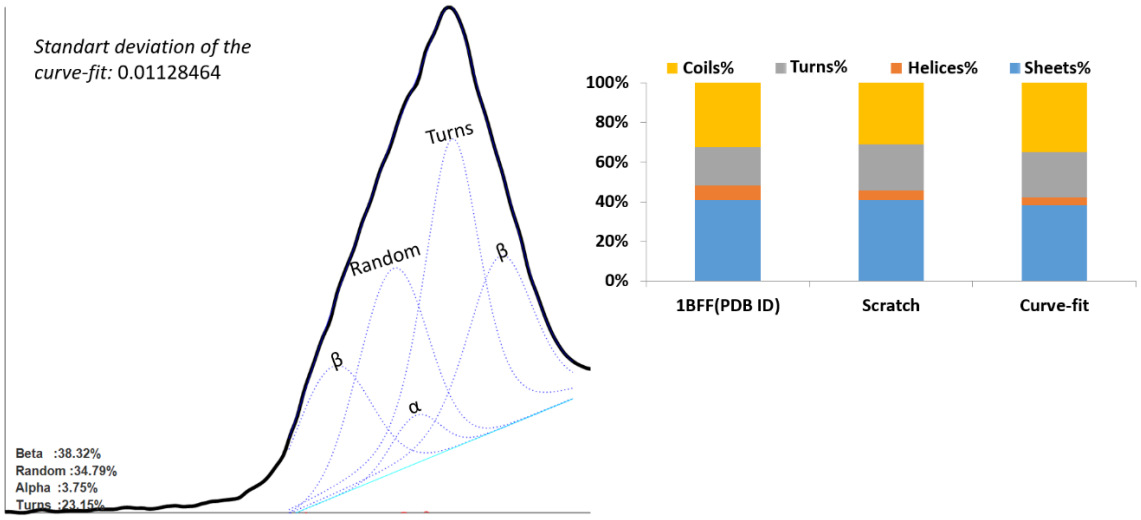

Figure 2. Curve-fit of bFGF ATR FTIR absorbance spectrum at 1600-1700 cm^{-1}, by using KineticsR12 software, through optimizing the parameter set according to the 1BFF (PDB ID) structure. Inset displays comparison of curve-fit results with the PDB structure 1BFF and with the prediction results of Scratch. Alpha (α) and beta (β) signs stand for (α-)helices and (β-)sheets.

Table 7. Comparison of curve-fit results with the bFGF (PDB structure 1BFF) and with the prediction results of Scratch.

bFGF structure-source	β-sheets%	α-helices%	Turns%	Coils%	*RMSD1	*RMSD2
1BFF (PDB ID)	41.09	6.98	19.38	32.56		
Scratch	41.09	4.65	23.26	31.01	2.39	
Curve-fit	38.32	3.75	23.15	34.79	3.05	2.39

*RMSD1 is to the 1BFF and RMSD2 is to the structure predicted by Scratch.

Computational protein-protein interaction results of FGFA and BSA to calculate 1FGFA/50BSA by taking 2% of the structure from the 1FGFA/1BSA complex and 98% of the structure from the BSA led only to a slight shift from the BSA structure. Besides, 1-to-1 interaction is an assumption, which may not be the case, and the dynamic stated most likely involves a distribution of interactions, ranging from no interaction to multiple interactions. Yet, 1FGFA/1BSA complex had significant change in the FGFA structure with an RMSD of 18.84 from the PDB (2AFG:A) structure, while it was not the case for BSA, with an RMSD of 2.82 from the PDB (3V03:A) structure. Secondary structure of FGFA in the complex was 43.41 (sheets), 2.33 (helices), 44.19 (turns), and 10.08 (coils) while that of the PDB (2AFG:A) structure was 37.86 (sheets), 8.57 (helices), 17.86 (turns), and 35.71 (coils). The highest difference in the turns and coils is indicating stabilization of the structure in the BSA-complexed

form. This result proves that modelling rather than a simple calculation of the structure in the protein mixtures is feasible, although requires evidently more time, but would be necessary when the proteins have relatively equal shares in the mixture, which was not the case in this study. In support, model by GRAMM also had the same secondary structure. Energy of the Hex model was lower, but this difference apparently did not alter the secondary structures in the two of the lowest energy models by Hex and GRAMM.

Optimised parameters were applied to the measurement result of the 1FGFA/50BSA. RMSD of the results of that curve-fit was high (data row-2, Table 8). This high RMSD was expected because the curve-fit parameters were not optimized for BSA. Model-based secondary structure inspection approaches developed by Goormaghtigh et al. [9,31] were utilized and RMSD from the 1FGFA/50BSA structure improved (data row-3 to 7, Table 8). Modifications improved the models, as observed through the RMSD values of model-1' and model-3' compared to model-1 and model-3, respectively (Table 8). This study cannot establish generalizability of the improvement due to the modified model, which awaits to be studied. However, the model works also for the bFGF (Table 9) and the modification of the model was obtained by improving the results of both bFGF and FGFA data. Results of model-1' had the lowest average RMSD (7.39) in the selected PDB structure (1BFF) and the calculated PDB-based structure (1FGFA/50BSA). Average RMSD of the structure percentages obtained through models (Table 8) from the structures of the PDB data in Table 1 are 10.78±1.70 for model-1, 9.23±1.56 for model-1', 10.58±1.53 for model-2, 7.81±1.80 for model-3, and 9.74±1.55 for model-3'. For comparison, in relation to model-1, average of the standard deviations of four secondary structures in Table 2 (direct valid) of Goormaghtigh et al. [9] is 5.78 and the RMSD calculated from the standard deviations is 6.03. While regarding model 2 and model-3, average of the standard deviations of four secondary structures in Table 1 of Goormaghtigh et al. [31] is 6.55 and the RMSD calculated from the standard deviations is 6.82. Higher RMSD values of our results was expected considering that the data was collected with a different device, with different spectral parameters, sampling method, and in different conditions. Previous approach utilizing curve-fit had no flexibility in the measured protein content while the models established by Goormaghtigh et al. [9,31] proved to be better in that case. Additionally, expected secondary structure was calculated through the PDB IDs, as described in the Extended Materials and Methods at the supplementary file. However, interaction of BSA with FGFA could have influenced the structures. It is also worth to mention that curve-fit is sensitive to overall spectral quality, but using a parameter set optimized with a different protein would not lead to a reliable outcome regardless of the spectral quality when an optimized and restricted set of parameters will be used during fit.

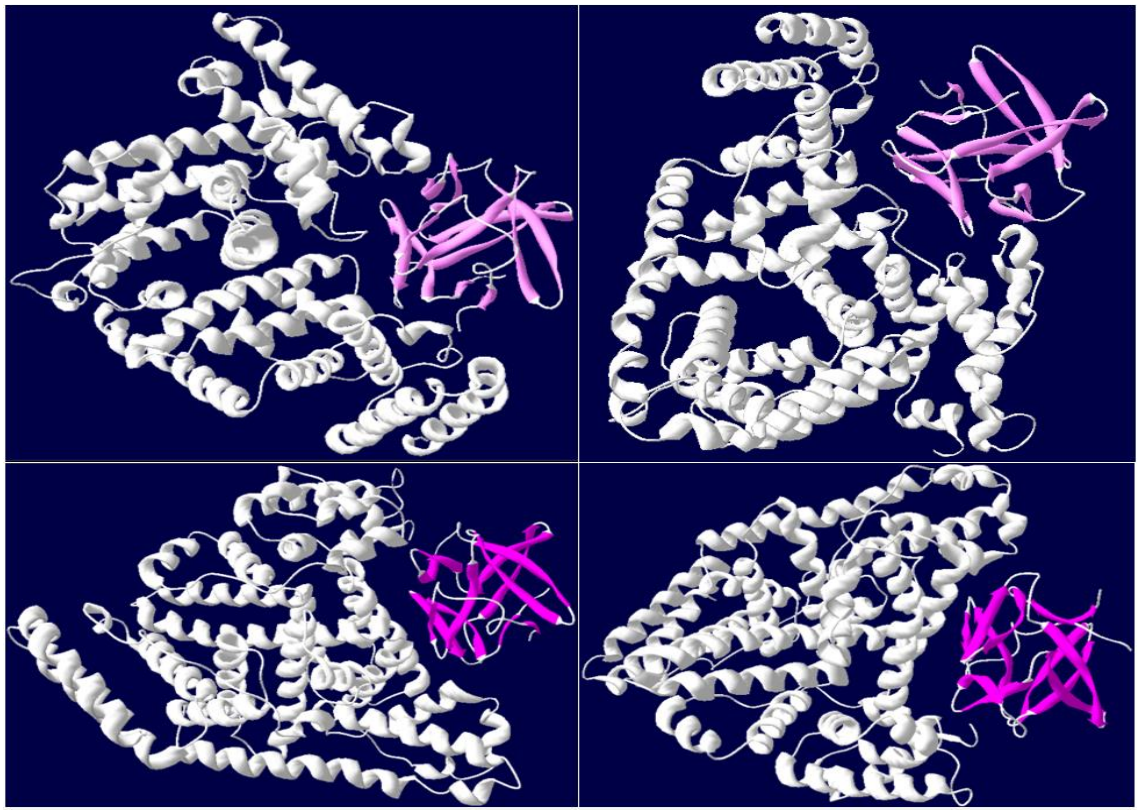

Figure 3. Two distinct views of the FGFA/BSA complexes modeled by Hex and GRAMM. Hex model views are on top and GRAMM model views at the bottom, where the large BSA protein is on the left while the small FGFA protein is on the right in all. Images are generated by Swiss-PdbViewer v4.1.0 at http://www.expasy.org/spdbv

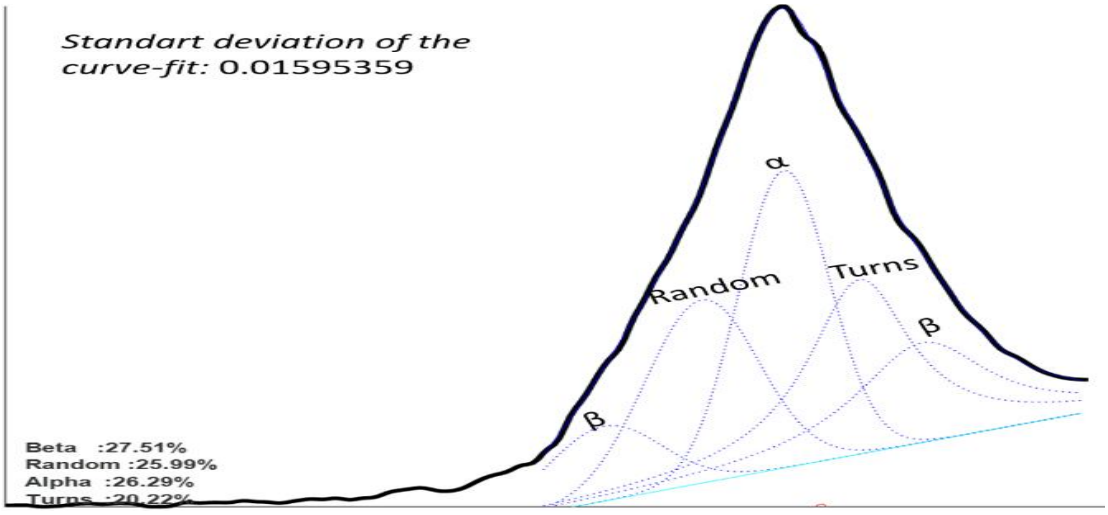

Figure 4. Curve-fit of 1600-1700 cm^{-1} region ATR FTIR absorbance spectrum of acidic fibroblast growth factor (FGFA) with 50 folds more BSA, by using Kinetics software, with the optimized parameter set. Alpha (α) and beta (β) signs stand for (α-)helices and (β-)sheets.

Table 8. Comparison of the experimental results (ATR FTIR data) with the calculated results from the docked complex and the PDB structures. Calculated results from the docked complex and the PDB structures included A chains in PDB files 2AFG (for FGFA) and 3V03 (for BSA) in the docked complex. The docked complex contributed as 1 part while the remaining 49 parts were contributed by BSA in the calculation. Comparison of the curve-fitted experimental results is displayed on top 2 data-rows. Below,

at data-rows 3–7, is the comparison of the structures obtained through applying models to the same ATR FTIR data, along with their RMSD to the data on the top row.

1FGFA/50BSA structure-source	**β-sheets%**	**α-helices%**	**Turns%**	**Coils%**	**RMSD**
***PDB**	0.75	72.26	9.28	17.71	
Curve-fit	27.51	26.29	20.22	25.99	27.47
Model-1	3.16	53.32	13.81	29.71	11.50
****Model-1'**	3.80	64.12	13.03	19.06	4.78
Model-2	0.00	59.88	15.85	24.28	7.75
Model-3	1.29	50.10	14.56	34.04	14.02
****Model-3'**	0.00	58.47	15.68	25.85	8.63

*A chains of PDBs with IDs 2AFG (for FGFA) and 3V03 (for BSA) used for docking, then 1FGFA/50BSA calculated with 1:49 docked complex and BSA

**Modified model

Table 9. Comparison of bFGF structures obtained through applying models to the same ATR FTIR data as the one curve-fitted for obtaining the optimised curve-fit parameters. RMSD is calculated with respect to the PDB structure 1BFF.

bFGF structure-source	**β-sheets%**	**α-helices%**	**Turns%**	**Coils%**	**RMSD**
1BFF (PDB ID)	41.09	6.98	19.38	32.56	
Model-1	26.21	16.60	15.33	41.86	10.21
***Model-1'**	32.63	20.67	14.76	31.95	8.38
Model-2	30.40	22.76	14.97	31.86	9.79
Model-3	33.05	11.43	13.61	41.92	7.17
***Model-3'**	30.86	20.81	14.74	33.60	8.92

*Modified model

In sum, when the results of the predictors are compared, secondary structure prediction of the bFGF by SOPM reveals the best results with only 1.45 RMSD from the 1BFF structure (Table 2). RMSD of both the SSpro8 and the SOPM prediction results are less than 2.5 in the prediction of bFGF secondary structure with the PDB ID: 1BFF (Table 2). RMSD of SOPM and SSpro8 predictions are within the range of those of the PDB-data in Table 1. In a study by Magnan and Baldi [41], it was found that the SSpro8 suite is "almost perfect" in estimating the secondary structure of proteins. Average RMSD (RMSDav) of SSpro8 predictions supports this conclusion (Table 3, RMSDav. 2.60). RMSDav is 2.89 in case of SOPM (Table 3), which is equally good. However, this result is probably due to the presence of a closely related protein variant among the proteins that were used to train the algorithm. Beyond these concerns, secondary structure predictions through primary sequences do not predict secondary structure changes upon interactions. Therefore, another type of web-based resource that can be utilized is docking, to generate protein-interaction model. Here we tested two such algorithms (Hex and GRAMM) as well. In relation, it can be noted here that chain-A of the PDB files were used with this purpose, as a proof-of-the-concept, but utilizing the other chains, or the other PDB structures, of the proteins under study could have an effect. Further, protein-protein interaction may not necessarily be one-to-one, like the one tested in this study, and the interactions may involve multiple proteins at different ratios, under a dynamic control of the protein crowding and the hydration level.

Deviations of the predicted values from the PDB data for secondary structure percentages are considerably lower than the residue-based accuracies. Therefore, predictors can be utilised to support the secondary structure analysis with the experimental techniques like FTIR. Web-based servers can assist in secondary structure analysis with the FTIR data to improve the quality of predictions [42] and/or to derive preliminary structural information, to serve as the basis of a reliable starting point during secondary structure analysis with FTIR. Although the secondary structure information can be derived through the FTIR spectra of the proteins, prior knowledge on the closely related structure is highly desired at

many instances for better outcomes. The need for prior knowledge is due to the complexity of the Amide I band, which is a large superposition of many smaller bands originating from secondary structures and side chain absorptions. After determining the number and position of composing bands, curve-fitting is applied to find the area of these bands. This analysis starts with an initial set of parameters defined by the user and the software fits the Amide I band by optimizing the individual band shapes and areas. However, results depend heavily on the initial parameter set (position, width, shape, and height). With similar fitting-errors, one can obtain different combination of component bands of the same Amide I region. Number of such different possible combinations under the same fitting conditions will be less as the error gets smaller (i.e., lower RMSD). Predictor results can be a good starting point at this point, once the primary structure of the protein is known. It will serve as a basis while setting the initial parameters and will also be a guide to judge whether curve-fitting yielded acceptable results. However, as implied above, judgement results will be depending on the selected tool and the inputs of the tool, i.e., the specific PDB files, if the tool would be performing docking.

Secondary structure predictors are reliable; however, their estimation is based solely on the a.a. sequence. Therefore, structure changes induced by buffer conditions or by effector molecules are not considered, as far as we know. Docking tools aid in these issues to some extent. However, experiments implement such changes during measurements, at different extents. Therefore, as an experimental method, FTIR can reveal such changes more realistically and hence becomes the instrument that uses the predictions of the predictor algorithms and carries that information to the level of structural determination at conditions imposing structure variations. FTIR can also be benefited by the predictors as a validation tool for the estimations of the predictors. Namely, there is a mutual benefit.

As a result, web-server based protein secondary structure algorithms are suitable as preliminary knowledge source when curve-fit parameters are aimed to be optimized. Success of web-servers can be deemed generalizable considering their undergoing tests with large protein sets before being released. Utilized primary sequence-based prediction tools could have advanced since they have been used in this work, and the evolution of their global performances was not within the scope of this work. As a result, secondary structure prediction tools and the docking tools can be used as supplementary tools for secondary structure determination through FTIR data of proteins and protein mixtures, respectively. Though not tested here, manual curve-fit can be utilized for secondary structure determination through FTIR data, as well as the methods established by Goormaghtigh et al. [9,31].

4. Conclusions

Web servers can assist experimental studies investigating unknown structures such that a prediction supporting the experimental results would be enforcing the findings, but the unsupported results would not necessarily falsify the experimental data. Apart from that, model-based prediction of 1FGFA/50BSA secondary structure from FTIR data proved feasible in this study. In the future, in addition to checking the broader performance of the modified models presented here, a curve-fit parameter set that yields <5 RMSD in different FGF structures deposited at the PDB and the 1FGFA/50BSA, as a protein mixture case, is aimed to be optimized.

ACKNOWLEDGEMENT

This work was supported by the Starting Research and Development Projects' Support of the Scientific and Technological Research Council of Turkey (TUBITAK) [214Z261]. We would like to acknowledge Prof. E. Goormaghtigh for providing KineticsR12 software running under MATLAB. We would like to thank Ibrahim Baran Adıgüzel for providing technical assistance.

Data availability

Raw data and processed data, wherever applicable, are available upon request.

References

[1] D.M. Byler, H. Susi, Examination of the secondary structure of proteins by deconvolved FTIR spectra, Biopolymers 25 (1986) 469–487.

[2] S. Krimm, J. Bandekar, Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins, Adv Protein Chem 38 (1986) 181–363.

[3] P. Haris, F. Severcan, FTIR spectroscopic characterization of protein structure in aqueous

and non-aqueous media, Journal of Molecular Catalysis B: Enzymatic 7 (1999) 207–221.

[4] J. Güldenhaupt, Y. Adiguzel, J. Kuhlmann, H. Waldmann, C. Kötting, et al., Secondary structure of lipidated Ras bound to a lipid bilayer, FEBS J 275 (2008) 5910–5918.

[5] Y. Adıgüzel, P.I. Haris, F. Severcan, Screening of proteins in cells and tissues by vibrational spectroscopy, in: Severcan F, Haris PI (Eds.), Vibrational Spectroscopy in Diagnosis and Screening, IOS Press, Amsterdam, 2012, 53–108.

[6] J.A. Hering, P.R. Innocent, P.I. Haris, An alternative method for rapid quantification of protein secondary structure from FTIR spectra using neural networks, Spectrosc. Int. J. 16 (2002) 53–69.

[7] M. Severcan, P.I. Haris, F. Severcan, Using artificially generated spectral data to improve protein secondary structure prediction from Fourier transform infrared spectra of proteins, Anal Biochem 332 (2004) 238–244.

[8] S. Navea, R. Tauler, A. de Juan, Application of the local regression method interval partial least-squares to the elucidation of protein secondary structure, Anal Biochem 336 (2005) 231–242.

[9] G. Goormaghtigh, J.M. Ruysschaert, V. Raussens, Evaluation of the information content in infrared spectra for protein secondary structure determination, Biophys. J. 90 (2006) 2946–2957.

[10] F. Korkmaz, S. Köster, O. Yildiz, W. Mäntele, In situ opening/closing of OmpG from E. coli and the splitting of β-sheet signals in ATR-FTIR spectroscopy, Spectrochim. Acta. A. Mol. Biomol. Spectrosc. 91 (2012) 395–401.

[11] M. Baldassarre, C. Li, N. Eremina, E. Goormaghtigh, A. Barth, Simultaneous Fitting of Absorption Spectra and Their Second Derivatives for an Improved Analysis of Protein Infrared Spectra, Molecules 20 (2015) 12599–12622.

[12] J.S. Kastrup, E.S. Eriksson, H. Dalboge, H. Flodgaard, X-ray structure of the 154-amino-acid form of recombinant human basic fibroblast growth factor. comparison with the truncated 146-amino-acid form, Acta Crystallogr Sect D 53 (1996) 160–168.

[13] M. Klähn, J. Schlitter, K. Gerwert, Theoretical IR spectroscopy based on QM/MM calculations provides changes in charge distribution, bond lengths, and bond angles of the GTP ligand induced by the Ras-protein, Biophys J 88 (2005) 3829–3844.

[14] V.A. Shashilov, I.K. Lednev, Advanced statistical and numerical methods for spectroscopic characterization of protein structural evolution, Chem Rev 110 (2010) 5692–5713.

[15] S. Sen-Britain, W. Hicks, R. Hard, G.A. Gardella Jr, The mechanism of secondary structural changes in Keratinocyte Growth Factor during uptake and release from a hydroxyethyl(methacrylate) hydrogel revealed by 2D Correlation Spectroscopy, arXiv 1808.03670 (2018) [cond-mat.soft].

[16] N.L. Benbow, S. Karpiniec, M. Krasowska, D.A. Beattie, Incorporation of FGF-2 into Pharmaceutical Grade Fucoidan/Chitosan Polyelectrolyte Multilayers, Mar Drugs 18 (2020) 531.

[17] S. Noji, T. Matsuo, E. Koyama, T. Yamaai, T. Nohno, et al., Expression pattern of acidic and basic fibroblast growth factor genes in adult rat eyes, Biochemical and Biophysical Research Communications 168 (1990)343–349.

[18] J. Engele, M. Churchill Bohn, Effects of acidic and basic fibroblast growth factors (aFGF, bFGF) on glial precursor cell proliferation: Age dependency and brain region specificity, Developmental Biology 152 (1992) 363–372.

[19] I. Tooyama, H.P.H. Kremer, M.R. Hayden, H. Kimura, E.G. McGeer,P.L. McGeer, Acidic and basic fibroblast growth factor-like immunoreactivity in the striatum and midbrain in Huntington's disease, Brain Research 610 (1993) 1–7.

[20] J. Smith, A. Yelland, R. Baillie, R.C. Coombes, Acidic and basic fibroblast growth factors in human breast tissue, European Journal of Cancer 30 (1994) 496–503.

[21] F. Coulier, P. Pontarotti, R. Roubin, H. Hartung, M. Goldfarb, D. Birnbaum, Of worms and men: an evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families, J Mol Evol 44 (1997) 43–56.

[22] M.R. Lozano, M. Redondo-Horcajo, M.Á. Jiménez, L. Zilberberg, P. Cuevas, et al.,

Solution Structure and Interaction with Basic and Acidic Fibroblast Growth Factor of a 3-kDa Human Platelet Factor-4 Fragment with Antiangiogenic Activity, J Biol Chem 276 (2001) 35723–35734.

[23] A. Tovchigrechko, I.A. Vakser, GRAMM-X public web server for protein–protein docking, Nucleic Acids Res 34 (2006) W310–W314.

[24] I.A. Vakser, Protein-protein docking: from interaction to interactome, Biophys J 107 (2014) 1785–1793.

[25] G. Macindoe, L. Mavridis, V. Venkatraman, M.-D. Devignes, D.W. Ritchie, HexServer: an FFT-based protein docking server powered by graphics processors, Nucleic Acids Res 38 (2010) W445–W449.

[26] S.J. Wodak, J. Janin, Computer analysis of protein-protein interaction, J Mol Biol 124 (1978) 323–342.

[27] D.W. Ritchie, S. Grudinin, Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry, J Appl Cryst 49 (2016) 158–167.

[28] D.W. Ritchie, Recent progress and future directions in protein-protein docking, Current Protein and Peptide Science 9 (2008) 1–15.

[29] J. Ollesch, E. Kuennemann, R. Glockshuber, K. Gerwert, Prion protein α-to-β transition monitored by Time-resolved Fourier Transform Infrared Spectroscopy, Applied Spectroscopy 61 (2007) 1025–1031.

[30] E. Goormaghtigh, V. Cabiaux, J.M. Ruysschaert, Determination of soluble and membrane protein structure by Fourier transform infrared spectroscopy, Subcell Biochem 23 (1994) 329–450.

[31] E. Goormaghtigh, R. Gasper, A. Bénard, A. Goldsztein, V. Raussens, 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. Biophys. Acta 1794 (2009) 1332–1343.

[32] M. Heinig, D. Frishman, STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins, Nucleic Acids Res 32 (2004) W500–W502.

[33] C. Geourjon, G. Deleage, SOPM: a self-optimized method for protein secondary structure prediction, Protein Eng 7 (1994) 157–164.

[34] K. Lin, V.A. Simossis, W.R. Taylor, J. Heringa, A simple and fast secondary structure prediction method using hidden neural networks, Bioinformatics 21 (2005) 152–159.

[35] L. Homaeian, L.A. Kurgan, J. Ruan, K.J. Cios, K. Chen, Prediction of protein secondary structure content for the twilight zone sequences, Proteins: Structure, Function, and Bioinformatics 69 (2007) 486–498.

[36] B. Rost, G. Yachdav, J. Liu, The PredictProtein server, Nucleic Acids Res 32 (2004) W321–W326.

[37] G. Deleage, B. Roux, An algorithm for protein secondary structure prediction based on class prediction, Protein Eng 1(1987) 289–294.

[38] R.D. King, M.J. Sternberg, Identification and application of the concepts important for accurate and reliable protein secondary structure prediction, Protein Sci 5 (1996) 2298–2310.

[39] G. Pollastri, D. Przybylski, B. Rost, P. Baldi, Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, Proteins 47 (2002) 228–235.

[40] A.E. Eriksson, L.S. Cousens, B.W. Matthews, Refinement of the structure of human basic fibroblast growth factor at 1.6 A resolution and analysis of presumed heparin binding sites by selenate substitution, Protein Sci 2(1993) 1274–1284.

[41] C.N. Magnan, P. Baldi, SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning, and structural similarity, Bioinformatics 30 (2014) 2592–2597.

[42] J.G. Lees, R.W. Janes, Combining sequence-based prediction methods and circular dichroism and infrared spectroscopic data to improve protein secondary structure determinations, BMC Bioinformatics 9 (2008) 24.