PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

The protein fluorescence and structural toolkit: Database and programs for the analysis of protein fluorescence and structural data

Chi Shen,[1] Rajiv Menon,[1] Dipanwita Das,[1] Nidhi Bansal,[1] Neha Nahar,[1] Neelima Guduru,[1] Stephen Jaegle,[1] Joan Peckham,[1] and Yana K. Reshetnyak[2]*

[1] Department of Computer Sciences and Statistics, University of Rhode Island, Kingston, Rhode Island 02881
[2] Department of Physics, University of Rhode Island, Kingston, Rhode Island 02881

ABSTRACT

Protein fluorescence is a powerful tool for studying protein structure and dynamics if we have a means to interpret the spectral data in terms of protein structural properties. Our previous research successfully provided this support through the development of individual software modules implementing the algorithms for fluorescence and structural analyses. Now we have integrated the developed software modules, introduced a new program for the assignment of tryptophan residues to spectral-structural classes, and created a web-based toolkit PFAST: protein fluorescence and structural toolkit: http://pfast.phys.uri.edu/. PFAST contains three modules: (1) FCAT is a fluorescence-correlation analysis tool, which decomposes protein fluorescence spectra to reveal the spectral components of individual tryptophan residues or groups of tryptophan residues located close to each other, and assigns spectral components to one of five previously established spectral-structural classes. (2) SCAT is a structural-correlation analysis tool for the calculation of the structural parameters of the environment of tryptophan residues from the atomic structures of the proteins from the Protein Data Bank (PDB), and for the assignment of tryptophan residues to one of five spectral-structural classes. (3) The last module is a PFAST database that contains protein fluorescence and structural data obtained from results of the FCAT and SCAT analyses.

Proteins 2008; 71:1744–1754.

Key words: *decomposition algorithms; tryptophan fluorescence; spectral-structural classes; discriminant analysis; classification score; scientific databases; scripted online database environment.*

INTRODUCTION

Fluorescence spectroscopy is a powerful tool for the investigation of protein structure, conformations, and dynamics, since the fluorescence properties of tryptophan residues vary widely depending on the tryptophan environment in a given protein. The major goal in the application of tryptophan fluorescence spectroscopy is to interpret the fluorescence properties in terms of structural parameters and predict structural changes in the protein. The problem in the analysis of the protein fluorescence spectra is their complex nature, since the overwhelming majority of proteins contain more than one tryptophan fluorophore, and therefore exhibit smooth spectra that contain more than one spectral component. We have developed methods for the mathematical analysis of the fluorescence spectra of multitryptophan proteins aimed for revealing the spectral components of individual tryptophan residues or groups of tryptophan residues located close to each other.[1–3] Among the spectral parameters, we considered are the position of maximum of the fluorescence spectra and the degree of quenching of tryptophan fluorescence by water-soluble quenchers, such as acrylamide, iodide, or others. To correlate the fluorescence properties of the tryptophan residues with the structural parameters, we have created an algorithm for the structural analysis of the tryptophan environment in 3D atomic structures of proteins from the Protein Data Bank (PDB).[4] The analysis of the spectral and structural characteristics of tryptophan residues allowed us to reveal a general correlation between the spectral and structural properties. It was demonstrated that tryptophan residues in proteins can be grouped into five discrete classes based on their spectral parameters, and that these classes are correlated well with the five discrete classes revealed based on the analysis of six structural properties of the environment of tryptophan residues.[4] We confirmed a previously-proposed

Grant sponsor: URI Cancer Prevention Center.
Chi Shen and Rajiv Menon contributed equally to this work.
*Correspondence to: Yana K. Reshetnyak, Department of Physics, University of Rhode Island, 2 Lippitt Rd, Kingston, RI 02881. E-mail: reshetnyak@mail.uri.edu
Received 21 August 2007; Accepted 8 October 2007
Published online 3 January 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21857

model of discrete states according to which tryptophan fluorophores belonging to various classes have different environments in proteins (see Table I).[5]

The methods of spectral and structural analysis were successfully applied to the study of folding kinetics and the structure of the membrane protein OEP16 and its single-tryptophan-containing mutants,[6] of the conformational changes of the 20S proteasome of rat natural killer cells induced by mono and divalent cations,[7] and of the interaction of the eosinophil cationic protein (ECP) with membrane to establish the tryptophan residues involved in those interactions.[8] The decomposition of the emission spectra of wild type (WT) and a mutant of the lecithin: cholesterol acyltransferase (LCAT) in absence and presence of substrates allowed us to reveal the conformation differences between WT and mutant forms of LCAT.[9] Analysis of fluorescence properties of isoforms of recombinant rat nucleoside diphosphate kinase (NDPK), which catalyze the transfer of γ-phosphate from nucleoside triphosphates to nucleoside diphosphates, led to the elucidation of the origin of unusual fluorescence (extremely high quantum yield) in NDPKα. The origin was the tyrosinate formation in the active center of the α-enzyme, which is crucial for the activity of the protein.[10–12] The spectral and structural analysis of skeletal myosin subfragment 1 (S1) enabled us to find tryptophan residues, the emission of which are sensitive to the binding of ATP with S1, and as a result, propose what kind of conformational changes in S1 structure are associated with the nucleotide binding.[13]

The spectral-structural analysis was demonstrated to be very powerful in the prediction of structural changes in proteins from fluorescence data, and therefore, extremely useful in the study of protein function. The package of spectral and structural analyses consisted of several programs written independently of one another in different programming languages and worked in a sequential mode. Therefore, there was a great demand for the implementation and design of a single, easy-to-use software package for automatically running programs. Here, we present the results of the implementation that package, which includes the previously developed programs for the spectral and structural analysis as well as the introduction of a new program for the statistical correlation analysis. Our implementation is realized as part of a web-based toolkit: protein fluorescence and structural toolkit (PFAST) designed to perform protein spectral, structural, and correlation analysis on-line. Along with the analyses programs, PFAST also contains the first database of protein fluorescence properties correlated with the structural features of the environment of tryptophan residues.

MATERIALS AND METHODS

PFAST consists of a protein fluorescence/structural database, fluorescence analysis programs, and structural analysis programs. The analysis programs are implemented in a variety of languages: Fortran, Visual Basic, and C#, and majority of them have been compiled as Windows programs. The goal of this work is to create a toolkit based on already existing programs, thus to avoid possible emulation issues for the existing Windows analysis programs we chose to use Windows as our operating system. However, in future we propose to rewrite the program codes and implement them using other operating systems (Linux).

Thus, for this phase of the project, we needed a Windows-based software framework that serves web pages for providing a user-interface, performs scripting for weaving together the analysis programs, and supports standard database operations. We chose the WAMP Server* software package because it is widely-used, mature, free, and meets all of our needs. It consists of the following preconfigured, open-source, software packages: Apache† web server, MySQL‡ database, and the PHP§ scripting system. Apache serves web pages to the Internet, which we use for providing a user-interface. It also handles the execution of PHP scripts, which we use for website scripting, accessing the database, preprocessing for the analysis programs, and actually executing the analysis programs. Finally, MySQL is the database engine for storing and retrieving protein fluorescence and structural data.

For the decomposition of the tryptophan spectral components, the SIMS and PHREQ algorithms are implemented in the Visual Basic programming language. The graphs for visualizing some of the FCAT data analysis results are generated using the JpGraph¶ graph-creating package. We chose JpGraph because it is written in PHP, free, easy-to-use, and can generate the kinds of graphs needed for our visualizations.

The structural analysis of the tryptophan local environment is implemented in programs written in Microsoft Visual Basic, Fortran, and Microsoft C#, tied together by the overarching PHP web script. The script processes user-specified PDB** data from the Protein Data Bank, and generates the appropriate intermediate files for processing by the analysis programs. The StatSoft (2006) STATISTICA†† data analysis software system was used to provide computer classification and discriminant functions.

The statistical-correlation analysis program (calculation of Mahalanobis distances, probabilities of tryptophan residues assignment to the spectra-structural classes, and classification scores) was implemented in Microsoft C#. This program uses the Gauss Jordan elimination method to

*http://www.wampserver.com.
†http://httpd.apache.org/.
‡http://www.mysql.com/.
§http://php.net/.
¶http://www.aditus.nu/jpgraph/.
**http://www.rcsb.org/.
††http://www.statsoft.com/.

Table I
Measured and Predicted (Based on the Analysis of Six Structural Parameters of Environment of Tryptophan Residues) Spectral Properties of Five Classes of Emitting Tryptophan Residues in Proteins

Spectral and structural parameters	Class A	Class S	Class I	Class II	Class III
The emission of tryptophan residues belonging to five classes[a] (nm) **calculated** as a result of spectral analysis.	308	321–325	330–333	341–344	346–350
The emission of tryptophan residues belonging to 5 classes[b] (nm) **predicted** from the structural analysis.	308	322.5 ± 4.6	331.0 ± 4.8	342.3 ± 3.3	348.0 ± 3.1
Acc—averaged value of the relative solvent accessibility of 9 atoms of indole ring of the tryptophan fluorophore.	1.9	0.8 ± 1.4	6.0 ± 3.6	14.8 ± 7.5	55.3 ± 15.9
Acc1-7—*averaged value of the relative solvent accessibility of 1 and 7 atoms of the tryptophan fluorophores.*	0.0	1.0 ± 2.2	11.2 ± 8.5	26.7 ± 19.1	71.1 ± 19.5
Den—packing density: the number of neighbor atoms at distance <7.5 Å from the indole ring.	138.3	148.3 ± 8.5	129.3 ± 9.1	109.3 ± 12.6	62.7 ± 18.8
B—*B*-factor: averaged value of the crystallographic *B*-factors of polar atoms in near and far layers normalized to the mean *B*-factor value of all Cα atoms in the crystal structure.	0.61	0.89 ± 0.17	1.11 ± 0.20	1.23 ± 0.32	1.54 ± 0.55
R—dynamic accessibility [$R = Acc \cdot B$]. Parameter, which reflects flexibility of the tryptophan environment.	0.9	0.7 ± 1.2	6.7 ± 4.0	18.2 ± 10.3	85.2 ± 30.9
A—averaged value of the relative polarity of environment: portion of the atoms of the polar groups amongst all the atoms around the tryptophan residue in near and far layers.	23.5	34.5 ± 5.8	39.3 ± 5.5	45.1 ± 7.4	65.5 ± 13.9

[a]The wavelengths of the most probable position of emission of tryptophan residues, which belong to various spectral classes, revealed from the analysis of the fluorescence spectra of 160 proteins.
[b]The wavelengths of the spectral positions of emission of tryptophan residues, which belong to various structural classes, predicted from the analysis of six structural parameters of environment of 137 tryptophan residues of 48 proteins from PDB.

calculate the inverse of the variance–covariance matrix used for the calculation of the probabilities of assignment.

RESULTS

To introduce a novel tool for the analysis of protein fluorescence spectra and correlate the data with the structural properties of tryptophan residues we have developed PFAST. PFAST is a web-based toolkit that integrates the original software packages for the protein fluorescence and structural analysis developed previously, and a newly introduced program for the statistical classification of tryptophan residues and their assignment to one of five spectral-structural classes. The system consists of three parts: the fluorescence-correlation and the structural-correlation analysis tools, FCAT, and SCAT, respectively, and a database of protein spectral and structural data (see Fig. 1).

FCAT: Fluorescence-correlation analysis tool

The fluorescence-correlation analysis tool is based on the previously developed programs for the decomposition of tryptophan fluorescence spectra into spectral components.[1] The revealing of the spectral components is an essential step in the analysis of the protein fluorescence properties, since only spectral properties representing individual tryptophan residues or a cluster of tryptophan residues considered as a single emitting unit, could be correlated with the structural properties of tryptophan environment. The FCAT tool combines two different mathematical approaches: SImple fitting procedure using the root-Mean-Square criterion (SIMS) and PHase-plot-based REsolution using Quenchers PHREQ.[1] SIMS employs the minimal least-square approach, while PHREQ uses an analytical pseudo-graphic solving technique. The detailed description of both algorithms can be found in Burstein *et al.*[1] The decomposition algorithms were tested on model compounds and single-tryptophan containing proteins. Recently, additional supportive evidence of the accuracy of the decomposition algorithms was obtained through the analysis of the correlation of the experimental spectra of single-tryptophan-containing mutants of multitryptophan proteins OEP16[6] and ECP.[8]

The input file for the fluorescence analysis contains the fluorescence spectra of protein measured without an external quencher of tryptophan fluorescence or with different concentrations of the quencher. Acrylamide, negatively charged iodide ions (I^-) or (NO_3^-) and positively charged cesium ions (Cs^+) can be used as quenchers of

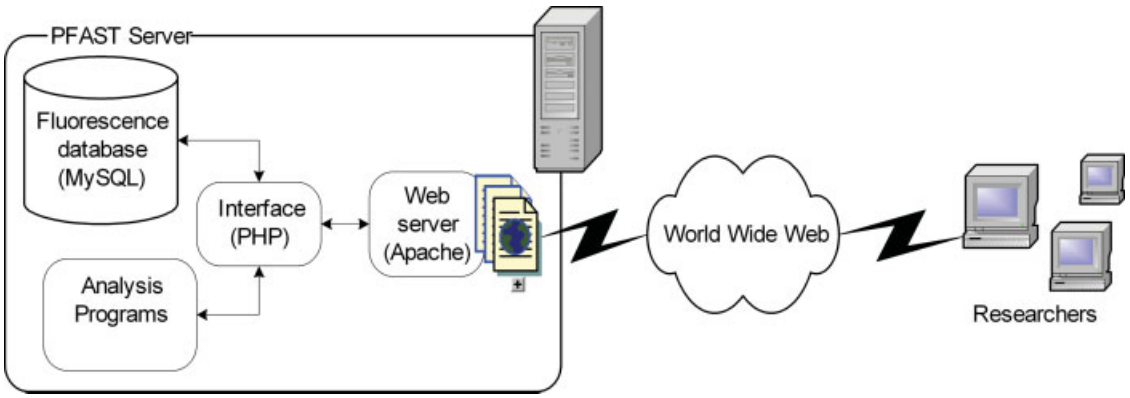

Figure 1
The PFAST web-site architecture diagram.

tryptophan fluorescence. The decomposition of fluorescence spectra measured in the presence of various quenchers should provide the same result, thus providing an additional control for the goodness of fit and stability of solution.

The following information is included in the input file for the decomposition analysis: the title, the quencher, and its concentrations used in the fluorescence experiment, the starting and ending wavelengths of the spectra, the step that was used to record the spectra and the values of intensities for each wavelength (fluorescence spectra). Each input file needs to include the spectrum of tryptophan in solution (recorded at the same setting of instrument as used for the measurement of protein spectra) for the calculation of the correction curve. The input files can be uploaded or created by filling out a web form and submitting it for analysis. The results of the calculation including the input file can be downloaded after analysis. The results are presented in the form of three ASCII files (results, graph-data, summary) and one graph file. The summary file contains the summarized information of the decomposition analysis: the positions of maxima of the spectral components in wavelengths and their contributions to the total spectrum in percents. The summary file also includes the assignment of the spectral components to the one of five spectral-structural classes, the goodness-of-fit, and the number of smoothing procedures used during the calculations. The result file contains more detailed information about the decomposition analysis: the values of activity for ionic quenchers, the information about spectral components in wavelength and frequency scales, the contribution (the area under the spectral component) of the components into total spectra measured at all quencher concentrations, and the relative and absolute values of Stern-Volmer constants for each spectral component. The graph-data file contains data for the construction of plots: the original experimental, smoothed and calculated theoretical spectra, the spectral components, the residuals, and the parameters for the Stern-Volmer plot calculated from the Stern-Volmer equation, and obtained as a result of the decomposition analysis. The graph-data file allows users to plot the graphs using their own style, while the Graph file provides already plotted results.

SCAT: Structural-correlation analysis tool

Structural analysis

The structural-correlation analysis tool is based on previously developed programs for the analysis of the environments of tryptophan residues from the atomic structures of proteins from the PDB and a new program for the assignment of tryptophan residue to one of five spectral-structural classes.[4] The SCAT module is fully

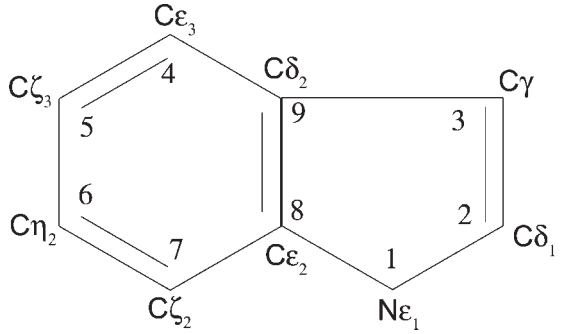

Figure 2
The schematic illustration of nine atoms of indole ring of tryptophan residue.

Table II
List of Proteins in the Training Set, their Codes, PDB Codes, and Crystallographic Resolution, Tryptophan Residue Position in Protein Sequence and Assignment of Tryptophan Residues to Spectral Components

	Protein	Protein codes	PDB-entries (resolution, Å)	Trp residue position	λm (nm)
1-Tryptophan containing proteins					
1	Albumin (human serum), N-form	ASH.N	1AO6 (2.5); 1BJ5 (2.5)	W214	344.8
2	L-Asparaginase (*E. coli* B)	ASP.N	3ECA (2.4)	W66	323.9
3	Azurin (*Pseudomonas aeruginosa*)	AZU	1JOI (2.05); 4AZU (1.9)	W48	307.9
4	Glucagon	GLG	1GCN (3.0)	W25	350.8
5	Protein G, streptococcal	GPS	1IGD (1.1)	W48	342.5
6	Telokin KRP, myosin light-chain kinase (chicken gizzard)	KRP	1TLK (2.8)	W75	332.4
7	Monellin (*Dioscoreophyllum cummensii*)	MON	1MON (1.7)	W3	340.3
8	Parvalbumin II (cod *Gadus morrhua*) Ca-form	PAC.CA	Model by M. Laberge	W102	326.5
9	Parvalbumin (whiting *Gadus merlangus*)	PAM	1A75 (1.9)	W102	317.9
10	Phospholipase A2 (bovine pancreas)	PLB	1UNE (1.5); 2BPP (1.8)	W3	352.3
11	Phospholipase A2 (porcine pancreas)	PLS	1P2P (2.6); 4P2P (2.4)	W3	349.4
12	RNAse T1 (*Aspergillus oryzae*)	RNT	9RNT (1.5)	W59	325.1
13	Vipoxin, protein A (*Vipera ammodytes* ammodyes); high-ionic strength (dimers)	VTA.HIS	3D model by B. Atanasov	W30	323.6
14	Vipoxin, protein A (*Vipera ammodytes* ammodyes); low-ionic strength (monomers)	VTA.LIS	1VPI (1.76)	W31	349.1
2-Tryptophan containing proteins					
15	α1-Antitrypsin (human)	A1AT	2PSI (2.9); 7API (3.0); 8API (3.1); 9API (3.0)	W194 W238	324.4 340.1
16	Neurotoxin II (cobra *Naja naja* oxiana venom)	NO2	1NOR (NMR)	W27 W28	344.4
17	HIV-1 protease	PRH.LIS	1HHP (2.7)	W6 W42	345.4
18	HIV-1 protease complex with pepstatin	PRH.PST	5HVP (2.0)	W6 W42	344.9
3-Tryptophan containing proteins					
19	Agglutinin (wheat germ)	AWG	7WGA (2.0); 9WGA (1.8)	W41 (Quenched) W107 W150	(349.9) 349.1 349.1
20	Ovalbumin (hen egg white)	OVH.LIS	1OVA (1.95)	W160 W194 W275	336.9 325.8 336.9
21	Phosphatase alkalaine (*Escherichia coli*)	PHA	1ALK (2.0)	W109 W220 W268	322.8 345.8 345.8
22	Vipoxin, complex protein A and protein B (*Vipera ammodytes* ammodyes)	VTC	1AOK (2.0)	W31 W220 (Quenched) W231	334.2 (334.2) 334.2
4-Tryptophan containing proteins					
23	G-Actin (rabbit skeletal muscule)	ACR.G	1ATN (2.8)	W79 (quenched) W86 W340 W356	(332.8) 320.9 320.9 332.8
24	α-Lactalbumin (cow milk)	LAB	1HFZ (2.3)	W26 W60 W104 W118	322.4 338.3 322.4 3338.3

interoperable with the PDB. The program downloads and parses the PDB file after the user provides the protein PDB code. The user is asked to select residues and chains (if there are more than one) to be used for the calculation. If several conformers are presented they can be analyzed separately. At the first step, the program generates an output file used in further calculation of the structural parameters of the environment of tryptophan residues in protein. The results of the calculation are presented in the form of 6 ASCII files (16-parameters, list of all atoms, list of polar atoms, result, result-table, and summary) available for download.

The detailed description of the calculated structural parameters can be found in Reshetnyak *et al.*[4] Briefly, the structural parameters are separately calculated for each of nine atoms of the indole ring of the tryptophan residue (see Fig. 2). The microenvironment of tryptophan is considered to consist of all the neighboring atoms of the protein, and the structure-defined solvent molecules located in near and far layers at a distance of

up to 5.5 and 7.5 Å from the indole atoms, respectively. The SCAT module outputs a complete list of all atoms and polar (nitrogen, oxygen, and sulfur) atoms in near and far layers in a form of two files: list of all atoms and list of polar atoms, respectively. The result file contains information about distances and orientations of carbon and polar atoms located up to 7.5 Å from each atom of the indole ring. The potential partners for hydrogen bonding with atoms of the indole ring (the interactions, which might occur in the excited state of the tryptophan fluorophores after the absorption of photons) are selected among the neighbor polar atoms based on the criteria of H-bonds formation and presented in file result-table.[14] The result-table also contains information about the probability of energy transfer (from ^1L_a states of potential donors to either ^1L_a or ^1L_b states of potential acceptors) between the indole rings of tryptophan residues with centers located at distances less than 12 Å. The 16-parametes output file includes 16 structural parameters, from which six structural parameters (see Table I) that correlate with tryptophan spectral properties, are calculated and used in statistical analysis for the classification of tryptophan residues and are output in a final "summary" file.

Classification analysis

The final part of the SCAT module is a statistical analysis that is used for the assignment of tryptophan residues to one of five spectral-structural classes. Two different approaches for the assignment such as calculation of (i) classification scores and (ii) probabilities of assignment are applied. Both approaches are very sensitive to a training set, which is used as a model for assignment of a new object (tryptophan residue) to spectral-structural classes. For the training set, we choose proteins containing no more than four tryptophan fluorophores for which the assignment of tryptophan residues to spectral components was obvious and straightforward. For this analysis, 42 tryptophan fluorophores of 24 proteins were taken in the training set (see Tables II and III).

We performed the discriminant analysis with the tryptophan residues from the training set to find the classification functions (CF_{ij}), where i denotes the respective class ($i = 1, \ldots, 5$) and j is the respective structural parameter ($j = 1, \ldots, 6$) (Table IV), which allows the construction of classification scores (S_i) for each class.

$$S_i = CF_i + CF_{1i} \cdot Acc + CF_{2i} \cdot Acc1-7 + CF_{3i} \cdot Den + CF_{4i} \cdot B + CF_{5i} \cdot R + CF_{6i} \cdot A \quad (1)$$

where *Acc*, *Acc1-7*, *Den*, *B*, *R*, and *A* are six structural parameters (see Table I). The classification scores are used to determine the most probable class to which a new object (tryptophan residue) belongs. A tryptophan residue belongs to the class for which it has the highest classification score.

To quantify the assignment of tryptophan residues to classes we introduced second classification approach that provides the values of the probabilities of the classification of new objects to five classes. The probability (P_i) is a function of three parameters: the Mahalanobis distances (MD_i) between new object assigned to the ith class and various classes centroids, within-classes covariance matri-

Table III
Six Structural Parameters of Environment of Tryptophan Residues of Proteins from the Training Set

Protein	*Acc*	*Acc1-7*	*Den*	*B*	*R*	*A*
Class A						
AZU W48	0.001	0	167	0.61	0.001	23.45
AZU W48[a]	0	0.0001	168	0.5	0	24
Class S						
ASP.N W66	1.78	7.15	147	0.75	1.35	38.25
PAC.CA W102	0	0	141	0.875	0	21.9
RNT W59	0	0	154	1.01	0	33.9
AX6 W343	0	0	154	0.88	0	34.5
OVH,LIS W194	4.38	2.01	144.3	0.675	2.9	38
PHA W109	0.23	1.045	154.5	0.545	0.11	34.9
VTA.HIS W30	1.66	3.125	136.8	0.825	1.37	33.4
A1AT W194	0	0	139	0.925	0	34
ACR W86	0	0	130	0.71	0	30.15
ACR W340	0	0	149	0.675	0	39.1
PAM W102	0.22	0	136	0.98	0.22	22
LAB W26	0	0	153	0.865	0	24.5
LAB W104	0.98	0	138	0.8	0.785	28.35
Class I						
KRP W75	0.33	0	136	1.41	0.465	35.8
ACR.G W356	3.69	12.185	133	1.04	3.85	34.25
VTC W231	8.85	10.4	133	1.17	10.305	43.8
VTC W31	5.39	17.85	136	0.945	5.095	36.15
LAB W118	11.5	22.65	123.5	0.835	9.04	42.8
LAB W60	2.66	8.925	121.8	1.02	2.665	31.35
Class II						
OVH.LIS W160	10.8	25.65	115	0.995	10.7	48.4
OVH.LIS W275	1.07	2.6	112.8	1.565	1.605	31.15
A1AT W238	20.5	65.25	102.5	0.89	18.25	39.9
MON W3	31.4	52.85	83	1.36	42.9	62.4
PHA W220	22.4	38.25	77.5	1.745	40.65	50.45
PHA W268	17.8	44.4	115.5	0.855	14.9	42.8
ASH.N W214	31.1	42	98	0.685	21.75	38.35
GPS W48	6.39	18.435	143	2.315	14.75	44.55
Class III						
NO2 W27	42	70.2	91	2.65	139.65	57.8
NO2 W28	39.5	61.3	85	2.105	114.8	49.75
PRH.LIS W6	61.5	74.7	78	0.95	58.25	62.8
PRH.LIS W42	39.7	86.6	69	1.315	52.1	50.05
PRH.PST W6	60.3	75	57	1.035	62.55	84.95
PRH.PST W42	34.9	81.85	78.5	1.825	45.9	60.05
AX6 W192	86.2	97.9	22	1.195	103.05	71.95
VTA.LIS W31	86.5	100	37	0.875	76	92.5
PLS W3	53.7	50.3	55	0.665	35.85	67.45
PLB W3	56.9	44.8	63	1.345	77.3	64.35
GLG W25	74.8	98.2	43	1.01	75.45	57.5
AWG W107	63	68.9	51	1.31	82.05	85.95
AWG W150	37.7	37.65	84	1.51	56.9	68.2

[a]AZU listed two times with similar parameters, since more than one object should be included in one class.

Table IV
Classification Functions (CF_{ij}) for Five Classes used for the Calculation of Classifications Scores

	Class A	Class S	Class I	Class II	Class III
CF_{i1} (for *Acc*)	1.887529	1.665856	1.660745	1.648479	1.896667
CF_{i2} (for *Acc1-7*)	0.116962	0.137027	0.205235	0.320241	0.349735
CF_{i3} (for *Den*)	1.749734	1.50127	1.400457	1.270009	1.225099
CF_{i4} (for *B*)	6.21771	11.41912	17.7545	24.44695	28.8079
CF_{i5} (for *R*)	−0.1984	−0.25385	−0.33244	−0.39137	−0.3135
CF_{i6} (for *A*)	0.241281	0.372331	0.431079	0.481138	0.571207
CF_i (constant)	−155.733	−120.833	−115.328	−111.483	−133.413

ces (S_i) calculated from the training set, and *a priori* probabilities q_i.

$$P_i = \frac{e^{\frac{D_i^2}{2}}}{\sum_{i=1}^{5} e^{\frac{D_i^2}{2}}} \tag{2}$$

where

$$D_i^2 = MD_i^2 + g_i' + g_i'' \tag{3}$$

$$g_i' = \ln |S_i| \tag{4}$$

$$g_i' = -2 \ln |q_i| \tag{5}$$

The Mahalanobis distance is a measure of distance in space, where variances and covariance within group is taken into account. The Mahalanobis distance (MD_i) between the tryptophan residue is represented by a multivariate vector X_{new}, and mean μ_i of the ith class of the training set ($i = 1, 2, \ldots 5$) and is defined as follows:

$$MD_i = \sqrt{(\mu_i - X_{\text{new}}) S_i^{-1} (\mu_i - X_{\text{new}})'} \tag{6}$$

The Mahalanobis distances are calculated in a two-dimensional space of roots (or canonical scores). First, we performed the discriminant and canonical analysis with the tryptophan residues from the training set to find the discriminant functions (DF_{ij}) (Table V), where j is the respective structural parameter ($j = 1, \ldots, 6$). This allows us to construct l number of canonical scores or roots (R_l) that are independent combinations of the six structural parameters (for the detailed description of the analysis, see Ref. 2):

$$R_l = DF_l + DF_{1l} \cdot Acc + DF_{2l} \cdot Acc1-7 + DF_{3l} \cdot Den + DF_{4l} \cdot B + DF_{5l} \cdot R + DF_{6l} \cdot A \tag{7}$$

Applying the sequential testing procedure[4,15] we found that two roots are the most significant ones, the rest are statistically nonsignificant and can be excluded from consideration. The calculation of roots leads to the reduction of the space dimension, i.e., conversion of the six-dimensional space of structural parameters into the two-dimensional space of roots. The roots are linear combinations of the structural parameters. Two roots were calculated for all tryptophan residues from the training using Eq. (7). In the next step, the Mahalanobis distances between the object and the centroids of five classes are computed in the two-dimensional space of roots. These distances are used for the calculation of the probabilities of the tryptophan assignments to classes.

A priori probabilities (q_i), which represent the natural occurrence of tryptophan fluorophores emitting at particular wavelengths were derived from the spectral distribution constructed previously.[2] Figure 3 presents the distribution of the occurrences of the positions of the maxima of more than 350 spectral components obtained as a result of the decomposition analysis of about 150 proteins. The distribution clearly indicates that tryptophan fluorophores of some classes are more frequent (the intensity is high for Classes II and S) than other classes (Class A). In other words, the probability of finding a

Table V
Discriminant Functions (DF_{ij}) Used for the Calculation of Two Most Significant Roots (or Canonical Scores)

	Root 1	Root 2
DF_{i1} (for *Acc*)	0.023049	−0.08415
DF_{i2} (for *Acc1-7*)	0.026581	0.0242
DF_{i3} (for *Den*)	−0.03878	−0.05765
DF_{i4} (for *B*)	2.208221	1.856157
DF_{i5} (for *R*)	−0.00819	−0.0484
DF_{i6} (for *A*)	0.026574	0.014981
DF_{i7} (constant)	−0.4562	6.108798

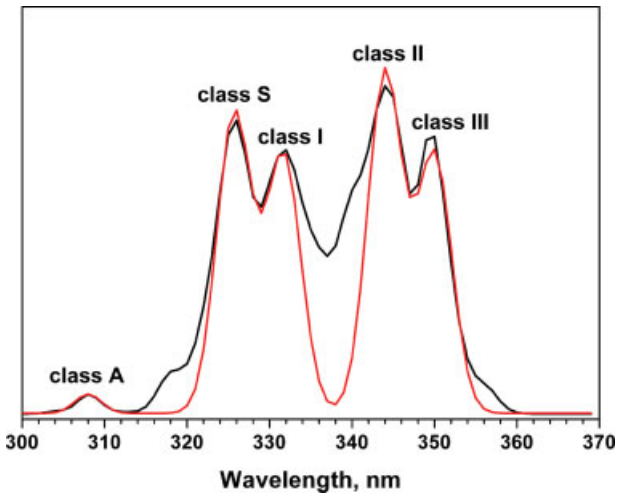

Figure 3

The distribution of occurrence of maximum positions of more than 350 spectral components obtained as a result of the decomposition analysis of about 150 proteins[2]—black line. Red-line is a best fit of the distribution of occurrence of maximum positions of spectral components by sum of five Gauss functions, which represent each of five classes.

tryptophan residue in a protein emitting at ~342 nm (Class II) is higher than the probability of finding a tryptophan residue emitting at 308–310 nm (Class A). To calculate *a priori* probabilities of the assignment of tryptophan fluorophores (objects) to five classes we fitted the distribution of the occurrence of the maximum position of the spectral components using the sum of five Gauss functions representing the five classes. The best fit was found for Gauss functions with the maxima positioned at 308.0, 325.7, 331.7, 344.0, and 350.0 nm. The area under the each Gauss functions was used to establish *a priori* probabilities, taking into account that the sum of all probabilities should be equal to 1 (Table VI).

The output summary file is generated as a result of the statistical analysis. It contains information about the classification scores, the Mahalanobis distances, and the probabilities of assignment of each of the tryptophan residues in protein to one of five spectral-structural classes.

PFAST database

PFAST contains not only a package of programs, but it is also the first database of protein spectral data and protein structural data, and their assignment to each other. The information in the PFAST database is stored in two major tables FCAT and SCAT, representing the fluorescence and structural data, respectively. Each table can be queried separately or together using links between them.

The FCAT table stores information about the protein name, source, experimental setting, and the conditions of fluorescence measurements; name of author who collected the data; the publication reference; and comments. The FCAT table includes the input file used in the calculations, and three output files generated by the FCAT tool. All files can be viewed or downloaded, and a graphical representation of the fluorescence decomposition results can be visualized. The links to the related experiments are available. The SCAT table contains information about the protein name and source, the PDB-code, the resolution of the crystal structure, the number of tryptophan residues in the protein, and the comments line. The SCAT table includes six output files generated by the SCAT tool. These files can be viewed or downloaded. The links to the related experiments are available.

The user can retrieve data from the database using protein name, the PFAST code, or the source. Submission of any new data to the database can occur only through contact with the system administrator. The data will be verified before submission.

DISCUSSION

Protein fluorescence is a powerful tool for the study of protein structure and conformation if we have a means to interpret the spectral data in terms of protein structural properties. Our previous research successfully provided this support through the development of individual software modules implementing the algorithms for fluorescence and structural analyses.[1–4] We have combined the developed modules, introduced a new one, and presented a web-based tool, PFAST (http://pfast.phys.uri.edu/pfast/) (Fig. 4 presents interaction diagram of PFAST web-site):

1. For the analysis of protein fluorescence spectra by revealing the spectral components of individual tryptophan residues, or groups of tryptophan residues located close to each other and having energy transfer among them. As a result of the FCAT module calculations, the fluorescence spectra can be decomposed and spectral components can be assigned to one of the five spectral-structural classes.
2. For the analysis of structural properties of tryptophan environment in proteins. As a result of the SCAT module, the structural properties of tryptophan residues

Table VI

Areas Calculated for Each of Five Gauss Functions Used for the Fitting of the Distribution of Occurrence of Position of Maximum of Spectral Components, the Areas Were Used to Calculate A Priori Probabilities (q_i)

	Class A	Class S	Class I	Class II	Class III	Total
Area under the Gauss functions	1.4	29	25	33	25	113.4
A priori probabilities, q_i	0.01	0.26	0.22	0.29	0.22	1.0

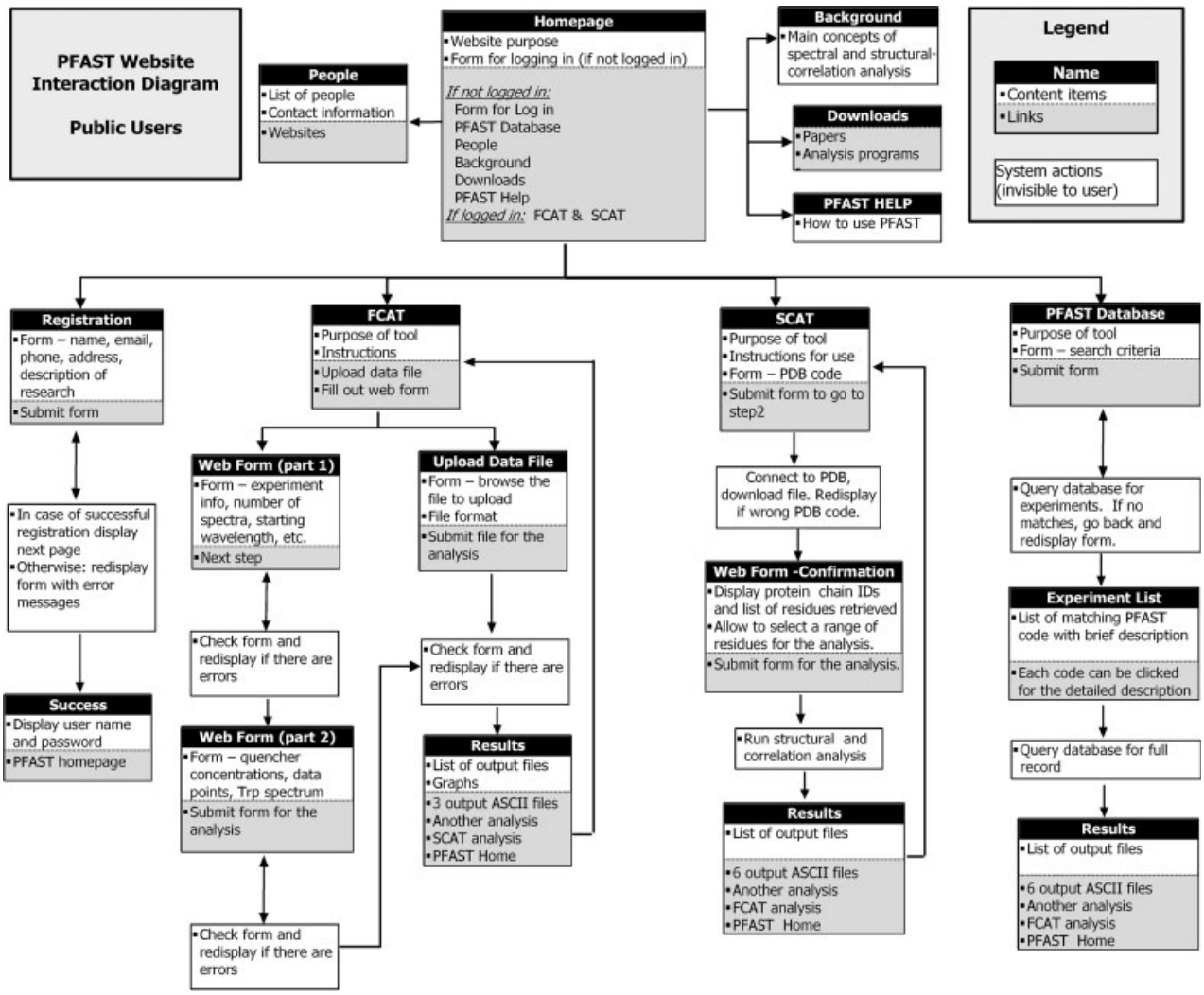

Figure 4

The PFAST web-site interaction diagram.

can be calculated, and the tryptophan residues can be assigned to one of the five spectral-structural classes.

3. The PFAST database contains the protein spectral and structural data obtained from the FCAT and SCAT analysis.

PFAST is a first toolkit that allows researchers to establish a correlation between protein fluorescence and structural parameters by assigning the tryptophan residues to one of five spectral-structural classes. The tryptophan residues, which belong to various classes, have different spectral and structural properties and, as a result, the different combinations of universal (dipole–dipole relaxation[16–23]) and specific (exciplex formation[24–28]) interactions could occur in the excited state of tryptophan fluorophores.

Class A contains tryptophan fluorophores deeply buried in protein matrix with nonpolar and nonflexible environment consisting mostly of atoms that involved in stabilization of elements of secondary structure of protein. No partners for H-bonding revealed for tryptophan residue of this class. As a result, the emission of tryptophan residues of this class is structured and has extremely short-wavelength position of maximum (308 nm) with zero accessibility of fluorophores to quenchers of tryptophan fluorescence. There are no specific or universal interactions in the excited state of tryptophan fluorophores of this class.

Class S Tryptophan residues have very similar to the fluorophores of Class A structural properties of environment: fluorophores also are deeply buried inside protein. The major difference between the structural properties of

tryptophans of Class A and S is higher polarity and flexibility of the microenvironment. Also, there are evident free partners for hydrogen bond formation nearby the fluorophores of Class S. The spectral properties of tryptophan residues of Class S is long-wavelength shifted to 322.5 ± 4.5 nm. Weak dipole–dipole interactions and exciplexes with 1:1 stoichiometry might be formed in the excited state of fluorophores of Class S.

Class I represents the fluorophores with the averaged maximum position of fluorescence at 331.0 ± 4.8 nm, the average width of spectrum $\Delta\lambda$ = 48–50 nm, and the relative Stern-Volmer constant values of about 10%. The Nε1 and Cζ2 atoms of the indole ring (where the most redistribution of electron density occurs after the absorption of photons[29]) are in contact with water-structured molecules. These atoms of indole ring could be good candidates for the formation of H-bonds in the excited state in 2:1 stoichiometry with surrounding protein or water molecules. The environment of tryptophan residues of Class I have lowered packing density that may result in an increase in frequency and/or amplitude of structural mobility of the environment favoring both hydrogen-bonded exciplex formation and dipole relaxation during the lifetime of the fluorophore's excited state. However, we assume that the environment of tryptophan residues of Class I is less flexible than water molecules and, therefore, time of dipole relaxation of environment is higher than the lifetime of fluorescence (ns), which precludes completing the relaxation-induced spectral shift during the excited state lifetime.

Class II contains the fluorophores with the averaged maximum position of fluorescence at 342.3 ± 3.3 nm, the average width of spectrum $\Delta\lambda$ = 53–55 nm, and the relative Stern-Volmer constant of 44%. The main feature of environment of tryptophan residues of this class is presence of the structured-water molecule near to indole ring. The time of dipole relaxation of structured-water is less than fluorescence lifetime, which results in dipole–dipole reorientation and exciplex formation in the excited state.

Class III Fluorophores of this class has spectral properties similar to the free tryptophan in solution: the averaged maximum position of fluorescence at 347.0 ± 3.1 nm, average width of spectrum $\Delta\lambda$ = 59–61 nm, and relative Stern-Volmer constant of about 76%. Tryptophan residues of this class are fully exposed to highly mobile water molecules that are enabled to complete relaxation during the excitation lifetime. As a result, the spectra position of tryptophan residues of Class III almost coincide with those of free aqueous tryptophan.

ACKNOWLEDGMENTS

We thank Prof. Edward Burstein and Victor Emelyanenko, Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences for the encouraging discussions, suggestions and comments. Also, we thank Steve Pellegrino, Physics Department, URI for the help with the server set up for PFAST.

REFERENCES

1. Burstein EA, Abornev SM, Reshetnyak YK. Decomposition of protein tryptophan fluorescence spectra into log-normal components. I. Algorithm of decomposition. Biophys J 2001;81:1699–1709.
2. Reshetnyak YK, Burstein EA. Decomposition of protein tryptophan fluorescence spectra into log-normal components. II. The statistical proof of discreteness of tryptophan classes in proteins. Biophys J 2001;81:1710–1734.
3. Burstein EA, Emelyanenko VI. Log-normal description of fluorescence spectra of organic fluorophores. Photochem Photobiol 1996; 64:316–320.
4. Reshetnyak YK, Koshevnik Y, Burstein EA. Decomposition of protein tryptophan fluorescence spectra into log-normal components. III. Correlation between fluorescence and microenvironment parameters of individual tryptophan residues. Biophys J 2001;81: 1735–1758.
5. Burstein EA, Vedenkina NS, Ivkova MN. Fluorescence and the location of tryptophan residues in protein molecules. Photochem Photobiol 1973;18:263–279.
6. Linke D, Frank J, Pope MS, Soll J, Ilkavets I, Fromme P, Burstein EA, Reshetnyak YK, Emelyanenko VI. Folding kinetics and structure of OEP16. Biophys J 2004;86:1479–1487.
7. Reshetnyak YK, Kitson RP, Lu M, Goldfarb RH. Conformational and enzymatic changes of 20S proteasome of rat natural killer cells induced by mono and divalent cations. J Struct Biol 2004;145:263–271.
8. Torrent M, Cuyás E, Carreras E, Navarro S, López O, de la Maza A, Nogués MV, Reshetnyak YK, Boix E. Topology studies on the membrane interaction mechanism of the eosinophil cationic protein. Biochemistry 2007;46:720–733.
9. Reshetnyak YK, Tchedre KT, Nair MP, Pritchard PH, Lacko AG. Structural differences between wild-type and fish eye disease mutant of lecithin: cholesterol acyltransferase. J Biomol Struct Dyn 2006; 24:75–82.
10. Orlov NY, Reshetnyak YK, Orlova TG, Orlov DN, Burstein EA, Ishijima Y, Kimura N. Protein fluorescence study of chimeras and tagged forms of recombinant rat nucleoside diphosphate kinases α and β. Biological Membranes 2003;20:53–59.
11. Orlov NY, Orlova TG, Reshetnyak YK, Burstein EA, Kimura N. Comparative study of recombinant rat nucleoside diphosphate kinases α and β by intrinsic protein fluorescence. J Biomol Struct Dyn 1999;16:955–968.
12. Orlov NY, Orlova TG, Reshetnyak YR, Burstein EA, Kimura N. Interaction of recombinant rat nucleoside diphosphate kinase with bleached bovine rod outer segment membranes: a possible mode of pH and salt effects. Biochem Mol Biol Int 1997;41:189–198.
13. Reshetnyak YK, Andreev OA, Borejdo J, Toptygin DD, Brand L, Burstein EA. The identification of tryptophan residues responsible for ATP-induced increase in intrinsic fluorescence of myosin subfragment 1. J Biomol Struct Dyn 2000;18:113–125.
14. McDonald I, Thronton J. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1995;238:777–793.
15. Mendoza JL, Markos VH, Gonter R. A new perspective on sequential testing procedures in canonical analysis: A Monte Carlo evaluation. Multivariate Behav Res 1978;13:371–382.
16. Mataga N, Kaifu Y, Koizumi M. The solvent effect on fluorescence spectrum change of solute–solvent interaction during the lifetime of excited solute molecule. Bull Chem Soc Jpn 1955;28:690–691.

17. Mataga N, Kaifu Y, Koizumi M. Solvent effects upon fluorescence spectra and the dipole-moments of excited molecules. Bull Chem Soc Jpn 1956;29:465–470.
18. Mazurenko YT, Bakhshiev NG. Effect of orientation dipole relaxation on spectral, time, and polarization characteristics of the luminescence of solutions. Opt Spectrosc 1970;26:490–494.
19. Vincent M, Gallay J, Demchenko AP. Solvent relaxation around the excited state of indole: analysis of fluorescence lifetime distributions and time-dependence spectral shifts. J Phys Chem 1995;99:14931–14941.
20. Toptygin D, Brand L. Spectrally- and time-resolved fluorescence emission of indole during solvent relaxation: a quantitative model. Chem Phys Lett 2000;322:496–502.
21. Lakowicz JR. On spectral relaxation in proteins. Photochem Photobiol 2000;72:421–437.
22. Toptygin D. Effects of the solvent refractive index and its dispersion on the radiative decay rate and extinction coefficient of a fluorescent solute. J Fluoresc 2003;13:201–219.
23. Nilsson L, Halle B. Molecular origin of time-dependent fluorescence shifts in proteins. Proc Natl Acad Sci 2005;102:13867–13872.
24. Lumry R, Hershberger M. Status of indole photochemistry with special reference to biological application. Photochem Photobiol 1978;27:819–840.
25. Hershberger MV, Lumry R, Verrall R. The 3-methylindole/*n*-butanol exciplexes evidence for two exciplex sites in indole compounds. Photochem Photobiol 1981;33:609–617.
26. Kavarnos GJ, Turro NJ. Photosensitization by reversible electron transfer: theories, experimental evidence, and examples. Chem Rev 1986;86:401–449.
27. Sengupta T, Basu S. Magnetic field effect on indole exciplexes: a comparative study. Spectrochim Acta A Mol Biomol Spectrosc 2004;60:1127–1132.
28. Stalin T, Rajendiran N. Effects of solvent, pH and β-cyclodextrin on the photophysical properties of 4-hydroxy-3,5-dimethoxybenzaldehyde: intramolecular charge transfer associated with hydrogen bonding effect. Spectrochim Acta A Mol Biomol Spectrosc 2005;61:3087–3096.
29. Callis PR. ^1L_a and ^1L_b transitions of tryptophan: applications of theory and experimental observations to fluorescence of proteins. Methods Enzymol 1997278:113–151.