Bioinformation
www.bioinformation.net

open access
Web Server

CFP: a web-server for constructing sequence-based protein conformational flexibility profiles

Igor B. Kuznetsov[1,*], Shalom Rackovsky[2]

[1] Cancer Research Center, Department of Epidemiology and Biostatistics, University at Albany, One Discovery Drive, Rensselaer, NY 12144, USA; [2] Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, One Gustave L Levy Place, New York, NY 10029, USA; Igor B. Kuznetsov – Email: ikuznetsov@uamail.albany.edu; *Corresponding author

Received September 20, 2009; Accepted October 14, 2009; Published October 19, 2009

Abstract:
Many proteins contain conformationally flexible segments that undergo significant changes in the backbone conformation or completely lack a well-defined conformation. Previously, we have developed the generalized local propensity (GLP), a quantitative sequence-based measure of the protein backbone flexibility. In this paper, we present the CFP (Conformational Flexibility Profile) web-server that constructs the GLP flexibility profile for a user-submitted sequence and uses this profile to identify segments with high backbone flexibility. The statistical significance of a flexible sequence segment is assessed using the discrete scan statistics based on the density of flexible residues observed in this segment.

Availability: CFP is publicly available at http://cfp.rit.albany.edu

Keywords: conformational variability, protein backbone, flexibility, local propensity, sequence

Background:
Many proteins contain conformationally flexible segments. These segments undergo significant changes in backbone conformation, or are completely disordered (lack a well-defined structure) **[1-3]**. A quantitative representation of the conformational flexibility of the protein backbone is important for many applications. Previously, we developed generalized local propensity (GLP), a quantitative sequence-based measure of backbone flexibility **[4]**. The GLP can be used to construct sequence-based protein flexibility profiles, and provides an objective numeric threshold for defining conformationally flexible segments **[5]**. For a given sequence position *k*, the GLP measures the width of the context-dependent distribution of backbone conformations accessible to this position, *glp(k)* (see references **[4-5]** for details). If *glp(k)* $\geq$ 1, it indicates that sequence position *k* is conformationally flexible.

Here, we present the CFP (Conformational Flexibility Profile) web server that constructs the GLP flexibility profile for a user-submitted sequence and uses this profile to identify segments with high conformational flexibility. Below is a brief outline of the steps implemented in CFP: The GLP flexibility profile is constructed for the query sequence and then smoothed using a sliding window of size W_1. Consecutive positions which have GLP above a threshold T_1 are merged into seed flexible segments. Each seed flexible segment is extended by adding extension windows of size W_2 until its average GLP drops below an extension threshold T_2. An extension window is added only if its average GLP is above a certain threshold T_3. This extension procedure is similar to that used in the SEG program **[6]**.

The extended flexible segments are reported in the final table. If the number of flexible residues observed in a given final flexible segment is unusually high (p-value < 0.05), then this segment is marked as statically significant. The significance of the number of flexible residues is estimated using the discrete scan statistic. This statistical procedure is the same as the one we previously implemented in the BIAS software to identify statistically significant clusters of user-specified amino acid types **[7-8]**. The web-server is publicly available at http://cfp.rit.albany.edu.

Methodology
Input:
The only mandatory input is the query protein sequence. All other input fields have default values that can be modified by advanced users, if desired. These input fields are described below. Instructions for each field and general information about the methodology and the output format can be found by clicking a corresponding help hyperlink on the input page.

Smoothing window size: The size of the sliding window (W_1) used to smooth the raw profile. High values of W_1 tend to reveal long flexible segments and mask the short ones. Lower values tend to reveal short segments.

GLP threshold for seed segments: The threshold T_1 used to identify seed flexible segments. Contiguous sequence positions that have values of the smoothed GLP profile above this threshold are merged into a seed flexible segment.

Extension threshold - Each seed segment with high flexibility is extended on both sides until its average GLP drops below this threshold (T_2).
Extension window threshold - The ends of a seed flexible segment are extended if the extension window has the average GLP above this threshold (T_3).

Extension window size: The size of the extension window (W_2).

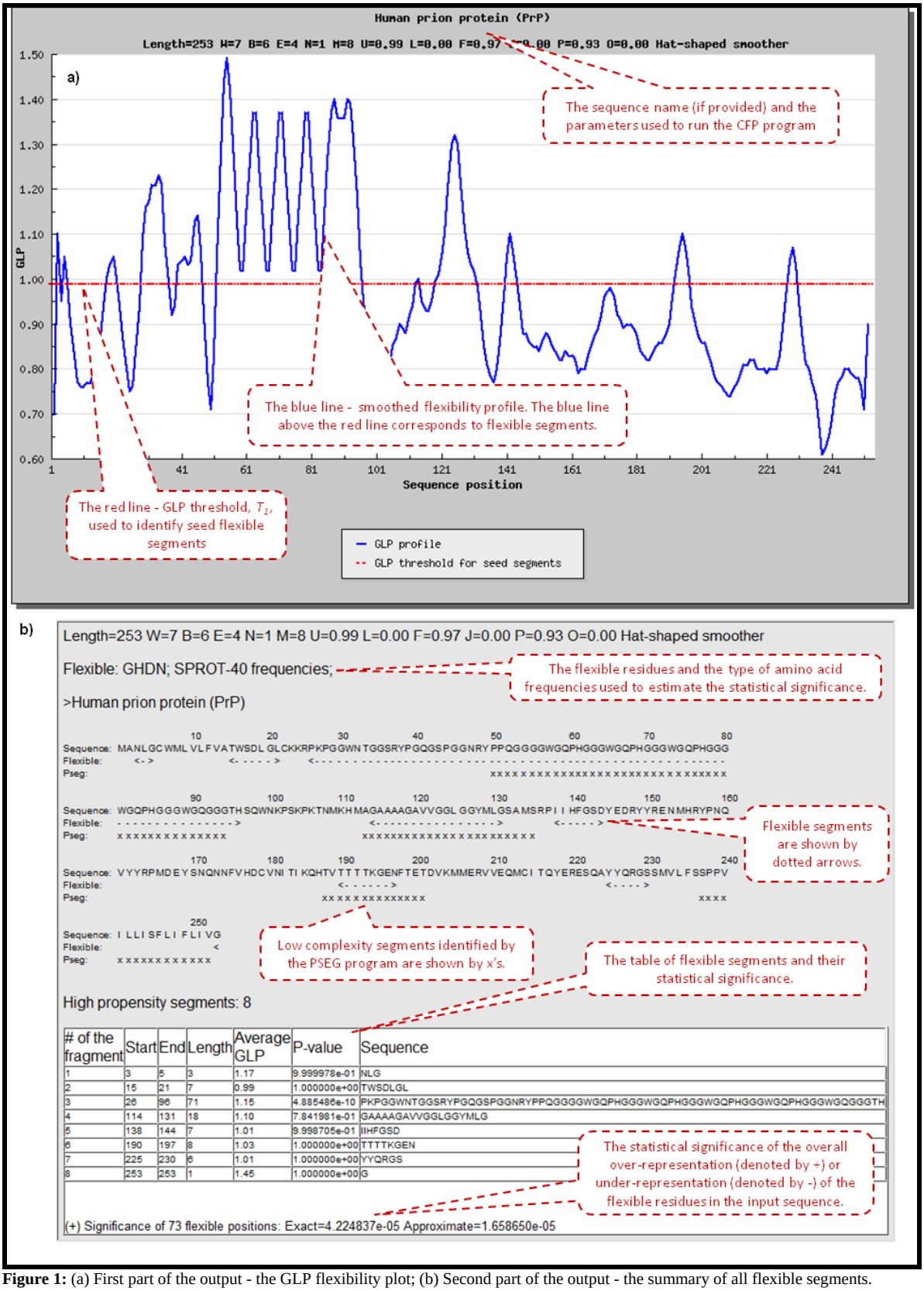
Hat-shaped local smoother: Positions in the center of the smoothing window contribute more to the smoothed GLP score than positions at the ends of the window.

Equal weights smoother: The smoothed GLP score is the unweighted average computed over all positions in the window.

Minimum seed segment: Seed segments with length smaller than this threshold are not extended.

Maximum separation between merged segments: Flexible segments separated by this or smaller number of positions are merged into one.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4 (5): 176-178 (2009)



# of the fragment	Start	End	Length	Average GLP	P-value	Sequence
1	3	5	3	1.17	9.999978e-01	NLG
2	15	21	7	0.99	1.000000e+00	TWSDLGL
3	26	96	71	1.15	4.885486e-10	PKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTH
4	114	131	18	1.10	7.841981e-01	GAAAAGAVVGGLGGYMLG
5	138	144	7	1.01	9.998705e-01	IIHFGSD
6	190	197	8	1.03	1.000000e+00	TTTTKGEN
7	225	230	6	1.01	1.000000e+00	YYQRGS
8	253	253	1	1.45	1.000000e+00	G

Figure 1: (a) First part of the output - the GLP flexibility plot; (b) Second part of the output - the summary of all flexible segments.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4 (5): 176-178 (2009)

Flexible residues: A set of flexible residues used in the scan statistics to estimate the statistical significance of flexible segments (G, H, D, N by default).

SWIS-PROT or PDB background frequencies: The amino acid frequencies of the SwissProt or Protein Databank are used to estimate the statistical significance.

X axis size, Y axis size: The size of X and Y axis of the plot in pixels.

Create a plot: Display the smoothed GLP profile in web-browser.

Create a text file: Save the raw and smoothed GLP profiles in a text file.

Output:
The CFP output consists of two parts. The first part shows the smoothed GLP plot of the input sequence **(Figure 1A)**. The second part shows the detailed information for every flexible segment found in the input sequence and the p-values that provide the estimates of the statistical significance **(Figure 1B)**. If the p-value for a given segment is less than 0.05, this segment has an unusually high density of residues with high degree of backbone flexibility.

Acknowledgement:
This work was supported by grant number R03LM009034 from the National Library of Medicine/NIH, and by grant number LM06789 from the National Library of Medicine/NIH.

References:
[1] V Alexandrov e*t al.*, *Protein Sci*, **14:** 633 (2005) [PMID: 15722444]
[2] M Kosloff, R Kolodny, *Proteins* **7:** 891 (2008) [PMID: 18004789]
[3] AK Dunker *et al.*, *BMC Genomics*, **9:** S1 (2009) [PMID: 18831774]
[4] IB Kuznetsov, S Rackovsky, *Protein Sci,* **12:** 2420 (2003) [PMID: 14573856]
[5] IB Kuznetsov, S Rackovsky, *Protein Sci,* **13:** 3230 (2004) [PMID: 15557265]
[6] JC Wootton, S Federhen, *Methods Enzymol,* **266:** 554 (1996) [PMID: 8743706]
[7] IB Kuznetsov, S Hwang, *Bioinformatics* **22:** 1055 (2006) [PMID: 16500936]
[8] IB Kuznetsov, *Bioinformatics* **24:** 1534 (2008) [PMID: 18480099]

Edited by P. Kangueane
Citation: Kuznetsov & Rackovsky, Bioinformation 4 (5): 176-178 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4 (5): 176-178 (2009)