Bioinformation
www.bioinformation.net

open access
Web Server

JUZBOX: A web server for extracting biomedical words from the protein sequence.

Paul Bobby, Seetharaman Balaji*, Variath Sathyanath, Santhosh J. Eapen

Indian Institute of Spices Research, Calicut, Kerala, India; S Balaji – Email: blast_balaji@rediffmail.com; Phone: +919036260732;
*Corresponding author: Present address: - Dept. of Biotechnology, Manipal Institute of Technology, Manipal University, Manipal, Karnataka, India.

Received May 25, 2009; Revised July 31, 2009; Accepted September 11, 2009; Published November 17, 2009

Abstract:
The recognition of gene/protein names in literature is one of the pivotal steps in the processing of biological literatures for information extraction or data mining. We have compiled a lexicon of biomedical words (conserved patterns/ potential motifs) which has the combination of only 20 alphabets of amino acids. The remaining 6 letters of the English alphabets (B, J, O, U, X, Z) are treated as invalid amino acid characters (to our context), We have jumbled the 6 letters for the sake of usage and convenience and termed as 'JUZBOX' and these characters were filtered in the biomedical lexicon. Undoubtedly, the generation of biomedical words from protein sequence using JUZBOX have applications specific for functional annotation.

Keywords: JUZBOX; biomedical words; lexicon;

Availability: JUZBOX is available freely at http://www.spices.res.in/juzbox

Background:
The exponential growth of large-scale molecular sequence data and of the PubMed scientific literature has prompted active research in biological literature mining and information extraction to facilitate genome/proteome annotation and improve the quality of biological databases **[1]**. Kranthammer *et al*. tagged gene and protein names in the biomedical text by substituting each letter of a name with a unique nucleotide sequence and finding "hits" in the literature using the BLAST algorithm **[2]**. Tanabe and Wilbur developed a gene/protein tagger, AbGene that identifies gene and protein names to extract single names, filtering for false negatives using a gene name dictionary and low frequency trigrams, identifying compound names using combinations of patterns found in known gene names and filtering for false negatives using contextual rules **[3]**. Motivated by the resource for protein literature mining - iProLINK (integrated Protein Literature INformation and Knowledge) **[1]** and other text mining methodologies like BioCreAtIvE **[4]**, we have used a simple approach to identify and determine conserved patterns or motifs in protein sequence which is essential for any biological sequence analysis. Hence, we have developed a web server 'JUZBOX' to extract conserved patterns (biomedical words) from the protein input sequence provided by the user.

Methodology:
We have compiled an array of lexicon of biomedical words (conserved patterns/ potential motifs) to find their biochemical functions in the protein sequence. The words were extracted and processed from UCMP Glossary© for biological terms and the New International Webster's Medical Dictionary © J. Radcliffe for medical terms. Besides, other technical and popular medical terms were also included to obtain a collection of over 2009 terms. We have processed the lexicon for the valid biomedical words, here valid words represents those words having the combination of 20 alphabets of amino acids. The remaining 6 letters of the English alphabets (B, J, O, U, X, Z) are treated as non-amino acid characters (because the amino acids representation with these alphabets are not so common). We jumble the 6 letters for the sake of usage and convenience and termed as 'JUZBOX' and these characters (in a biomedical word) are treated as invalid characters (to our context). Here we describe an approach which is based on the generation of valid biomedical terms from any protein sequences that are characterized by the commonly used 20 single letter amino acid character codes. We also have subjected the valid biomedical terms manually for pattern and motif analysis to find the conservation as well as the biological significance. The overall methodology is illustrated in the following figure (**Figure 1**).

Web interface:
The current version of JUZBOX is a web based tool and it was developed by using PHP, a server side scripting language.

Program input:
The user can input the protein sequence by pasting in FASTA/ raw format (**Figure 2**) and choose the lexicon options (1) PROSITE and (2) Biomedical words. The first option displays only the true positive hits, i.e. those valid biomedical words have hits in the PROSITE database and the second option displays all possible combination of valid biomedical words from the given sequence.

Program output:
In the result page **(Figure 3)**, if the user has selected the first option, it lists all the true positive hits, i.e. those words which have hits in the PROSITE database. The second option (Biomedical words) lists out all valid biomedical words from the given sequence input.

Applications:
JUZBOX has been applied to evaluate the potential motifs from the protein sequence and found some interesting motifs. These motifs are conserved in major domains which are of biochemical interest, for example the pattern 'DNA' (i.e. Aspartic acid 'D' followed by Asparagine 'N' followed by Alanine 'A'). This pattern is conserved in more than thousand sequences in the PROSITE database and it has been found in the SH2 and SH3 domains having varied functional regions which correspond to the secondary structure. The main focus is to efficiently extract the relevant structural and functional information from biomedical words. This provides the foundation for future knowledge-discovery tools able to identify previously undiscovered associations of the conserved pattern, its corresponding motif and functional sites. Undoubtedly, the generation of biomedical words from protein sequence have applications specific for functional annotation and it is the only way to cope with the tsunami of sequences produced in this field.

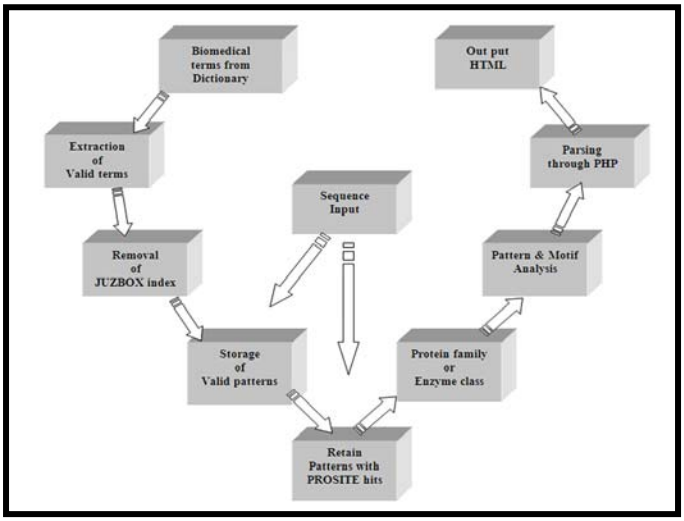


Figure 1: Flowchart illustrating the methodology (in anti-clock wise direction)

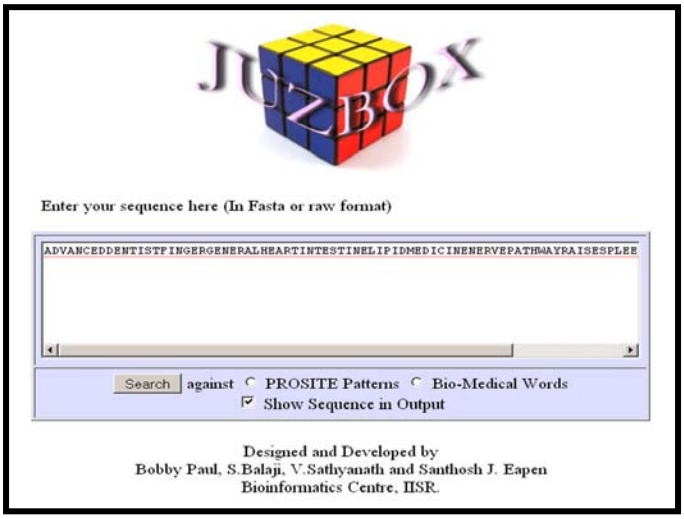
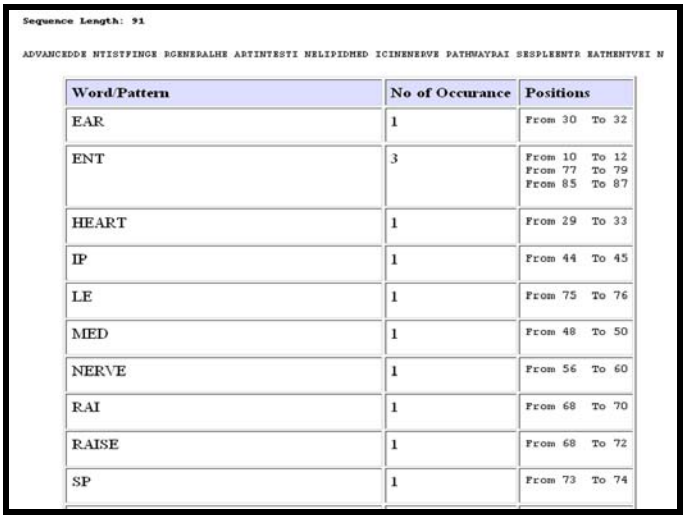


Figure 2: JUZBOX sequence input.



Sequence Length: 91

ADVANCEDDE NTISTFINGE RGENERALHE ARTINTESTI NELIPIDMED ICINENERVE PATHWAYRAI SESPLEENTR EATMENTVEI N

Word/Pattern	No of Occurance	Positions
EAR	1	From 30 To 32
ENT	3	From 10 To 12 From 77 To 79 From 85 To 87
HEART	1	From 29 To 33
IP	1	From 44 To 45
LE	1	From 75 To 76
MED	1	From 48 To 50
NERVE	1	From 56 To 60
RAI	1	From 68 To 70
RAISE	1	From 68 To 72
SP	1	From 73 To 74

Figure 3: JUZBOX result page.

Future development:
We are developing the output options in more attracting form by providing the meaning for each biomedical word found in the sequence as well as the biological meaning for those hits like active site, ion binding site, glycosylation site, phosphorylation site, myristylation sites etc.

Acknowledgement:
This work was supported by Distributed Information Sub Centre for Bioinformatics, Indian Institute of Spices Research, Calicut, Kerala, India.

References:
[1] ZZ Hu *et al.*, *Computational Biology and Chemistry*, **28**: 409 (2004) [PMID: 15556482]
[2] M Krauthammer *et al.*, *Gene* **259**:245 (2000) [PMID: 11163982]
[3] L Tanabe, WJ Wilbur, *Bioinformatics* **18**:1124 (2002) [PMID: 12176836]
[4] L Hirschman *et al.*, *BMC Bioinformatics*, **6:**S1 (2005) [PMID: 15960821]

Edited by P. Kangueane
Citation: Paul *et al.*, Bioinformation 4(5): 179-181 (2009)