A Text-Based Framework for Prediction of Protein Subcellular Localization: SherLoc & EpiLoc

Ren Li
renli@udel.edu
Fanchao Meng
fcmeng@cis.udel.edu

Computer and Information Sciences Department
University of Delaware

April 29, 2014

Outline

Motivation

- Protein Subcellular Localization Problem: Locations => Functions
- Experimental methods?
  - E.g. Based on green fluorescent proteins or immunolocalization [1]
  - **Pros:** Accurate
  - **Cons:** Slow, Labor-intensive, Only effective for proteins readily expressible
- Scenarios we are facing… [1]
  - Large number of little-known proteins
  - Not being expressed yet
  - What to do now?
- Computational inference!
  - Based on sequence data [2] [3] [4] [5]
  - Based on text data [1]

Motivation (cont.)

- Text-Based Methods
  - Explicitly extract location statements from the literature
    - ➢ Not prediction! [1]
  - Classifiers using text-features
    - Earlier works may also need explicit information such as the protein's name [1] [7]
    - Predictive but depend on an explicit text source, and the performance is unknown! [1]
  - SherLoc + EpiLoc

An Informal Example of Text-Based Inference

Task:

Determine if a person is a good professor or not?

As known Dr. Shatkay is a good professor:

She always works late at night

She's patient on answering questions

Her office is Smith 416

Her students are all outstanding

Almost all CIS professors are in Smith Hall

She has done a lot of research in Bioinfo

She is good at BioInfo and Database

She is smart

Should I take the Database class next semester? People told the teacher is not bad

Most of CIS professors are fine

She always makes notes very clarified

Her homework is reasonable and helpful

An Informal Example of Text-Based Inference (cont.)

Task:

Determine if a person is a good professor or not?

Get a pattern?

*She always **works late at night***

*She's **patient** on **answering questions***

*Her **office** is **Smith 416***

*Her **students** are all **outstanding***

*Almost all **CIS professors** are in **Smith Hall***

*She has done **a lot of research** in **Bioinfo***

*She is **good at BioInfo** and **Database***

*She is **smart***

*Should I take the **Database class** next **semester**? People told the **teacher** is **not bad***

*Most of **CIS professors** are **fine***

*She always makes **notes** very **clarified***

*Her **homework** is **reasonable** and **helpful***

An Informal Example of Text-Based Inference (cont.)

Get a pattern?

Person → Terms → Good Professor

Terms
Office
Smith, 416
CIS, Professor, Fine
Smart
Note, Clarified
Student, Outstanding
Work, Late at night
A lot of, Research
Good at, BioInfo, Database
Class, Semester
Teacher, Not bad
Patient, Answer, Question
Homework, Reasonable, Helpful

An Informal Example of Text-Based Inference (cont.)

Question:

Is Dr. McCoy a good professor?

Comments:

She comes to Smith every day

She gave us many useful handouts

Even though she's always busy, she would like to answer our questions anytime

She's quite productive in her research area

Her classes and homework are very accessible and helpful

She always replies mails at midnight

Her research mainly on AI and IR

She is very smart when talking about Math

She got an Award in CIS deptarment

She got a high score of teaching on ratemyprofessor.com

Priscilla and Zhou are all top Ph.D students in her lab

She always makes notes very clarified

An Informal Example of Text-Based Inference (cont.)

Question:

Is Dr. McCoy a good professor?

Comments:

*She comes to **Smith** every day*

*She gave us many **useful handouts***

*Even though she's always busy, she would like to **answer** our **questions anytime***

*She's quite **productive** in her **research** area*

*Her **classes** and **homework** are very **accessible** and **helpful***

*She always **replies mails** at **midnight***

*She is an **expert** mainly in **AI** and **IR***

*She is very **smart** when talking about Math*

*She got an **Award** in **CIS** deptartment*

*She got a **high score** of **teaching** on ratemyprofessor.com*

*Priscilla and Zhou are all **top** Ph.D **students** in her lab*

*She always explain her **notes** very **clearly***

An Informal Example of Text-Based Inference (cont.)

Match the pattern?

Pattern Terms	Terms for Dr. McCoy
Office	
Smith, 416	Smith
CIS, Professor, Fine	CIS, Award
Smart	Smart
Note, Clarified	Note, Clear
Student, Outstanding	Student, Top
Work, Late at night	Reply Mail, Midnight
A lot of, Research	Productive, Research
Good at, BioInfo, Database	Expert, AI, IR
Class, Semester	Class
Teach, Not bad	Teach, High Score
Patient, Answer, Question	Anytime, Answer, Question
Homework, Reasonable, Helpful	Homework, Accessible, Helpful

Totally Match

Synonym

Same Category

An Informal Example of Text-Based Inference (cont.)

Make an inference?

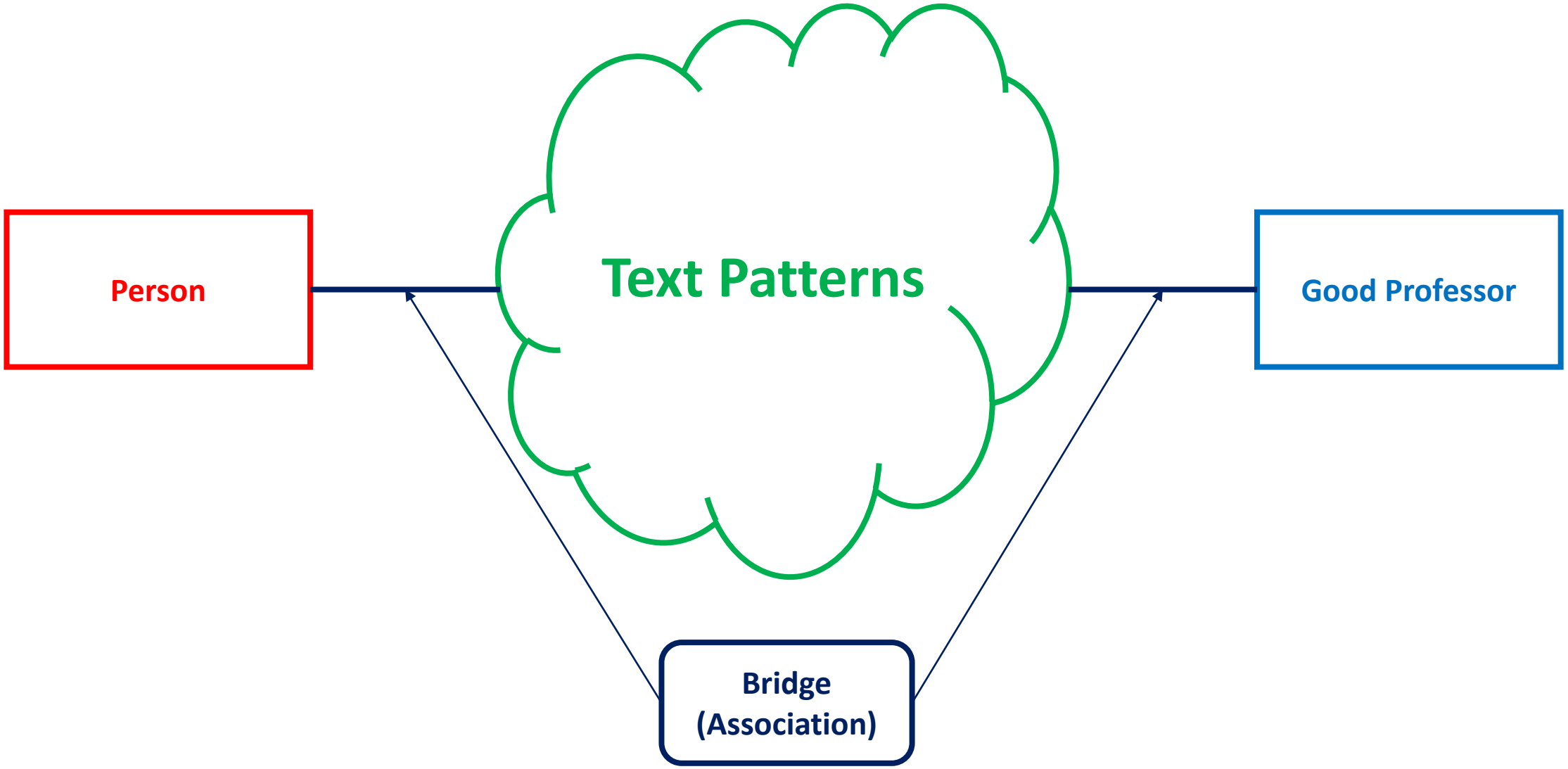
Dr. McCoy →

Pattern Terms	Terms for Dr. McCoy
Office	
Smith, 416	Smith
CIS, Professor, Fine	CIS, Award
Smart	Smart
Note, Clarified	Note, Clear
Student, Outstanding	Student, Top
Work, Late at night	Reply Mail, Midnight
A lot of, Research	Productive, Research
Good at, BioInfo, Database	AI, IR, Math
Class, Semester	Class
Teach, Not bad	Teach, High Score
Patient, Answer, Question	Anytime, Answer, Question
Homework, Reasonable, Helpful	Homework, Accessible, Helpful

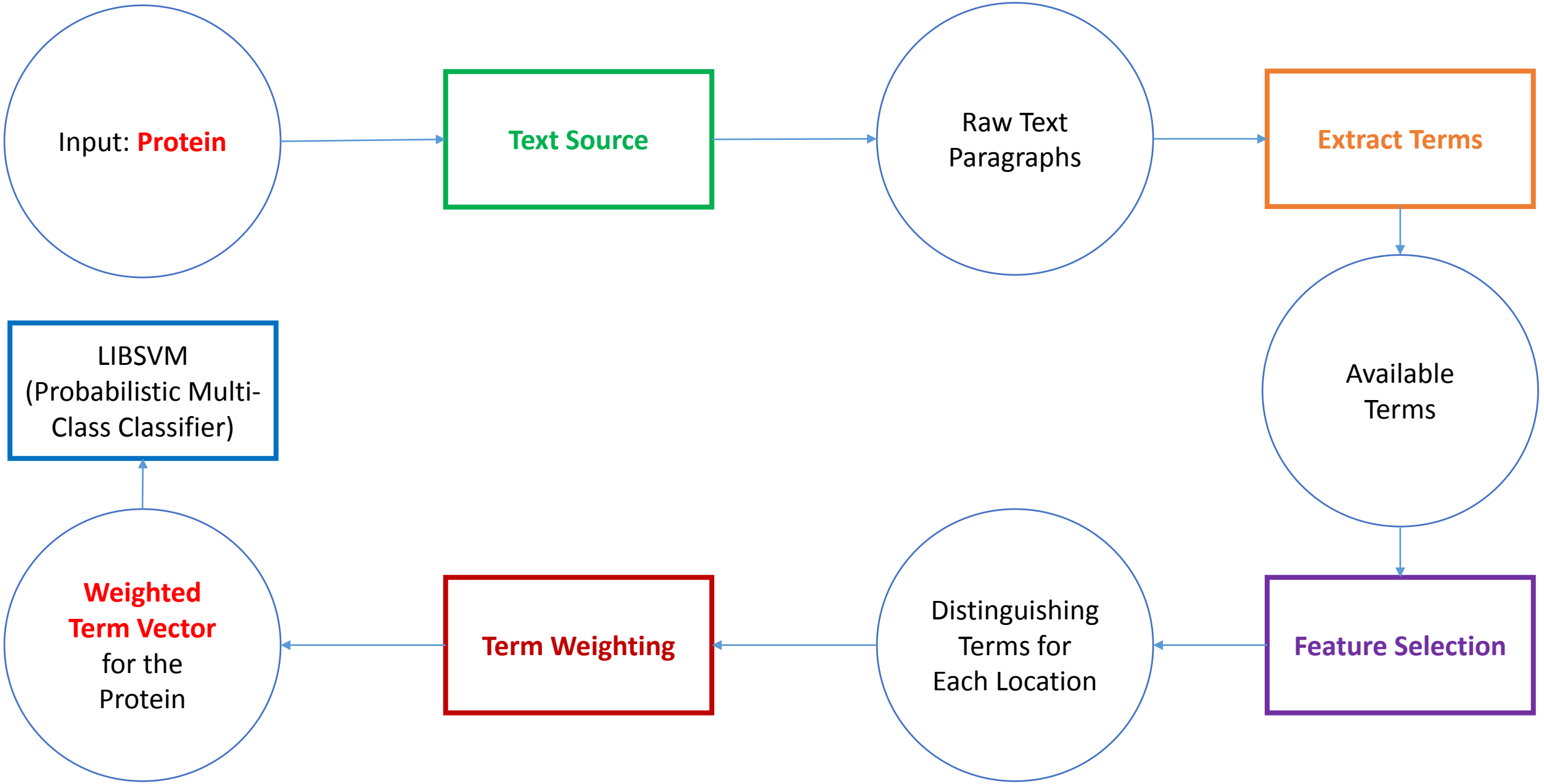
→ **Good Professor**

An Informal Example of Text-Based Inference (cont.)

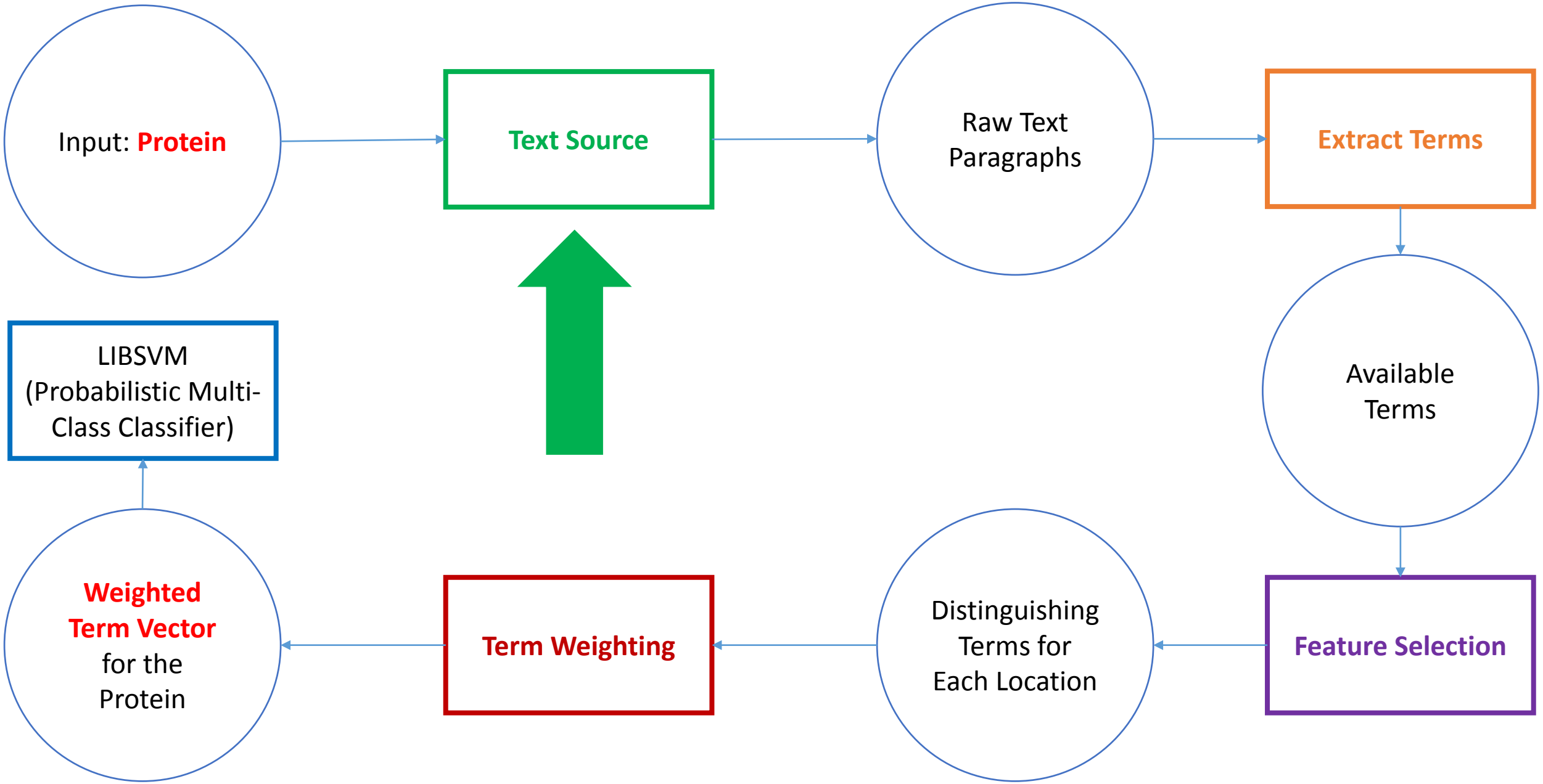
What we are doing here?



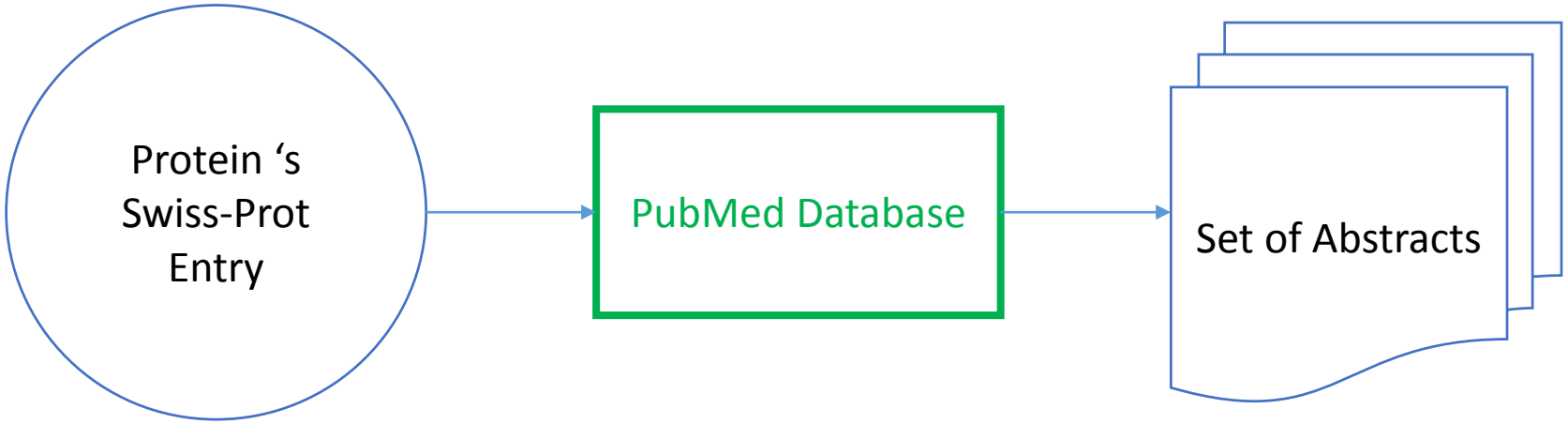
General View of EpiLoc



Text Source


Input: Protein
Text Source
Raw Text Paragraphs
Extract Terms
Available Terms
Feature Selection
Distinguishing Terms for Each Location
Term Weighting
Weighted Term Vector for the Protein
LIBSVM (Probabilistic Multi-Class Classifier)

Primary Text Source



1. Abstracts associated with proteins from three or more subcellular locations are excluded [1]

2. A set of authoritative abstracts determined by Swiss-Prot curators [1]

3. The abstracts do not typically discuss localization [1]

Primary Text Source (cont.)

ExPASy Bioinformatics Resource Portal

Documents

UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot is the manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB).
It is a high quality annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions.
Since 2002, it is maintained by the UniProt consortium and is accessible via the UniProt website.

List of UniProtKB/Swiss-Prot (reviewed) entries.
Download - UniProt FTP sites.
Statistics.

Additional information:

- Why is UniProtKB composed of 2 sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL?
- Biocuration in UniProt.
- How do we manually annotate a UniProtKB entry?
- UniProt manual annotation program.
- UniProt general documentation.
- FAQ.

Around UniProtKB/Swiss-Prot

NextProt - Exploring the universe of human proteins.
Viralzone - A portal to viral UniProtKB/Swiss-Prot entries.
HAMAP - **H**igh-quality **A**utomated and **M**anual **A**nnotation of **P**roteins.
SwissVar - A portal to Swiss-Prot diseases and variants.
UniPathway - A metabolic door to UniProtKB/Swiss-Prot.
Enzyme - A repository of information relative to the nomenclature of enzymes.

Swiss-Shop - A service that allows you to automatically obtain (by email) new UniProtKB/Swiss-Prot entries relevant to your field(s) of interest.

Protein Spotlight - One month, one protein.
Proteomics tools

Contact the UniProt helpdesk

http://www.uniprot.org/contact

Primary Text Source (cont.)

UniProt › UniProtKB Downloads · Contact · Documentation/Help

Search | Blast | Align | Retrieve | ID Mapping *

Search in: Protein Knowledgebase (UniProtKB) Query: P32267 Search Advanced Search » Clear

1 - 25 of 544,996 results for reviewed:yes in UniProtKB sorted by protein names ascending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% Download

Page 1 of 21,800 | Next »

Results Customize

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
P32267	ASIA_BPT4	★	10 kDa anti-sigma factor	asiA	Enterobacteria phage T4 (Bacteriophage T4)	90
P82445	CWP37_TOBAC	★	10 kDa cell wall protein		Nicotiana tabacum (Common tobacco)	7
Q6F8P5	CH10_ACIAD	★	10 kDa chaperonin	groS groES ACIAD2839	Acinetobacter baylyi (strain ATCC 33305 / BD413 / ADP1)	96
B7GY35	CH10_ACIB3	★	10 kDa chaperonin	groS groES ABBFA_000811	Acinetobacter baumannii (strain AB307-0294)	96
B7I619	CH10_ACIB5	★	10 kDa chaperonin	groS groES AB57_3080	Acinetobacter baumannii (strain AB0057)	96
B0VSP4	CH10_ACIBS	★	10 kDa chaperonin	groS groES ABSDF0817	Acinetobacter baumannii (strain SDF)	96
A1TKQ6	CH10_ACIAC	★	10 kDa chaperonin	groS groES Aave_0946	Acidovorax citrulli (strain AAC00-1) (Acidovorax avenae subsp. citrulli)	96
B0VDR5	CH10_ACIBY	★	10 kDa chaperonin	groS groES ABAYE0822	Acinetobacter baumannii (strain AYE)	96
B7J562	CH10_ACIF2	★	10 kDa chaperonin	groS groES AFE_0543	Acidithiobacillus ferrooxidans (strain ATCC 23270 / DSM 14882 / NCIB 8455) (Ferrobacillus ferrooxidans (strain ATCC 23270))	96
A3N121	CH10_ACTP2	★	10 kDa chaperonin	groS groES APL_1013	Actinobacillus pleuropneumoniae serotype 5b (strain L20)	96
A6VM39	CH10_ACTSZ	★	10 kDa chaperonin	groS groES Asuc_0663	Actinobacillus succinogenes (strain ATCC 55618 / 130Z)	96
B3H1P5	CH10_ACTP7	★	10 kDa chaperonin	groS groES APP7_1070	Actinobacillus pleuropneumoniae serotype 7 (strain AP76)	96
A0KGL0	CH10_AERHH	★	10 kDa chaperonin	groS groES AHA_0859	Aeromonas hydrophila subsp. hydrophila (strain ATCC 7966 / NCIB 9240)	97
B5EN20	CH10_ACIF5	★	10 kDa chaperonin	groS groES Lferr_0699	Acidithiobacillus ferrooxidans (strain ATCC 53993) (Leptospirillum ferrooxidans (ATCC 53993))	96
O68308	CH10_AERSA	★	10 kDa chaperonin	groS groES	Aeromonas salmonicida	97
A4SR81	CH10_AERS4	★	10 kDa chaperonin	groS groES ASA_3432	Aeromonas salmonicida (strain A449)	97
A3M837	CH10_ACIBT	★	10 kDa chaperonin	groS groES A1S_2665	Acinetobacter baumannii (strain ATCC 17978 / NCDC KC 755)	96
B6ENX1	CH10_ALISL	★	10 kDa chaperonin	groS groES VSAL_I0018	Aliivibrio salmonicida (strain LFI1238) (Vibrio salmonicida (strain LFI1238))	95
P94165	CH10_ACTPL	★	10 kDa chaperonin	groS groES mopB	Actinobacillus pleuropneumoniae (Haemophilus pleuropneumoniae)	96
A1W3W9	CH10_ACISJ	★	10 kDa chaperonin	groS groES Ajs_0700	Acidovorax sp. (strain JS42)	96
P46399	CH10_AGGAC	★	10 kDa chaperonin	groS groES	Aggregatibacter actinomycetemcomitans (Actinobacillus actinomycetemcomitans) (Haemophilus actinomycetemcomitans)	96
Q0ACQ8	CH10_ALKEH	★	10 kDa chaperonin	groS groES Mlg_0020	Alkalilimnicola ehrlichii (strain ATCC BAA-1101 / DSM 17681 / MLHE-1)	95
B9K1Y7	CH10_AGRVS	★	10 kDa chaperonin	groS groES Avi_5829	Agrobacterium vitis (strain S4 / ATCC BAA-846) (Rhizobium vitis (strain S4))	98
P31295	CH10_ALLVI	★	10 kDa chaperonin	groS groES	Allochromatium vinosum (Chromatium vinosum)	96
B4S113	CH10_ALTMD	★	10 kDa chaperonin	groS groES MADE_1019630	Alteromonas macleodii (strain DSM 17117 / Deep ecotype)	96

Page 1 of 21,800 | Next »

Primary Text Source (cont.)

References

« Hide 'large scale' references

[1] **"The asiA gene of bacteriophage T4 codes for the anti-sigma 70 protein."**
Orsini G., Ouhammouch M., Le Caer J.-P., Brody E.N.
J. Bacteriol. 175:85-93(1993) [PubMed] [Europe PMC] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA].
Strain: D.

[2] **"Bacteriophage T4 genome."**
Miller E.S., Kutter E., Mosig G., Arisaka F., Kunisawa T., Ruger W.
Microbiol. Mol. Biol. Rev. 67:86-156(2003) [PubMed] [Europe PMC] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].

[3] **"The asiA gene product of bacteriophage T4 is required for middle mode RNA synthesis."**
Ouhammouch M., Orsini G., Brody E.N.
J. Bacteriol. 176:3956-3965(1994) [PubMed] [Europe PMC] [Abstract]
Cited for: FUNCTION.

[4] **"Conserved regions 4.1 and 4.2 of sigma(70) constitute the recognition sites for the anti-sigma factor AsiA, and AsiA is a dimer free in solution."**
Urbauer J.L., Adelman K., Urbauer R.J., Simeonov M.F., Gilmore J.M., Zolkiewski M., Brody E.N.
J. Biol. Chem. 276:41128-41132(2001) [PubMed] [Europe PMC] [Abstract]
Cited for: STRUCTURE BY NMR.

[5] **"T4 AsiA blocks DNA recognition by remodeling sigma70 region 4."**
Lambert L.J., Wei Y., Schirf V., Demeler B., Werner M.H.
EMBO J. 23:2952-2962(2004) [PubMed] [Europe PMC] [Abstract]
Cited for: STRUCTURE BY NMR IN COMPLEX WITH HOST RPOD, FUNCTION.

+ Additional computationally mapped references.

Primary Text Source (cont.)

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Display Settings: Abstract

Send to:

J Bacteriol. 1993 Jan;175(1):85-93.

The asiA gene of bacteriophage T4 codes for the anti-sigma 70 protein.

Orsini G[1], Ouhammouch M, Le Caer JP, Brody EN.

Author information

Abstract

The anti-sigma 70 factor of bacteriophage T4 is a 10-kDa (10K) protein which inhibits the sigma 70-directed initiation of transcription by Escherichia coli RNA polymerase holoenzyme. We have partially purified the anti-sigma 70 factor and obtained the sequence of a C-terminal peptide of this protein. Using reverse genetics, we have identified, at the end of the lysis gene t and downstream of an as yet unassigned phage T4 early promoter, an open reading frame encoding a 90-amino-acid protein with a predicted molecular weight of 10,590. This protein has been overproduced in a phage T7 expression system and partially purified. It shows a strong inhibitory activity towards sigma 70-directed transcription (by RNA polymerase holoenzyme), whereas it has no significant effect on sigma 70-independent transcription (by RNA polymerase core enzyme). At high ionic strength, this inhibition is fully antagonized by the neutral detergent Triton X-100. Our results corroborate the initial observations on the properties of the phage T4 10K anti-sigma 70 factor, and we therefore propose that the gene which we call asiA, identified in the present study, corresponds to the gene encoding this T4 transcriptional inhibitor.

PMID: 8416914 [PubMed - indexed for MEDLINE] PMCID: PMC196100 Free PMC Article

Publication Types, MeSH Terms, Substances, Secondary Source ID

LinkOut - more resources

Primary Text Source (cont.)

Problems:

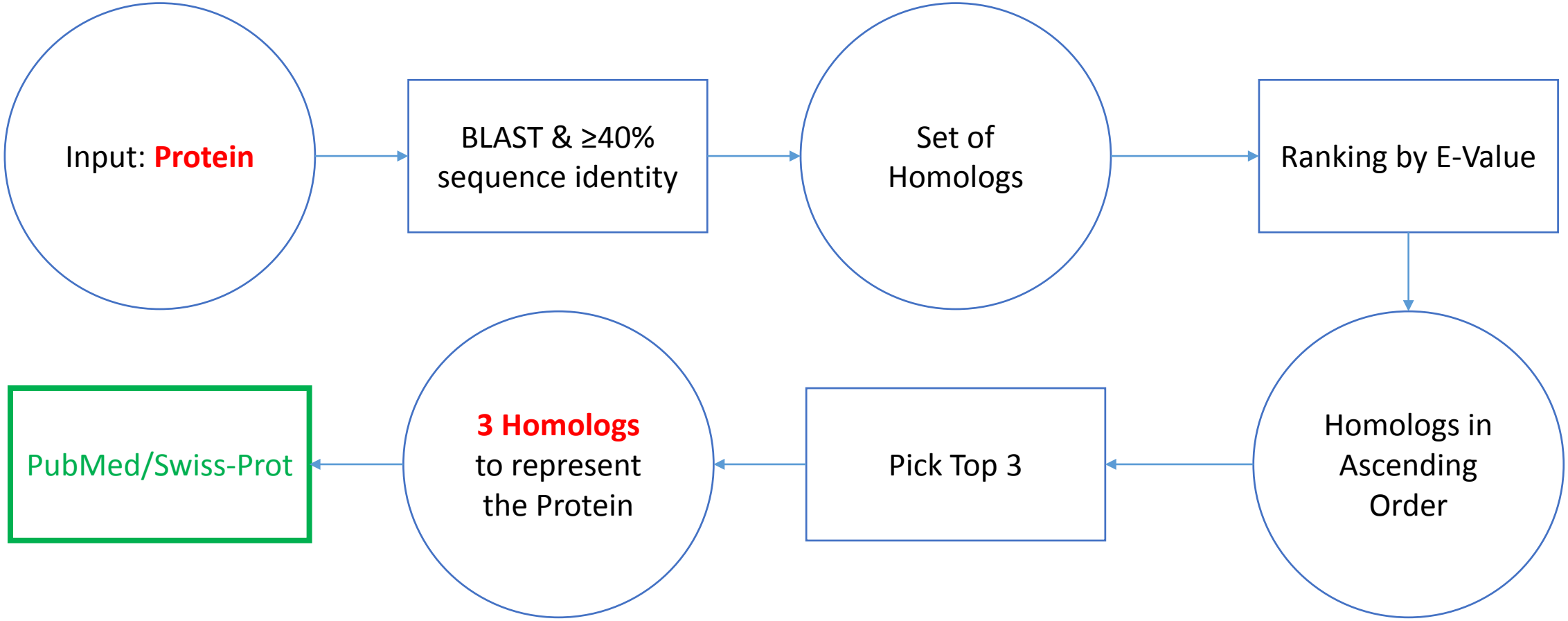
- What if a protein does not have PubMed identifiers in its Swiss-Prot entry?
- What if it even does not have a Swiss-Prot entry (newly discovered one)?

HomoLoc Text Source

- **Motivation**
  - No Cappuccino today? Go with a Latte!

E-Value: Expectation of by-chance match

- **How it be done?**



HomoLoc Text Source (cont.)

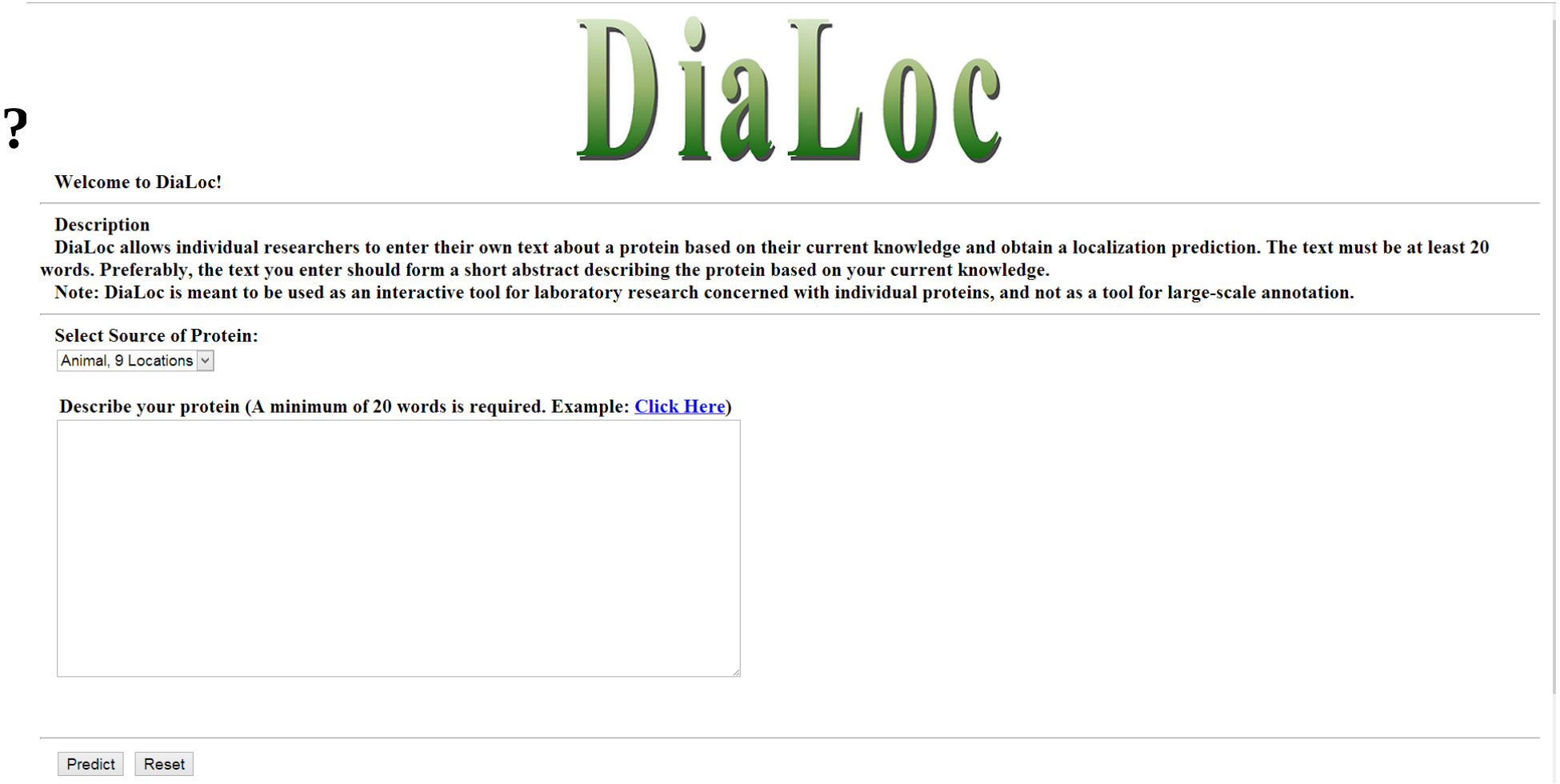
Problems:

- What if the protein is too new to find homologs and little info about it in PubMed or Swiss-Prot?

DiaLoc Text Source

- **Motivation**
  - The researchers should know their proteins very well, so they could describe them.

- **How it be done?**



PubLoc Text Source

- **Motivation**
  - No PubMed references in Swiss-Prot entries ≠ No PubMed abstracts

- **How it be done?**

PubLoc Text Source (cont.)

UniProt › UniProtKB

Search | Blast * | Align | Retrieve | ID Mapping *

Search in: Protein Knowledgebase (UniProtKB)

Query: Search | Advanced Search » | Clear

P32267 (ASIA_BPT4) Reviewed, UniProtKB/Swiss-Prot

Last modified February 19, 2014. Version 76. History...

Clusters with 100%, 90%, 50% identity | Documents (1) | Third-party data

Names · Attributes · General annotation · Ontologies · Interactions · Sequence annotation · Sequences · References · Cross-refs · Entry info · Documents · Customize order

Names and origin

Protein names	*Recommended name:* **10 kDa anti-sigma factor** *Alternative name(s):* 10 kDa RNA polymerase-associated protein Audrey Stevens' inhibitor
Gene names	Name:**asiA**
Organism	**Enterobacteria phage T4 (Bacteriophage T4)** [Reference proteome]
Taxonomic identifier	10665 [NCBI]
Taxonomic lineage	Viruses › dsDNA viruses, no RNA stage › Caudovirales › Myoviridae › Tevenvirinae › T4likevirus ›
Virus host	Escherichia coli [TaxID: 562]

Protein attributes

Sequence length	90 AA.
Sequence status	Complete.
Protein existence	Evidence at protein level

PubLoc Text Source (cont.)

NCBI Resources How To Sign in to NCBI

PubMed Home More Resources Help

PubMed Advanced Search Builder

YouTube Tutorial

(((10 kDa anti-sigma factor) OR 10 kDa RNA polymerase-associated protein) OR Audrey Stevens' inhibitor) OR asiA

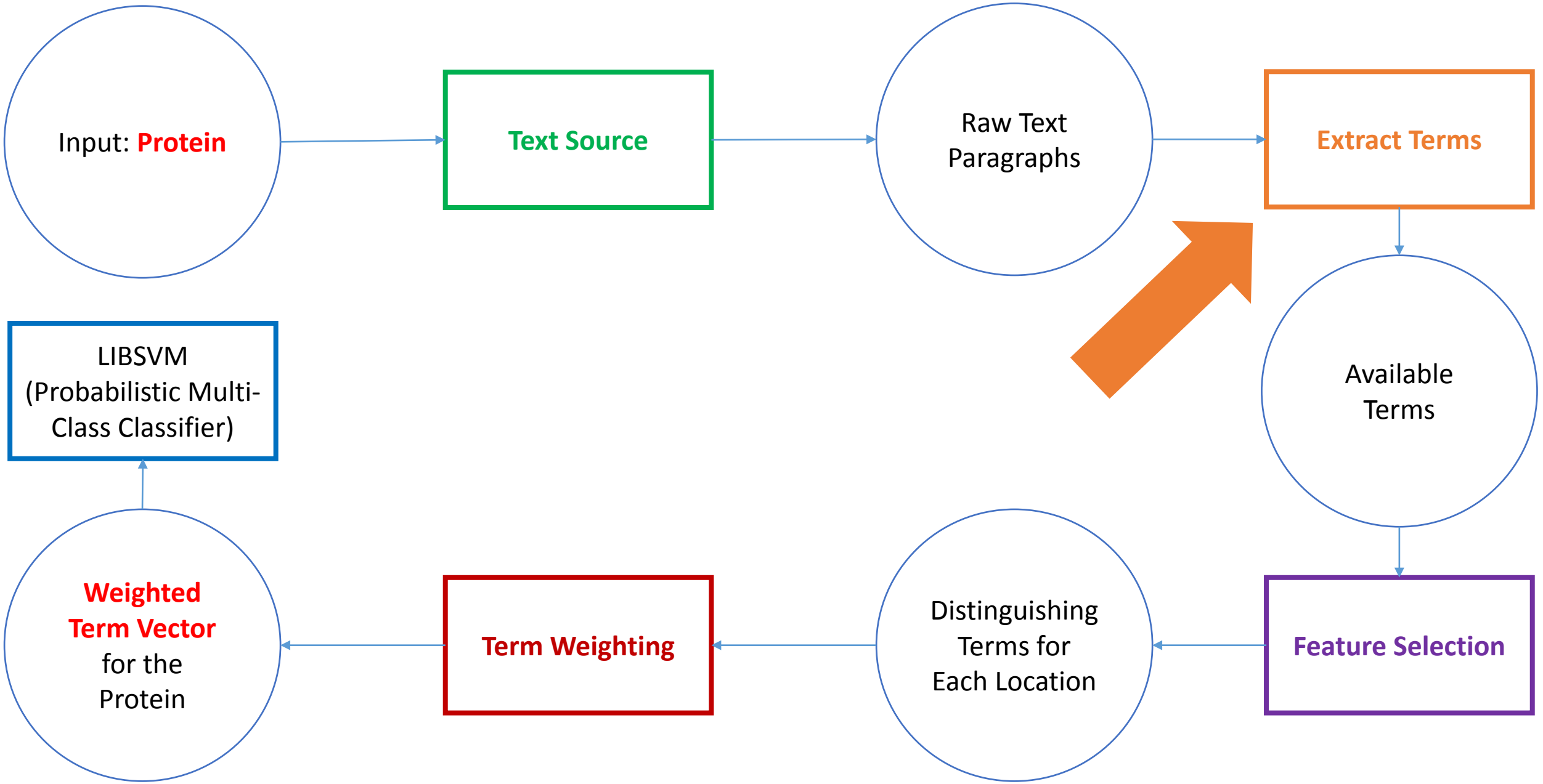
Edit Clear

Builder

	All Fields	10 kDa anti-sigma factor	Show index list
OR	All Fields	10 kDa RNA polymerase-associated protein	Show index list
OR	All Fields	Audrey Stevens' inhibitor	Show index list
OR	All Fields	asiA	Show index list

Search or Add to history

Extract Terms



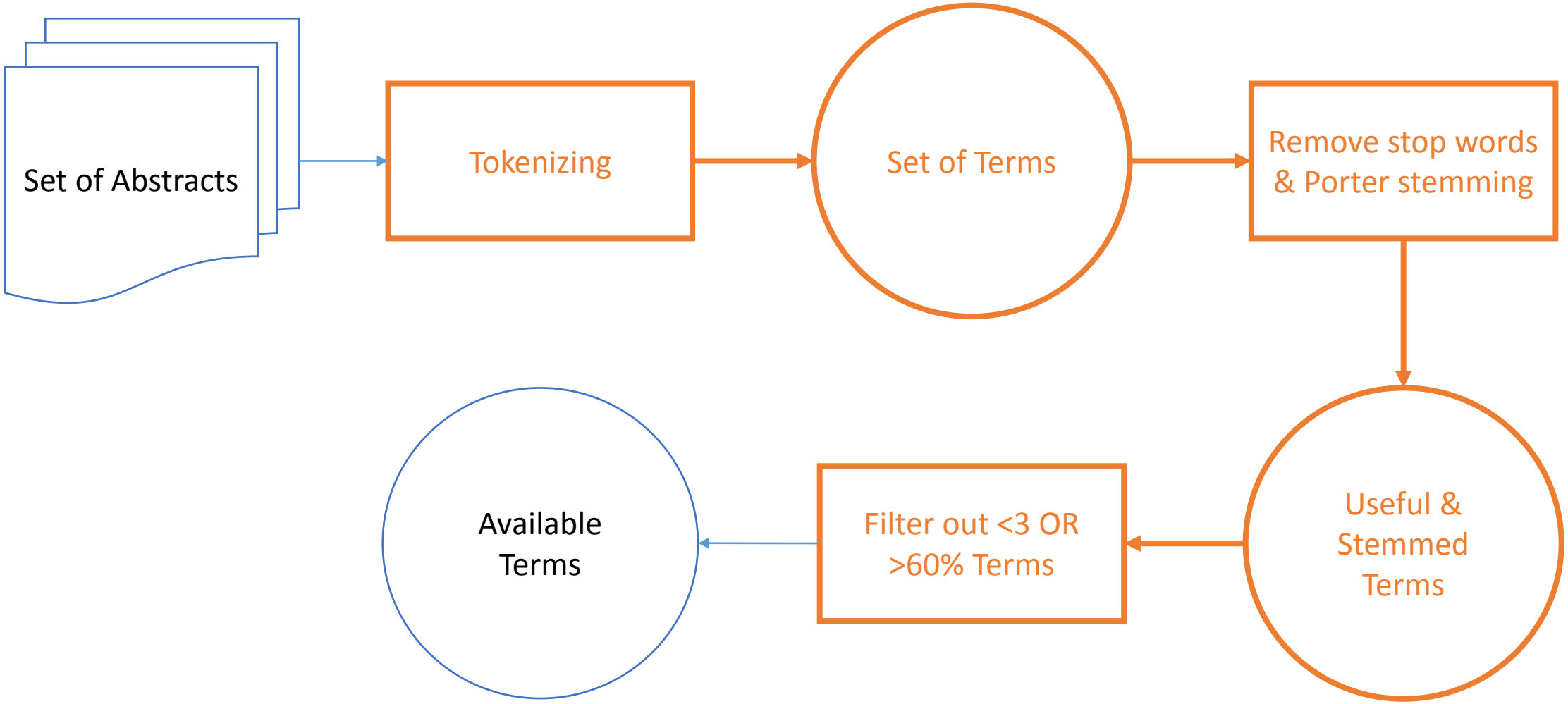
Extract Terms (cont.)

- **Motivation**
  - Use a N-dimensional term vector to represent a protein so that more convenient to do classification

- **How it be done?**

Extract Terms (cont.)

- **How it be done?**



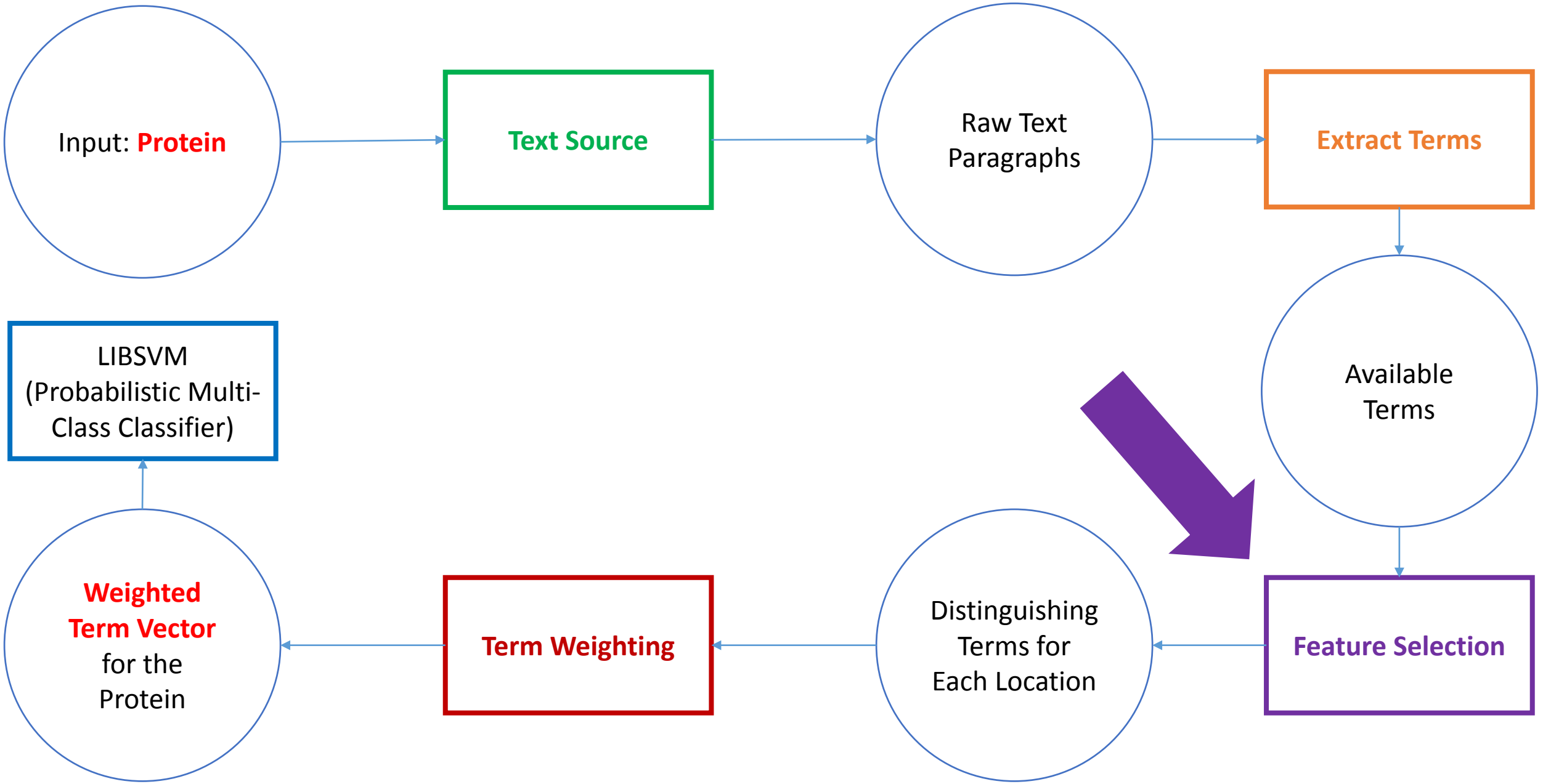
[1] [10]

Extract Terms (cont.)

Problems:

- The number of resulting terms typically exceeds 20,000 [1]
- Some of them might not be so distinguishing in terms of location

Feature Selection



Feature Selection (cont.)

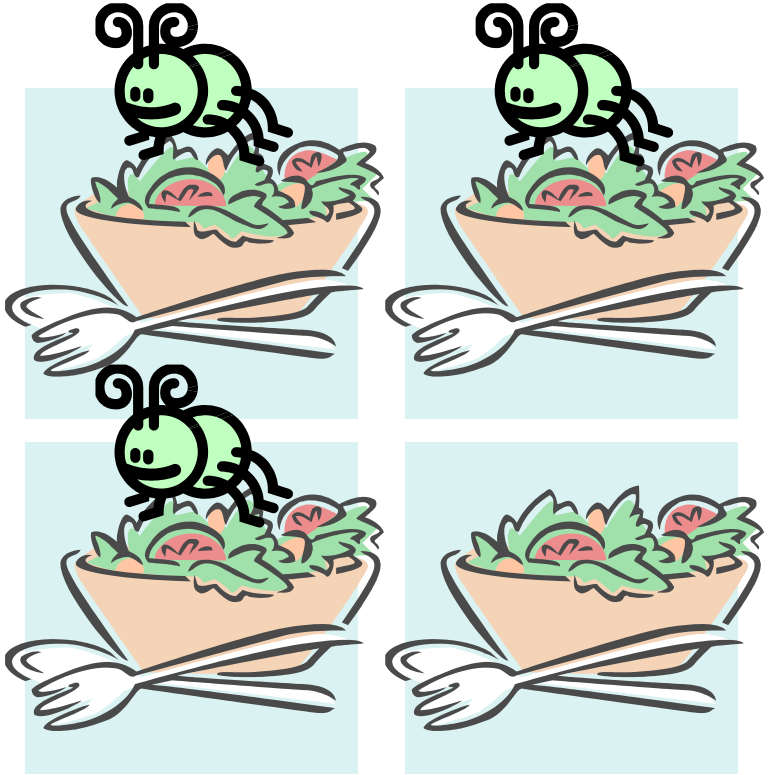
- **Motivation**
  - Dimension reduction would lower computational complexity. [1]
  - A given term might not be distinguishing to all locations. [1]

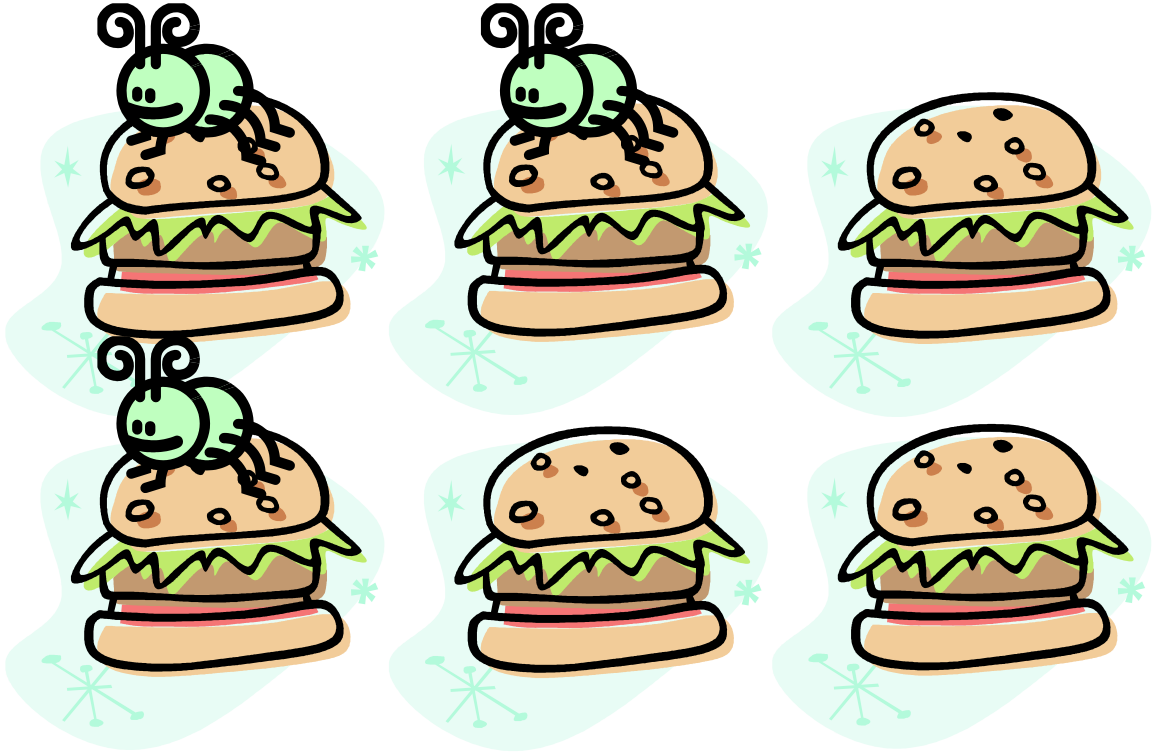
- **How it be done?**
  - Z-Test Method [1] [11]

Feature Selection (cont.)

- **Z-Test Method**
  - Motivation

    Which one is the bugs' favorite? Salad or Burger?





Feature Selection (cont.)

- **Z-Test Method**
  - Terminology [1][12]
    - t: A term
    - p: A protein
    - L: A location
    - $p \in L$: p is localized to L
    - D_p: A set of abstracts associated with p
    - P_L: A set of protein known to be localized to L
    - D_L: The set of abstracts associated with L (all abstracts associated with the proteins localized to L)
    - $Pr(t|L)$: The probability of term t to be associated with location L

$Pr(t|L) = \Pr(t \in d | d \in L)$, where d is an abstract associated with L

Maximum Likelihood Estimate of

$$Pr(t|L) \approx \frac{\#\ of\ abstracts\ d \in D_L\ such\ that\ t \in d}{|D_L|}$$

Feature Selection (cont.)

- **Z-Test Method**
  - In our case:

    Bug = Term

    Salad = Abstracts associated with location 1

    Sandwich = Abstracts associated with location 2

  - Z-Score:

$$Z_{L,L'}^{t} = \frac{\Pr(t|L) - \Pr(t|L')}{\sqrt{\bar{P}\cdot(1-\bar{P})\cdot\left(\frac{1}{|D_L|}+\frac{1}{|D_{L'}|}\right)}}, \text{ where } \bar{P} = \frac{|D_L|\cdot\Pr(t|L)+|D_{L'}|\cdot\Pr(t|L')}{|D_L|+|D_{L'}|}$$

$|Z_{L,L'}^{t}|$ means the significance of difference

Threshold = 1.96

Confidence $\geq$ 95%

Feature Selection (cont.)

- **Z-Test Method**
  - Effectiveness
    - Comparison: odds ratio (OR), Chi-square (χ^2), mutual information (MI), information gain (IG), and Entropy method [13]
    - Scoring: a combination of location-specific scores
      - ➢OR, IG: sum of scores
      - ➢MI, χ^2: max score
      - ➢Entropy: based on the difference between Shannon information and the maximum attainable information
    - Measure: overall accuracy achieved by classifiers based on each method
    - Data Sets: plant and animal proteins of the MultiLoc dataset
    - Experimental Setting:
      - ➢the same text pre-processing
      - ➢five-fold cross-validation
      - ➢possible number of selected terms (500 - 4000)

Feature Selection (cont.)

- **Z-Test Method**
  - Effectiveness
    - Experimental Results: [1][14]

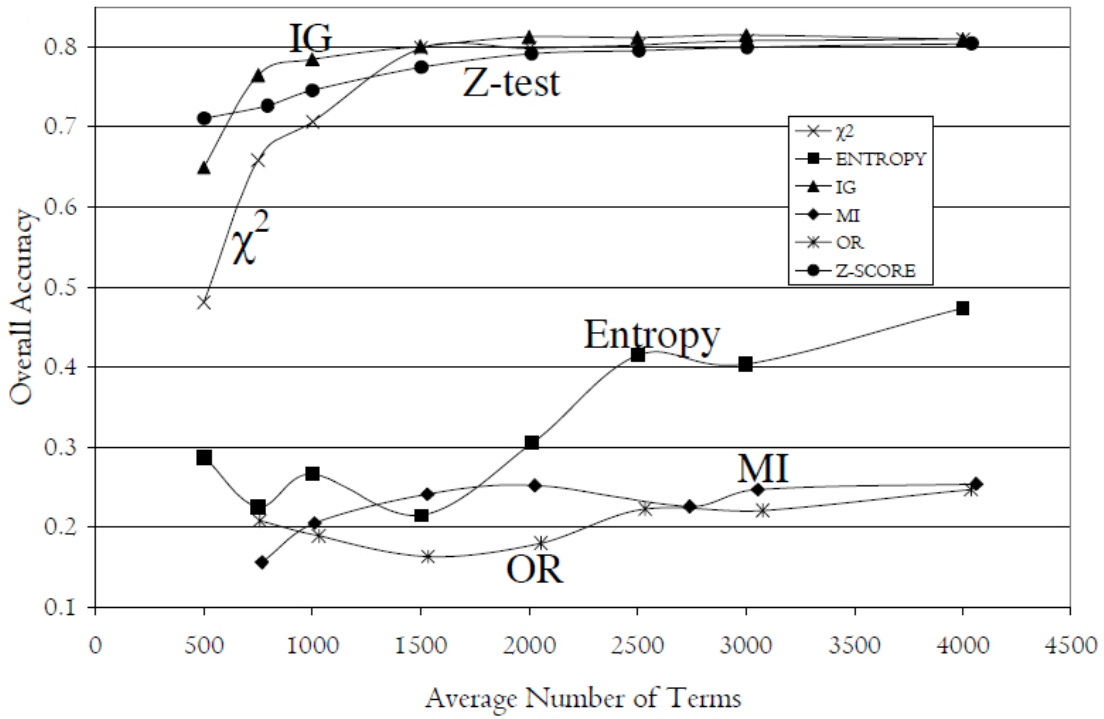
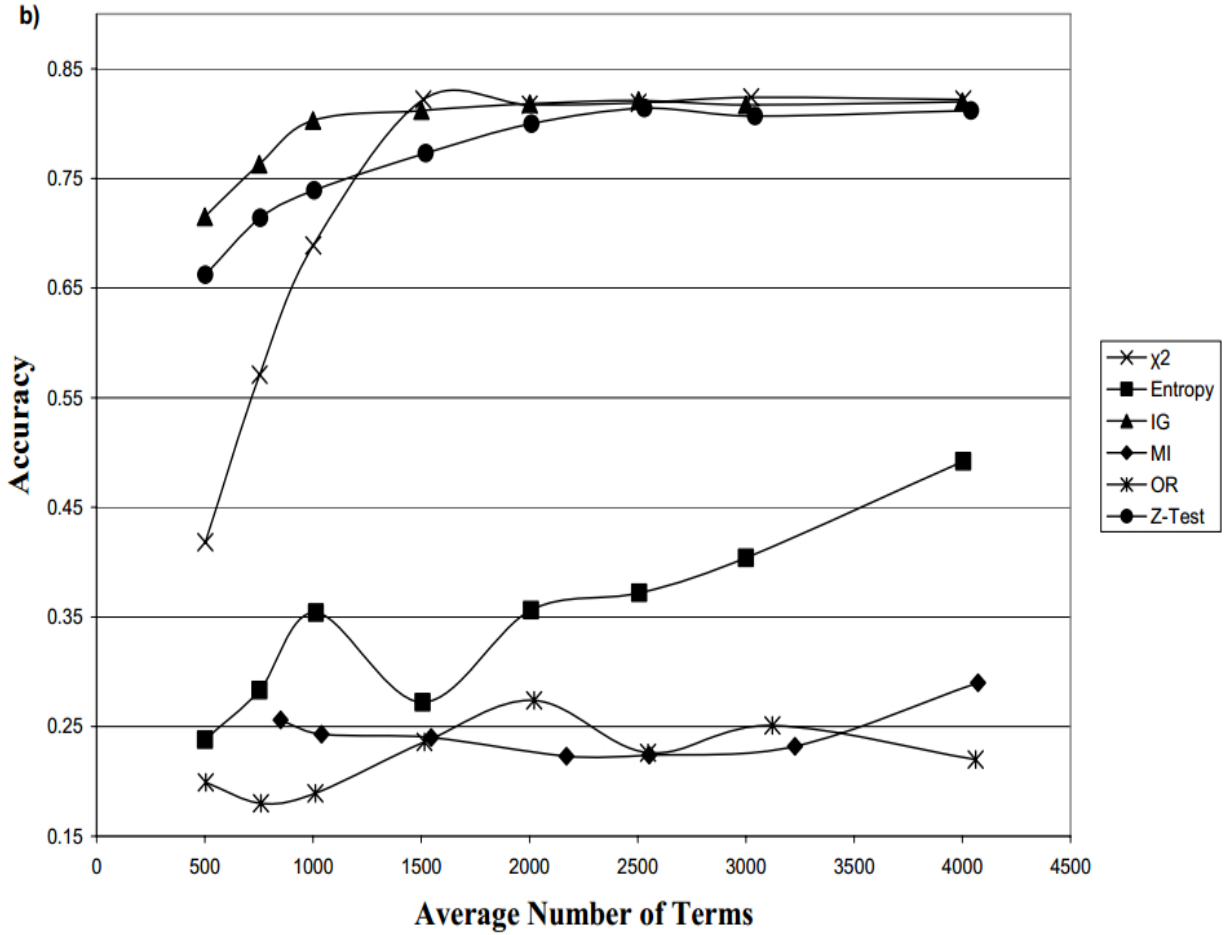


Figure 1. Accuracy of the classifiers (for plant proteins), based on different feature selection methods, as a function of the average number of selected terms (features).



Feature Selection (cont.)

- **Z-Test Method**
  - <u>What we are doing here?</u>

    Compare proportions! The greater the proportion, the greater the significance.
  - <u>Make a hypothesis:</u>

    H_0: p1 = p2
  - <u>What we are going to do?</u>
    - ➢If one of them is more significantly different from the other, we first reject this hypothesis.
    - ➢Second, evaluate the significance.
  - <u>Why it works?</u>
    - ➢If the hypothesis is true, then it should not be much likely to observe a counterexample.
    - ➢Hence, if we did see a counterexample, then it means the hypothesis may not be of a great confidence.

Feature Selection (cont.)

- **Z-Test Method**
  - Statistic Analysis: [1][11]
    - $Pr(t|L) \approx \frac{\#\ of\ abstracts\ d \in D_L\ such\ that\ t \in d}{|D_L|}$ means the proportion of abstracts containing the term t.
    - We are comparing $p_1 = Pr(t|L)$ and $p_2 = Pr(t|L')$
    - Both of them are of the binomial distribution
    - $\mu_{P_1-P_2} = p_1 - p_2$, $\sigma^2_{P_1-P_1} = \frac{p_1 \cdot (1-p_1)}{|D_L|} - \frac{p_2 \cdot (1-p_2)}{|D_{L'}|}$, where P_1, P_2 are corresponding random variables, and $P_1 - P_2$ is also a random variable.
    - Then $Z = \frac{(P_1-P_2)-\mu_{P_1-P_2}}{\sqrt{\sigma^2_{P_1-P_1}}}$ is a standard normal variable. (By Central Limit Theorem)
    - If H_0 is true, then $p_1 = p_2 = p$.
    - Then $Z = \frac{P_1-P_2}{\sqrt{p \cdot (1-p) \cdot (\frac{1}{|D_L|} + \frac{1}{|D_{L'}|})}}$.
    - We use pooled estimate of the proportion $\hat{p} = \frac{|D_L| \cdot Pr(t|L) + |D_{L'}| \cdot Pr(t|L')}{|D_L| + |D_{L'}|}$ to compute p.
    - Then we get $Z^t_{L,L'}$.

Feature Selection (cont.)

- **Z-Test Method**
  - Confidence Interval: [11]
    - ➢ α is the significance level (e.g. $\frac{\alpha}{2} = 0.025$ means $100 \cdot (1-\alpha)\% = 95\%$ confidence)

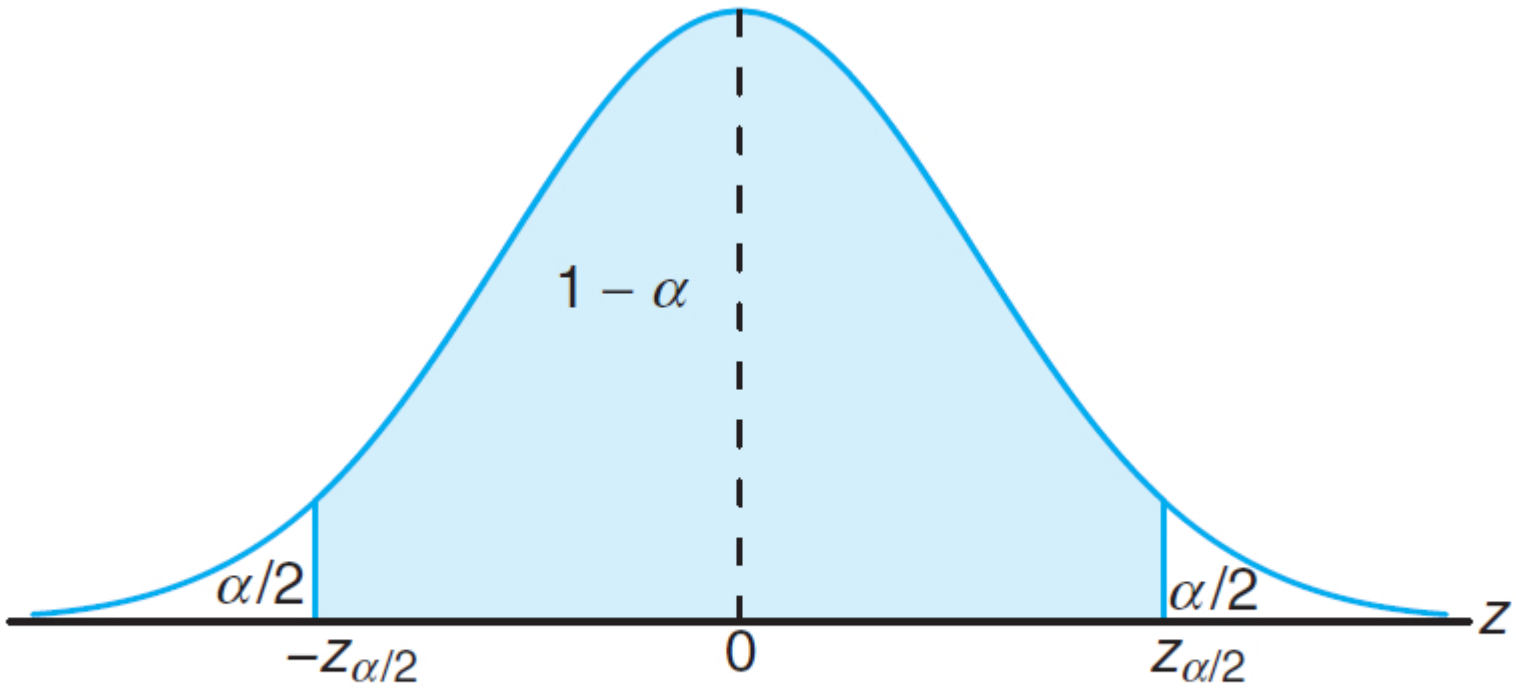


Figure 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Feature Selection (cont.)

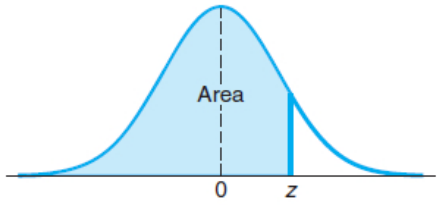


Table A.3 Areas under the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559

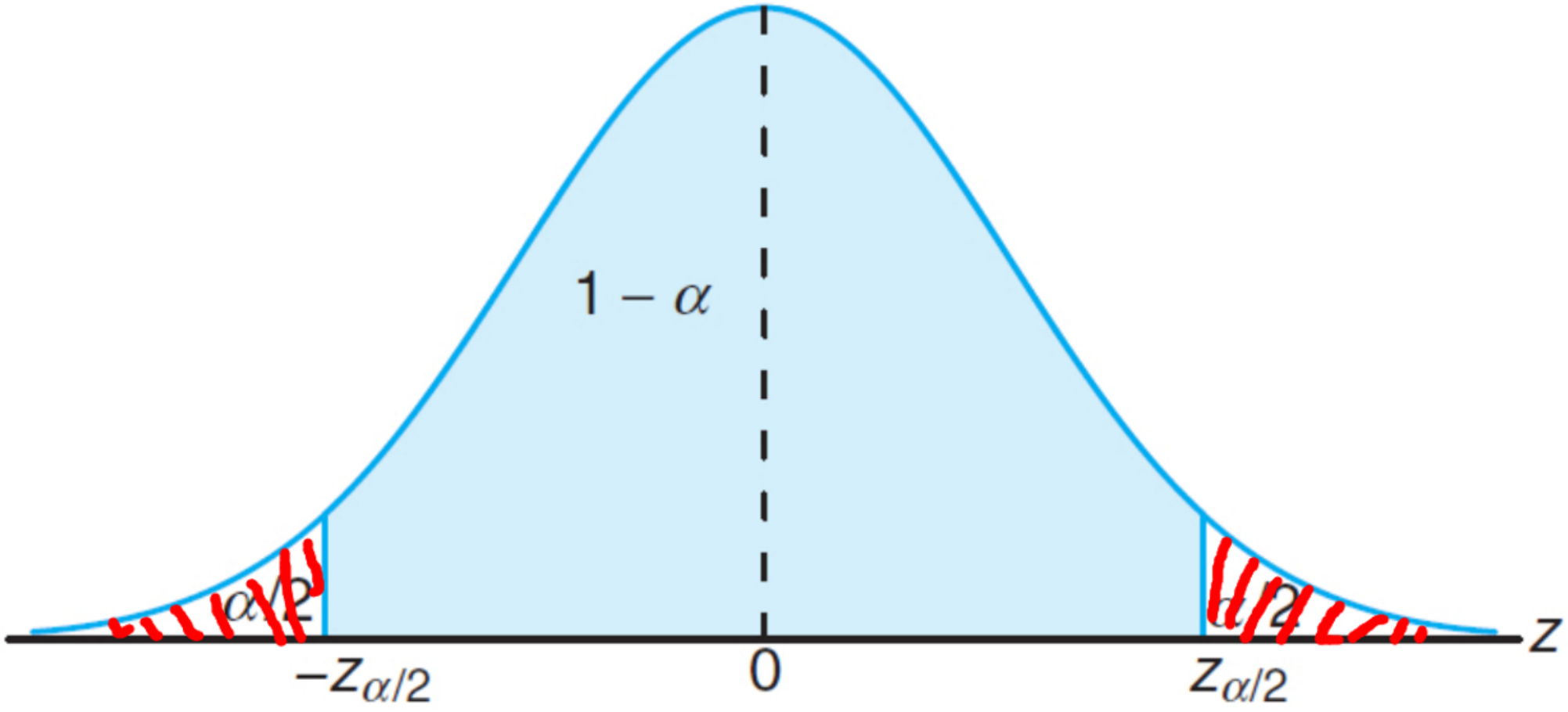
Feature Selection (cont.)

Table A.3 (continued) Areas under the Normal Curve

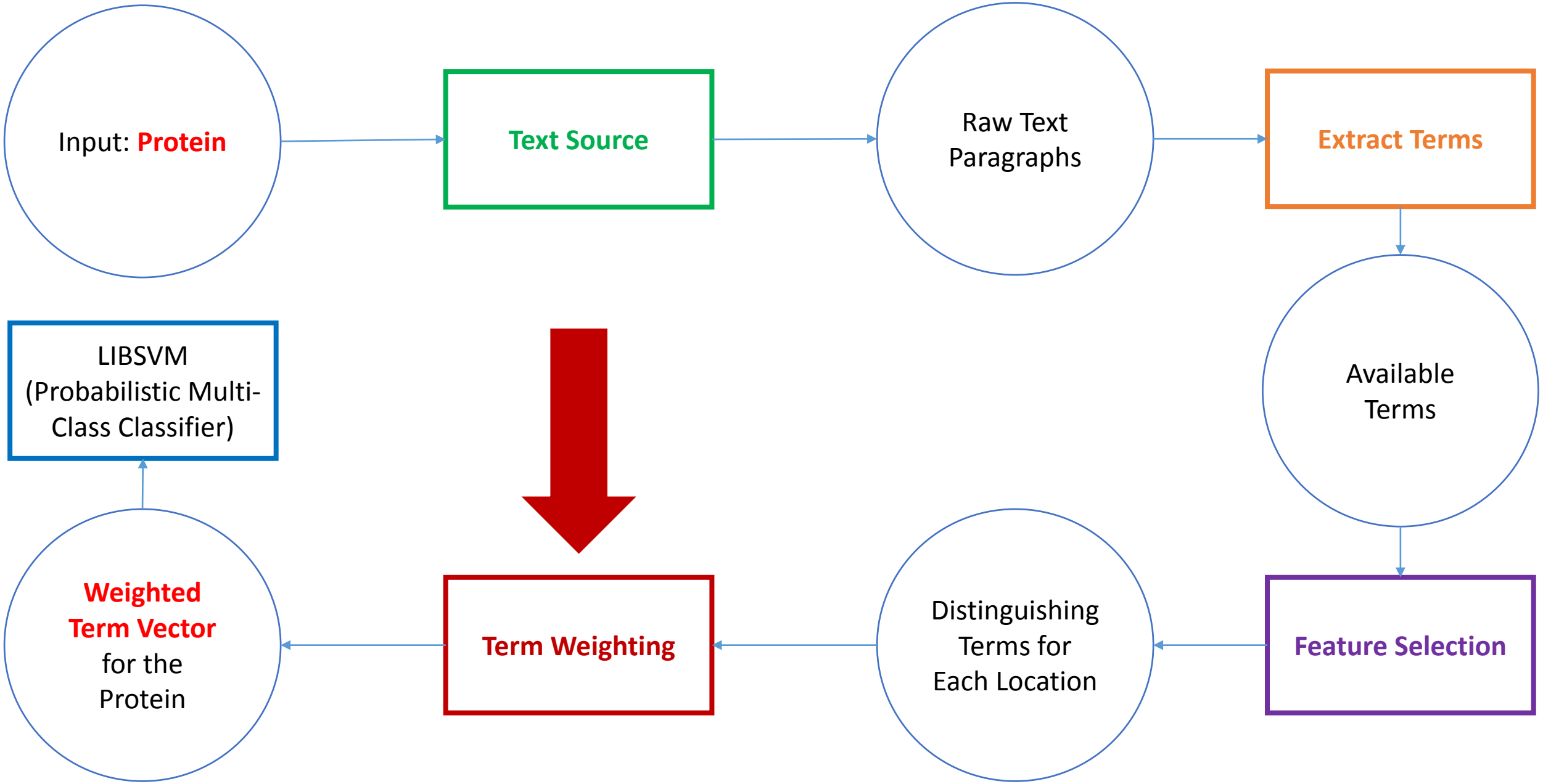
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857

Feature Selection (cont.)

- **Z-Test Method**
  - Reject H_0:
    - If $Z > 1.96$ cases are observed, then H_0 is rejected, and Z will mean the significance of difference.



Term Weighting



Term Weighting (cont.)

- **Motivation**
  - Different terms may contribute various significance in terms of describing the protein.
  - How to measure this difference? Weights!

- **How it be done?**
  - $W_{t_i}^{p} = \frac{\#\ of\ times\ t_i\ occurs\ in\ D_p}{\sum_{j \in T_N}(\#of\ times\ t_j\ occurs\ in\ D_p)}$
  - p: Protein
  - N: # of distinguishing terms
  - t_i: Term at ith position, $1 \leq i \leq N$
  - T_N: The set of distinguishing terms
  - D_p: The set of abstracts known to be associated with protein p

[1]

Term Weighting (cont.)

- **HomoLoc Weighting**
  - To reflect the degree of homology [1]

- **How it be done?**
  - $W_{t_i}^{p} = \frac{\sum_{h \in H}(\#\ of\ times\ t_i\ occurs\ in\ D_h)\cdot(\%\ identity\ of\ h)}{\sum_{h \in H}\sum_{j \in T_N}(\#of\ times\ t_j\ occurs\ in\ D_h)\cdot(\%\ identity\ of\ h)}$
  - p: Protein
  - h: A homolog
  - H: A set of homologs
  - N: # of distinguishing terms
  - t_i: Term at ith position, $1 \leq i \leq N$
  - T_N: The set of distinguishing terms
  - D_h: The set of abstracts known to be associated with h

[1]

Experimental Results for EpiLoc

Table 3. Prediction performance of TargetP and EpiLoc on the TargetP dataset, for both plant and non-plant proteins.

Loc.	TargetP	EpiLoc	TargetP	EpiLoc
	Plant (*Sens Spec MCC*)		**Non-Plant** (*Sens Spec MCC*)	
ch	0.85 **0.69** **0.72**	**0.92** 0.53 0.68	N/A	
mi	0.82 **0.90** 0.77	**0.89** 0.81 **0.82**	0.89 0.67 0.73	**0.92** **0.84** **0.86**
SP	**0.91** **0.95** **0.90**	0.89 0.84 0.80	**0.96** **0.92** **0.92**	0.93 0.86 0.84
OT	**0.85** 0.78 0.77	0.84 **0.95** **0.78**	**0.88** **0.97** **0.82**	**0.88** 0.95 0.81
Acc	0.853 (±0.035)	**0.862 (±0.004)**	0.900 (±0.007)	**0.901 (±0.006)**
Avg	0.856 (n/a)	**0.883 (±0.001)**	0.907 (n/a)	**0.908 (±0.003)**

Experimental Results for EpiLoc (cont.)

Table 4. Prediction performance of PLOC and EpiLoc on the animal proteins of the PLOC dataset. Specificity and MCC values were not available for PLOC, hence only its sensitivity is listed and compared with our sensitivity values.

		PLOC Dataset (Animal)										
	Loc.	***go***	***cs***	***ly***	***er***	***pe***	***Mi***	***ex***	***cy***	***pm***	***nu***	**Acc/Avg**
PLOC	(*Sens*)	0.15	0.59	0.62	0.47	0.25	0.57	**0.78**	**0.72**	**0.92**	**0.90**	**0.796 (± 0.009)**/ 0.579 (± 0.021)
EpiLoc	(*Sens*)	**0.76**	**0.84**	**0.89**	**0.72**	**0.85**	**0.79**	0.74	0.53	0.79	0.81	0.743 (±0.002)/ **0.773 (±0.0012)**
	(*Spec*)	0.51	0.32	0.32	0.30	0.55	0.85	0.68	0.63	0.85	0.90	
	(*MCC*)	0.62	0.51	0.53	0.45	0.68	0.80	0.66	0.50	0.78	0.80	

Experimental Results for EpiLoc (cont.)

Table 5. Prediction performance of MultiLoc, EarlyText (our basic text-based system used in earlier work[8,16]), EpiLoc and HomoLoc on the animal[d] proteins of the MultiLoc dataset.

	MultiLoc Dataset (Animal)											
Loc.	**MultiLoc**			**EarlyText**			**EpiLoc**			**HomoLoc**		
	(Sens Spec MCC)											
go	0.71	0.43	0.53	0.86	0.40	0.57	0.88	0.62	0.73	**0.90**	**0.72**	**0.80**
ly	0.69	0.36	0.48	0.75	0.32	0.47	**0.86**	0.39	0.57	0.85	**0.49**	**0.63**
er	0.68	0.56	0.60	0.74	0.48	0.58	0.74	0.59	0.65	**0.77**	**0.67**	**0.71**
pe	0.71	0.31	0.44	**0.93**	0.60	0.74	0.90	**0.77**	**0.82**	0.80	0.69	0.74
mi	**0.88**	0.82	**0.83**	0.80	0.79	0.77	0.82	0.82	0.80	0.79	**0.84**	0.80
ex	0.79	**0.83**	0.77	0.76	0.78	0.72	0.80	0.82	0.77	**0.83**	**0.83**	**0.79**
cy	0.67	**0.85**	0.68	0.51	0.77	0.53	0.68	0.79	0.65	**0.72**	0.80	**0.67**
pm	0.73	0.90	0.76	0.80	0.91	0.81	0.85	0.90	0.84	**0.89**	**0.91**	**0.87**
nu	0.82	0.73	0.73	0.84	0.71	0.73	0.84	0.81	0.80	**0.87**	**0.84**	**0.83**
Acc	0.746 (± 0.01)			0.725 (±0.007)			0.792 (±0.008)			**0.812 (± 0.010)**		
Avg	0.741 (± 0.025)			0.775 (±0.015)			0.818 (±0.005)			**0.822 (± 0.005)**		

Reference

[1] Brady, Scott, and Hagit Shatkay. "EpiLoc: a (working) text-based system for predicting protein subcellular location." Pacific Symposium on Biocomputing. Vol. 13. 2008.

[2] Emanuelsson, Olof, et al. "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence." Journal of molecular biology 300.4 (2000): 1005-1016.

[3] Höglund, Annette, et al. "MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition." Bioinformatics 22.10 (2006): 1158-1165.

[4] Nakai, Kenta, and Minoru Kanehisa. "A knowledge base for predicting protein localization sites in eukaryotic cells." Genomics 14.4 (1992): 897-911.

[5] Park, Keun-Joon, and Minoru Kanehisa. "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs." Bioinformatics 19.13 (2003): 1656-1663.

Reference (cont.)

[6] Craven, Mark, and Johan Kumlien. "Constructing biological knowledge bases by extracting information from text sources." ISMB. Vol. 1999. 1999.

[7] Stapley, B. J., L. A. Kelley, and M. J. Sternberg. "Predicting the sub-cellular location of proteins from text using support vector machines." Proceedings of the 7th Pacific Symposium on Biocomputing. 2002.

[8] BLAST, Gapped. "PSI-BLAST: a new generation of protein database search programs Altschul." Stephen F: 3389-3402.

[9] Brenner, Steven E., Cyrus Chothia, and Tim JP Hubbard. "Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships." Proceedings of the National Academy of Sciences 95.11 (1998): 6073-6078.

[10] Porter, Martin F. "An algorithm for suffix stripping." Program: electronic library and information systems 14.3 (1980): 130-137.

Reference (cont.)

[11] Walpole, Ronald E., et al. Probability and statistics for engineers and scientists. Vol. 5. New York: Macmillan, 1993.

[12] Höglund, Annette, et al. "Significantly improved prediction of subcellular localization by integrating text and protein sequence data." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 2005.

[13] Nair, Rajesh, and Burkhard Rost. "Inferring sub-cellular localization through automated lexical analysis." Bioinformatics 18.suppl 1 (2002): S78-S86.

[14] Brady, Scott. Text-based prediction of protein subcellular location. Diss. MSc Thesis, School of Computing, Queen’s University, Kingston, ON, 2007.