# Named Entity Recognition and Classification in Biomedical Domain


- Ajas Mohammed Jansher B080437CS
- Binu Jasim T B080273CS
- Febin A Rasheed B080119CS
- Muneeb TH B080438CS

# Problem Definition

In this project, we propose a machine learning approach for Named Entity Recognition and Classification(NERC) that can retrieve named entities in biomedical text and classify them into certain predefined classes or as Others, if they don’t belong to any of these classes.

The Classes are

- Protein
- DNA
- RNA
- Cell Type

# Introduction – Named Entity Recognition(NER)

- General NER – classify Proper nouns into classes such as Person, Place, Organization, etc..
- Biomedical Domain – classes such as proteins, genes, names of diseases etc…
- Useful in Information Extraction (IE) tasks, Data Mining etc…

# Related Works

- GENIA 3.0 JNLPBA tagged training set which contains around 2000 abstracts of biomedical texts – by Tsuji of University of Tokyo.

- Maximum Entropy models (ME) as in Jon Patrick et al [2] and Hidden Markov Models (HMMs) as in Zhou Guodong Et al.[1].

- Jon Patrick et al.[2] used part of speech tagging features and bigram features, achieved an f-measure of 68%

- Zhou et.al.[1] proposed features, including orthographic, morphological,part-of-speech and semantic trigger features, and achieved an f-measure of 66.5%

# Sample Input

While specific constitutive binding to the peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional, specific binding. Understanding the monocyte-specific function of the peri-kappa B factor may ultimately provide insight into the different role monocytes and T-cells play in HIV pathogenesis.

# Proposed Output

While specific constitutive binding to the <DNA> peri-kappa B site </DNA> is seen in <cell> monocytes </cell>, stimulation with phorbol esters induces additional , specific binding . Understanding the monocyte-specific function of the <protein> peri-kappa B factor </protein> may ultimately provide insight into the different role <cell> monocytes <cell> and <cell> T-cells <cell> play in HIV pathogenesis.

# Features

- Orthographic Features
- Morphological Features
- Head Noun Features
- Sequential Features
- Dictionary Features

# Orthographic Features

| | |
|---|---|
| allCaps | HIV |
| GreekLetter | alpha, beta, kappa |
| ATCG sequence | CCATG |
| IsDigit | 78- |
| allSmall | protein |
| Hyphen | - |
| Roman Letter | I II ii |
| capsAndDigit | MEK1 |
| initCapDigit | Am80 |
| initCapLower | Ctx |
| twoCaps | FasL |

# Morphological Features

| $F_m$ Name | Prefix/Suffix | Example |
|---|---|---|
| Protein | -nase | kinase |
| | -factor | kappa-binding factor |
| | activator | DNA-binding transcription activator |
| | -receptors | IFNgamma receptors |
| cell type | -ytes | monocytes, leukocytes |
| | -cells | HeLa cells |
| | -lines | human MM cell-lines |
| protein | STAT- | STAT1s |
| DNA | -gene | interleukin gene |
| | -site | NF-kappa B site |

Table 3.2: Morphological features($F^m$)

# Head Noun Features

| Class | Head Nouns |
|---|---|
| PROTEIN | kinase, binding, interleukin activator, protein, interferon receptor, ligand, subunit antibody, complex |
| DNA | DNA, X-chromosome, breakpoint alpha, promoter, cDNA binding, motif, chromosome promoter, element |
| Cells | Lymphocyte, macrophage monocyte, neutrophils |
| RNA | RNA, transcripts |

# Dictionary Features

- We created a dictionary of commonly occurring English words that are always tagged as Others (O)

- Examples:
- A , The, An
- About, After, again
- Air all along
- also and another
- As at with etc.

# Training Set – GENIA JNLPBA

| | |
|---|---|
| IL-2 | B-DNA |
| gene | I-DNA |
| expression | O |
| and | O |
| NF-kappa | B-protein |
| B | I-protein |
| activation | O |
| through | O |
| CD28 | B-protein |
| requires | O |
| reactive | O |
| oxygen | O |
| production | O |

IL2 gene is DNA

NF-kappa B is Protein

CD28 is Protein

# IOB Tagging

- Example : “alpha-globin  promoter”

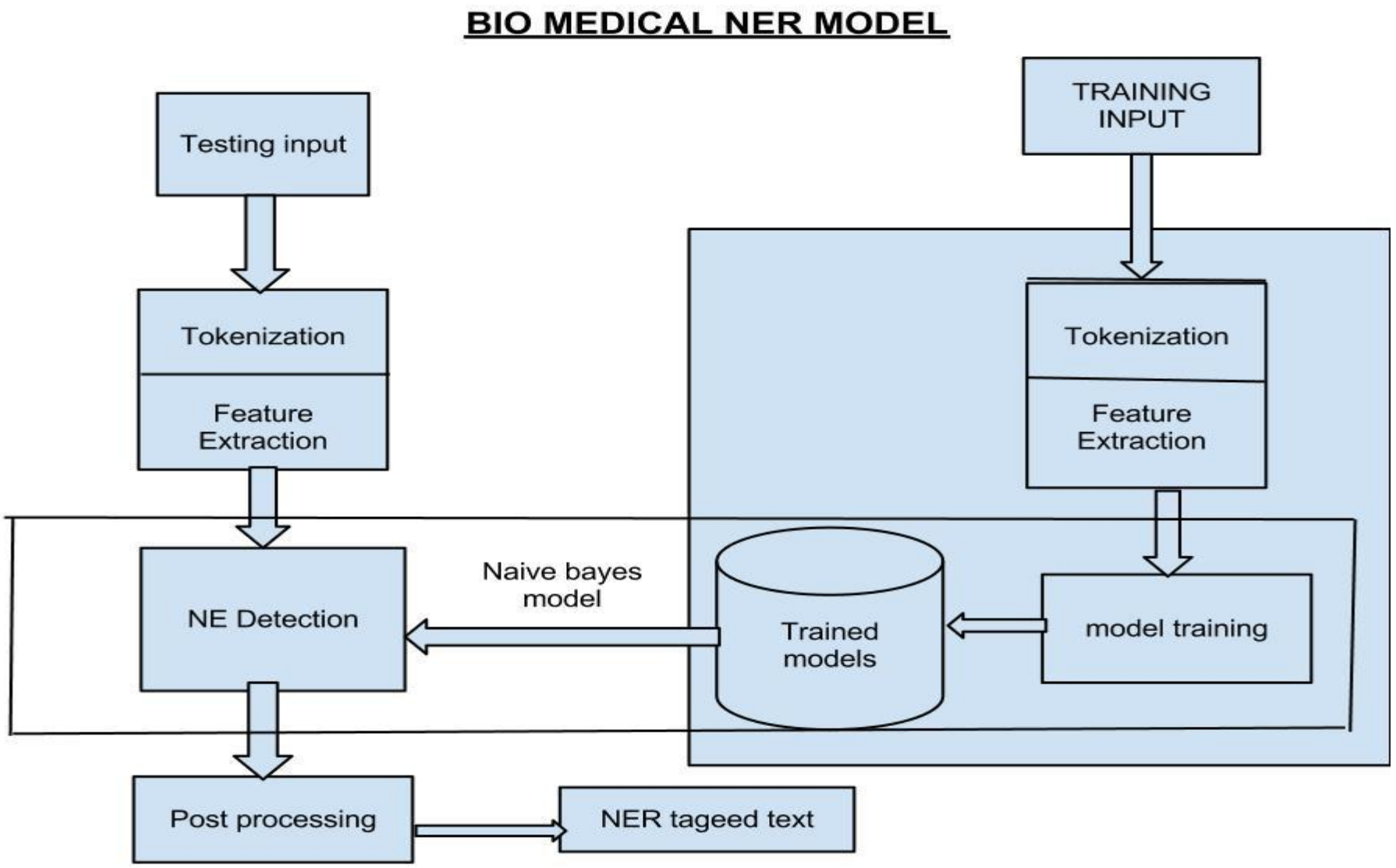
“alpha-globin” <B-DNA>
“promoter”  <I-DNA>

- Where B – Beginning of an entity
- I - Inside an entity
- O - Others

# Sequential Features

- "I-protein always comes after a B-protein or an I-protein“
- "B-protein only comes at the beginning of an entity name“
- "I-protein is mostly followed by I-proteins and Other"

# Design Diagram



# Trained Model

| Feature List | Protein | DNA | …… |
|---|---|---|---|
| Greek Letter | 0.03 | 0.02 | …… |
| hyphen | 0.02 | 0.015 | ……. |
| allSmall | 0.004 | 0.001 | |
| ……. | ……. | ……. | |

# Testing - illustration

**peri-kappa** – Features present are

1. Greek letter – kappa
2. All small
3. Hyphen …

Probability(protein) = 0.03 * 0.02 * 0.004 *…..
Probability(dna) = 0.2 * 0.015 * 0.001 *…..

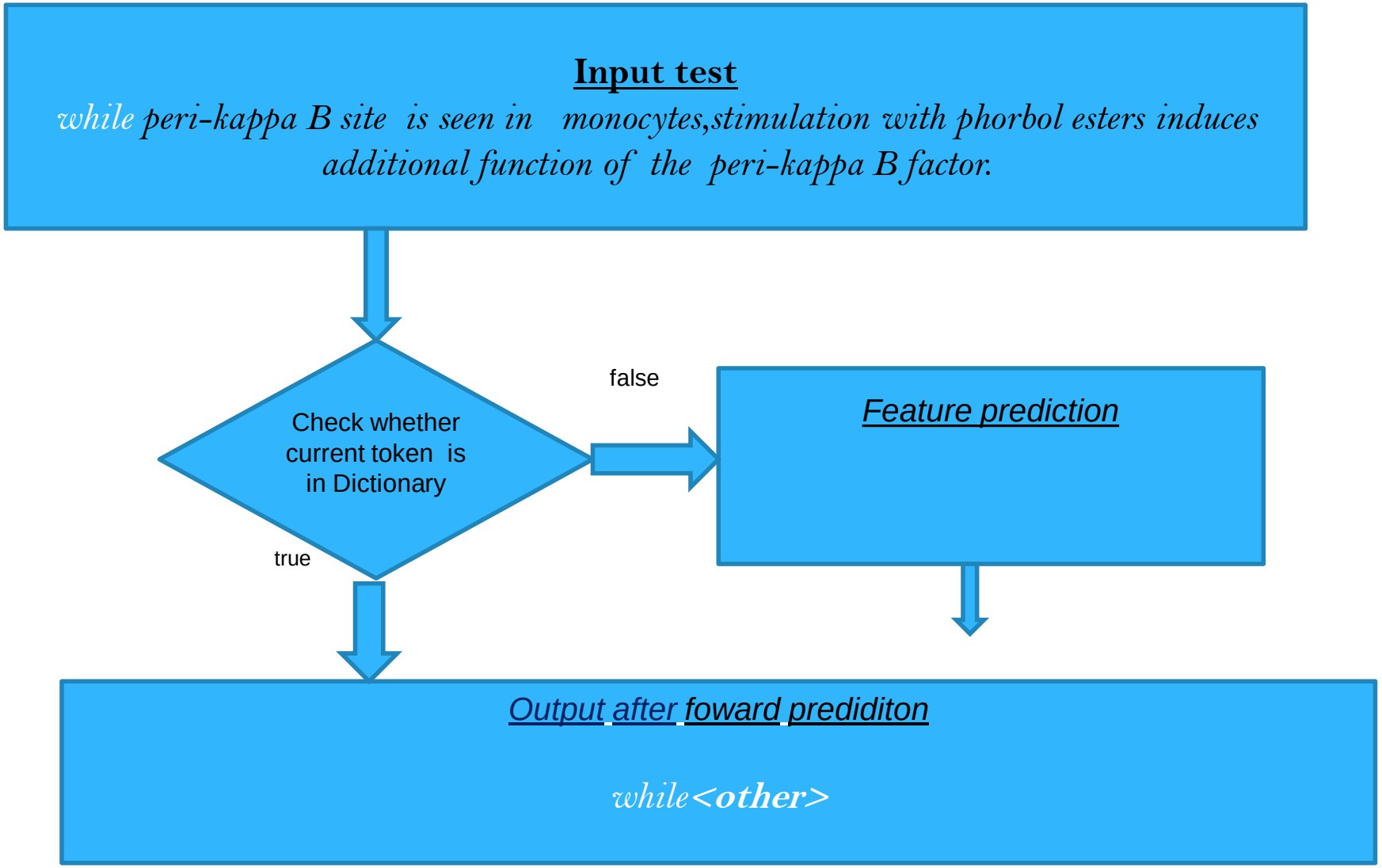
As prob(protein) > all other classes it's choosen as the class of peri-kappa.

# Illustration of a text - Testing

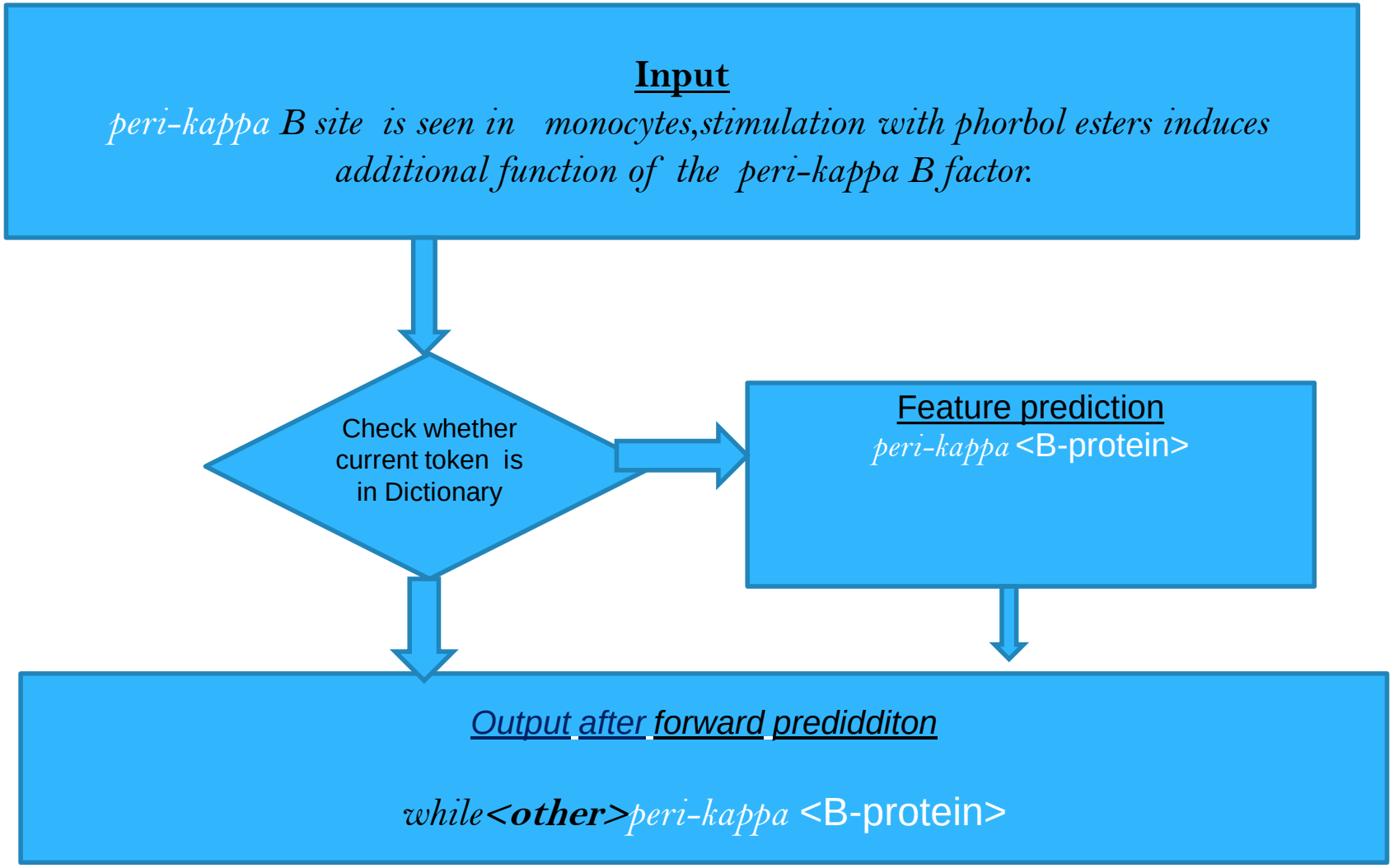
- **Test Input**

While peri-kappa B site is seen in monocytes,stimulation with phorbol esters induces additional function of the peri-kappa B factor

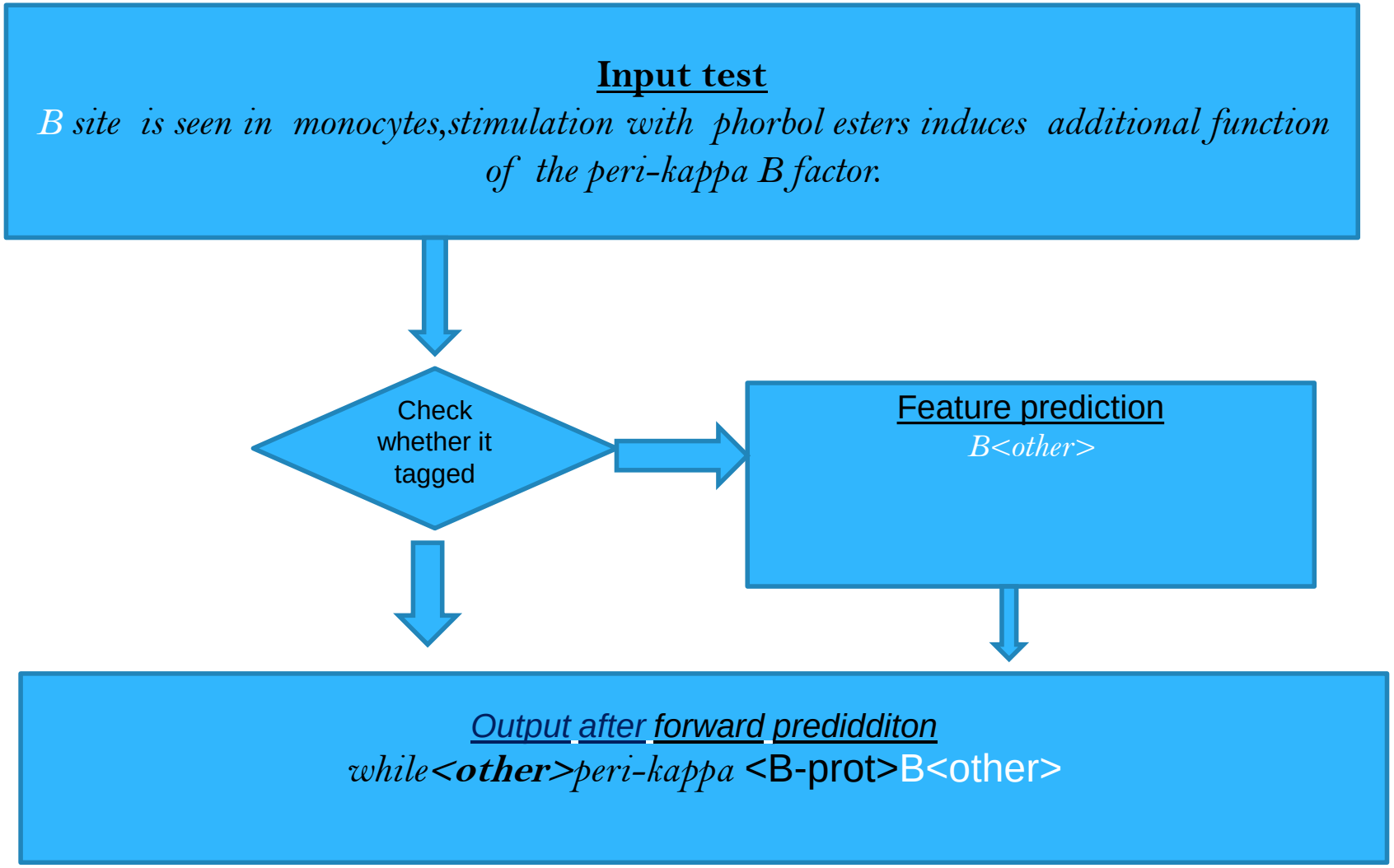
# Illustration of prediction

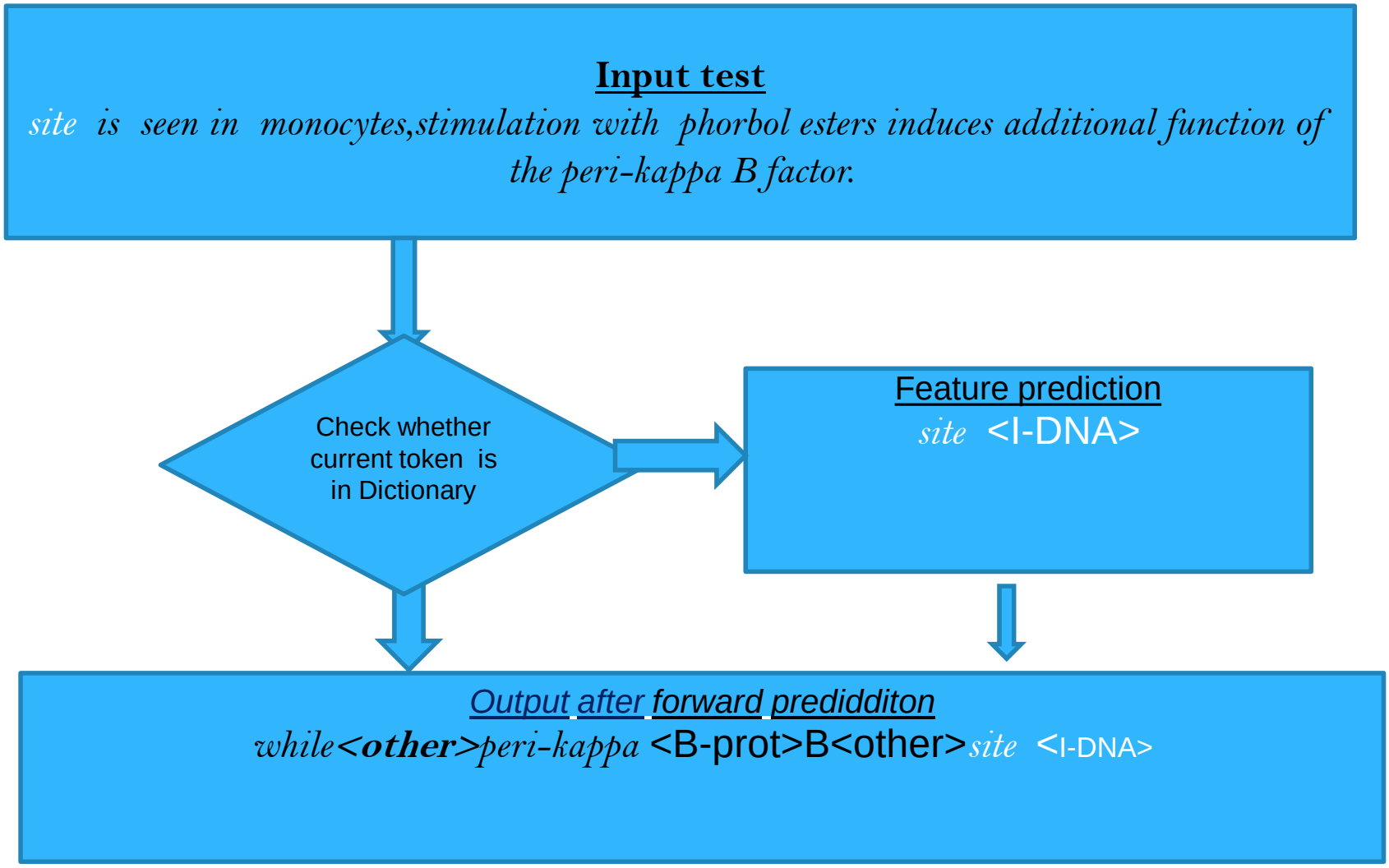


# In forward direction

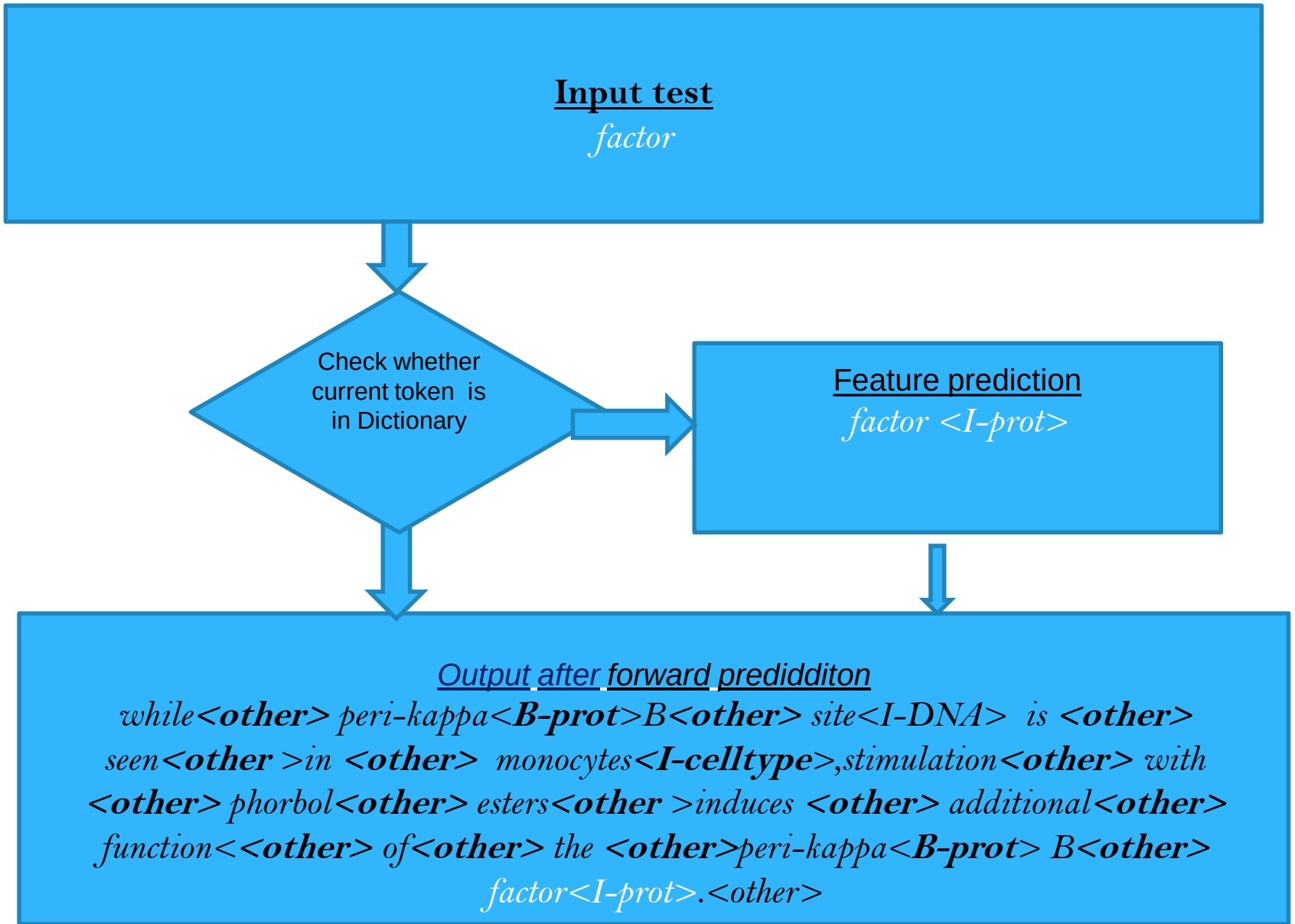


# Without Sequential Features




Input test
site is seen in monocytes,stimulation with phorbol esters induces additional function of the peri-kappa B factor.
Check whether current token is in Dictionary
Feature prediction
site <I-DNA>
Output after forward predidditon
while<other>peri-kappa <B-prot>B<other>site <I-DNA>

# Forward prediction(…last step)



# Backward Processing

- Begin the prediction from the end.
- backward processing with prediction using preceding words.

(Also considering other features such as orthographic, morphological features etc.)

Eg:- **peri-kappa B factor** is a protein

**peri-kappa B site** is a DNA

# Backward prediction

**Input test**

*While peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional function of the peri-kappa B factor .<other>*

Feature prediction

*factor<I-prot>.<other>*

*Output after backward predidditon*

*While peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional function of the peri-kappa B factor<I-prot>.*

# Backward prediction

**Input test**

*While peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional function of the peri-kappa B factor<I-prot>*

Feature prediction

*B **<I-prot>** factor<I-prot>.*

Output after forward predidditon

*While peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional function of the peri-kappa B**<I-prot>** factor<I-prot>.*

# Backward prediction(conti…)

**Input test**

*While peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional function of the peri-kappa B* ***<I-prot>***

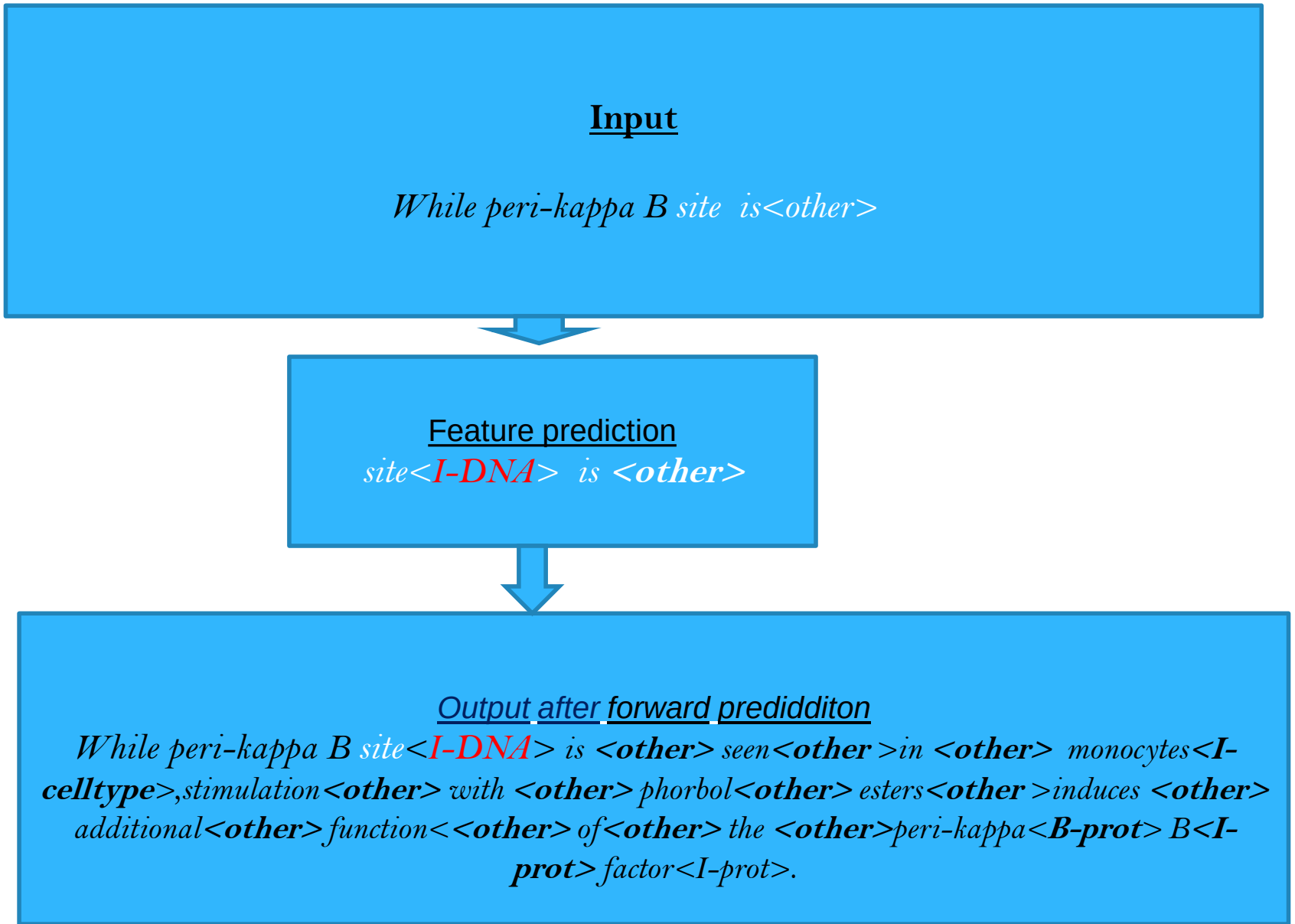
Feature prediction

*peri-kappa<****B-prot****> B* ***<I-prot>***

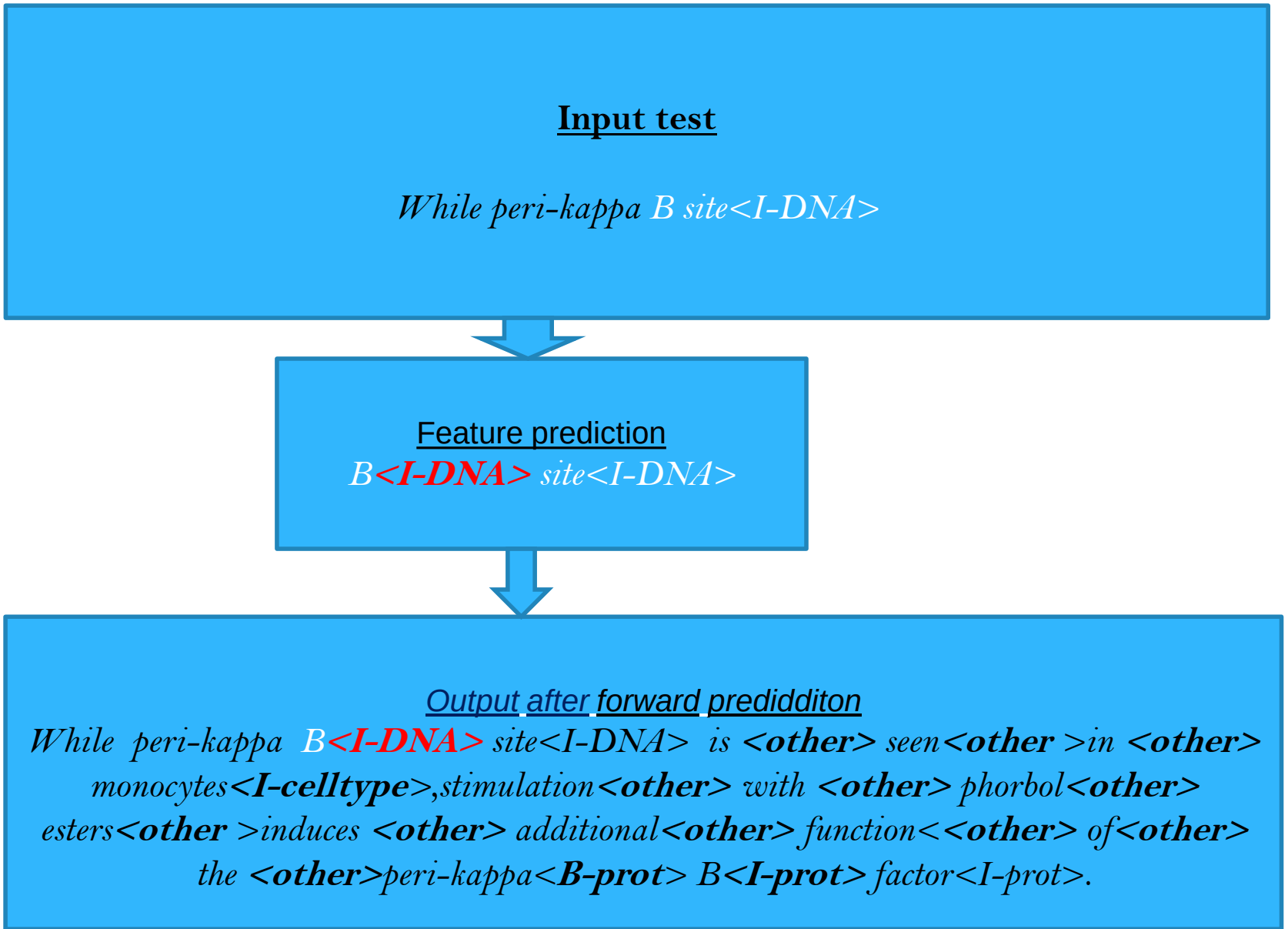
Output after forward predidditon

*While peri-kappa B site is seen in monocytes, stimulation with phorbol esters induces additional function of the peri-kappa<****B-prot****> B****<I-prot>*** *factor<I-prot>.*

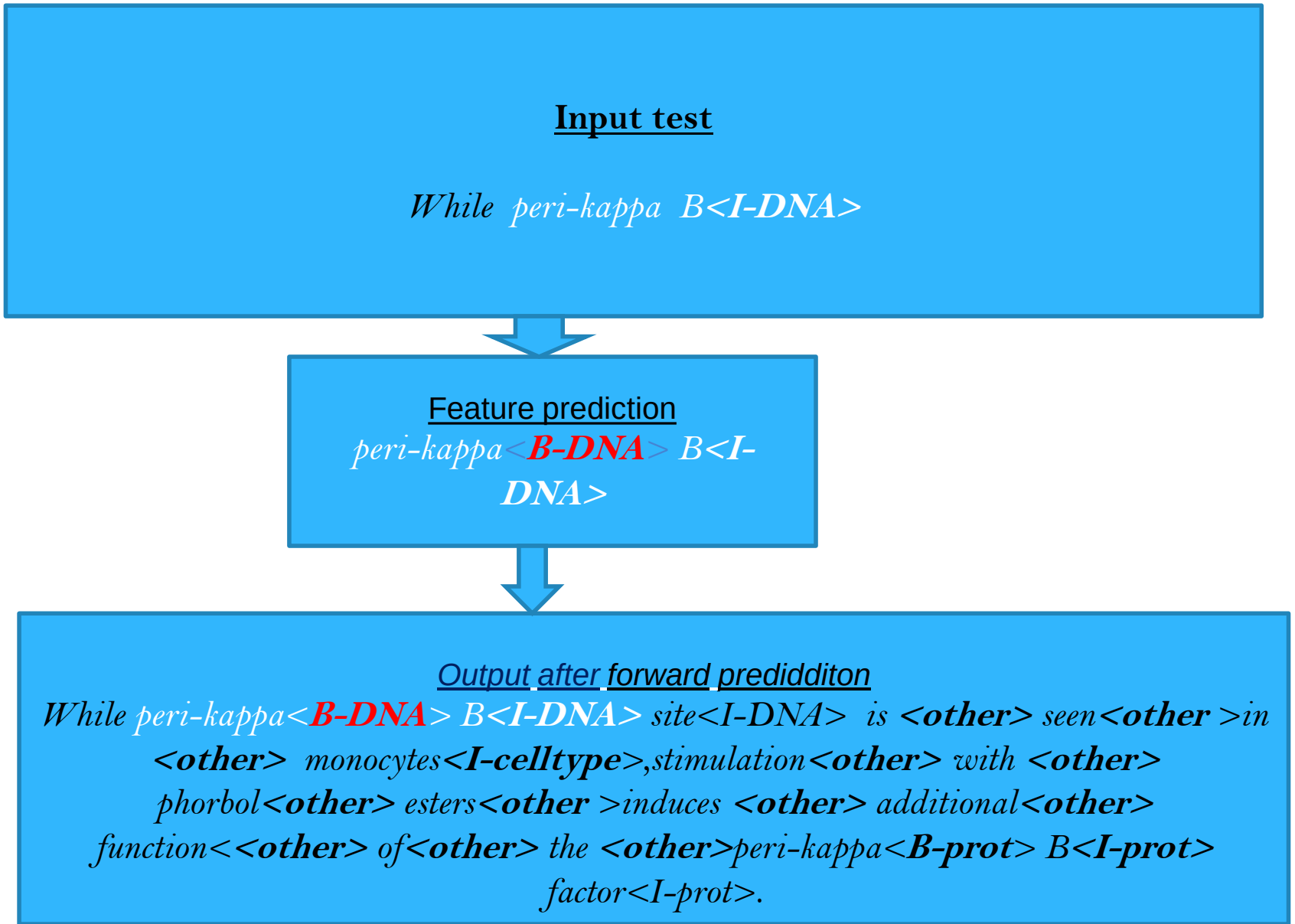
# Backward prediction(…middle)



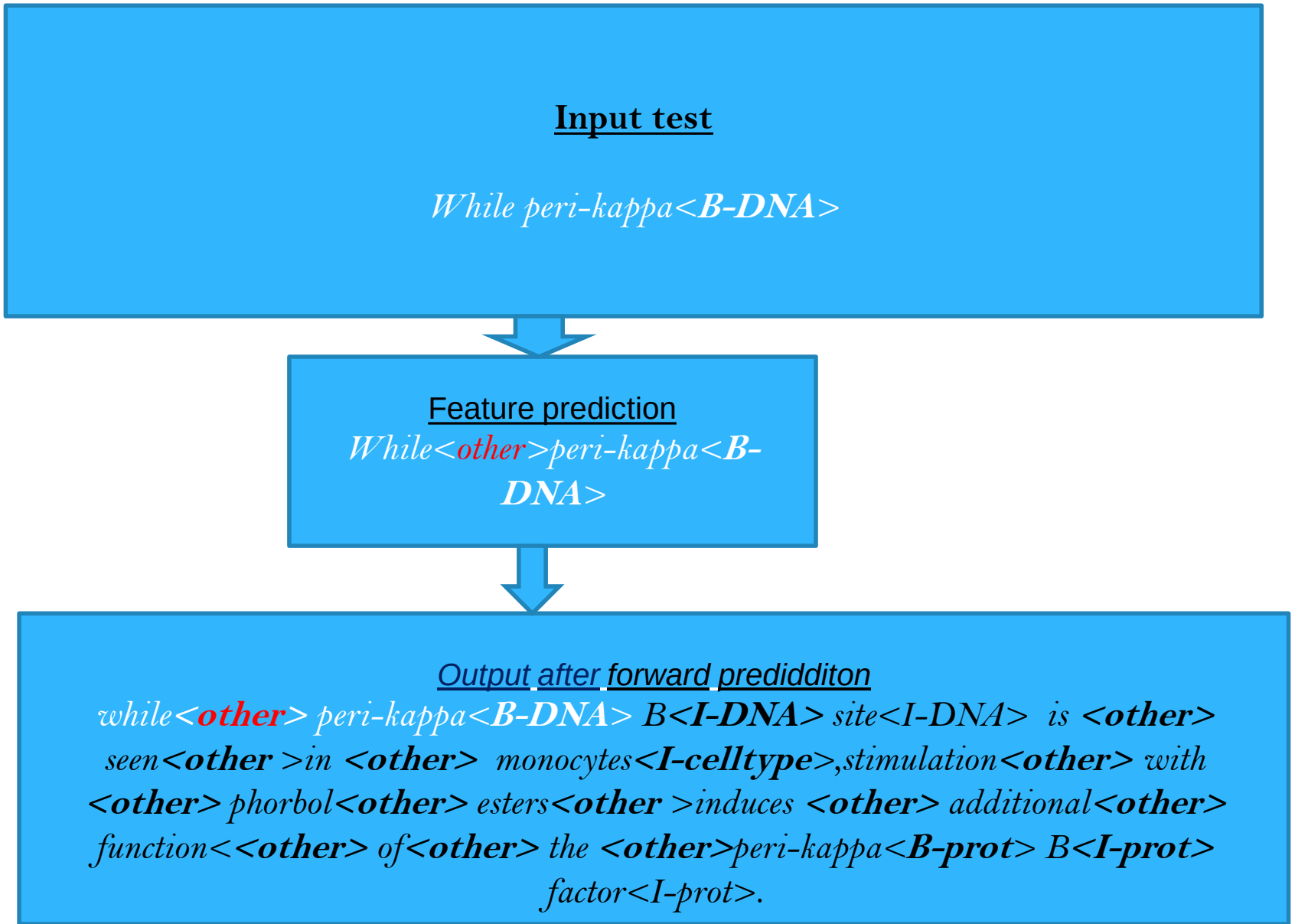
# Backward prediction(…middle)



# Backward prediction(…middle)



# Backward prediction(…middle)



# The Results - Testing

- 1. **Input**: peri-kappa B factor

  **Output** : peri-kappa<B-protein> B<I-protein> factor <I-protein>

- 2. **Input**: peri-kappa B site

  **Output**: peri-kappa<B-DNA> B<I-DNA> site<I-DNA>

# Proposed o/p vs Experimental o/p

1. While specific constitutive binding to the <DNA> peri-kappa B site </DNA> is seen in <cell> monocytes </cell>, stimulation with phorbol esters induces additional , specific binding . Understanding the monocyte-specific function of the <protein> peri-kappa B factor </protein> may ultimately provide insight into the different role <cell> monocytes <cell> and <cell> T-cells <cell> play in HIV pathogenesis.

2. While specific constitutive binding to the <DNA > peri-kappa B site </DNA > is seen in <Cell > monocytes, </Cell > stimulation with phorbol esters induces additional, specific binding. Understanding the <Cell > monocyte-specific </Cell > function of the <Protein > peri-kappa B factor </Protein > may ultimately provide insight into the different role <Cell > monocytes </Cell > and <Cell > T-cells </Cell > play in <DNA> HIV pathogenesis </DNA > .

# Performance Metrics

- Precision $= \frac{Number\ of\ True\ Positives\ (Correct\ Predictions)}{Number\ of\ \ Total\ Predicted(Correct+Wrong)}$
- Recall $= \frac{Number\ of\ True\ Positives\ (Correct\ Prediction)}{Actual\ Number\ of\ an\ Entity\ in\ the\ Testing\ Set}$
- F-measure $= \frac{2 * Precision * Recall}{Precision+Recall}$ (Harmonic Mean)

# Without using sequential features – Forward Processing.

| Entity | precision | Recall | F-measure |
|---|---|---|---|
| Protein | 0.48 | 0.67 | 0.56 |
| DNA | 0.37 | 0.38 | 0.375 |
| RNA | 0.59 | .27 | .37 |
| Cell | 0.53 | .42 | .47 |
| Total | 0.48 | .54 | .51 |

# Back Processing using Sequential Features

| Entity | Precision | Recall | F-measure |
|---|---|---|---|
| Protein | 0.48 | 0.76 | .59 |
| DNA | 0.53 | .54 | .535 |
| RNA | 0.64 | 0.35 | 0.46 |
| Cell | 0.65 | 0.52 | 0.58 |
| Total | .53 | .64 | .58 |

# A comparison to other systems

| | Precision | Recall | F-measure |
|---|---|---|---|
| Jon Patrick et al.[2] | 0.70 | 0.67 | 0.68 |
| Zhou et al.[1] | 0.73 | 0.69 | 0.71 |
| Our Experimental System | 0.53 | 0.65 | 0.58 |

# Observations

1. Many RNA and Cells have low recall because of lack of training examples.

2. Proteins and DNAs have comparable Recall to state of the art systems, but their precision is comparatively lower because many Others are classified as Protein or DNAs. (due to some features present in them like

   eg:- IL-4 promoter which is tagged as Others in Genia testing set, but as DNA in our system, because it appears in many DNA names.

# Observations

1. When corrected with some rules (like Others before B-entity is changed to that entity type) don’t contribute much to the total f-measure (only 1% improvement)

2. The Efficiency of Recognition is 68% while efficiency of classification 58%.

3. Efficiency of Boundary Detection is 40%. That is an entity is taken as correct only when all tokens in that entity is classified correctly.

# Conclusion and Future Work

- We gathered together a large number of features
- implemented using a Naive Bayes Model from without using ready made tool kits.
- We also used a dictionary of 2000 most frequent words
- we will pursue ways for incorporating incremental training in our system
- Collecting more features as well as adding contextual features is also a major aim in our future work.
- We haven’t  used POS tagging features, instead used Back processing for IOB tagged text.
- Achieved performance close to 60%.

# Reference


- [1] Zhang Jie, Shen Dan, Zhou GuoDong, Su Jian and Tan Chew Lim. “Enhancing HMM-based Biomedical Named Entity Recognition by Studying Special Phenomena.” *Journal of Biomedical Informatics,* ***Special Issue on Natural Language Processing in Biomedicine: Aims, Achievements and Challenge. 37(6). 411-422. 2004****.*

- [2] Jon Patrick and Yefeng Wang. “Biomedical Named Entity Recognition System.” Sydney Language Technology Research Group. School of Information Technologies, University of Sydney. ***The Tenth Australian Document Computing symposium(ADCS 2005) 12 December 2005.***

# Reference


- [3] David Nadeau, “A survey of named entity recognition and classication,” Satoshi Sekine National Research CouncilCanada New York University. ***Special issue of Lingvistic Investigationes. 30(1) pp. 3-26.***

- [4] ZHOU GuoDong SU Jian, “Exploring Deep knowledge Resources in Biomedical Name Recognition,” Institute for Infocomm Research 21 Heng Mui Keng Terrace Singapore. ***In Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 473-480.***