GaBNER: A Generalized Architecture of Text Engineering (GATE) based Named Entity Recognition System for Protein Kinases

Gurinder P. S. Gosal[1], Shubangi Sharma[2], Neeraj Sharma[1]

[1]*Department of Computer Science, Punjabi University, Patiala*

[2]*Department of Computer Engineering, University College of Engineering, Punjabi University, Patiala*

ABSTRACT

Protein kinases play a very important role in cellular function and disease such as cancer. They are one of the very few families that have been extensively studied both from the basic and clinical point of view. The resultant mammoth growth in the availability of literature related to protein kinases pose many challenges for kinase researchers in analyzing the overwhelming information. For the present work, the focus is to provide the researchers a system by which protein kinase entities could be mined from the text automatically and presented to them for further annotation and usage. GaBNER: A Generalized Architecture of Text Engineering (GATE) based Entity Recognition System for Protein Kinases, has been developed as a rule based system that automatically extracts the protein kinase information from kinase text. The evaluation shows that the overall recall is 94% and precision is 90% for GaBNER. The F-measure metric following the strict measure (exact match of entities considered only) gives value of 92%. The GaBNER is reproducible and plan is there to soon made it available in the public domain at GitHub- a web-based hosting service for version control.

KEYWORDS: Bio-NLP, Information extraction, Text mining, NER, Generalized Architecture of Text Engineering, GATE, JAPE grammar.

1 INTRODUCTION

There has been an exponential growth in the availability of biological literature in the recent years thereby amassing a huge wealth of domain knowledge [1]. This fact could be gauged from the corresponding increase in the size of MEDLINE[1], the most widely used bibliographic database of biomedical citations, created by the US National Library of Medicine (NLM) [2]. Most of the scientific data in the public domain is available in the unstructured format often as natural language texts. This is true as well for the biological domain where natural language text is used to describe and communicate new discoveries, with scientific papers constituting a considerable fraction of it.

There are plenty of bibliographical resources corresponding to biological literature available in the public domain, with majority of these having the size of enormity. The gigantic volume of information in the bibliographical resources put the researchers in a dire need of having some automated means for organizing and managing efficiently the knowledge. Biological knowledge management is becoming familiar and important concept for the life scientists as it is evident from its plenty of coverage in the scientific publications [3]. The effective management of the biological knowledge, when the data is available in textual form, requires processing of the natural language texts.

[1] *http://www.nlm.nih.gov/pubs/factsheets/medline.html*

Bio-NLP is the name given to the application of natural language processing in the biological or biomedical domain. Bio-NLP has been expanding its reach to a variety of biomedical texts while supporting a wide range of tasks such as document classification, named-entity recognition, relation extraction, event extraction, ontology development, summarization, question answering, literature-based discovery. One of the key elements in the processing of natural language is text mining and most of the text mining systems make use of the methods and tools of NLP. Researchers define **text mining** as a knowledge extracting method to extract useful and previously unknown information from a document or set of texts by identifying facts inherent and inexplicit in the data [4]. **Biomedical text mining**, the term some practitioners use as synonymous with Bio-NLP, has been defined as the application of the automated methods of text mining for using the enormous amount of knowledge available in the biomedical literature [5].

The biomedical publications, like the other types of unstructured texts, have to confront with the issues such as complexity, ambiguity, synonymy and polysemy. Many surveys and works have shown that the general-purpose text and data mining tools do not perform that well for biological text as they perform for other types of texts mainly due to the highly specialized nature of biomedical domain [4-6]. The years since 1998 have seen an explosion in work in biomedical text mining (Bio-NLP) of both clinical text and the biomedical literature [7]. The text mining systems can have a wide scope ranging from simple tasks of recognizing named entities and their categorization to more complex tasks of summarizing, question answering, and processing non-textual texts. Information extraction (IE), often considered as an initial processing step for most of the text mining systems, extracts explicitly stated facts or structured facts from unstructured or semi-structured text and often becomes the basis of other biomedical text mining applications. For processing unstructured text, three major subtasks of information extraction have been reported in literature: Named Entity Recognition (NER), Relation Extraction (RE) and Event Extraction (EE).

Named entity recognition (NER) is a task that tries to find entities in the text and classifies these entities into some predefined classes. Biological Named Entity Recognition (BER) is a subfield of NER which recognizes the predefined categories of entities in the biological domain, such as the names of proteins, genes, drugs, or diseases, when the text under consideration is biological text. There has been a growing interest in the task of named entity recognition in the text mining area and a lot of research has been done in this direction since 1996, when the 6th Message Understanding Conference (MUC6) was organized [8]. Particularly, a lot of progress has been made in the biomedical domain with greater emphasis on identifying the domain-specific entities. The recognition of biological entities in text by identifying the key concepts of interest allows for further extraction of relationships and other complex information, and hence enabling more advanced text-mining tasks to be performed.

Protein kinases play a very important role in cellular function and disease such as cancer. They are one of the very few families that have been extensively studied both from the basic and clinical point of view. This has led to a mammoth growth in the availability of literature related to protein kinases in the public domain and researchers have difficult time in analyzing the overwhelming information. There have been many reported works in the literature that deal with the extraction of biological entities, such as genes, proteins, disease, drugs, from unstructured text. However, there is lack of works that aim for the recognition of entities focusing specifically on the protein kinases.

For the present work, our focus is to provide the researchers a mechanism by which protein kinase entities could be mined from the text automatically and presented to them for further annotation and usage. This semantic knowledge might not be present in the text itself and could be augmented from outside domain knowledge resources such as ontology. The presented work, GaBNER: A Generalized Architecture of Text Engineering (GATE) based Entity Recognition System for Protein Kinases, relates to the development of a system that automatically extracts the protein kinase information from biomedical text.

2 TEXT PROCESSING FRAMEWORKS

Due to growing importance of NLP research field, there have been many initiatives in past years to provide the language engineers with some kind of a development environment for language engineering. A lot of effort has been made to create the resources for major languages and similar attempts have been going on for developing similar resources for less widely-used languages. The need has been observed that resources and tools are reusable as well as compatible in terms of their representation due to the cost and efforts involved in their creation and the potential for linking multi-lingual and multi-modal language data [9].

OpenNLP[2] provides a set of tools to do many of the typical tasks in linguistic processing pipelines such as sentence detection, tokenization, part-of-speech tagging, phrase chunking, syntactic parsing, named entity recognition, and co-reference resolution [10]. Due to growing need to process unstructured information, specifically multilingual, natural language text, coupled with IBM Research's investment in NLP, led to development of middleware architecture for natural language processing known as Unstructured Information Management Architecture (**UIMA)** [11]. UIMA framework, a data management system, that supports pipelined applications, is now being maintained by the Apache Software Foundation after the Java based framework was accepted as Apache Incubator project in 2006 and has been standardized by OASIS in 2009.

Generalized Architecture of Text Engineering (GATE) framework provides a platform to develop a number of successful applications for various natural language processing tasks, such as, information extraction and at the same time also helps to develop and annotate corpora and perform evaluations on the applications developed [12]. GaBNER has been developed using the GATE framework and it is discussed in detail in the next section.

2.1 General Architecture for Text Engineering (GATE)

University of Sheffield has developed a leading natural language processing and information extraction platform dubbed as General Architecture for Text Engineering (**GATE**)[3]. Originally released in 1996, GATE is the architecture that usually divides the natural language process into the elements of software systems as various types of component, known as resources. Individual processing steps in the GATE pipeline are performed by *processing resources* (PRs) such as lemmatisers, translators, generators, parsers or speech recognizers. The GATE platform provides these different processing layers as standardized building blocks for creating sophisticated NLP applications. Using GATE platform's user-friendly interface (**Figure 1**), developers can use tools and linguistic databases easily and launch different processes on the same text and compare the results or they can run the same module on different text collections and investigate the differences [13, 14].

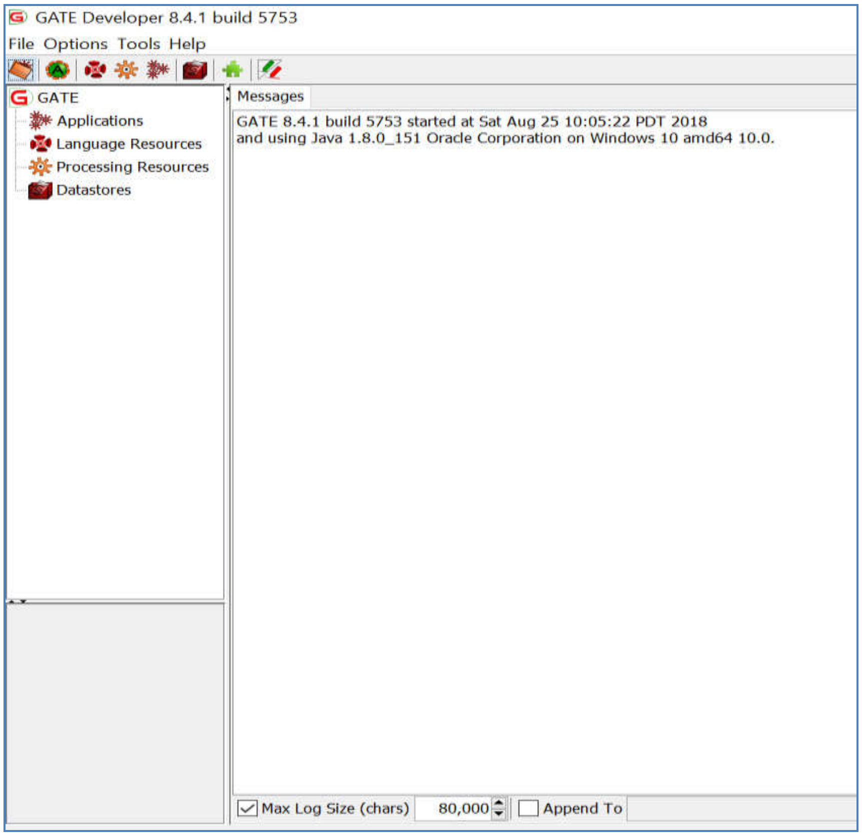
GATE Application Programming Interface (API) provides programmatic access to deploy applications developed within GATE outside its Graphical User Interface (GUI). We can use the reusable modules, the document and annotation model, and the visualization components independent of the development environment. The GATE framework makes a distinction between data, algorithms, and ways of visualizing them.

The components in GATE are one of three types:

- **Language Resources** (LRs) represent entities such as lexicons, corpora or ontologies;
- **Processing Resources** (PRs) represent entities that are primarily algorithmic, such as parsers, generators or language modelers;

[2] http://opennlp.sourceforge.net
[3] http://gate.ac.uk

- **Visual Resources** (VRs) represent visualization and editing components that participate in GUIs.



```
GATE 8.4.1 build 5753 started at Sat Aug 25 10:05:22 PDT 2018
and using Java 1.8.0_151 Oracle Corporation on Windows 10 amd64 10.0.
```

☑ Max Log Size (chars) 80,000 □ Append To

Figure 1: The user-friendly interface of GATE Developer Tool

Collectively, all resources are known as **CREOLE** (a Collection of REusable Objects for Language Engineering), which can be seen as a library of processing resources, language resources and data structures for general usage. GATE has a very good feature in form of Java Annotation Pattern Engine (JAPE) language for doing finite state processing over annotations based on regular expressions. JAPE works on given pattern rules which describe a pattern of annotations and their features in the input document, and a resultant action to do matching of pattern [15]. GATE is distributed with an information extraction (IE) system called "A Nearly-New IE (ANNIE)" that relies on finite state algorithms and the JAPE language. ANNIE consists of the following main processing resources: tokeniser, sentence splitter, POS tagger, gazetteer, name entity tagger (based on JAPE), orthomatcher and co-reference resolver [13].

The use of the frameworks for NLP tasks in a specific domain, by applying different NLP techniques and using ontologies, is the area which is being explored extensively now a day. Our focus is to have an application set-up as discussed here for named entity recognition in the specific domain of protein kinases which are discussed in the next section.

3 PROTEIN KINASE

Protein kinases are kinase enzymes that add phosphate groups to substrate proteins (process called phosphorylation) and as a consequence effectuate usually a functional change of the target protein (substrate) by changing the enzyme activity, cellular location, or association with other proteins, as shown in **Figure 2** [16]. The kinases may regulate majority of cellular pathways, especially they are very prominent in signal transduction. Signal transduction begins with a signal to a receptor and ends with a change in the cell behavior.

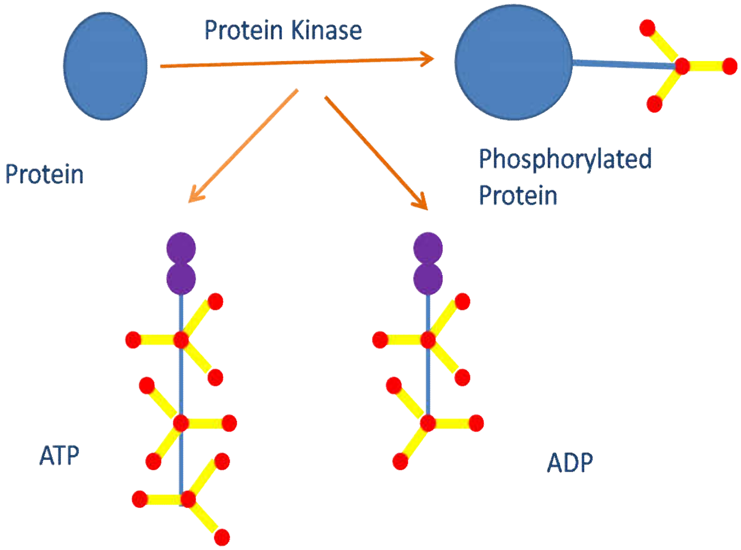


Figure 2: Protein kinase phosphorylation (adapted from [17]).

There has been a non-redundant set of 538 human protein kinase genes identified based on a comprehensive approach using human genome analysis in KinBase [18]. This collection in KinBase is made up of published human genome sequences as well as of other sequence databases and also including directed cloning and sequencing of individual genes. The set of protein kinase genes in KinBase includes most human members of the eukaryotic protein kinase super family, and many atypical kinases and almost all human protein phosphorylation [18-19]**.**

The collection of eukaryotic protein kinases is categorized into **twelve groups**. The classification of protein kinases into groups by KinBase is shown in **Figure 3**. There are 538 protein kinase genes in human genome and these form approximately 2% of all human genes, however, a huge percentage (up to 30%) of human proteins may be altered by kinase activity [18].

Protein kinase family is one of the largest and most functionally diverse gene families that are implicated in many types of human cancers. This protein family is one of the very few families that have been extensively studied both from the basic and clinical point of view.

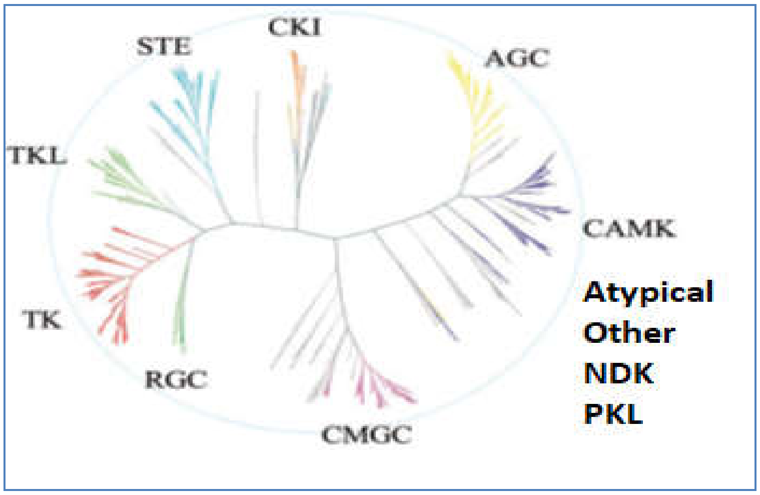


Figure 3: The Protein kinases classification in groups by KinBase (Atypical, Other, NDK, PKL shown in addition to groups in the original figure) [18].

4 DEVELOPMENT OF GaBNER PIPELINE

The text mining pipeline for named entity recognition (GaBNER) has been implemented in the GATE framework. GATE is the architecture that usually divides the natural language process into elements of software systems as different types of components, known as resources. Individual processing steps in the GATE are performed by *processing resources* (PRs), such as lemmatisers, tokenizers, POS taggers, parsers. Our GATE based pipeline GaBNER performs the main functions of a **gazetteer based matching** and **grammar-based recognition** of key entities of protein kinase in the text. The recognized entities have been normalized to standard terms by using synonyms to be further annotated with full names and other relevant information about entities that have been obtained from KinBase [18, 20]. The components of the GaBNER pipeline are briefly explained below:

Document Reset Processing Resource: This is a standard GATE component and it removes all the annotation sets except the ones that are specified to be retained.

English Tokenizer: This is another standard GATE component and splits the input text in tokens like words, punctuation characters.

Gazetteer: A gazetteer is the lookup list of entities. The gazetteer lists are used for basic matching of the entities in the GABNER pipeline. The gazetteer lists used for the pipeline include a list of protein kinase with canonical names as listed in KinBase, a list of protein kinase synonyms and so on.

Sentence Splitter: This is another standard GATE component and splits the input text into sentences according to punctuation. This module is required for the POS tagger and other modules.

POS Tagger: This tagger is used to produce a part-of-speech tag as an annotation on each word or symbol.
Jape Transducer: Java Annotation Pattern Engine (JAPE) works on given pattern rules that describe a pattern of annotations and their features in the input document, and a resultant action to do matching of pattern [21]. Transducer JAPE rules utilize these annotations along with annotations from other processing resources to further identify patterns/entities from the text. The transducer JAPE grammar rules are defined in GABNER for finding protein kinase annotations. The **Figure 4** shows the components of the GaBNER pipeline in the execution order.

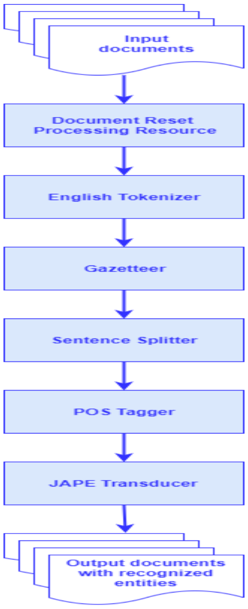


Figure 4: GaBNER entity recognition pipeline

GaBNER has two important enabling components for its pipeline that have been customized for the recognition task. The gazetteer lists (lexicons) have been prepared for key data from KinBase [18, 20] for basic matching of the entities. Also the rules embedded as JAPE grammar have been hand crafted to extend the recognition procedure beyond basic matching.

4.1 Gazetteers for Pipeline

The **KinBase** [18, 20] has been used for preparing different gazetteer lists for basic matching of the entities in the NLP pipeline used. We have prepared the gazetteer lists, for protein kinase with canonical names as listed in KinBase and for protein kinase synonyms. **Table 1** shows an extracted view of three out of different gazetteer lists prepared from knowledge contained inside *ProCDIO*.

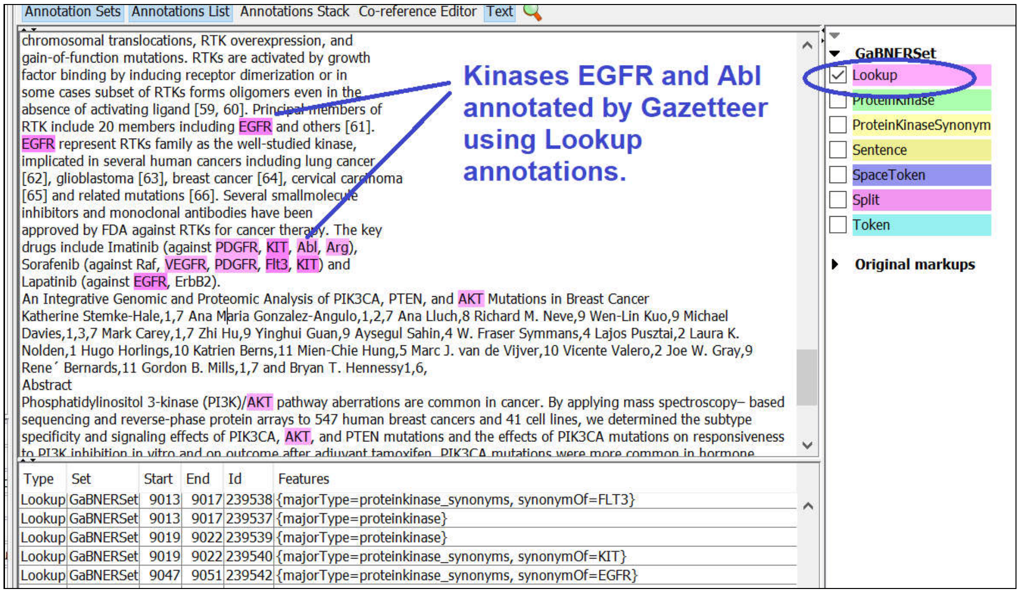
Table 1: The excerpt from two gazetteers used in GaBNER

Protein Kinase Gazetteer List	Protein Kinase Synonym Gazetteer List
ANKRD3	v-abl&synonymOf=ABL1
ANPA	ABL2&synonymOf=ABL2
ANPB	ABLL&synonymOf=ABL2
ARAF	ARG&synonymOf=ABL2
ATM	ACK&synonymOf=ACK
ATR	ACK1&synonymOf=ACK
AURA	FLJ44758&synonymOf=ACK
AURB	FLJ45547&synonymOf=ACK
AURC	p21cdc42Hs&synonymOf=ACK
AXL	TNK2&synonymOf=ACK
BARK1	ACTRII&synonymOf=ACTR2
BARK2	ACVR2&synonymOf=ACTR2
BAZ1A	ACVR2A&synonymOf=ACTR2
BAZ1B	ACTR-IIB&synonymOf=ACTR2B
BCKDK	ACVR2B&synonymOf=ACTR2B
BCR	hActR-IIB&synonymOf=ACTR2B
BIKE	Hs.23994&synonymOf=ACTR2B
BLK	ADCK1&synonymOf=ADCK1
BLVRA	FLJ39600&synonymOf=ADCK1
BMPR1A	LOC57143&synonymOf=ADCK1
BMPR1B	AARF&synonymOf=ADCK2
BMPR2	ADCK2&synonymOf=ADCK2
BMX	………………
…………	

The tokens generated in the GaBNER pipeline have been matched with these ontology-based gazetteers and we get the **Lookup** type annotations that annotate the tokens based on protein kinases, protein kinase synonyms. For example, kinases EGFR and ABL are Lookup annotations based on gazetteers as shown in GATE GUI window in **Figure 5**.

4.2 JAPE Grammar Rules

The annotations in the Lookup further subjects to certain grammar-rules, written in JAPE [21] that are included in the GATE pipeline as **Jape Transducer** processing resource. These grammar rules have been crafted and used to refine and enhance the basic annotations provided by gazetteers and also add certain other important features for these annotations. *For example*, basic kinase annotation is added synonym with a feature "*synonymOfKinase*", by rules written in some of these grammars. One excerpt of one such grammar rule written in JAPE is shown in ***Figure 6*.**



chromosomal translocations, RTK overexpression, and gain-of-function mutations. RTKs are activated by growth factor binding by inducing receptor dimerization or in some cases subset of RTKs forms oligomers even in the absence of activating ligand [59, 60]. Principal members of RTK include 20 members including EGFR and others [61]. EGFR represent RTKs family as the well-studied kinase, implicated in several human cancers including lung cancer [62], glioblastoma [63], breast cancer [64], cervical carcinoma [65] and related mutations [66]. Several smallmolecule inhibitors and monoclonal antibodies have been approved by FDA against RTKs for cancer therapy. The key drugs include Imatinib (against PDGFR, KIT, Abl, Arg), Sorafenib (against Raf, VEGFR, PDGFR, Flt3, KIT) and Lapatinib (against EGFR, ErbB2).
An Integrative Genomic and Proteomic Analysis of PIK3CA, PTEN, and AKT Mutations in Breast Cancer
Katherine Stemke-Hale,1,7 Ana Maria Gonzalez-Angulo,1,2,7 Ana Lluch,8 Richard M. Neve,9 Wen-Lin Kuo,9 Michael Davies,1,3,7 Mark Carey,1,7 Zhi Hu,9 Yinghui Guan,9 Aysegul Sahin,4 W. Fraser Symmans,4 Lajos Pusztai,2 Laura K. Nolden,1 Hugo Horlings,10 Katrien Berns,11 Mien-Chie Hung,5 Marc J. van de Vijver,10 Vicente Valero,2 Joe W. Gray,9 Rene´ Bernards,11 Gordon B. Mills,1,7 and Bryan T. Hennessy1,6,
Abstract
Phosphatidylinositol 3-kinase (PI3K)/AKT pathway aberrations are common in cancer. By applying mass spectroscopy– based sequencing and reverse-phase protein arrays to 547 human breast cancers and 41 cell lines, we determined the subtype specificity and signaling effects of PIK3CA, AKT, and PTEN mutations and the effects of PIK3CA mutations on responsiveness to PI3K inhibition in vitro and on outcome after adjuvant tamoxifen. PIK3CA mutations were more common in hormone

Kinases EGFR and Abl annotated by Gazetteer using Lookup annotations.

Type	Set	Start	End	Id	Features
Lookup	GaBNERSet	9013	9017	239538	{majorType=proteinkinase_synonyms, synonymOf=FLT3}
Lookup	GaBNERSet	9013	9017	239537	{majorType=proteinkinase}
Lookup	GaBNERSet	9019	9022	239539	{majorType=proteinkinase}
Lookup	GaBNERSet	9019	9022	239540	{majorType=proteinkinase_synonyms, synonymOf=KIT}
Lookup	GaBNERSet	9047	9051	239542	{majorType=proteinkinase_synonyms, synonymOf=EGFR}

Figure 5: A snapshot of Lookup annotations based on gazetteers in GaBNER pipeline

```
/*  pkname.jape
*  This jape file is protein kinase jape file (under main jape file) for GaBNER
*  $Id: pkname.jape $    */


Phase: ProteinKinasePhase
Input: Lookup   // The Input Annotations list contains a list of all the annotation types you want to use for matching
// on the LHS of rules in that grammar phase
Options: control = appelt // applet option - In the appelt style, which rule to apply is selected in the following order:
//longest match, explicit priority, rule defined first
Rule: ProteinKinaseSynonym
Priority: 80
(      {Lookup.majorType == proteinkinase_synonyms}      ):pks -->
{
```

```
//:pks.ProteinKinase = {kind="ProteinKinase",rule = "ProteinKinaseSynonym",
synonymOf=}
// create an annotation set consisting of all the full name annotations of organisms
try{
gate.AnnotationSet pkSet = (gate.AnnotationSet)bindings.get("pks");
gate.Annotation pkAnnot = pkSet.iterator().next();
gate.FeatureMap features = Factory.newFeatureMap();
try{
features.put("synonymOfKinase", pkAnnot.getFeatures().get("synonymOf"));
features.put("rule", "ProteinKinaseSynonym");
outputAS.add(pkSet.firstNode(), pkSet.lastNode(), "ProteinKinaseSynonym",
features);
} catch (Exception ex) {
System.out.println( "Exception: " + ex );     }
} catch (Exception ex) {
System.out.println( "Exception: " + ex );   }
}
```

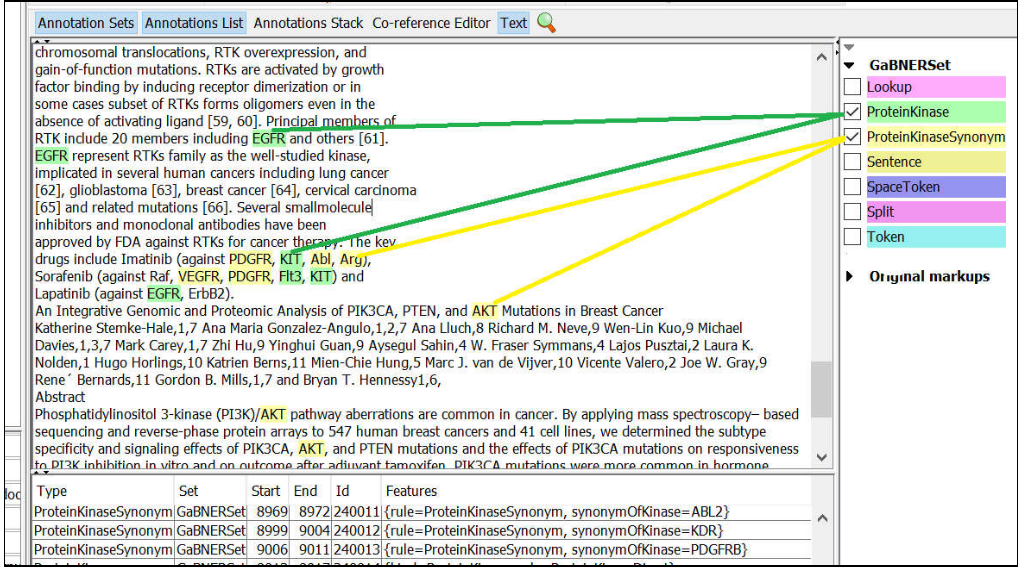
Figure 6: An excerpt of JAPE grammar rule for "*SynonymOfKinase*"

As an application scenario, GaBNER was run on corpora of 100 documents related to protein kinases obtained manually from bio-medical online sources such as, PubMed (*http://www.ncbi.nlm.nih.gov/pubmed/*). These documents as a corpus when put into the GaBNER pipeline run automatically using Java coded software and the output is the annotated documents. This annotated corpus is automatically stored into GATE data store and can be saved as Gate XML format (**Figure 7**).

```
………………..
</Annotation>
<Annotation Id="21541" Type="ProteinKinaseSynonym" StartNode="45423"
EndNode="45427">
<Feature>
  <Name className="java.lang.String">rule</Name>
  <Value className="java.lang.String">ProteinKinaseSynonym</Value>
</Feature>
<Feature>
  <Name className="java.lang.String">synonymOfKinase</Name>
  <Value className="java.lang.String">FASTK</Value>
</Feature>
</Annotation>
………….
```

Figure 7: An excerpt of XML file annotated by GaBNER (here text segment "FASTK" has been shown annotated)

GaBNER annotates the documents by recognizing the named entities of kinases organizing these entities under "Annotation Set" placed under "GaBNERSet". A document annotated in such a way for ProteinKinase" and "ProteinKinasSynomym annotation sets is shown in **Figure 8**.



chromosomal translocations, RTK overexpression, and
gain-of-function mutations. RTKs are activated by growth
factor binding by inducing receptor dimerization or in
some cases subset of RTKs forms oligomers even in the
absence of activating ligand [59, 60]. Principal members of
RTK include 20 members including EGFR and others [61].
EGFR represent RTKs family as the well-studied kinase,
implicated in several human cancers including lung cancer
[62], glioblastoma [63], breast cancer [64], cervical carcinoma
[65] and related mutations [66]. Several smallmolecule
inhibitors and monoclonal antibodies have been
approved by FDA against RTKs for cancer therapy. The key
drugs include Imatinib (against PDGFR, KIT, Abl, Arg),
Sorafenib (against Raf, VEGFR, PDGFR, Flt3, KIT) and
Lapatinib (against EGFR, ErbB2).
An Integrative Genomic and Proteomic Analysis of PIK3CA, PTEN, and AKT Mutations in Breast Cancer
Katherine Stemke-Hale,1,7 Ana Maria Gonzalez-Angulo,1,2,7 Ana Lluch,8 Richard M. Neve,9 Wen-Lin Kuo,9 Michael
Davies,1,3,7 Mark Carey,1,7 Zhi Hu,9 Yinghui Guan,9 Aysegul Sahin,4 W. Fraser Symmans,4 Lajos Pusztai,2 Laura K.
Nolden,1 Hugo Horlings,10 Katrien Berns,11 Mien-Chie Hung,5 Marc J. van de Vijver,10 Vicente Valero,2 Joe W. Gray,9
Rene´ Bernards,11 Gordon B. Mills,1,7 and Bryan T. Hennessy1,6,
Abstract
Phosphatidylinositol 3-kinase (PI3K)/AKT pathway aberrations are common in cancer. By applying mass spectroscopy– based
sequencing and reverse-phase protein arrays to 547 human breast cancers and 41 cell lines, we determined the subtype
specificity and signaling effects of PIK3CA, AKT, and PTEN mutations and the effects of PIK3CA mutations on responsiveness
to PI3K inhibition in vitro and on outcome after adjuvant tamoxifen. PIK3CA mutations were more common in hormone

Type	Set	Start	End	Id	Features
ProteinKinaseSynonym	GaBNERSet	8969	8972	240011	{rule=ProteinKinaseSynonym, synonymOfKinase=ABL2}
ProteinKinaseSynonym	GaBNERSet	8999	9004	240012	{rule=ProteinKinaseSynonym, synonymOfKinase=KDR}
ProteinKinaseSynonym	GaBNERSet	9006	9011	240013	{rule=ProteinKinaseSynonym, synonymOfKinase=PDGFRB}

Figure 8: A snapshot of GATE GUI window showing entities annotated under *protein kinases* and *protein kinase synonyms* annotation sets

5 EVALUATION

The standard metrics for evaluation i.e. *precision*, *recall* and *F-measure* have been applied to evaluate the performance of entity recognition pipeline GaBNER in detecting the entities in the kinase related text by using corpora of manually annotated full-text documents. Keeping in view the large scale of overall system and furthermore intensive effort involved in manually annotating these kinase-related documents, we restricted the corpora of manually annotated gold standard documents to three in this component's evaluation.
The **Table 2** shows the evaluation statistics obtained by executing the GaBNER pipeline on manually annotated corpora, for recognition of the named entity "ProteinKinase"**.** This statistic is based on the strict criteria of F-measure which means that only exact matches are accounted for and partial matches are discarded while considering an entity match correct.

Table 2: Entity recognition evaluation results by GaBNER in terms of recall, precision and F-measure

Annotation	Recall	Precision	F-Measure (strict)
ProteinKinase	**0.9451**	**0.8959**	**0.9198**

We see from the evaluation metrics result that the overall recall is ~**94%** and the value of precision is ~**90%**. The F-measure metric following the strict measure (exact match of entities considered only) gives value of ~**92%**. WE observe that the recall, precision and F-measures for "ProteinKinase" named entity is very high very much on the expected lines and the reason for that could be attributed to the fact that all the 538 canonical kinases and most of their synonyms have been well represented in the lexicons.

6 CONCLUSION

Recognizing the key entities from the unstructured biomedical text and classifying these entities into some predefined classes has been considered as a fundamental task in text mining based on which many text mining. Protein kinases are a large family of proteins that play a very significant role in cancer disease. Annotating text by recognizing kinase related entities in full-text articles can play an important role in mining the large amount of literature related to protein kinases.

GaBNER is an effort to provide a mechanism in which the relevant key entities in the text get mined and further utilized in different advanced applications or tasks. The pipeline in itself performs the functions of a gazetteer based matching and grammar-based recognition of basic entities of protein kinase in the text. Once annotated by usage of different gazetteer lists, the rules have been manually crafted in the GaBNER pipeline using JAPE grammar language to extend the coverage of entity recognition. The performance of the developed system has been evaluated on the prescribed metrics and shows a F-measure of 92%. The outputs from system could greatly help the researchers to focus specifically on relevant literature and thus saving huge amount of time and resources of theirs.

REFERNCES

1. Hunter, L., & Cohen, K. B. (2006). Biomedical language processing: perspective what's beyond PubMed?. *Molecular cell*, *21*(5), 589.
2. MEDLINE, the U.S. National Library of Medicine (NLM) premier bibliographic database. 30 July 2018, date last accessed]; Available from: *(https://www.nlm.nih.gov/bsd/medline.html)*
3. Antezana, E., Kuiper, M., & Mironov, V. (2009). Biological knowledge management: the emerging role of the Semantic Web technologies. Briefings in bioinformatics, 10(4), 392-407.
4. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., & Verspoor, K. (2014). Biomedical text mining: State-of-the-art, open problems and future challenges. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics* (pp. 271-300). Springer Berlin Heidelberg.
5. Chapman, Wendy W. Cohen, K. Bretonnel et al. (2009). Current issues in biomedical text mining and natural language processing. Journal of Biomedical Informatics, Volume 42, Issue 5, 757 – 759.
6. Zweigenbaum, Pierre and Demner-Fushman, Dina: Advanced Literature-Mining Tools. Springer New York. (2009) 347{380}.
7. Verspoor, K., Cohen, K. B., Mani, I., & Goertzel, B. (2006). Linking natural language processing and biology: towards deeper biological literature analysis. Brooklyn, New York: Association for Computational Linguistics, 3-5.
8. Grishman, R., & Sundheim, B. (1996, August). Message Understanding Conference-6: A Brief History. In *COLING* (Vol. 96, pp. 466-471).2q
9. Ide, N., & Suderman, K. (2009, August). Bridging the gaps: interoperability for GrAF, GATE, and UIMA. In *Proceedings of the Third Linguistic Annotation Workshop* (pp. 27-34). Association for Computational Linguistics.
10. Wilcock, G. (2009, September). Linguistic Processing Pipelines: Problems and Solutions. In *Book of Abstracts GSCL Workshop: Linguistic Processing Pipelines* (p. 7).
11. Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3-4), 327-348.
12. Cunningham H,Maynard D, Bontcheva K, Tablan V (2002) GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th anniversary meeting of the association for computational linguistics (ACL'02)
13. Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, *36*(2), 223-254.
14. Wilks, Y. GATE-a General Architecture for Text Engineering Hamish Cunningham.
15. Bank, M., & Schierle, M. (2012). A Survey of Text Mining Architectures and the UIMA Standard. In *LREC* (pp. 3479-3486).
16. Protein Kinases. 30 January 2018, date last accessed]; Available from: http://en.wikipedia.org/wiki/Protein_kinase.
17. Wikipedia. Protein Kinase Phosphorylation. 30 January 2018, date last accessed; Available from: http://en.wikipedia.org/wiki/Protein_kinase.
18. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., & Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, *298*(5600), 1912-1934.
19. Subramani, S., Jayapalan, S., Kalpana, R., & Natarajan, J. (2013). HomoKinase: a curated database of human protein kinases. ISRN Computational Biology, 2013.
20. *kinase.com*. 31 January 2018, date last accessed; Available from: http://kinase.com/about/Acknowledgements.html
21. Cunningham, H., Maynard, D., & Tablan, V. (1999). JAPE: a Java annotation patterns engine.