NTreceptorDB: a Database of Polymorphisms and Disease-Gene Associations in Behavioral Disorders

Aliyu Kabir Musa[1], Ekrem Varoğlu[1], Bahar Taneri[2]

[1] Department of Computer Engineering, Eastern Mediterranean University
Famagusta North Cyprus.
{Aliyukabir.Musa, Ekrem.Varoglu}@Emu.edu.tr

[2] Department of Biological Sciences, Eastern Mediterranean University
Famagusta North Cyprus.
Bahar.Taneri@Emu.edu.tr

Abstract. Genetic variation in neurotransmitter receptors have been shown to be associated with various behavioral and mental disorders. Presentation of disease-neurotransmitter receptor relationship in a comprehensive manner would prove useful for future experimental as well as computational work. In this study, we approach the relationship of neurotransmitter receptors to mental and behavioral disorders from a biomedical text mining perspective. To this extent, we collected an initial set of known neurotransmitter receptors and mental disorders, and built an association by automated literature and text mining methods using Support Vector Machines. NTreceptorDB, a database that enables users to analyze association between neurotransmitter receptor and mental disorder data is developed. Abstracts available in NTreceptorDB show biomedical evidence for neurotransmitter receptor-disease association and are linked to PubMed. In particular, NTreceptorDB covers specific polymorphisms in neurotransmitter receptor genes that are associated with diseases.

1 Introduction

Rapid accumulation of high-throughput biomedical data presents opportunities and at the same time challenges for data integration and interpretation. The main goal of the post genome era is to further elucidate the role of genetics in human health and diseases [1]. The current amount of biomedical literature regarding the identification of disease genes is rapidly increasing. One of the main challenges researchers in this domain face is that most of the relevant information is buried in the articles, in form of unstructured text. It is clear that text mining models are essential for handling large amount of information that is available only in unstructured textual form. In recent years, mining relations between genes and diseases in the text has become a major aim for researchers [2].

In this paper, we investigate the genetic variations in neurotransmitter receptors that are associated with certain behavioral disorders. This study focuses on finding the relationships between neurotransmitter receptors and behavioral disorders particularly from experimentally validated data reported in biomedical literature.

Genetic variation in neurotransmitter receptors have been shown to be implicated in behavioral variations across individuals in a given population and in various behavioral disorders [3]. Allelic variation in synaptic neurotransmission has shown to be implicated in various behavioral and neurological disorders including depression [4], alcoholism [5], drug dependence [5], and bipolar disorder [6]. Abnormalities in the production of or functioning of certain neurotransmitter receptors have been linked with a number of diseases. Imbalances in neurotransmission can result in depression, anxiety and other mood disorders [7]. Malfunctional neurotransmitter receptors, such as glutamate and dopamine receptors, have been shown to underlie major brain pathologies [8] [9].

To the best of our knowledge, to date the majority of the proposed biomedical systems does not focus particularly on the gene-disease relationship associated with neurotransmitter receptors and behavioral or mental disorders. Therefore, our work is unique in itself. For the first time, it provides a specific source on neurotransmitter receptors and their associated disease conditions.

2 Methods

We use state-of-the-art text mining methodologies in order to extract associations between neurotransmitter receptors and behavioral disorder from biomedical documents indexed in the NCBI's PubMed [10]. In this study, we develop and employ computational tools to detect neurotransmitter receptor-disease association on a large scale, from accumulating biomedical literature data. Figure 1 shows an overview of the text mining pipeline used in this study.

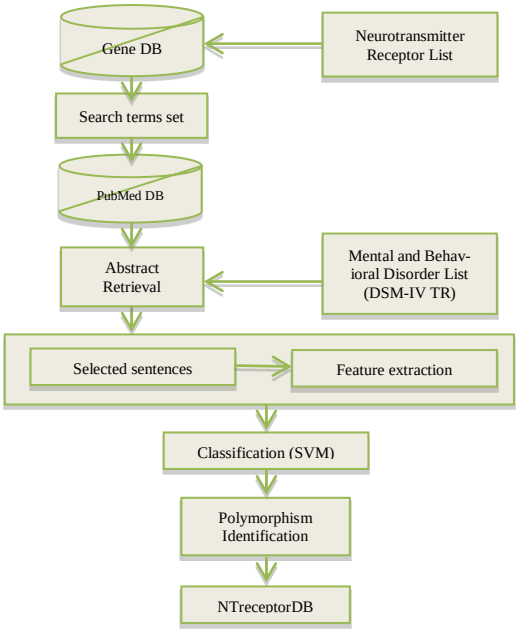


Fig. 1. An overview of Text Mining Pipeline.

2.1 Dataset Generation

In their 2008 article, Iwama and Gojobori published an extensive review on different categories of neurotransmitter receptors [11]. Building upon the original list provided in this article, we generate a comprehensive search term set for neurotransmitter receptors. A comprehensive search term set is generated for each neurotransmitter receptor provided in Iwama and Gojobori's list by using Gene DB of NCBI [12] since a gene may appear in an abstract by its symbol, name, and alias or even by its description.

Using the generated list, each term in the search term set generated for a particular gene is submitted to the PubMed database to find the PMIDs of abstracts associated with that gene.

In order to have a comprehensive list of mental and behavioral disorders we referred to the Diagnostic and Statistical Manual of Mental Disorders Text Revision (DSM-IV TR) [13]. This is a standard list used to describe and identify the types and thresholds of mental illness. We extract the list of diseases from DSM-IV TR and then we list the mental disorders that have been generally considered to be associated with a neurotransmitter receptor. Finally, we manually filtered the disorder list that we use as our search terms.

We identified candidate association sentences from the articles retrieved by choosing those sentences with a co-occurrence of a neurotransmitter receptor and a mental disorder. Our assumption is that a sentence that describes a relationship between a neurotransmitter receptor and mental disorder should contain at least one neurotransmitter receptor and at least one mental disorder.

2.2 Feature Extraction

2.2.1 Bag of Words (BOW)

Bag-of-words (BOW) feature extraction is the process of transforming what is essentially a list of words into a feature vector that can be utilized by a classifier. Many classifiers use a dictionary style feature where text is transformed into a form of dictionary. The existence of each word from a corpus in the dictionary is marked as a '1' in the feature vector, when the binary representation is used. Following the work of Mooney and Bunescu [14], in this study, 'all words between 2 entities', '3 words preceding the left entity' and '3 words following the right entity' of every sentence are used to form the BOWs. Here, each entity describes a neurotransmitter receptor or a mental disorder in the sentence. Fig.2 shows the bag-of-words feature we constructed for the sentence with PubMed ID: 20732371,

"*The **CCKB receptor** plays an important role in **anxiety** and gastric acid secretion*".

The words in the sentence between these entities are ‘receptor’, ‘plays’, ‘an’, ‘important’, ‘role’, and ‘in’. Among these words ‘receptor ‘and ‘in’ are not likely to directly suggest an association between neurotransmitter receptor CCKB and anxiety disorder but the phrase ‘plays an important role’ clearly shows the relationship between them. Thus, the words in the bag-of-words between this pair give sufficient information to identify their relationship. In addition, the left word ‘the’ and the right words ‘and’, ‘gastric’ and ‘acid’ are used in the BOW representation.

Left: “*The”*
Middle: “*receptor plays an important role in”*
Right: “*and gastric acid”*

Fig. 2. Example of BOW feature extraction from a sentence.

The motivation is based on the observation that the shortest path between the entities usually captures the necessary information to identify their relationship.

2.2.2 Association Words

Association words are often used as a domain specific feature in order to extract associations between entities [15]. Here, the assumption is that sentences containing interaction words are more likely to describe an association between the entities. A list of interaction words that consists of 30 verb root words was gathered from the retrieved articles [15]. The presence of any interaction words in a candidate sentence is marked as an entry in the feature vector representation.

2.2.3 Lexical Features

Grammatical functions of the words in the textual data are known as lexicons. The lexical feature used in this thesis is part-of-speech (POS) tags. POS tag of a word describes if it is a noun, adjective, preposition etc. in the sentence.

2.3 Classification Using SVMs

Support Vector Machines (SVM) is a supervised machine learning approach which has recently been used in many text classification and text mining problems including the biomedical domain [1]. In this study, SVM^{light} [16] is used throughout the experiments with a linear kernel setting. The feature vectors are composed of numerical values. The dimension of the feature vector is 12401 corresponding to all the word stems belonging to words represented in the BOWs of different sentences, the association words and POS tags feature.

The train and test data sets used in this thesis are constructed manually by annotating randomly selected sentences from the set of abstracts retrieved. The training set contains 570 annotated sentences with 479 positive and 91 negative sentences respectively. The test data on the other hand consist of 100 sentences with 55 positive and 45 negative samples. The performance of the proposed method measured using 3-fold cross validation has found to achieve a 91.25% precision and 98.59% recall, resulting in 94.78% F-score and 53.78% precision and 78.77% recall, resulting in 62.89% F-score on the test data. The difference in performance on the cross validation experiments and test data can be attributed to the lack of a "negation identification module". Presence of negative associations may be causing many false positives and we are currently working on the integration of this module to our text mining pipeline to solve this problem.

2.4 Polymorphism Identification

A rule-based tagger is used to identify polymorphisms in sentences. The following rules are applied by the tagger on every sentence in the test set that includes the keywords "polymorphism", "genotype", "allele", "SNP", "intron", "exon", or "mutation".

(i) Since many polymorphism events are described in text in the close vicinity of the keywords "polymorphism", or "genotype" check if the 3 preceding and following 1 word (window size of -3, +1) after the keywords "polymorphism", or "genotype" contain an alphanumeric string. The window size is decided experimentally by obtaining results on windows in the range -5, +5 and testing the results on a set of 150 random For a comprehensive set of experiments, all possible window size combinations in the min-max range has been tested.

(ii) Check the occurrence of slashes (/), dash (-), greater than (>) and brackets characters because most of the polymorphism notation in the sentences were observed to contain these characters.

(iii) Check for denotations of nucleotides and amino acids in the sentences (i.e. C267T and Ser9Gly).

(iv) Check for SNP denotations shown as "rs#" (i.e. rs7412).

(v) Check for alphanumeric symbols directly preceding or following the keyword "allele" (window size -1,+1).

(vi) Capture keywords "intron" and "exon" followed by a Arabic or Roman digit (i.e. intron 2, exon III) as well as references to position (i.e. second intron).

(vii) Use the initial list of neurotransmitters to post process and filter out the symbols that fall within our defined window size.

(viii) Concatenate single or capitalized characters found within the strings in order to solve blank space problem in polymorphism representation in the sentences.

The set of rules were tested on 150 sentences that contain at least one the keywords and an F-score of 84.08% was achieved.

3 Results

3.1 Major Findings

By applying text mining methods, we investigate the association between 1337 unique neurotransmitter receptor symbols and 465 unique mental and behavioral disorders. In total, we retrieved 835691 abstracts and analyzed 4642 sentences. These led to the identification of 1517 unique gene-disease association pairs. A brief overview of the sample associations from the database is provided in Table 1.

Table 1. Number of associations for specific neurotransmitter receptor-disease pairs.

	Schizophrenia	Anxiety	Alcohol Dependence
DRD3 receptor	124	8	2
5-HT1A receptor	93	21	16
MGLU5 receptor	32	55	-
MOR receptor	-	14	-

3.2 Polymorphisms and Difference in Disease Association

There has been conflicting evidence reported in the biomedical literature relevant to genetic variations in neurotransmitter receptors and their associations as susceptibility genes for certain diseases. It has been experimentally shown that different polymorphisms of a given gene could either be implicated in a disease state or could be irrelevant for that particular disease (Table 2). It is of particular importance to mine the specific polymorphisms in the neurotransmitter receptors genes yielding to association with diseases. Our text mining tools include a polymorphism identification module, which mines the specific genetic variations as illustrated in Table 3.

Table 2. Polymorphisms of neurotransmitter receptors and difference in disease association.

Sentences	PubMed ID
"Dopamine receptor D1 gene -48A/G polymorphism is associated with bipolar illness but not with schizophrenia in a Polish population."	249316

Table 3. Polymorphism sentences.

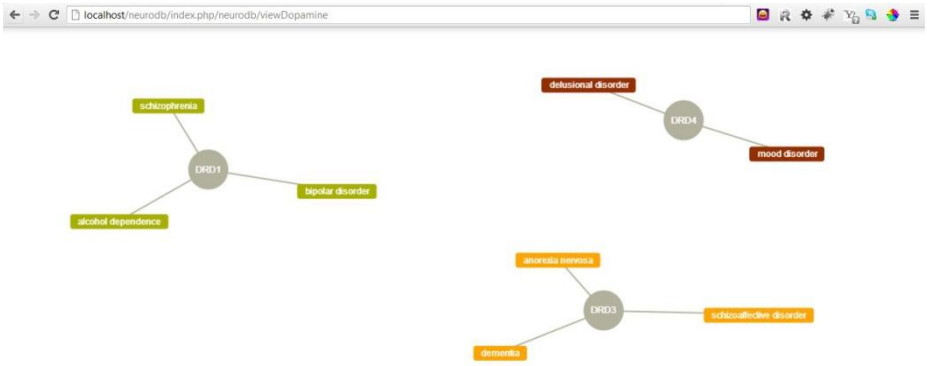
Sentences	PubMed ID
"An association between a polymorphism of the 5-HT receptor (5-HT2A) gene promoter (-1438G/A) and anorexia nervosa has been reported."	641576
"Dopamine receptor D1 gene -48A/G polymorphism is associated with bipolar illness but not with schizophrenia in a Polish population."	249316
"Association analysis between the C516T polymorphism in the 5-HT2A receptor gene and schizophrenia."	641576

3.3 NTReceptorDB Web Interface

In general, the experimental studies linking neurotransmission and various behavioral disorders are done on a single gene level for a single disorder or on a group of genes and proteins. Hence, biomedical data specific to this field is not present in a comprehensive, publicly accessible database. Our results from biomedical literature text mining presented in this study provide a centralized, comprehensive source documenting neurotransmitter receptors implicated in several diseases states.

Abstracts available in NTreceptorDB show biomedical evidence for neurotransmitter receptor-disease association and are linked to the PubMed database. Hence, NTreceptorDB serves as a public tool for analysis of the relationship between neurotransmitter receptors and mental or behavioral disorders. NTreceptor database was constructed containing neurotransmitter receptor-disease interaction data based on our biomedical literature work. NTReceptorDB is accessible through http://projects.emu.edu.tr/ NTreceptorDB.

The usefulness of the web-interface in NTreceptorDB is demonstrated below in Figure 3. This figure is a snapshot from NTreceptorDB showing the retrieval of the association of various behavioral diseases with different dopamine receptors, DRD1, DRD3 and DRD4. The legend in the figure shows the relationship that contains information between the two entities that is linked to the PubMed database. The information is also linked to the NCBI's Entrez Gene DB by using the Gene DB IDs in the search query for validation. Figure 4 shows a sample polymorphism output of NTreceptorDB. In particular in this example, users are able to visualize the 3 different polymorphisms within the same gene (5-HT2A), leading to three different diseases, namely alcohol abuse, anorexia nervosa and schizophrenia.



DRD1	DRD3	DRD4
schizophrenia	anorexia nervosa	mood disorder
alcohol dependence	Dementia	delusional disorder
bipolar disorder	schizoaffective disorder	-

Fig. 3. NTreceptorDB retrieval of disease data for a subset of dopamine receptors.

Fig. 4. NTreceptorDB: polymorphisms of 5-HT2A and their disease relationships.

4 Conclusion

We present a new approach to predict neurotransmitter receptor and mental disease associations based on literature data by employing text mining analysis. We retrieve a large number of relevant articles from PubMed and identify the disease associations of given neurotransmitter receptors. This work is unique in the sense that it focuses particularly on the neurotransmitter receptor-behavioral disorder association as opposed to the general gene-disease associations. This is the first dataset of its kind, specifically focusing on this group of genes and disorders.

Our results provide a centralized, comprehensive source documenting neurotransmitter receptors implicated in several diseases states. It is evident from the data that a given neurotransmitter receptor is implicated in several different diseases as illustrated in Figure 3. NTreceptorDB and the associated user friendly web-interface would enable storage of and access to the relevant neurotransmitter receptor–disease data, which is validated using the PubMed database. End users would be able to view annotations, search for biological data, validate links across resources, and create new information resources to capture new concepts as they arise. Main benefit of this work relies in the originality of the dataset and its comprehensive presentation in a publicly accessible platform, specifically providing molecular data on gene polymorphism level. These features would facilitate further research in the field on a large scale, in addition would provide healthcare professionals with a valuable biomedical source.

References

1. Ozgur A., Vu T., Erkan G., and Radev. D.R.: Identifying gene-disease associations using centrality on a literature mined gene interaction network. Bioinformatics, Volume 24, Number 13, pp. i277-i285, 2008.

2. Chun H. W. et al.: Extraction of gene-disease relations from Medline using domain dictionaries and machine learning, in Proc. the Pacific Symposium on Biocomputing, vol. 11, pp. 4-15, 2006.
3. Schosser A., Fuchs K., Scharl T.: Interaction between Serotonin 5-Ht2a Receptor Gene and Dopamine Transporter (Dat1) Gene Polymorphisms Influences Personality Trait of Persistence in Austrian, World Journal of Biological Psychiatry, Informahealthcare.com, Vol. 11, pp. 417-424, 2010.
4. Jokela M., Keltikangas-Jarvinen L.: Serotonin Receptor 2a Gene and the Influence of Childhood Maternal Nurturance on Adulthood Depressive Symptoms - Archives of General, Am Med Assoc, vol. 64, pp. 3, 2007.
5. Enoch M. A.: The Role of Gabaa Receptors in the Development of Alcoholism Pharmacology: Biochemistry and Behaviour, ELSEVIER, vol. 90, pp. 94-104, 2008.
6. Nakic M., Krystal J.H.: Neurotransmitter Systems in Bipolar Disor-der. Bipolar Disorder: Clinical Wiley Online Library, 2010.
7. Dalley J.W.: Dopamine Receptors in the Learning, Memory and Drug Reward Circuitry: Seminars in Cell and Developmental Biology, ELSEVIER, vol. 20, pp. 403-410, 2009.
8. Cha J. H., Kosinski C. M., Kerner J. A., Alsdorf S. A., Mangi-arini L., Davies S. W., Penney J. B., Bates G. P., Young A. B.: Altered Brain Neurotransmitter Receptors in Transgenic Mice Expressing a Portion of an Abnormal Human Huntington Disease Gene, PNAS, vol. 95, pp. 6480-6485, 1998.
9. Fool B. Le, Gallo A, Strat Y. Le, Lu L., Gorword L.: Genetics of Dopamine Receptors and Drug Addiction: A Comprehensive Review, Behavioural Pharmacology, vol. 20, pp. 1-17, 2009.
10. NCBIs PubMed: http://www.ncbi.nlm.nih.gov/pubmed.
11. Iwama H., Gojobori T.: Identification of Neurotransmitter Receptor Genes Under Significantly Relaxed Selective Constraint By Orthologous Gene Comparisons Between Humans And Rodents, Mol. Biol. E., vol. 19, pp. 1891-1901, 2008.
12. NCBI's GeneDB: http://www.ncbi.nlm.nih.gov/gene/.
13. American Psychiatric Association. Appendix I: Outline for cultural formulation and glossary of culturebound syndromes. In Diagnostic and statistical manual of mental disorders (4th ed., text rev.).
14. Mooney R. J., Bunescu R.,: Mining knowledge from text using information extraction, ACM SIGKDD Explorations Newsletter, v.7 n.1, p.3-10, 2005.
15. Bhardwaj N. and Lu H.: Correlation between gene expression profiles and protein protein interactions within and across genomes. Bioinformatics, 21(11):2730–2738, 2005.
16. Joachims T.: Advances in Kernel Methods-Support Vector Learning. Cambridge, MA, USA: MIT-Press; Making Large-Scale SVM Learning Practical, 1999.