Phylogeny, Predicton and Analysis of Sickle Cell Anaemia Using Data Mining Tools from Chhattisgarh State, India

Aakanksha Sharma and Dowluru Kaladhar

December 4, 2021

PHYLOGENY, PREDICTON AND ANALYSIS OF SICKLE CELL ANAEMIA USING DATA MINING TOOLS FROM CHHATTISGARH STATE, INDIA

Aakanksha Sharma[1,*] and DSVGK Kaladhar[2]

Department of Computational Biology, Atal Bihari Vajpayee University, Bilaspur (C.G.)

*Corresponding Author: Aakanksha Sharma; email:Aakanshasharma490@gmail.com; Phone: 7415408366

ABSTRACT

Sickle-Cell Anemia is a genetically inherited blood disorder that transmits millions of people around the world. A phylogenetic tree was constructed and the gene of Sickle cell Anemia from Human is found related to the *M. cynomolgus* Beta-globin. Sickle cell trait was now observing in some regions of Chhattisgarh state of India. Data collection from 156 patients in nine regions (Bilaspur, Pendra, Bilha, Sipat, Jairamnagar, Belgahana, Kota, Takhatpur and Mungeli) has been collected in September 2018. There are more number of Sickle cell people in Bilaspur(CG) followed by Pendra, Bilha, Sipat, Jairamnagar, Belgahana, Kota, Takhatpur and Mungeli. Based on Linear Regression analysis, Females, age with 34 and blood as major cells from pendra region is predicted test positive. Based on the data mining results of Sickle-Cell Anemia disease dataset using WEKA software, BayesNet and Adbaboost M1 classifier provides highest accuracy 80.52% and 80.12% respectively, compared with NaiveBayes, Bagging, J48, Random forest, Random tree and CART classifiers.

Key words: Sickle Cell Anemia, Data Mining, Chhattisgarh

INTRODUCTION

Sickle-Cell Anemia is an inherited blood disorder that is common among people from ancestors that are migrated and present in sub-Sahara Africa and Spain (Halberstein1999; Herrera and Garcia-Bertrand, 2018). About 2 million Americans that are belongs to sickle cell trait were carry genes to offsprings every year. This gene transfer may be about 3,000 times greater than the naturally occurring mutation rates that are calculated for man (Allison 1954; Jeremiah 2006). Some of the people from Chhattisgarh state in India are showing Sickle-Cell Anemia. A few of the symptoms that are caused by sickle-cell anemia include bone damage, eye damage, lung blockage, stroke, infections, and delayed growth (Serjeant 1997).

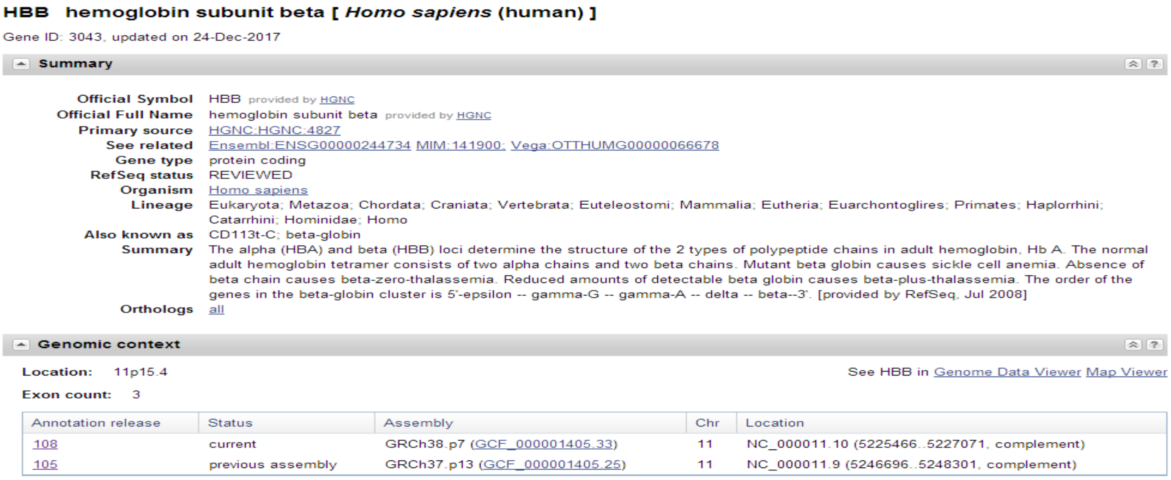
There is only temporary treatment for sickle cell anemia and permanent cure is not there. Basic treatment can be done heavily by taking upon pain killers and oral or intravenous fluids to reduce pain (Adams 2001). A recessive gene mutation change from glutamic acid to valine (GAG → GTG) at the sixth position on the 146 amino acid beta globin (HbB) of protein sequence located at the 15.5 region of chromosome 11 in haemoglobin formation leads to Sickle-Cell Anemia (Ashley-Koch et al., 2000).

Data mining (or data discovery) is the machine learning process of analyzing and predicting data collected by researchers in many fields. Data mining techniques are mainly applying in healthcare sectors gene expressions correlation studies, data collection and analysis, diagnosis and treatment predictions, etc (Tomar and Agarwal, 2013).

METHODOLOGY

Gene retrival and analysis

Complete genome (GenBank: NC_000011.10) was retrieved from the NCBI database.



HBB hemoglobin subunit beta [*Homo sapiens* (human)]

Gene ID: 3043, updated on 24-Dec-2017

Summary

Official Symbol	HBB provided by HGNC
Official Full Name	hemoglobin subunit beta provided by HGNC
Primary source	HGNC:HGNC:4827
See related	Ensembl:ENSG00000244734 MIM:141900; Vega:OTTHUMG00000066678
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	CD113t-C; beta-globin
Summary	The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'. [provided by RefSeq, Jul 2008]
Orthologs	all

Genomic context

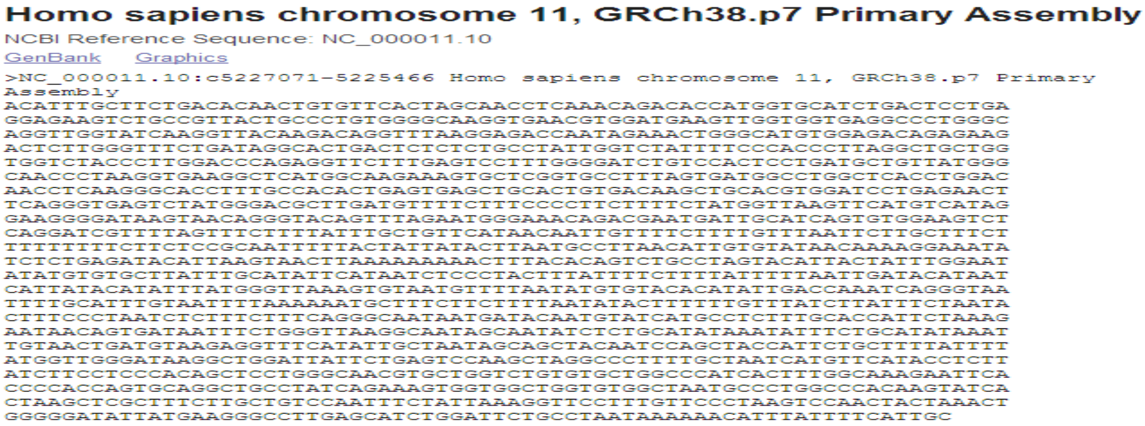
Location: 11p15.4

See HBB in Genome Data Viewer Map Viewer

Exon count: 3

Annotation release	Status	Assembly	Chr	Location
108	current	GRCh38.p7 (GCF_000001405.33)	11	NC_000011.10 (5225466..5227071, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (5246696..5248301, complement)

Figure 1: Summary of Retrieved sequence



Homo sapiens chromosome 11, GRCh38.p7 Primary Assembly

NCBI Reference Sequence: NC_000011.10

```
>NC_000011.10:c5227071-5225466 Homo sapiens chromosome 11, GRCh38.p7 Primary
Assembly
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGACAGAGAAG
ACTCTTGGGTTTCTGATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGG
TGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACT
TCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTTTTCTATGGTTAAGTTCATGTCATAG
GAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCT
CAGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTTGTTTAATTCTTGCTTTCT
TTTTTTTTCTTCTCCGCAATTTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAAAGGAAATA
TCTCTGAGATACATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACTATTTGGAAT
ATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTTAATTGATACATAAT
CATTATACATATTTATGGGTTAAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAA
TTTTGCATTTGTAATTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTTGTTTATCTTATTTCTAATA
CTTTCCCTAATCTCTTTCTTTCAGGGCAATAATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAG
AATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCTCTGCATATAAATATTTCTGCATATAAAT
TGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAGCTACAATCCAGCTACCATTCTGCTTTTATTTT
ATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACCTCTT
ATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGC
```

Figure 2: Retrieved sequence

Construction of Phylogenetic tree

The sequence has submitted to BLASTN and the related sequences are retrieved. The seqences are submitted to MEGA software and the phylogenetic tree has been constructed.

Data collection

Sickle-Cell Anemia Data collection from Sickle cell Institute,Genetic lab, Department of Biochemistry, Pt. J.N.M. Medical Collage,Raipur (Chhattisgarh).

Data Mining

Weka and Orange softwares are used to conduct analysis and predictions from the data collected that was related to Sickle-Cell Anemia.

RESULTS AND DISCUSSION

A phylogenetic tree was constructed and the gene of Sickle cell Anemia from Human is found related to the *M. cynomolgus* Beta-globin (Figure 3).

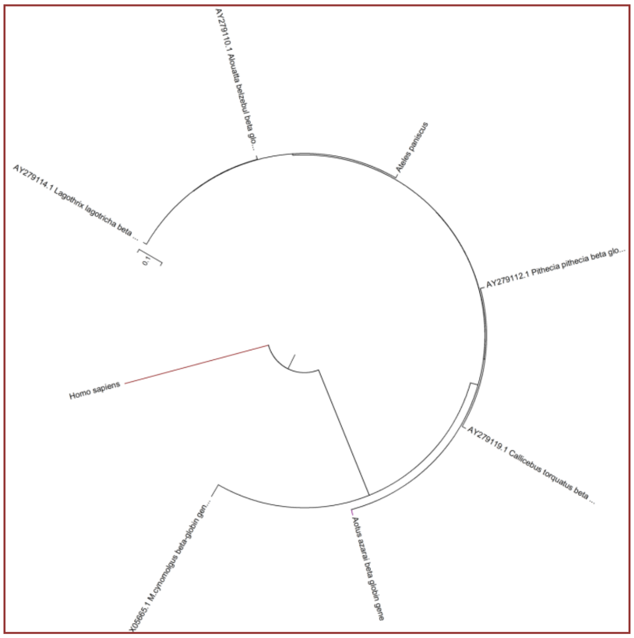
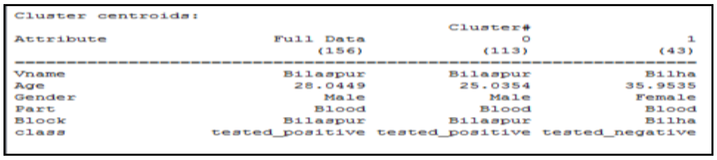

AY279110.1 Alouatta belzebul beta glo...
Ateles paniscus
AY279114.1 Lagothrix lagotricha beta ...
0.1
AY279112.1 Pithecia pithecia beta glo...
Homo sapiens
AY279119.1 Callicebus torquatus beta ...
X05665.1 M.cynomolgus beta-globin gen...
Aotus azarai beta globin gene

Figure 3: Phylogenetic tree using MEGA for HBB Gene

The data mining using Simple Means with cluster centroids shown that Bilaspur has more number of patients with Sickle cell anemia with 25 years, Male and Blood as tested Positive (Figure 4).



```
Cluster centroids:
                                              Cluster#
Attribute                    Full Data               0                1
                                 (156)           (113)             (43)
=========================================================================
Vname                         Bilaspur        Bilaspur            Bilha
Age                            28.0449         25.0354          35.9535
Gender                            Male            Male           Female
Part                             Blood           Blood            Blood
Block                         Bilaspur        Bilaspur            Bilha
class                  tested_positive tested_positive  tested_negative
```

Figure 4: Simple Means

The data distribution result and attribute statistics has been shown in Figure 5 and 6.

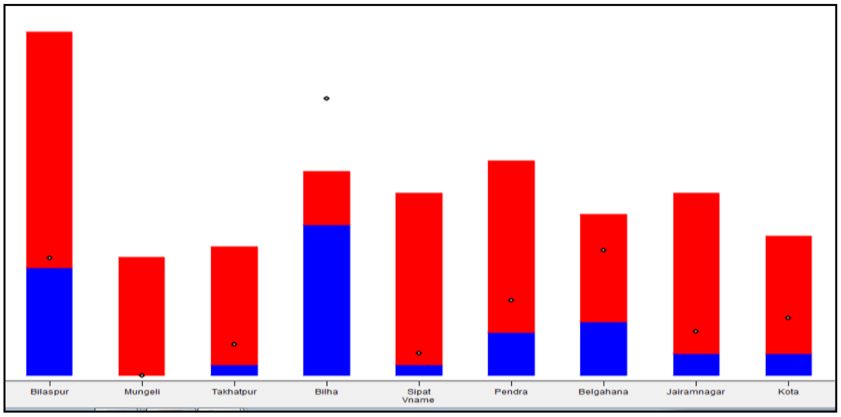
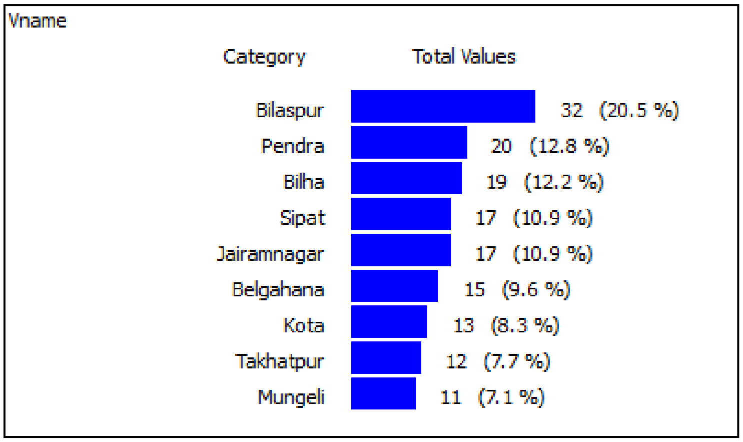


Figure 5: Data Distribution



Vname

Category	Total Values
Bilaspur	32 (20.5 %)
Pendra	20 (12.8 %)
Bilha	19 (12.2 %)
Sipat	17 (10.9 %)
Jairamnagar	17 (10.9 %)
Belgahana	15 (9.6 %)
Kota	13 (8.3 %)
Takhatpur	12 (7.7 %)
Mungeli	11 (7.1 %)

Figure 6: Attribute statistics

The linear regression result from orange software has shown in Figure 7. Orange software using sickle cell anemia data set and form distribution, linear regression, Attributes statistics and shown diseases highly infected region Bilaspur i.e., 32 and minimum in Mungeli i.e. 11

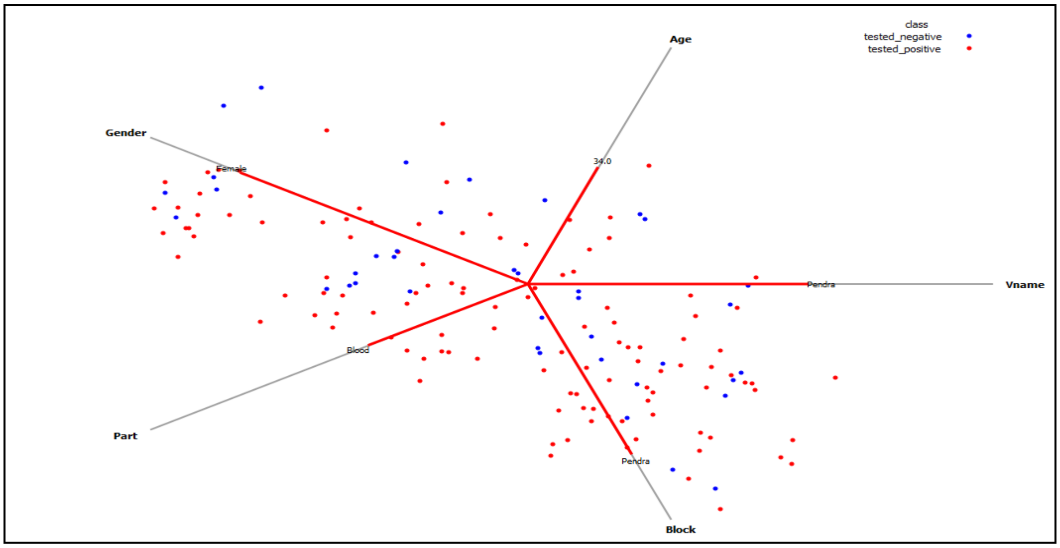
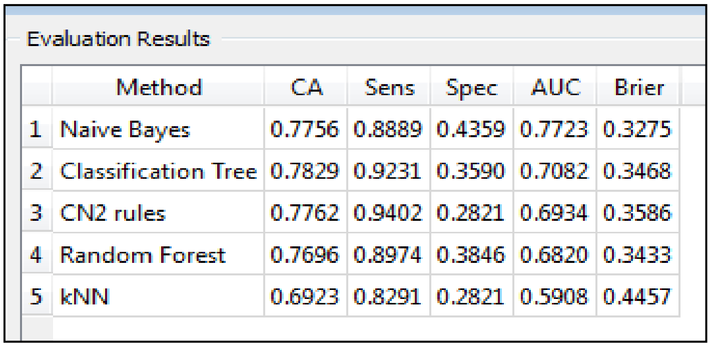


Figure 7: Orange software using Linear Regression Result

The data evaluation in orange software has shown good results with Classification Tree followed by CN2 rules and Naïve Bayes (Figure 8).



Evaluation Results

	Method	CA	Sens	Spec	AUC	Brier
1	Naive Bayes	0.7756	0.8889	0.4359	0.7723	0.3275
2	Classification Tree	0.7829	0.9231	0.3590	0.7082	0.3468
3	CN2 rules	0.7762	0.9402	0.2821	0.6934	0.3586
4	Random Forest	0.7696	0.8974	0.3846	0.6820	0.3433
5	kNN	0.6923	0.8291	0.2821	0.5908	0.4457

Figure 8: Orange software using Test Learner Result

The results obtained from weka with the given dataset classified into two classes i.e. patients with sickle cell anemia and without sickle cell Anemia using various data mining techniques (Table 1). The accuracy to predict the Sickle cell anemia disease using different techniques is shown in different figures. Based on the results demonstrated, Bays Net and Adbaboost M1 classifier provides highest accuracy 80.52% and 80.12% to predict the diseases. The performance of the algorithm is calculated using the equation for Total Accuracy and Random

Accuracy. Here, True positive and True Negative, False positive and False Negative parameters are taken to evaluate the equation and Random tree has72.42% shown lowest accuracy.

Table 1: Classification for Sickle cell anemia Dataset in WEKA software

Algorithm	**Correctly Classified**	**Time Taken(Seconds)**
Bays Net	**80.52%**	0.02 sec.
Naive Bayes	79.16%	0 sec.
Naive Bays Simple	79.84%	0.02 sec.
Naive Bays Updatable	79.84%	0 sec.
Adbaboost M1	**80.12%**	0.02 sec.
Bagging	79.41%	0.03 sec.
J48	75.64%	0.02 sec.
J48 Graft	75.64%	0.02 sec.
Random forest	74.25%	0.05 sec.
Random tree	72.42%	0 sec.
CART	76.92%	0.09 sec.
User Classifier	10 fold cross validation	12.96 sec.

CONCLUSION

In the medical field accuracy in prediction of datasets of the diseases of living systems is the most important factor. In the analysis of data mining techniques and tools Bays Net Classifier gives 99.87% of highest accuracy using WEKA tool. In future the sickle cell anemia can be prevented using gene analysis, machine learning methods and previous history of the diseases.

References

1. Halberstein, R. A. (1999). Blood pressure in the Caribbean. *Human biology*, *71*(4), 659.
2. Herrera, R. J., & Garcia-Bertrand, R. (2018). *Ancestral DNA, Human Origins, and Migrations*. Academic Press.

3. Allison, A. C. (1954). Protection afforded by sickle-cell trait against subtertian malarial infection. *British medical journal*, *1*(4857), 290.
4. Jeremiah, Z. A. (2006). Abnormal haemoglobin variants, ABO and Rh blood groups among student of African descent in Port Harcourt, Nigeria. *African health sciences*, *6*(3), 177-181.
5. Serjeant, G. R. (1997). Sickle-cell disease. *The Lancet*, *350*(9079), 725-730.
6. Adams, R. J. (2001). Stroke prevention and treatment in sickle cell disease. *Archives of neurology*, *58*(4), 565-568.
7. Ashley-Koch, A., Yang, Q., & Olney, R. S. (2000). Sickle hemoglobin (Hb S) allele and sickle cell disease: a HuGE review. *American journal of epidemiology*, *151*(9), 839-845.
8. Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241-266.