A Review of Vocabularies for a Subset of the Gastric Cancer Literature for Years 2020-2021

Jeff Cromwell, PhD

Mathematical Learning Space Research Portfolio Fall Semester 2021 : Vocabulary Review Series

To be Updated with Additional Vocabularies: November 2021

1 Abstract

Each year thousands of scientific articles on Gastric Cancer are published that represents the necessity of adding vocabulary terms to the gastric cancer researcher's rules of formation and separation. Here a brief review of the vocabulary is presented based on public available databases and a few algorithmic designs to assist the researcher in vocabulary formation and memorization based on N=8823 abstracts from the Pubmed database with 2307 genes and 82066 vocabulary words. Using algorithms in a recursive manner can provide additional segmentation and increase the vocabulary and grammar based on two and three term combinations with spatial reasoning, analogies and metaphors. A literature review based on the conclusions from the set of abstracts with the first two entries based on a 5 letter search term glyco was also presented with an algorithm. Here the algorithms and the tables together generate demonstrate potential opportunities for adding additional value in algorithm development and direction in gastric cancer research.

2 Introduction

In the study of the pathways for Gastric Cancer, the following two figures provide guidance into the list of genes. [1] [2] The first figure is the integrated cancer pathway for homo sapiens.

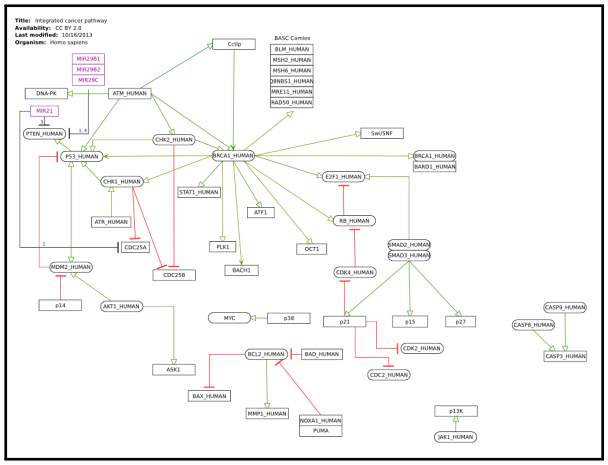

Figure 1: (a) Wang C, Bian Z, Wei D, Zhang JG; "miR-29b regulates migration of human breast cancer cells."; Mol Cell Biochem, 2011 PubMed Europe PMC Scholia. Gabriely G, Wurdinger T, Kesari S, Esau CC, Burchard J, Linsley PS, Krichevsky AM; "MicroRNA 21 promotes glioma invasion by targeting matrix metalloproteinase regulators."; Mol Cell Biol, 2008 PubMed Europe PMC Scholia. (b) Meng F, Henson R, Wehbe-Janek H, Ghoshal K, Jacob ST, Patel T; "MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer."; Gastroenterology, 2007 PubMed Europe PMC Scholia. (c) Zhao JJ, Lin J, Lwin T, Yang H, Guo J, Kong W, Dessureault S, Moscinski LC, Rezania D, Dalton WS, Sotomayor E, Tao J, Cheng JQ; "microRNA expression profile and identification of miR-29 as a prognostic marker and pathogenetic factor by targeting CDK6 in mantle cell lymphoma."; Blood, 2010 PubMed Europe PMC Scholia [1]

Table 1 has the RNA, protein and gene product with ID Symbol Category Name and Misc for Figure 1.

*The Mathematical Learning Space Research Portfolio
Email address: http://mathlearningspace.weebly.com/ (Jeff Cromwell, PhD)

ID	Symbol	Category	Name	Misc
1	AKT1_HUMAN	Protein	P31749 (Uniprot-TrEMBL)	
2	ASK1	Protein	4217 (Entrez Gene)	
3	ATF1	GeneProduct	ENSG00000123268 (Ensembl)	
4	ATM_HUMAN	Protein	Q13315 (Uniprot-TrEMBL)	
5	ATR_HUMAN	Protein	Q13535 (Uniprot-TrEMBL)	
6	BACH1	GeneProduct	ENSG00000156273 (Ensembl)	
7	BAD_HUMAN	Protein	Q92934 (Uniprot-TrEMBL)	
8	BARD1_HUMAN	Protein	Q99728 (Uniprot-TrEMBL)	
9	BAX_HUMAN	Protein	Q07812 (Uniprot-TrEMBL)	
10	BCL2_HUMAN	Protein	P10415 (Uniprot-TrEMBL)	
11	BLM_HUMAN	Protein	P54132 (Uniprot-TrEMBL)	
12	BRCA1_HUMAN	Protein	P38398 (Uniprot-TrEMBL)	Phosphorylated
13	CASP3_HUMAN	Protein	P42574 (Uniprot-TrEMBL)	
14	CASP8_HUMAN	Protein	Q14790 (Uniprot-TrEMBL)	
15	CASP9_HUMAN	Protein	P55211 (Uniprot-TrEMBL)	
16	CDC25A	GeneProduct	ENSG00000164045 (Ensembl)	
17	CDC25B	GeneProduct	ENSG00000101224 (Ensembl)	
18	CDC2_HUMAN	Protein	P06493 (Uniprot-TrEMBL)	Oncogene
19	CDK2_HUMAN	Protein	P24941 (Uniprot-TrEMBL)	
20	CDK4_HUMAN	Protein	P11802 (Uniprot-TrEMBL)	
21	CHK1_HUMAN	Protein	O14757 (Uniprot-TrEMBL)	
22	CHK2_HUMAN	Protein	O96017 (Uniprot-TrEMBL)	Phosphorylated
23	Cctlp	Protein		
24	DNA-PK	Metabolite		
25	E2F1_HUMAN	Protein	Q01094 (Uniprot-TrEMBL)	
26	JAK1_HUMAN	Protein	P23458 (Uniprot-TrEMBL)	
27	MDM2_HUMAN	Protein	Q00987 (Uniprot-TrEMBL)	Oncogene
28	MIR21	Rna	ENSG00000284190 (Ensembl)	
29	MIR29B1	Rna	ENSG00000283797 (Ensembl)	
30	MIR29B2	Rna	ENSG00000284203 (Ensembl)	
31	MIR29C	Rna	ENSG00000284214 (Ensembl)	
32	MMP1_HUMAN	Protein	P03956 (Uniprot-TrEMBL)	
33	MRE11_HUMAN	Protein	P49959 (Uniprot-TrEMBL)	
34	MSH2_HUMAN	Protein	P43246 (Uniprot-TrEMBL)	
35	MSH6_HUMAN	Protein	P52701 (Uniprot-TrEMBL)	
36	MYC	GeneProduct	ENSG00000136997 (Ensembl)	
37	NOXA1_HUMAN	Protein	Q86UR1 (Uniprot-TrEMBL)	
38	OCT1	Protein	5451 (Entrez Gene)	
39	P53_HUMAN	Protein	P04637 (Uniprot-TrEMBL)	
40	PLK1	GeneProduct	ENSG00000166851 (Ensembl)	
41	PTEN_HUMAN	Protein	P60484 (Uniprot-TrEMBL)	
42	PUMA	Protein	27113 (Entrez Gene)	
43	Q8NBS1_HUMAN	Protein	Q8NBS1 (Uniprot-TrEMBL)	
44	RAD50_HUMAN	Protein	Q92878 (Uniprot-TrEMBL)	
45	RB_HUMAN	Protein	P06400 (Uniprot-TrEMBL)	
46	SMAD2_HUMAN	Protein	Q15796 (Uniprot-TrEMBL)	Tumor supressor
47	SMAD3_HUMAN	Protein	P84022 (Uniprot-TrEMBL)	
48	STAT1_HUMAN	Protein	P42224 (Uniprot-TrEMBL)	
49	Swi/SNF	Protein		
50	p13K	Protein		
51	p14	Protein		
52	p15	Protein	1030 (Entrez Gene)	
53	p21	Protein	1026 (Entrez Gene)	
54	p27	Protein	1027 (Entrez Gene)	
55	p38	Protein		

Table 1

Figure 2 has the Wikipathway for Gastric Cancer Homo Sapiens. [1]

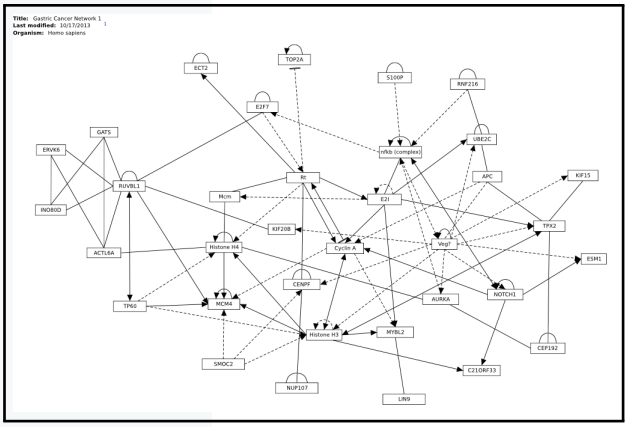

Figure 2: (a) Fan B, Dachrut S, Coral H, Yuen ST, Chu KM, Law S, Zhang L, Ji J, Leung SY, Chen X; "Integration of DNA copy number alterations and transcriptional expression analysis in human gastric cancer."; PLoS One, 2012 PubMed Europe PMC Scholia [1]

Table 2 has list of gene product with the ID Symbol Category Name and Misc for Figure 2.

ID	Symbol	Category	Name	Misc
1	ACTL6A	GeneProduct	86 (Entrez Gene)	
2	APC	GeneProduct	324 (Entrez Gene)	
3	AURKA	GeneProduct	6790 (Entrez Gene)	
4	C21ORF33	GeneProduct	8209 (Entrez Gene)	
5	CENPF	GeneProduct	1063 (Entrez Gene)	
6	CEP192	GeneProduct	55125 (Entrez Gene)	
7	Cyclin A	GeneProduct	8900 (Entrez Gene)	
8	E2F7	GeneProduct	144455 (Entrez Gene)	
9	E2I	GeneProduct		
10	ECT2	GeneProduct	1894 (Entrez Gene)	
11	ERVK6	GeneProduct	64006 (Entrez Gene)	
12	ESM1	GeneProduct	11082 (Entrez Gene)	
13	GATS	GeneProduct	352954 (Entrez Gene)	
14	Histone H3	GeneProduct	3020 (Entrez Gene)	
15	Histone H4	GeneProduct	121504 (Entrez Gene)	
16	INO80D	GeneProduct	54891 (Entrez Gene)	
17	KIF15	GeneProduct	56992 (Entrez Gene)	
18	KIF20B	GeneProduct	9585 (Entrez Gene)	
19	LIN9	GeneProduct	286826 (Entrez Gene)	
20	MCM4	GeneProduct	4173 (Entrez Gene)	
21	MYBL2	GeneProduct	4605 (Entrez Gene)	
22	Mcm	GeneProduct		
23	NOTCH1	GeneProduct	4851 (Entrez Gene)	
24	NUP107	GeneProduct	57122 (Entrez Gene)	
25	RNF216	GeneProduct	54476 (Entrez Gene)	
26	RUVBL1	GeneProduct	8607 (Entrez Gene)	
27	Rt	GeneProduct		
28	S100P	GeneProduct	6286 (Entrez Gene)	
29	SMOC2	GeneProduct	64094 (Entrez Gene)	
30	TOP2A	GeneProduct	7153 (Entrez Gene)	
31	TP60	GeneProduct		
32	TPX2	GeneProduct	22974 (Entrez Gene)	
33	UBE2C	GeneProduct	11065 (Entrez Gene)	
34	Veg	GeneProduct		
35	nfkb (complex)	GeneProduct		

Table 2

Based on the R programming language [1000] and APIs with pubmed [1001] and algorithms developed with grep and regular expressions, a review of vocabulary was performed to gain insight into vocabulary formation and possibilities in research direction.

2.1. Algorithms

Algorithms 1, 2 and 3 inputs are based on the PubMed abstract and title files with a txt and csv extensions and wikipathways files as well as keywords and table numbers. The output of each is translated into a TeX format (i.e. table tags) for the pdf and png graphics format. [?800][?801]

```
algorithm.1<-function(x,y,keywords)
{
  require(pubmed.mineR)
  L=readabs(x)
  L2 = cleanabs(L)
  L3=gene_atomization(L)
  L4=word_atomizations(L)
  L5=as.data.frame(read.csv(y))
  n<-length(keywords)
  for(i in 1:n)
  {
    keys<-grep(keywords[i],as.character(L5[,2]))
    L6<-NULL;L7<-NULL
    for(j in 1:length(keys)){
      L6[j]<- stringr::str_c(as.character(L5[keys[j],2]), "
            \\","\\","cite{key",j,"}")
      L7[j]<- stringr::str_c("\\","\\","bibitem[",j,"]{key",j,"}
            ",as.character(L5[keys[j],2]))


    }
  }
    L8<-unlist(strsplit(as.character(L5[,2])," "))
  return(list(L3,L4,L5,L6,L7,L8))

}
algorithm.2(url,k)
{
URL1 <- url
result <- URL1 %>%
  html %>%
  html_nodes("table")
Genes<-html_table(result[k])
return(Genes)

}
algorithm.3<-function(x,y,keyword1)
{

ABS<-pubmed.mineR::readabs(x)
GC.Conclusions<-Find_conclusion(ABS)
GC.Conclusions.keys<-grep(keyword1,GC.Conclusions)
L5=as.data.frame(read.csv(y))
Intro<-NULL; Bib<-NULL
for(i in 1:length(GC.Conclusions))
{

  Intro[i]<-stringr::str_c(GC.Conclusions[GC.Conclusions.keys][[i]][1]," ",
                           GC.Conclusions[GC.Conclusions.keys][[i]][2]," ",
                           "\\","cite{key",i,"} ", "\\","\\")

  Bib[i]<-stringr::str_c("\\","bibitem[",i,"]","{key",i,"}",
  as.character(L5[GC.Conclusions.keys[i],2])," \\","\\")
```

```
}

write(Intro,file=stringr::str_c("Literature_Review_",keyword1,".txt"))
write(Bib,file=stringr::str_c("Bib_",keyword1,".txt"))
}
```

3 Vocabulary Review

Let X be a NxK matrix such that N is the number of articles in a database and K is the number of features of each article in the the database. This matrix will be referred to as a data table where the columns are tableware with food and drink and the rows the seats at the table. Any subset of X such as $V_{n,k}$ is a view of the table with $k \leq K$ and $n \leq N$. This is common terminology in database design and used throughout the text.

Consider a set of N=8823 abstracts from the Pubmed database [2] with 2307 genes and 82066 vocabulary words. Table 1 has an example of the top 100 Genes based on Gene Symbol, Name, and frequency.

3.1. Gene Vocabulary

Table 3, 4, 5, 6 and 7 have the Gen Symbol, Name and Frequencies based on greater than 20 in frequency with a 125 genes in total.

	Gene_symbol	Genes	Freq
1	GC	group-specific component (vitamin D binding protein)	8858
2	T	T, brachyury homolog (mouse)	726
3	ESD	esterase D	330
4	HR	hair growth associated	206
5	TP53	tumor protein p53	158
6	CDH1	cadherin 1, type 1, E-cadherin (epithelial)	142
7	SRC	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)	126
8	HP	haptoglobin	117
9	STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	105
10	EGFR	epidermal growth factor receptor	97
11	PC	pyruvate carboxylase	96
12	MET	met proto-oncogene (hepatocyte growth factor receptor)	94
13	NHS	Nance-Horan syndrome (congenital cataracts and dental anomalies)	89
14	TG	thyroglobulin	80
15	PTEN	phosphatase and tensin homolog	78
16	KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	73
17	CXCR4	chemokine (C-X-C motif) receptor 4	71
18	APC	adenomatous polyposis coli	70
19	FGFR2	fibroblast growth factor receptor 2	59
20	ARID1A	AT rich interactive domain 1A (SWI-like)	56
21	CD44	CD44 molecule (Indian blood group)	53
22	ADC	arginine decarboxylase	51
23	AFP	alpha-fetoprotein	51
24	PGC	progastricsin (pepsinogen C)	50
25	ERAS	ES cell expressed Ras	49

Table 3

	Gene_symbol	Genes	Freq
1	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	48
2	PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha	47
3	THBS1	thrombospondin 1	46
4	EZH2	enhancer of zeste homolog 2 (Drosophila)	45
5	MUC1	mucin 1, cell surface associated	45
6	AR	androgen receptor	44
7	NRP1	neuropilin 1	44
8	YAP1	Yes-associated protein 1	43
9	HMGB1	high mobility group box 1	39
10	GIP	gastric inhibitory polypeptide	38
11	KLF5	Kruppel-like factor 5 (intestinal)	38
12	TRG	T cell receptor gamma locus	38
13	GCA	grancalcin, EF-hand calcium binding protein	37
14	ATAD2	ATPase family, AAA domain containing 2	36
15	CRP	C-reactive protein, pentraxin-related	35
16	E2F1	E2F transcription factor 1	35
17	FOXP3	forkhead box P3	35
18	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	35
19	CLDN18	claudin 18	34
20	FAP	fibroblast activation protein, alpha	34
21	SALL4	sal-like 4 (Drosophila)	34
22	SOX9	SRY (sex determining region Y)-box 9	34
23	ATM	ataxia telangiectasia mutated	33
24	HGF	hepatocyte growth factor (hepapoietin A; scatter factor)	33
25	LGR5	leucine-rich repeat containing G protein-coupled receptor 5	33

Table 4

As an example of further exploration, consider number 25 LGR5 a leucine-rich repeat containing G protein-coupled receptor 5 with a frequency of 33 from 228 abstracts from 2020-2021 presented in Table 5.

	Gene_symbol	Genes	Freq
1	LGR5	leucine-rich repeat containing G protein-coupled receptor 5	356
2	RSPO2	R-spondin 2	24
3	LGR4	leucine-rich repeat containing G protein-coupled receptor 4	22
4	OLFM4	olfactomedin 4	20
5	LGR6	leucine-rich repeat containing G protein-coupled receptor 6	14
6	CD151	CD151 molecule (Raph blood group)	10
7	NLRP3	NLR family, pyrin domain containing 3	10
8	BMI1	BMI1 polycomb ring finger oncogene	9
9	CD44	CD44 molecule (Indian blood group)	9
10	EGF	epidermal growth factor	9
11	RNF43	ring finger protein 43	9
12	CD24	CD24 molecule	8
13	YAP1	Yes-associated protein 1	8
14	FAP	fibroblast activation protein, alpha	7
15	METTL3	methyltransferase like 3	7
16	RSPO1	R-spondin 1	7
17	ZNF25	zinc finger protein 25	7
18	AGR2	anterior gradient 2 homolog (Xenopus laevis)	6
19	ANKRD22	ankyrin repeat domain 22	6
20	DCC	deleted in colorectal carcinoma	6
21	GC	group-specific component (vitamin D binding protein)	6
22	GREM1	gremlin 1, DAN family BMP antagonist	6
23	SOCS2	suppressor of cytokine signaling 2	6
24	TNC	tenascin C	6
25	WWOX	WW domain containing oxidoreductase	6

Table 5

	Gene_symbol	Genes	Freq
1	MLH1	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	33
2	CDX2	caudal type homeobox 2	32
3	BRCA2	breast cancer 2, early onset	31
4	COL1A1	collagen, type I, alpha 1	30
5	MUC6	mucin 6, oligomeric mucus/gel-forming	30
6	TFF1	trefoil factor 1	30
7	COL10A1	collagen, type X, alpha 1	29
8	HMGA1	high mobility group AT-hook 1	29
9	MMP9	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	29
10	PSCA	prostate stem cell antigen	29
11	RNF43	ring finger protein 43	29
12	TC2N	tandem C2 domains, nuclear	29
13	FH	fumarate hydratase	28
14	FOXK1	forkhead box K1	28
15	MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	28
16	ZEB1	zinc finger E-box binding homeobox 1	28
17	METTL3	methyltransferase like 3	26
18	VCAN	versican	26
19	COL1A2	collagen, type I, alpha 2	25
20	PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)	25
21	BCL2	B-cell CLL/lymphoma 2	24
22	COL5A2	collagen, type V, alpha 2	24
23	E2F2	E2F transcription factor 2	24
24	ENO1	enolase 1, (alpha)	24
25	FN1	fibronectin 1	24

Table 6

	Gene_symbol	Genes	Freq
1	GPX8	glutathione peroxidase 8 (putative)	24
2	BRCA1	breast cancer 1, early onset	23
3	C3	complement component 3	23
4	S100A16	S100 calcium binding protein A16	23
5	ADAMTS12	ADAM metallopeptidase with thrombospondin type 1 motif, 12	22
6	ANTXR1	anthrax toxin receptor 1	22
7	ARL4C	ADP-ribosylation factor-like 4C	22
8	CD4	CD4 molecule	22
9	CS	citrate synthase	22
10	PRRX1	paired related homeobox 1	22
11	ASPN	asporin	21
12	C2	complement component 2	21
13	CMTM6	CKLF-like MARVEL transmembrane domain containing 6	21
14	FGF2	fibroblast growth factor 2 (basic)	21
15	HGD	homogentisate 1,2-dioxygenase	21
16	KIF23	kinesin family member 23	21
17	LAMA4	laminin, alpha 4	21
18	LATS2	LATS, large tumor suppressor, homolog 2 (Drosophila)	21
19	MDM2	MDM2 oncogene, E3 ubiquitin protein ligase	21
20	PGR	progesterone receptor	21
21	PTX3	pentraxin 3, long	21
22	RUNX3	runt-related transcription factor 3	21
23	TFF2	trefoil factor 2	21
24	THBS2	thrombospondin 2	21
25	TRIM37	tripartite motif containing 37	21

Table 7

In the gene vocabulary, Table 8 has the index for each of the genes based on the first letter in the sequence.

N/L	Freq
a	134
A	41
b	28
B	30
c	257
C	57
d	46
D	23
e	54
E	25
f	86
F	20
g	78
G	23
h	64
H	7
i	73
I	4
j	5
J	3
k	40
K	18
l	59
L	8
m	131
M	10
n	73
N	32
o	13
O	2
p	221
P	20
Q	1
r	53
R	47
s	162
S	59
t	133
T	30
u	25
U	6
v	43
V	1
w	3
W	11
x	3
X	6
Y	5
z	23
Z	2

Table 8

Table 9 has the filter with the single letter a for the gene vocabulary.

1	adenomatous polyposis coli	arginine decarboxylase
2	alpha-fetoprotein	androgen receptor
3	ataxia telangiectasia mutated	anthrax toxin receptor 1
4	asporin	activating transcription factor 3
5	aurora kinase A	angiotensin I converting enzyme (peptidyl-dipeptidase A) 2
6	anaplastic lymphoma receptor tyrosine kinase	aldehyde dehydrogenase 2 family (mitochondrial)
7	aurora kinase B	alpha thalassemia/mental retardation syndrome X-linked
8	acyl-CoA thioesterase 7	apolipoprotein C-I
9	acyl-CoA binding domain containing 3	adenosine deaminase, RNA-specific
10	acylglycerol kinase	alanine-glyoxylate aminotransferase 2-like 1
11	annexin A2	aquaporin 9
12	autophagy related 7	activin A receptor, type IIA
13	aryl hydrocarbon receptor interacting protein	angiomotin like 1
14	anoctamin 9	apolipoprotein C-II
15	absent in melanoma 2	aldo-keto reductase family 1, member C3
16	autophagy related 12	axin 2
17	angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	aryl hydrocarbon receptor
18	aldo-keto reductase family 1, member B10 (aldose reductase)	aldehyde dehydrogenase 3 family, member A2
19	angiopoietin 2	anillin, actin binding protein
20	amyloid beta (A4) precursor protein-binding, family B, member 2	argininosuccinate lyase
21	acid phosphatase 5, tartrate resistant	acyl-CoA synthetase long-chain family member 5
22	adenylate cyclase 3	anoctamin 1, calcium activated chloride channel
23	acidic (leucine-rich) nuclear phosphoprotein 32 family, member E	aquaporin 5
24	activating transcription factor 4 (tax-responsive enhancer element B67)	alkB, alkylation repair homolog 1 (E. coli)
25	annexin A4	aquaporin 1 (Colton blood group)

Table 9

4 English Word Vocabulary

The same Algorithm was applied to the vocabulary with word ID , Name and Frequency for the N abstracts with the expected results of the two word keyword filter Gastric Cancer applied for Table 10.

	Words	Freq
32627	cancer	21121
80624	was	18735
44775	gastric	15853
80823	were	15828
64395	patients	14836
44985	gc	8858
42571	expression	7500
79309	university	6656
33746	cells	6438
15285	0	6236
33679	cell	5955
70107	results	5585
74969	study	5226
78411	tumor	4938
48120	hospital	4865
75492	survival	4670
16442	1	4201
28323	analysis	4145
31008	between	4033
77970	treatment	3990
34333	china	3741
56268	medical	3724
75435	surgery	3678
35176	clinical	3589
46180	group	3528

Table 10

Based on the 82066 words, a three letter sequence bio was used to filter with a return of 355 words with a subset presented in Table 11.

1	biomarkers	biological	biomarker	biology
2	biopsy	bioinformatics	biomedical	biopsies
3	microbiota	biochemistry	antibiotic	microbiology
4	bioinformatic	microbiome	biotechnology	antibiotics
5	biostatistics	cbioportal	biochemical	biotherapy
6	bioactive	dysbiosis	biopharma	biobank
7	biosynthesis	biologic	biomedicine	biosciences
8	probiotics	bioavailability	biodistribution	biologically
9	probiotic	biosimilars	biomdica	biocompatibility
10	biosimilar	bio	bioengineering	biomaterials
11	biogenesis	biopsied	bioelectrical	biologics
12	bioactivity	bioavailable	biomolecular	biopharm
13	biosci	bioscience	biotech	microbiological
14	mrbiom	bio-medico	biomâ©dicas	biomass
15	radiobiology	(ibiomed)	biocompatible	bioinformation
16	biolâ³gicas	bioluminescence	bioreactor	biosanitaria
17	biotechnologies	fisabio-isabial	microbiology/key	nanobiotix
18	pathobiology	bioactivities	biocartis	bioenergetics
19	biofilm	biologia	biomedicas	biometrics
20	biomimetic	bionanocomposites	biontech	bioprinting
21	biopsy-proven	biotec	dysbiotic	jacobio
22	non-antibiotic	xenobiotics	(5)biostatistics	antibiotic-treated
23	apobiologix	biodonostia	bioengineered	biofluid
24	biofluids	biokinetic	biol	biologãa
25	biologie	biomolecules	biophysics	biopolymers

Table 11

Based on the 82066 words, a two letter sequence ep was used to filter with a return of 1311 words with a subset presented in Table 12.

1	epiblast	epic	epicardiectomy	epicardium
2	epicentre	epicertin	epicure	epidemiol
3	epidemiologists	epidemiologists)	epidermis	epigenomically
4	epigenomics	epimers	epimutations	episodios
5	episome	epistemonikos	epiteliais	epitelial
6	epithelial-	epithelial-cadherin	epithelial-driven	epithelial-myofibroblast
7	epithelial-stromal	epithelial/mesenchymal	epithelia mesenchymal	epithelialization
8	epithelioid/mixed	epitheliotropic	epithelium-either	epithetical
9	epiunit	epiunit-instituto	epiya-a	epiya-bc
10	epiya-d)	epiya-motifs	epizyme	epk
11	epoc1201	epochs	epoetin-alfa	epoxylathyrane
12	epp85-181rdb	epp85-181rnov	epr	epsilometer
13	epsteinbarr	etanercept	extra-hepatic	factor-dependent
14	farnesoid-x-receptor)	fc-epsilon	epithelial	fgfr4-dependent
15	depressed	folate-dependent	forceps-assisted	foxm1-dependent
16	gadd45a-dependent	gagnep	gaplinc-depleted	gastro-epiploic
17	gastro-hepatic	gastroentero-hepatologist	gastroentero-hepatology	gender-independent
18	gene-epigenetics	geo-epidemiological	gep-nec	gfral/ret-dependent
19	glycoepitope	glycoepitopes	glycolysis-dependent	glycopeptide
20	glycopeptidomics	gradepro	h1-receptor	hemihepatectomy
21	hep-g2	hepâ¡tico	heparin-binding	hepatic/biliary
22	hepaticogastrostomy	hepaticojejunal	hepatopancreatobiliaryâ surgery	hepatocellular)
23	hepatoduodenale	hepatogastric	hepatokine	hepatol
24	hepatoma-derived	hepatomaderived	hepatomesenteric	hepatopancreatoduodenectom
25	hepatoprotection	hepatorenal	hepatotoxity	heppar1

Table 12

5 Literature Review

As example, the keyword with the 5 letter glyco was used to generate the following subset N=86 titles for citation and the biblography with the first two sentences in the conclusions a subset of the abstracts. Three titles from that subset as an example are:

CONCLUSIONS: FOXO4 has an important role in the regulation of glycolysis in Gastric Cancer with disruption of the HIF-1 α -FOXO4-LDHA axis and demonstrates promise as a therapeutic strategy for Gastric Cancer. [50]
CONCLUSIONS: Cancer cells with a predominant glycolytic pathway, metabolomic analyses(i,e. hypoxic conditions) enable the global metabolism profiling and inhibiting the supply of nicotinamide adenine dinucleotide phosphate by glycolysis blocking as a potential treatment strategy for cancer as well as cystine blockade by salazosulfapyridine.[54]
CONCLUSIONS: Safranal, a plant secondary metabolite of saffron, has promising pharmacological properties for the management of Alzheimer's disease. The druglike attributes of safranal, (a) Lipinski's rule of five; (b) optimum lipophilicity; (c) high permeability; (d) low blood-to-plasma ratio; (e) less to moderate propensity to interact with P-glycoprotein (P-gp) or breast cancer-resistant protein (BCRP) transporters and (f) high plasma protein binding as common to vitro and ex vivo marketed drug models. [83]

[50] Hypoxia-induced FOXO4/LDHA axis modulates gastric cancer cell glycolysis and progression

[54] Metabolic Profiling of Human Gastric Cancer Cells Treated With Salazosulfapyridine

[83] Description of Druglike Properties of Safranal and Its Chemistry behind Low Oral Exposure

6 Conclusions

In this brief review of the development of a gastric cancer vocabulary, a set of N=8823 abstracts from the Pubmed database with 2307 genes and 82066 vocabulary words was examined with an algorithm that generated a report from a set of views (e.g. 11 in total) from a single table . This report of a collection of gene views with an overall frequency of greater than 20 separated into groups of 25 for possible future analysis as well as tables of a three and two letter match making pattern for building tables of words for vocabulary building and repeating the same pattern and different levels of inquiry. A literature review based on the conclusions with the first two entries based on a 5 letter search term glyco was also presented with an algorithm.

7 References

[1] Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, Miller RA, Digles D, Lopes EN, Ehrhart F, Dupuis LJ, Winckers LA, Coort SL, Willighagen EL, Evelo CT, Pico AR, Kutmon M. WikiPathways: connecting communities Nucleic Acids Research, (2021), Volume 49, Issue D1, 8 January 2021, Pages D613–D621, doi.org/10.1093/nar/gkaa1024 PMID:33211851

[2] Wikipedia contributors. "PubMed." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 6 Oct. 2021. Web. 9 Nov. 2021.

[800] Cromwell, JB. (2021). The Mathematical Learning Space Research Portfolio:Main and Reader Pages. mathlearningspace.weebly.com. October 27,2021.

[801] Cromwell, JB. (2021). MathematicalLearningSpace. https://github.com/MathematicalLearningSpace October 27, 2021.

[1000] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[1001] Jyoti Rani, S.Ramachandan and Ab. Rauf Shah (2014). Text mining of PubMed abstracts. R package version 1.0.5