Analysis of Embryonic Lethal ASD Biology via BioNER

Background

Recent advances in genomic technologies have started to uncover the genetic architecture associated with autism spectrum disorder (ASD). ASD gene encoded proteins have been largely discovered in biological pathways related to synaptic function, transcriptional regulation, and chromatin modification. Experiments using mouse models have led to the observation of genes which are essential in development. These essential genes are required for the continued development of the embryo and can result in embryonic, perinatal, or neonatal lethality upon homozygous knockout. Autism-associated genes have shown significant enrichment for essential genes in homozygous knockout models.

AutDB + IMPC

~1,300 genes in the Human Gene Module
295 genes have annotated genetic mouse models

~ 6,400 homozygous knockout mouse models
613 lethal before E12.5
24 are ASD genes
(P = 2.247e-5)

Data Acquisition

58 Lethal Gene Symbols → 57 UniProt Accessions (from 53 gene symbols) → 522 PMIDs (355 picked up by PubTator)

BioNER Models

PubTator

Synthesis of many SOTA models

GeneTUKit - gene mentions

1. CRF trained on BioCreAtIvE II Gene Mention Recognition Task
2. Dictionary based recognition from Entrez-Gene Database
3. ABNER (A Biomedical Named Entity Recognizer) ~Settles, 2005

GenNorm and SR4GN– gene normalization, species assignment
Heuristics:

1. Prefix
2. Co-occurring word
3. Focus species of document

+ 2 rules for focus species
- Title vs abstract mention
- Frequency in Linnaeus corpus

+ 1 rule for empty species
- Infers based on words strongly correlated to a species ('cohort' for human, 'ferment' for yeast, etc.)

Dnorm – diseases
Pairwise learning to rank
- NCBI disease and MEDICdatabase

tmVar – variants
500 manually curated abstracts - CRF

Table 1. Text-mining tools used for pre-annotating bio-entities in PubMed articles

Bio-entity	Text-mining tool	Nomenclature	F_1 score (%)
Gene (mention)	GeneTUKit	N/A	82.97
Gene (normalization)	GenNorm	NCBI Gene	92.89
Disease	DNorm	MEDIC	80.90
Species	SR4GN	NCBI Taxonomy	85.42
Chemical	A dictionary-based lookup approach	MeSH	53.82
Mutation	tmVar	NCBI dbSNP (rs#) or tmVar normalized forms	93.98

ScispaCy

CNN model
- Trained on MedMentions ~ over 4,000 abstracts and over 350,000 linked mentions
- 3 million concepts from UMLS 2017

Model	Precision	Recall	F1
en_core_sci_sm	69.22	67.19	68.19
en_core_sci_md	70.44	67.56	68.97

Wei, Chih-Hsuan, Hung-Yu Kao, and Zhiyong Lu. "PubTator: a web-based text mining tool for assisting biocuration." Nucleic acids research 41.W1 (2013): W518-W522.
Neumann, Mark, et al. "ScispaCy: fast and robust models for biomedical natural language processing." arXiv preprint arXiv:1902.07669 (2019).

Results

PubTator

entity	term	count	entity	term	count
Species	mouse	283	Chemical	calcium	35
Species	mice	208	Chemical	tyrosine	30
Species	human	152	Chemical	iron	26
Gene	LIS1	121	Chemical	5-HT	19
Gene	ErbB4	78	Chemical	lysine	19
Gene	AMBRA1	52	Chemical	Ser	17
Gene	Brd4	49	Chemical	serine	16
Gene	MEF2	47	Chemical	steroid	13
Gene	GRIP1	46	Disease	cancer	30
Gene	frataxin	45	Disease	lissencephaly	20
Gene	MEF2C	42	Disease	infection	15
Gene	Keap1	40	Disease	tumor	12
Gene	FGFR1	35	Disease	diabetes	11

Scispacy

entity	count	entity	count	entity	count
Expression procedure	183	Development Lot	66	Complex	47
Phosphorylation	140	Receptor Tyrosine-Protein Kinase ErbB-4, human	66	member	47
Homo sapiens	134	in vitro	65	Identified	46
Proteins	132	AMBRA1 gene	61	Clone Cells	45
Activation action	124	Cells	58	CDISC Findings Class	44
protein expression	120	Data call receiving device	57	Fibroblast Growth Factor	43
Genes	112	Adult	56	NCOA2 gene	43
CASP14 gene	104	Mechanism (attribute)	56	Test Result	43
Classical Lissencephaly	90	Induce (action)	55	DYRK1A gene	42
Binding action	88	Mammals	55	Regulation	42
Mus	88	mutant	55	RNA, Messenger	42
Increased	87	Scientific Study	54	Murine	41
in vivo	80	Transcription, Genetic	54	Protein Isoforms	41
Mathematical Operator	77	Neurons	53	Protein Overexpression	41
Embryo	75	receptor	53	Associated with	40
Levels (qualifier value)	72	BRD4 gene	49	Autophagy	40
Reduced	72	Brain	48	Microtubules	40
Social Interaction	71	Histopathologic Grade differentiation	48	Sequence - TransmissionRelationshipTypeCode	40
TEK gene	69	Mutation	48	Amino Acids	39
DNA, Complementary	68	Activities	47	Functional	39

Discussion

Despite the granular differences in entity recognition in these models, they both picked up on similar frequent terms found in the abstracts. While the PubTator model showed semantic specificity when matching the term to its label, the ScispaCy model was able to show the same level of specificity at the level of the label. This is because of the large size UMLS annotation vocabulary compared to the limited selection of 9 entities arising from the PubTator model. Some semantically equivalent terms had very different counts such as "lissencephaly" at 20 from PubTator and "classical lissencephaly" at 90 from Scispacy. In contrast, gene mention counts were more consistent across models such as AMBRA1 at 52 from PubTator and 61 from ScispaCy and BRD4 at 49 from both models. While the newer ScispaCy model has finer granularity in its entities, it would be good to test these models on the same labeled data set to see if the margin between F1 scores are comparable to what is published. It is possible that despite PubTator's lower resolution in terms of label semantics, the mixture of models and their features specific to biomedical business logic make PubTator a more thorough model for biomedical NER.