ALS and Cognitive Computing: IBM Watson for Drug Discovery identifies novel RNA binding proteins altered in ALS

Nadine Bakkar[1], Tina Kovalik[1], Scott Spangler[2], Alix Lacoste[3], Ileana Lorenzini[1] ,Lucas Vu[1], Kyle Sponaugle[1], Philip Ferrante[1], Elenee Argentinis[3], Rita Sattler[1] and Robert Bowser[1]

Gregory W. Fulton ALS Center, Barrow Neurological Institute[1] and St Joseph's Hospital and Medical Center, Divisions of Neurology and Neurobiology, Phoenix, AZ; IBM Research - Almaden[2], San Jose, CA; IBM Watson Health[3], New York, NY

Abstract

Introduction: Various RNA binding proteins (RBPs) are altered in amyotrophic lateral sclerosis (ALS), with mutations in 11 RBPs causing familial ALS. There are 1,455 RBPs in the human genome, therefore other RBPs may also be linked to ALS. **Methods**: We used IBM-Watson for drug discovery (IBM-Watson) to propose new RBPs linked to ALS (ALS-RBP). The cognitive capabilities of IBM-Watson enable it to extract features from published literature to identify new connections between entities of interest. IBM-Watson analyzed published abstracts describing known ALS-RBPs, and applied that analysis to all RBPs. To test IBM-Watson's performance, we first restricted its knowledge to abstracts prior to 2012 and asked it to rank-order RBPs by disease probability. Three ALS-RBPs (Matrin-3, GLE1, and ARHGEF28) have been discovered since 2012. **Results**: Matrin-3 was the top candidate in this retrospective analysis, with both ARHGEF28 and GLE1 ranking in the top 10%. Having shown the validity of IBM-Watson's rankings, we expanded the analysis to all current abstracts. Of the top 50-ranked genes, 5 are altered in ALS, albeit not mutated (RBM45, MTHFSD, SMN2, EWSR1 and hnRNPA3). Also included within the top 10 proteins were hnRNPU, SRSF2, SYNCRIP and CAPRIN1. To validate Watson's rankings, we examined the subcellular distribution of these RBPs in ALS and control post-mortem tissues. We identified altered levels and distribution of the cytoplasmic isoform (Q1) of Syncrip/hnRNPQ in ALS cerebellum. This isoform interacts with Staufen and SMN in neuronal RNA granules. In cerebellum, Q1-Syncrip showed increased nuclear immunoreactivity in Purkinje neurons of SALS (n=7) compared to controls (n=6). Interestingly, Purkinje cells from C9orf72-ALS patients exhibited increased cytoplasmic granular staining for Syncrip (n=4). No changes were observed for other Syncrip isoforms in cerebellum. **Conclusions**: Overall, our approach using IBM-Watson to mine scientific literature to find new ALS-RBPs has led to exciting findings and may further aid efforts to understand RBP-mediated pathology.

IBM-Watson Model Validation:

Retrospective Analysis

IBM-Watson analysis was performed on all abstracts published prior to 2012. For this analysis, 3 RBPs published after 2012 were omitted from the training set consisting of RBPs linked to ALS. IBM-Watson then ranked all candidate RBPs by similarity to the training set.

Training Set
TARDBP
FUS
ATXN2
ANG
RBM45
SETX
HNRNPA2B1
HNRNPA1
TAF15

Entity	Score (GD)	Rank
MATR3	0.00204078	1
NUPL2	0.00181635	2
SRSF2	0.0017781	3
SYNCRIP	0.00175763	4
HNRNPU	0.00174455	5
RBM6	0.00161879	6
IGHMBP2	0.00154716	7
HNRNPA3	0.00154361	8
HNRNPC	0.00153549	9
HNRNPM	0.00151568	10
- - -		
ARHGEF28	3.95E-04	89
GLE1	3.85E-04	165

The top hit, MATR3, was subsequently shown to be linked to familial forms of ALS in 2014. hnRNPA3 was also linked to ALS pathology. Mutations in the two other RBPs subsequently linked to familial ALS, ARHGEF28 and GLE1 ranked in the top 10% of all RBPs.

Leave-One Out Validation (LOO)

A leave-one-out cross-validation (LOO) of the training set was also performed to validate the performance of the model. The analysis was ran multiple times, with one training set entity omitted per iteration and added to the candidate list instead. Candidate RBPs were rank-ordered again to see how they rank in this model. The statistical significance of this analysis (Fisher exact test: (1469,48,11,7) = 7.62E-9) indicates that the training set genes analyzed have predictive power over each other compared to the candidate set.

IBM-Watson Analysis

IBM-Watson analysis was performed on all abstracts published prior to 2015. For this analysis, only RBPs mutated in ALS were used as the training set. IBM-Watson then ranked all candidate RBPs by similarity to the training set.

Training Set
TARDBP
FUS
ATXN2
ANG
SETX
HNRNPA2B1
HNRNPA1
TAF15
MATR3
GLE1
ARHGEF28

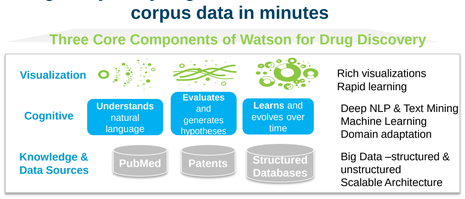
Protein	Score (GD)	Rank (GD)
HNRNPU	0.00291413	1
SYNCRIP	0.0027466	2
RBM45	0.00267976	3
RBMS3	0.00249365	4
SRSF2	0.0024586	5
HNRNPH2	0.00225497	6
NUPL2	0.00215157	7
CAPRIN1	0.00210949	8
RBM6	0.00191485	9
MTHFSD	0.00191011	10

5 proteins that ranked in the top 5% of the Watson analysis have already been shown to be altered in ALS: RBM45, MTHFSD, hnRNPA3, SMN2, and EWSR1, ranked at 3, 10, 18, 63 and 66, respectively.
SRSF2 (SC-35) and hnRNPH1 have been previously shown to colocalize with C9orf72 RNA foci in cerebellum of C9 patients (Cooper-Knock 2014 and 2015). Since this analysis has been performed, Caprin1 has been shown to localize to TDP-43 and Fus inclusions in patients carrying TDP43 or Fus mutations (Blokhuis AM. 2016-Acta Neuropathol).

IBM-Watson for Drug Discovery

Watson for Drug Discovery enables breakthrough insights by analyzing millions of articles and other corpus data in minutes

Three Core Components of Watson for Drug Discovery



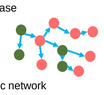
Watson's Predictive Methodology - Overview

1 Generate predictive model

Train Watson based on a set of entities that embody the target concept

Analyze millions of scientific text documents

Create a semantic fingerprint for each protein, drug or disease



Build semantic network

2 Evaluate model
Retrospective Analysis

Rank candidate entities based on model

Protein	Rank
NL	1
AX	2
FF	3
…	…
IR	10
IX	11
MN	12
AH	13
…	…

Recently discovered RBPs rank high

3 Prioritize top predictions

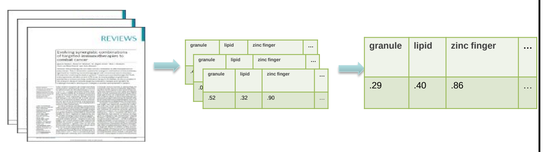
RNA-binding proteins

Use diverse criteria to further prioritize top predictions, such as specific biological relationships extracted from text

Ranking for Semantic Similarity with Graph diffusion

1) Create a text "fingerprint" for every entity

1) All abstracts that mention a specific RBP — *2) Text vectors of those abstracts* — *3) Average text vector for RBP*



2) Create a distance matrix relating each RBP to every other RBP

Distance matrix connecting each entity to every other entity

Average vector for RBP 1

granule	lipid	zinc finger	…
.29	.40	.86	…

Average vector for RBP 2

granule	lipid	zinc finger	…
.29	.40	.86	…

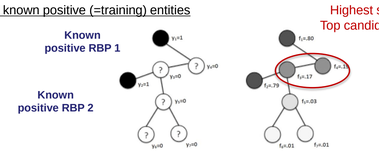
	RBP 1	RBP 2	RBP 3	…	RBP N
RBP 1	0	0.277	0.691	…	0.574
RBP 2	0.277	0	0.356	…	0.731
RBP 3	0.691	0.356	0	…	0.117
…	…	…	…	0	…
RBP N	0.574	0.731	0.177	…	0

3) Ranking algorithm

Label known positive (=training) entities

Diffuse label through the network

Highest score = Top candidate RBP

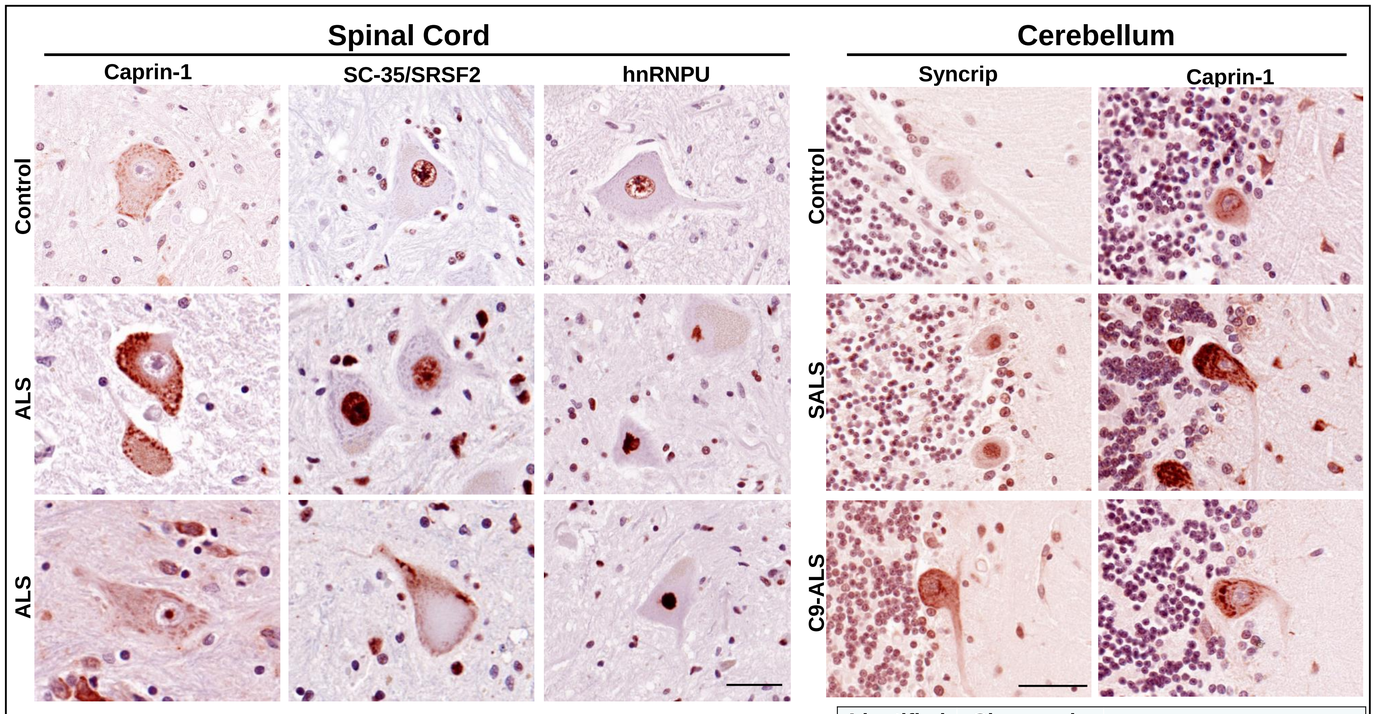


Protein	Rank
HNRNPU	1
SYNCRIP	2
RBM45	3
RBMS3	4
SRSF2	5
	6
HNRNPH2	
NUPL2	7
CAPRIN1	8
…	…

Spangler et al. KDD 2014. Example of graph diffusion on a network. Given a set of edges and labeled nodes, information is diffused to find additional candidate nodes for the annotation in question.

Use a machine learning approach to determine the similarity of entities to the overall training entities

Results



Post-mortem spinal cords, cerebellum or frontal cortex (for SC-35) from SALS, C9orf72-ALS or non-neurological disease control were stained for the specified antibodies (Caprin-1: Proteintech, 15112-1; SC-35: Abcam ab11826; or hnRNPU (Abcam ab10297), Syncrip (HPA041275)). Biotinylated secondary antibodies were used and the signal was developed with Novared peroxidase substrate kit (Vector Labs) and images were captured at 40x. Representative images are shown for various antibodies/brain regions, and results for altered proteins are summarized in the table.

Identified genes	Changes in ALS (Y/N)	Changes observed
hnRNPU	Y	Increased expression+ nuclear inclusions
Syncrip	Y	Increased expression in cerebellum+inclusions
RBMS3	Y	Intense staining in cerebellum
SC-35	Y	Some inclusions in spinal cord
hnRNPH2	N	N/A
NUPL2	Y	Expression in astrocytes
Caprin1	Y	Increased expression and/or granule size
RBM6	N/D	-

Experimental Design

We posed a question related to RBPs altered in ALS to IBM-Watson. There are currently 1455 RBPs in the genome, and so far, mutations in 11 RBPs have been shown to cause familial ALS. Given the role of RNA dysregulation in both familial and sporadic ALS, we aimed to identify further RBPs altered in ALS using IBM-Watson's cognitive abilities. As a proof of concept, we started with a retrospective analysis using known RBPs linked to ALS through 2012 as a training set and asked IBM-Watson to predict other RBPs linked to ALS based on their similarity to the training set. As shown below, IBM-Watson successfully identified a number of proteins that were subsequently shown to be altered in ALS. We then proceeded to perform the full prospective analysis using known RBPs known to be mutated in ALS through 2015 as the training set to predict other RNA binding proteins linked to ALS. From the table ranking all other RNA binding proteins, we explored the expression and subcellular distribution of the top 10 in tissue samples from ALS and control subjects. For the tissue analysis, we used cerebellum, and spinal cord from 10 ALS, 5 non-neurologic disease controls, 4 C9orf72 ALS, and 2 frontotemporal dementia cases. We similarly examined RNA expression levels of these RBPs in post-mortem spinal cords from ALS and non-neurologic disease controls, as well their tissue distribution in induced pluritpotent cells (iPS) from ALS and controls.

Conclusions

- Using IBM-Watson, we have identified multiple RNA-binding proteins that show altered expression, altered distribution or form inclusions in ALS compared to non-neurological disease controls. We have not identified mutations in either of these proteins.
- Use of IBM-Watson to mine scientific literature is a novel unique approach to find new ALS-RBPs and has led to exciting findings that may further aid efforts to understand RBP-mediated pathology in ALS.

*We would like to acknowledge The Target ALS Human Postmortem Tissue Core for the tissues used in this study.