Introduction

Our goal is to programmatically build a model of neuroendocrine aging, finding sentences in article abstracts where the subjects and objects of the sentences are nouns in our vocabulary - and the predicates of the sentence, relating the subjects and the objects, are verbs in our vocabulary. We've developed software to retrieve article abstracts from PubMed, filter out sentences that do not contain verbs in our chosen vocabulary, parse sentences to examine their phrase structure and lastly extract relations (i.e. subject-predicate-object triples) by pattern matching on the phrase structure of parsed sentences. In addition to the software that allows us to extract relations from article abstracts to form our model of neuroendocrine aging, we publish an interactive graph visualization of our model, accessible from your web browser.

Prior to starting this project, we had been building a model of neuroendocrine aging by hand: we read article abstracts ourselves, finding relations between two nouns in sentences and then adding them to our model. This method of building a model of neuroendocrine aging appears ineffective to us, as articles about neuroendocrine aging are published at an ever-increasing rate.

Previous work in the relation extraction and semantic web communities, such as Carnegie Mellon's Read the Web [3] project and the University of Mannheim's DBpedia [9] project, inspired us to find a method for constructing our model of neuroendocrine aging programmatically. Neither project provided us with precisely the software tools we needed to retrieve article abstracts and extract relations from them, so we felt it was necessary to develop our own. We indeed believe that our relation extraction software has enabled us to construct a model of neuroendocrine aging programmatically and has hence helped us study neuroendocrine aging more effectively, though there is still much room for improvement with regards to the accuracy and performance of our software.

Methods

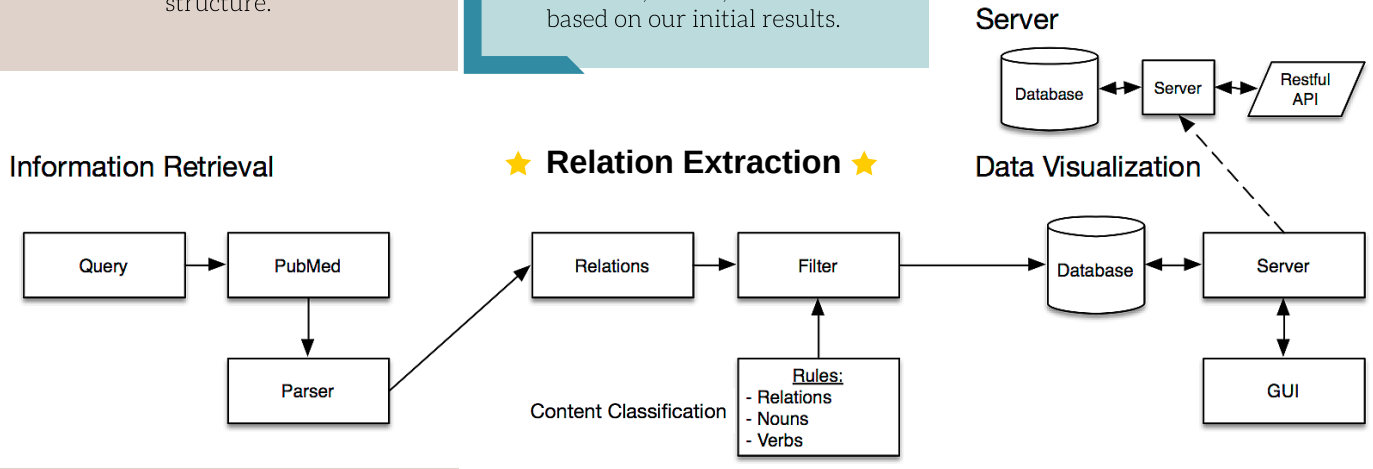
We search for articles on PubMed with queries containing nouns from our neuroendocrine aging vocabulary. The National Center for Biotechnology Information (NCBI) provides with the Entrez Programming Utilities [5], allowing us to search for articles on PubMed in our software, obtaining a list of article unique identifiers as a response. We chose to uniformly sample articles about each noun in our vocabulary, after searching for each noun in our vocabulary & obtaining a list of article unique identifiers for each noun. Then, we download article abstracts, after sampling from the list of article unique identifiers that we have obtained; the Entrez Programming Utilities lets us fetch article abstracts as well. After downloading article abstracts, we tokenize them into sentences, using a small software module provided by David Hall [8]. We then filter out sentences that do not contain a word stem of one of the verbs in our vocabulary; we do not want to extract relations from sentences that we do not need to, since it takes considerable time to extract relations from sentences. Lastly, we extract relations from the remaining sentences that contain verbs from our vocabulary and, finally, we write them to a sink - our relational database.

Figure 2.0: Process Diagram

We extracted relations from sentences by first parsing them and then pattern matching on their phrase-structure. We chose to use software called the Berkeley Parser [6] to take sentences and produce a tree, describing their phrase-structure.

We queried PubMed articles using a control set of biology terms in which semantic properties of the terms were known, extracting relations from abstracts relevant to our queries only. We then defined rules for classifying types for relations, nouns, and verbs based on our initial results.

We take the results from text mining and attempt to find a more meaningful way to represent data. We choose to put this data in a graph, and equipped with analytical tools, we might find new trends and explore additional patterns.



- Assign each word in the sentence a part-of-speech tag
- Group words into phrases
- Group phrases into clauses
- Clauses relate a subject noun to an object noun with a predicate verb

noun types

substrate
structure
process

relation types

Type 1:
substrate-substrate
Type 2:
substrate-structure
Type 3:
substrate-process

- GUI interactivity; web app
- Database storage of relations
- Object access through requests to server, using the SigmaJS library to help render graphs within the browser

Sentences can contain multiple clauses, and each clause relates a subject noun to an object noun with a predicate verb; the Penn Treebank Project [11] provides a description for each tag used when parsing sentences and annotating them for their phrase-structure. It took a considerable amount of effort to define our own patterns for extracting relations from parsed sentences; we could not find existing software that (a) did what we wanted and (b) was free to use and modify.

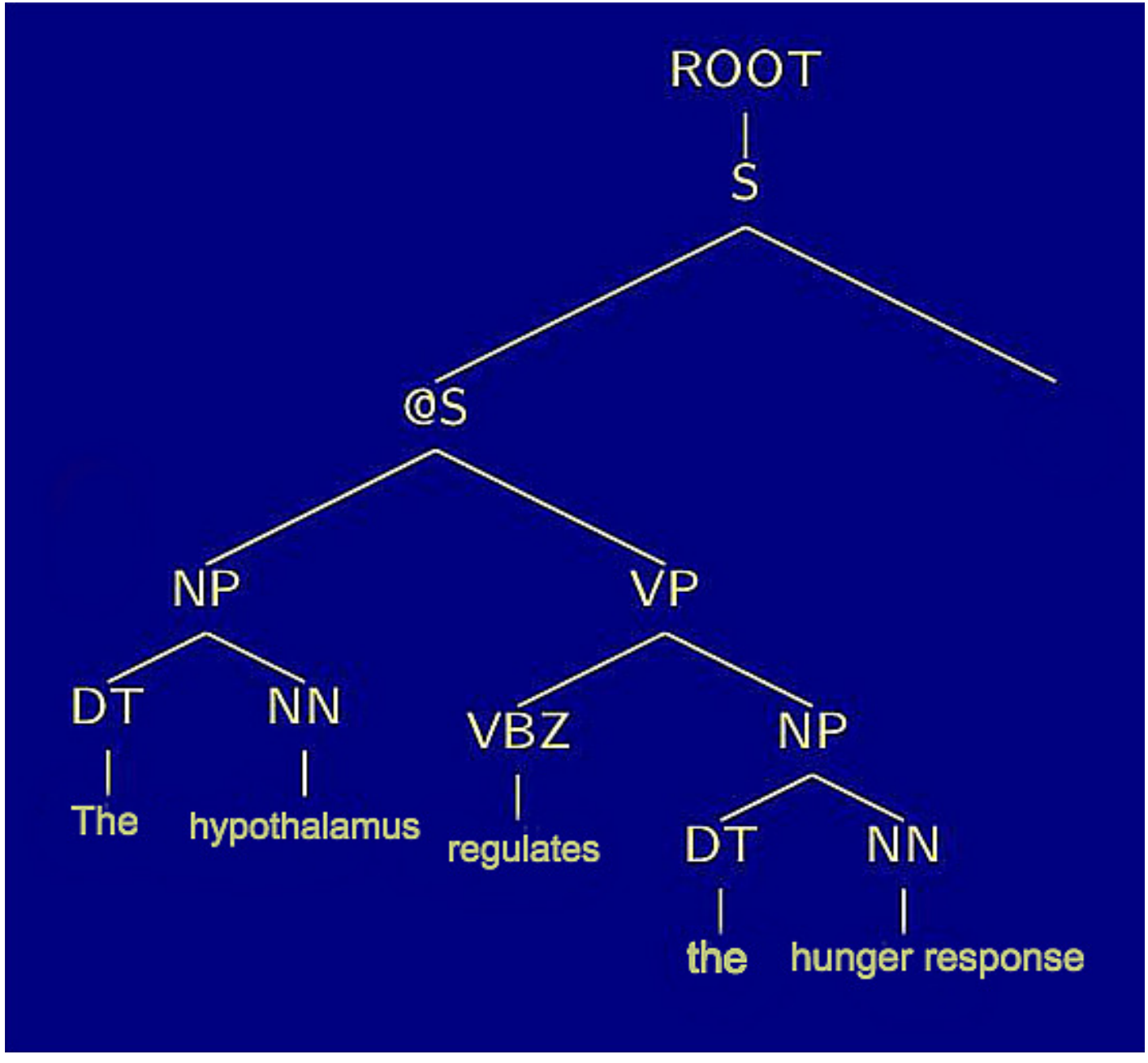
Tag	Meaning	Example
DT	Determiner	the
IN	Preposition	of
JJ	Adjective	blue
RB	Adverb	quickly
CC	Coordinating Conjunction	and
NN	Singular Noun	monkey
NNS	Plural Noun	monkeys
VB	Base Verb	fall
VBZ	Singular Present Verb	falls
VBD	Past Tense Verb	fell
VBN	Past Participle Verb	fallen
VBG	Gerund Verb	falling
...	...	...

Tag	Meaning	Example
NP	Noun Phrase	the woman
VP	Verb Phrase	calls the man
PP	Prepositional Phrase	from the store
ADVP	Adverb Phrase	quickly and quietly
ADJP	Adjective Phrase	blue and red
CONJP	Conjunctive Phrase	as well as
...	...	...

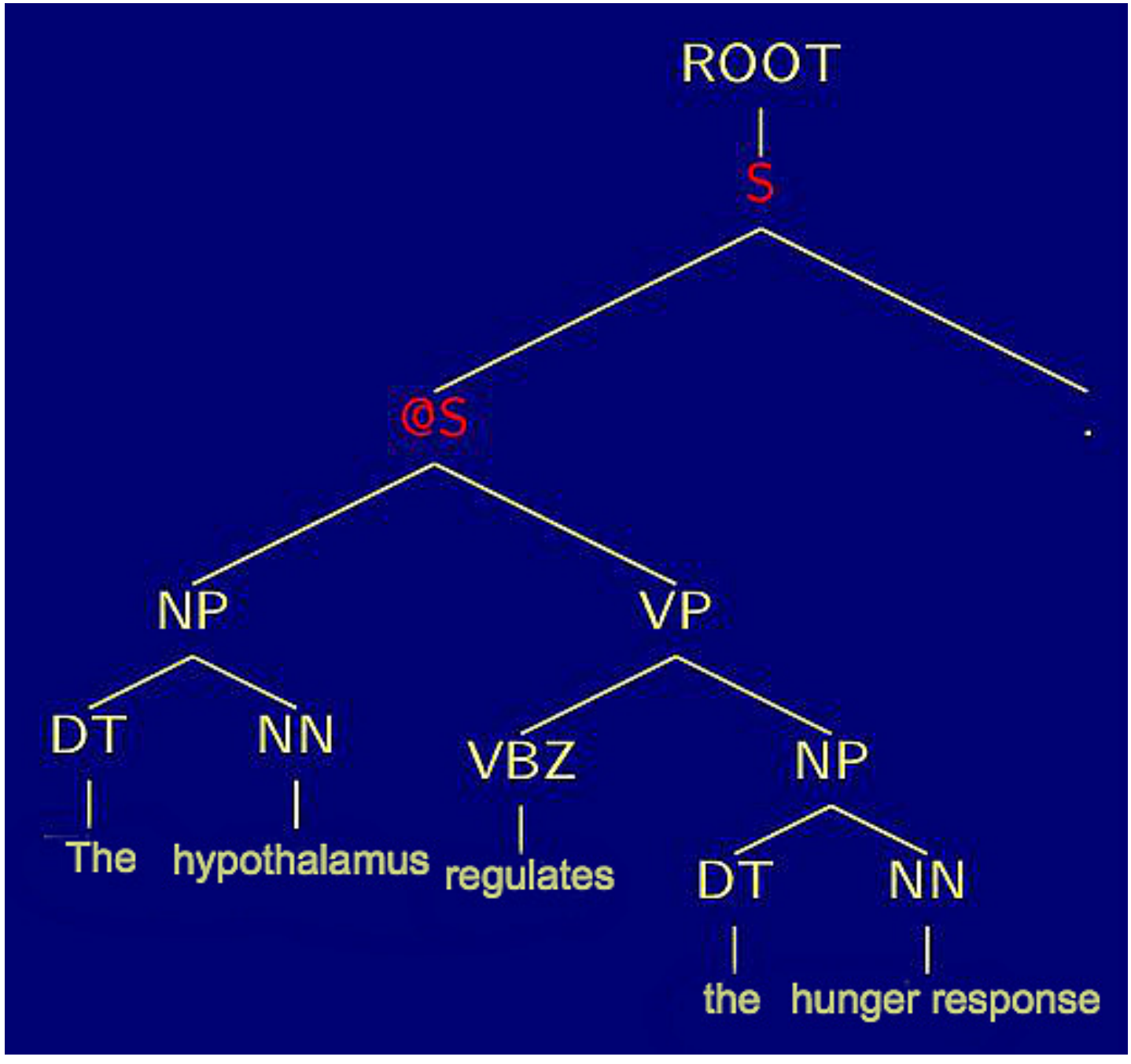
Tag	Meaning	Example
S	Declarative Clause	the dog walks
SBAR	Conjunction + Clause	that the dog walks
...	...	...

Figure 2.1: Penn Treebank Tags

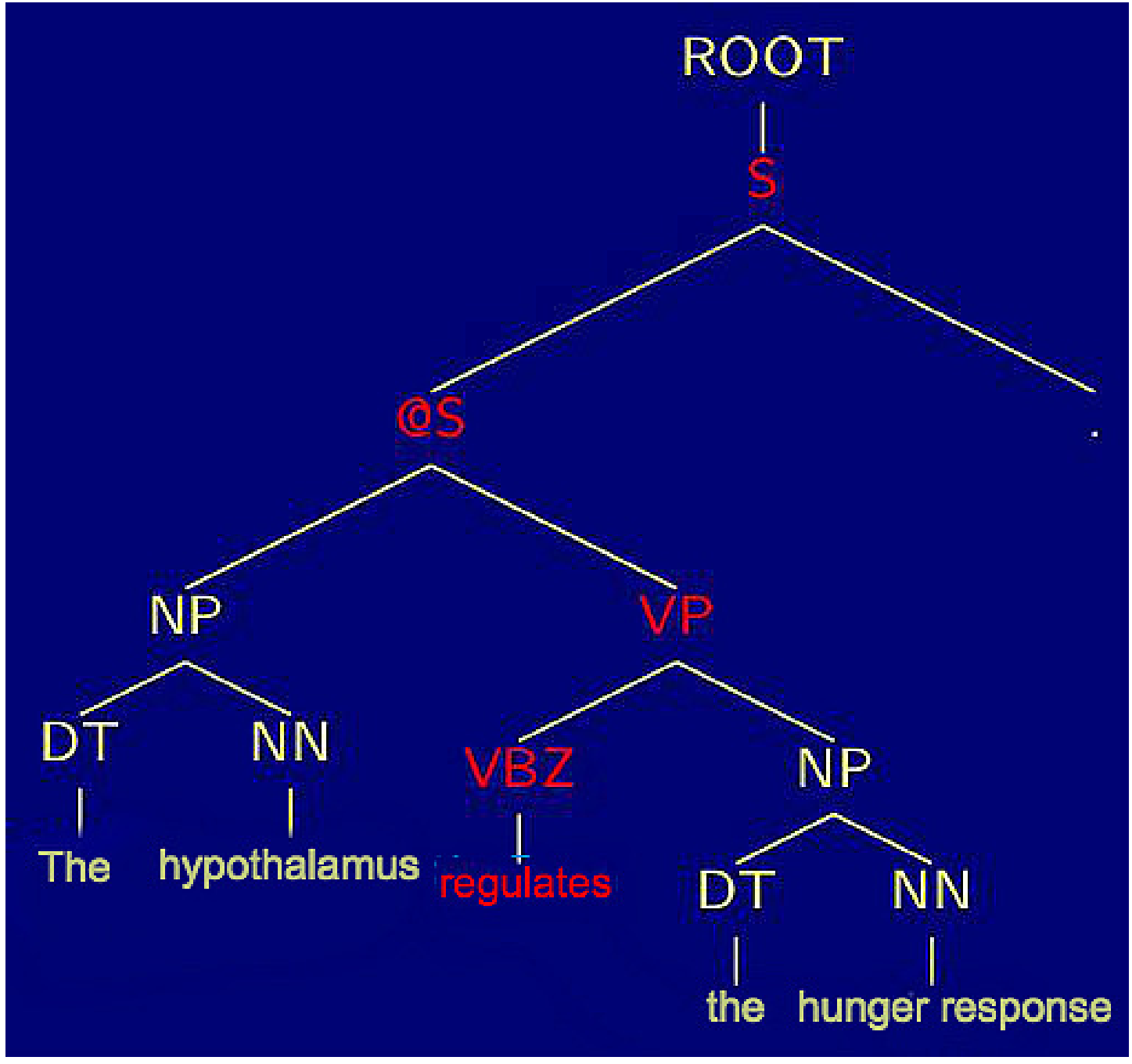
Figure 2.2 (A - F): Example of Relation Extraction **(A)**



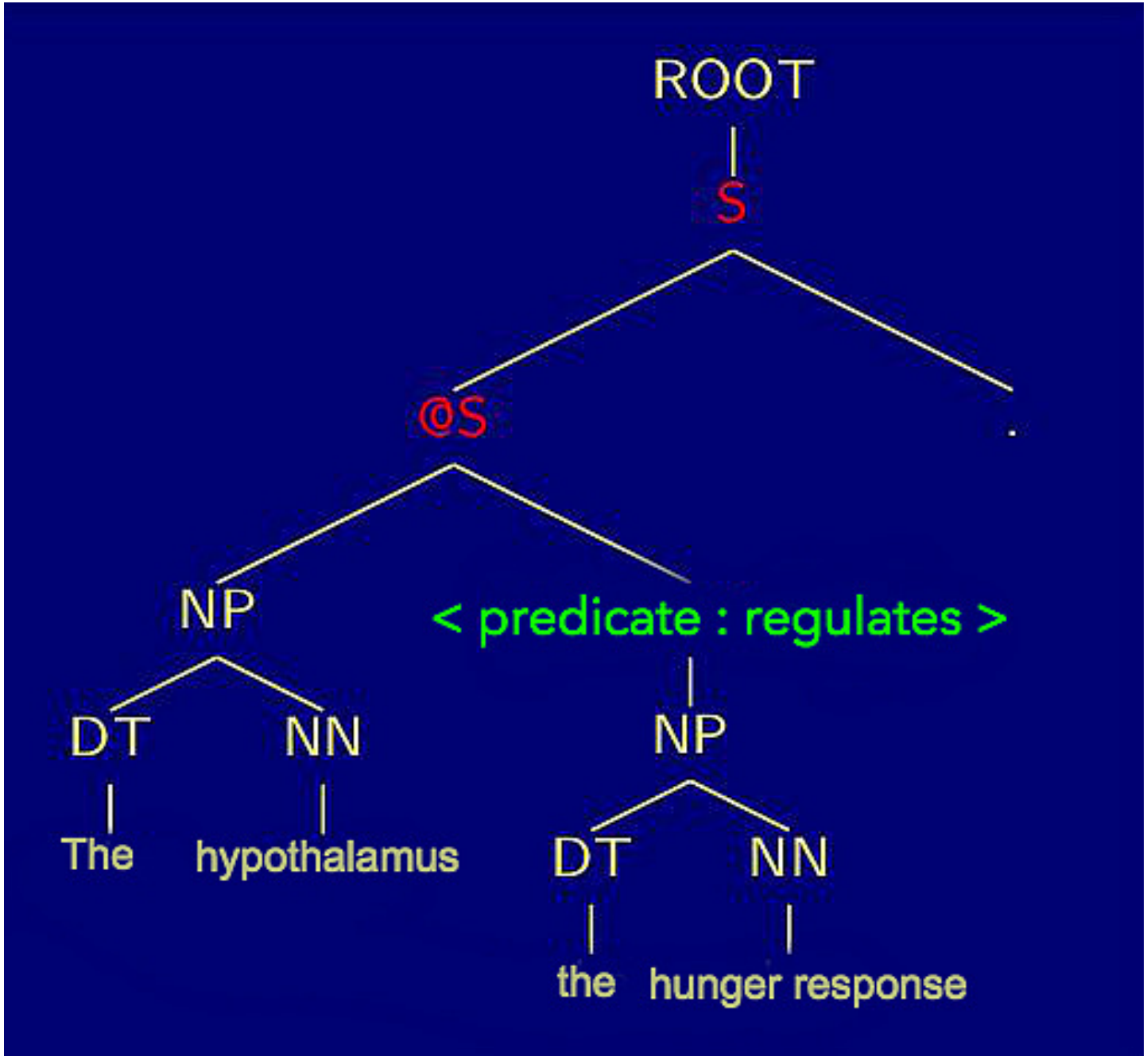
(B)


ROOT
S
@S
.
NP
VP
DT
NN
VBZ
NP
The
hypothalamus
regulates
DT
NN
the
hunger
response

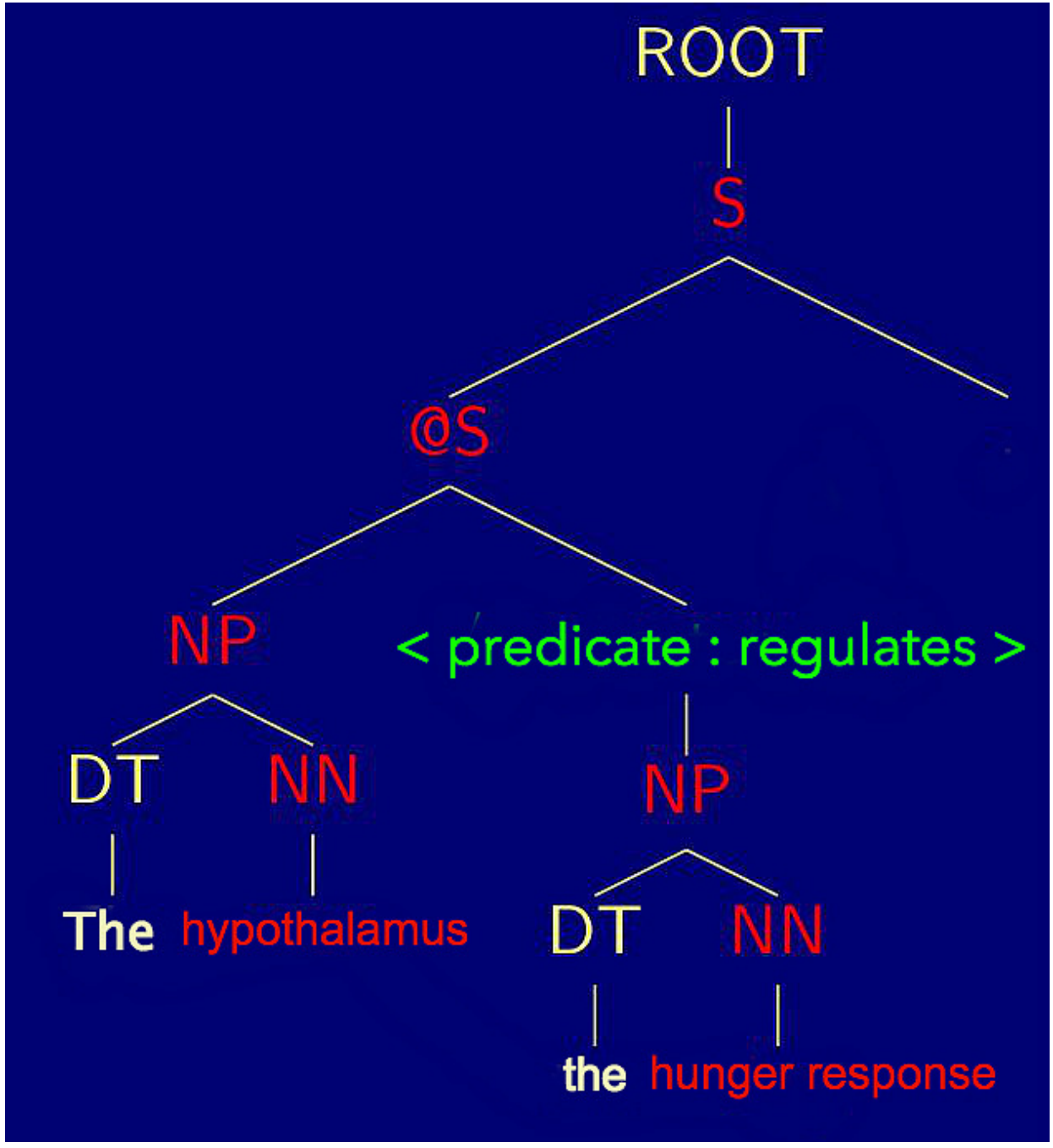
(C)


ROOT
S
@S
NP
VP
DT
NN
VBZ
NP
The
hypothalamus
regulates
DT
NN
the
hunger response

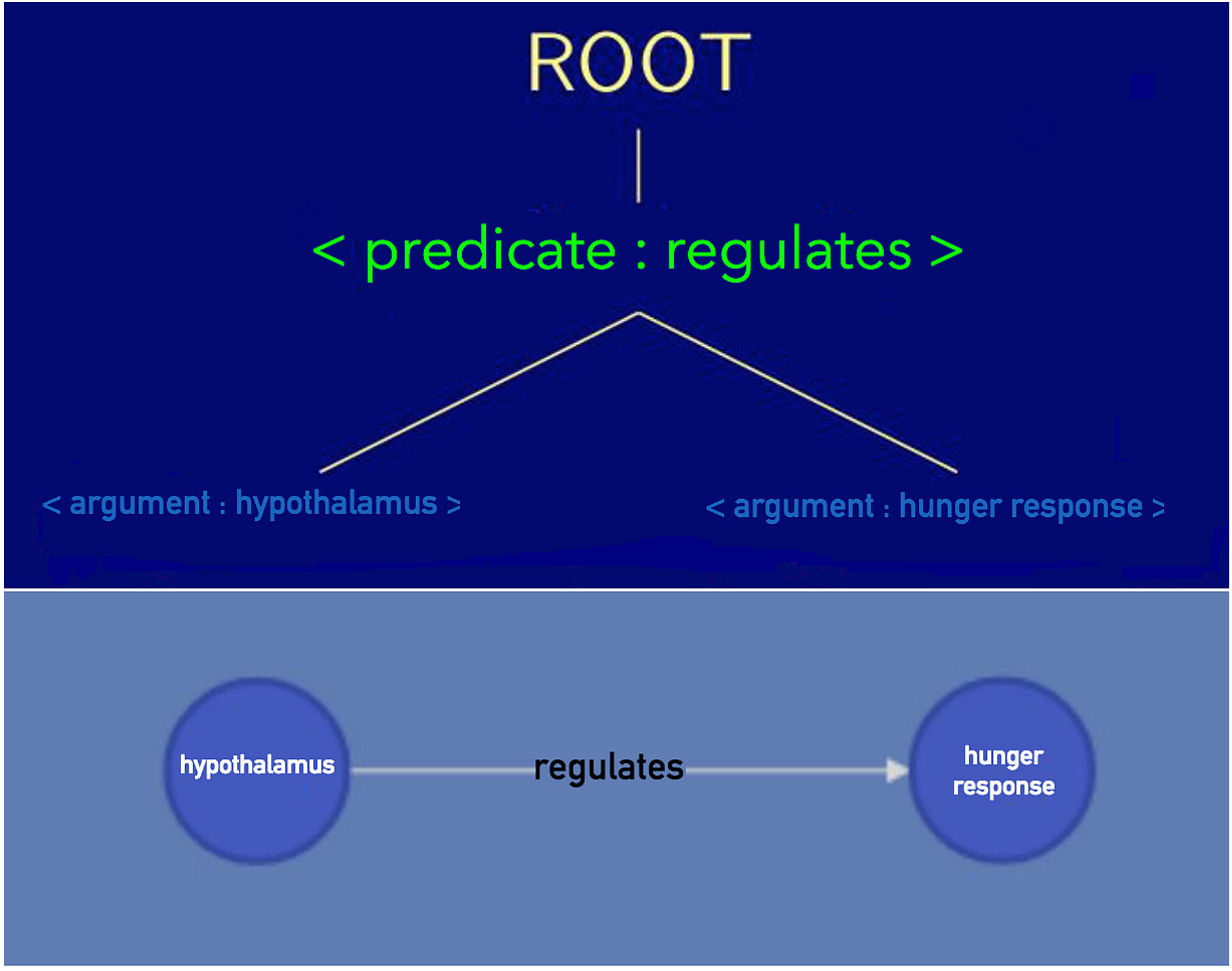
(D)


ROOT
S
@S
.
NP
< predicate : regulates >
DT
NN
NP
The
hypothalamus
DT
NN
the
hunger response

(E)


ROOT
S
@S
NP
< predicate : regulates >
DT
NN
NP
The
hypothalamus
DT
NN
the
hunger response

(F)


ROOT
< predicate : regulates >
< argument : hypothalamus >
< argument : hunger response >
hypothalamus
regulates
hunger response

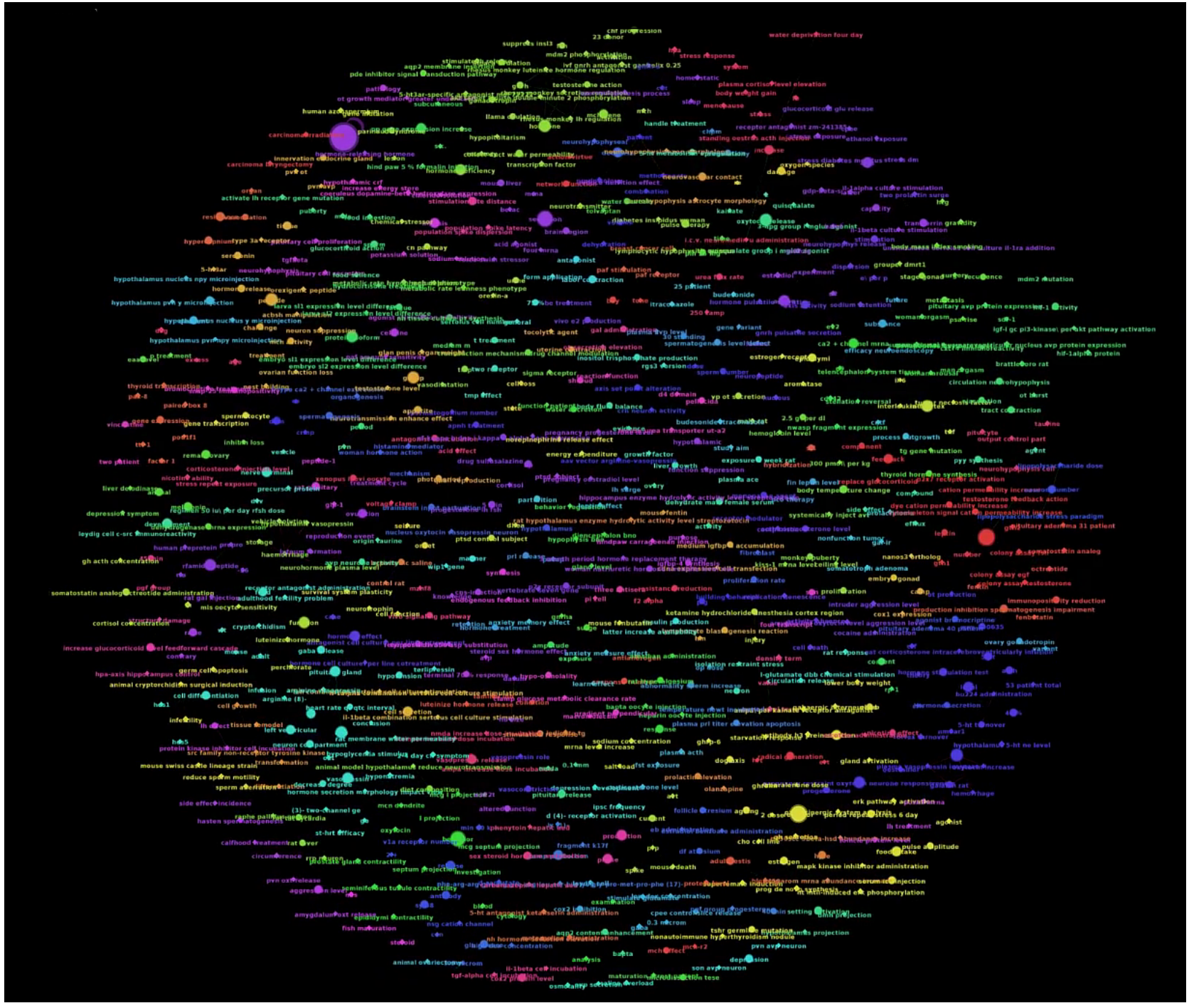


Figure 2.3: Biological Knowledge Network

Mining Biological Knowledge

We focused on low level interactions in a biological system, defining relations from bottom-up. This was chosen due to the fact that biochemically quantifiable relations in biology provide a viable source for data manipulation as we can ensure that each object in our database and subsequent graph can be altered numerically (i.e. substrate concentration). For instance, we can apply functions with real-life parameters (i.e. concentration of X, cell volume, telomere length, or blood pressure) for setting initial start variables that modulate the impact of a biological object on its relational neighbours. This may give insight to the effect of how a small change in the properties of a single object could have on a higher level system.

noun types

substrate	structure	process
elastin	hypothalamus	food intake
dopamine	pineal gland	obesity
ghrelin	neuron	bone formation
leptin	brain tumour	cancer
insulin	splenocyte	apoptosis
progesterone	d1 dopamine receptor	stress
prolactin	amygdala	hunger response
oxytocin	pituitary gland	vasoconstriction
elastase	neuroendocrine system	angiogenesis

Table A: Example Noun Types

Sample of noun subjects and noun objects queried in our initial dataset. Relations extracted either contained a query noun as a subject or object, or came from PubMed article abstracts that contained at least a single instance of the query noun in the body of the text.

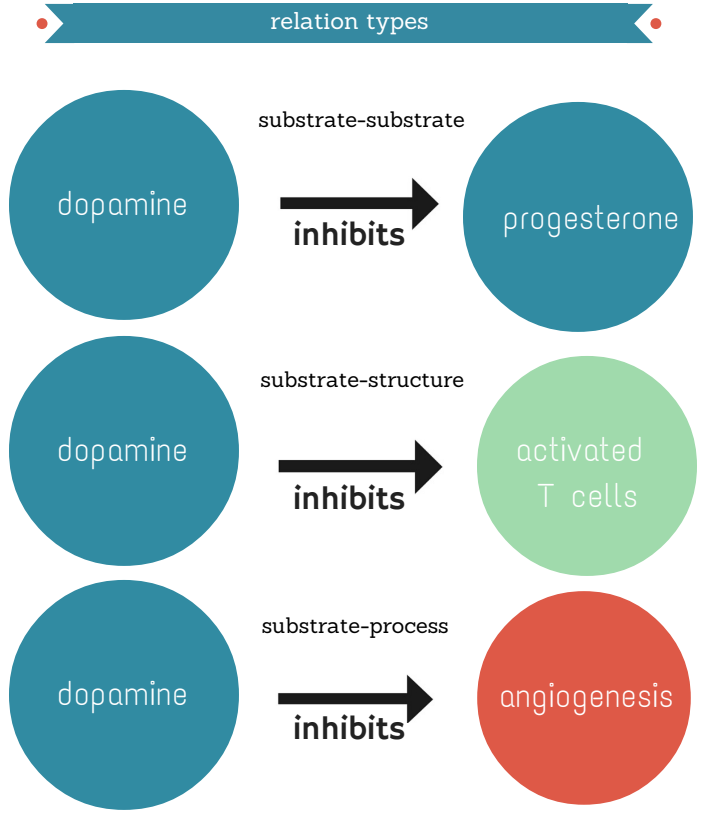


Figure 2.4: Example Relation Types

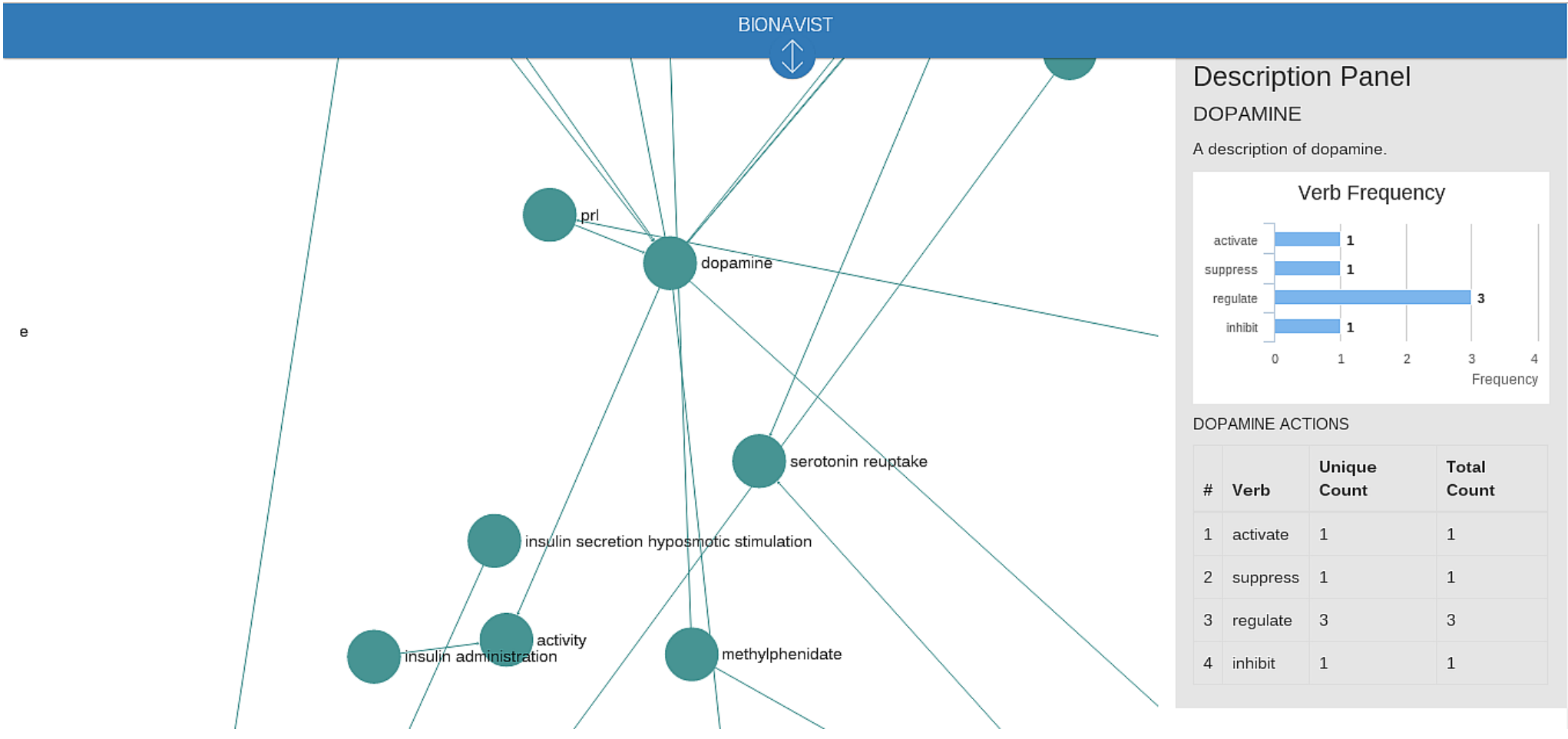


Figure 2.5: Web Application Screenshot I

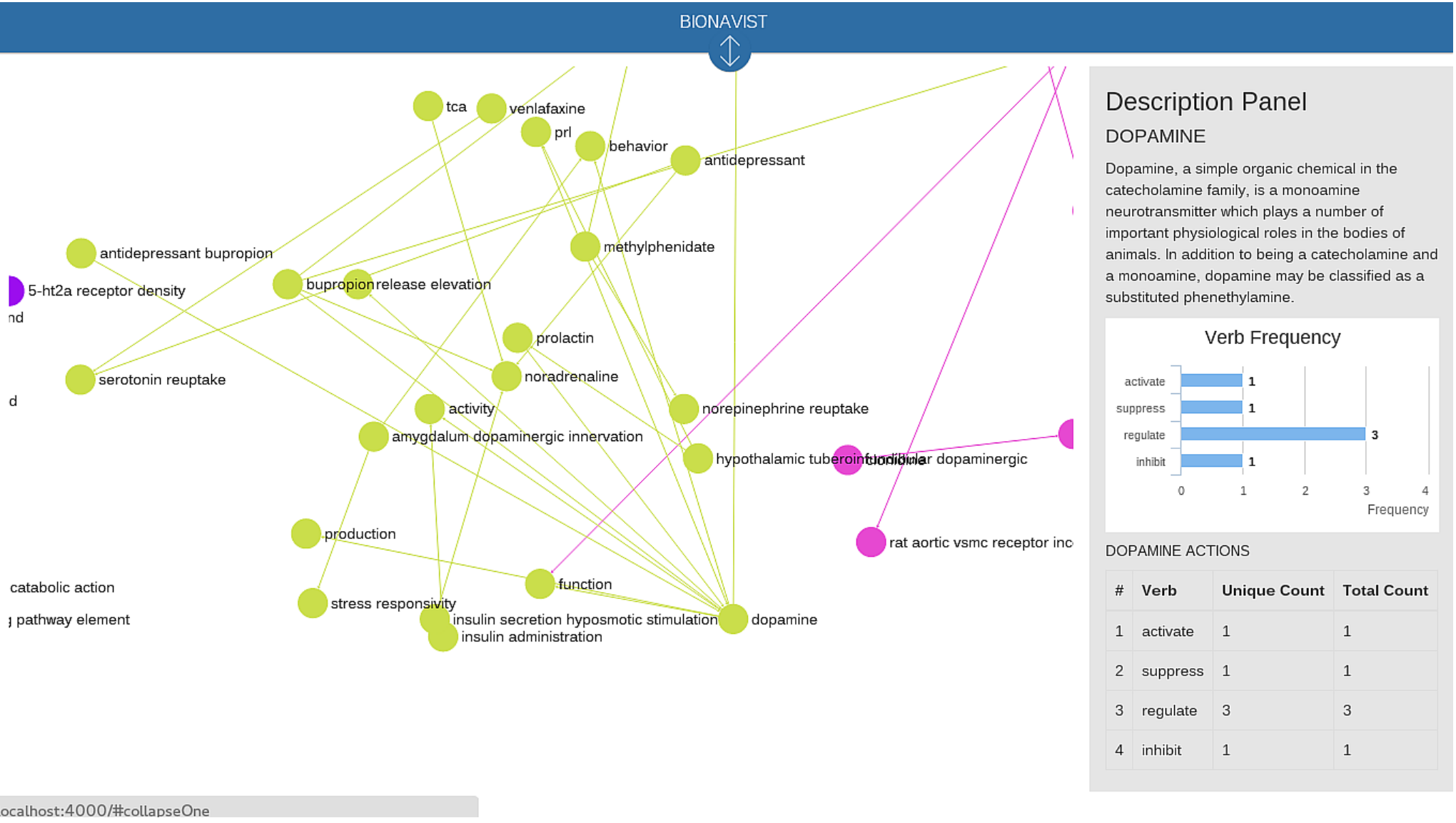


Figure 2.6: Web Application Screenshot II

Results

We extracted relations from article abstracts for about four weeks, in our first attempt to use our software to programmatically construct a model of neuroendocrine aging; in total, we extracted 59,254 relations from nearly 12,000 articles, with a failure rate of roughly 53%. Then, we took the relations we extracted from article abstracts - our model - and exported them to a GEXF graph file [7]. This allowed us to view & analyze our model using the Gephi graph visualization software [1], and also present it in our interactive graph visualization viewable in your web browser. We use the Chinese Whispers clustering algorithm [2] to group nouns (that is, nodes on our graph) using its graph properties like the weight and degree of its edges. After clustering, it appears as though words of the similar meaning or have strong relations with each other are closer together, much like how we categorize nouns by colour on the model of neuroendocrine aging that we originally built by hand. Though, we need to make further comparisons between our programmatically-constructed models of neuroendocrine aging and our hand-constructed model of neuroendocrine endocrine, in order to determine precisely how well a clustering algorithm like Chinese Whispers correctly clusters words with similar meaning together.

There are a total of 27,004,753 articles published about the 178 nouns chosen for our aging vocabulary between January 1st, 1970 and February 20th, 2015 - we computed this value. By building our software with the technology known as scalaz-stream [4], we were able to fetch and extract relations from multiple articles in parallel. In our case, we were able to take advantage of the 64 cores and 192 GB RAM offered by our host machine, named high-fructose-corn-syrup [10]. However, we still were only able to extract relations from a small percentage of these article abstracts in the four week period that we ran our software for - with a high rate of failure when extracting relations from sentences because we did not have patterns defined for sentences of particular phrase-structures.

After querying nouns from our vocabulary to obtain relations from PubMed abstracts, we filtered with tuples of noun types that make up our three relation classes using items from our vocabulary of nouns. We then were able to abstract verb instances from these relations that are commonly used, keeping track of the ones that are used more in a particular relation class.

In the last step of analyzing our relation classification method, we took verbs and nouns (Figure 2.4, Table A & B) from each relation type and queried them (Table B) as a seed tuple to obtain a wildcard object-noun or subject-noun to test the confidence of our relation definitions against the entire set of relations extracted (Table C). The purpose was to evaluate the validity of our definitions of what a significant relation would be, for our model of neuroendocrine aging.

Relation Class	Example Noun Tuples: [nounObject, nounSubject]	Example Noun+Verb Tuples: [noun (object or subject), verb]
substrate—substrate	[dopamine, prolactin] [oxytocin, leptin] [elastase, progesterone]	[dopamine, inhibit] [oxytocin, activate] [elastin, stimulate]
substrate—structure	[dopamine, activated T cell] [neuron, ghrelin] [hypothalamus, leptin]	[neuron, increase] [activated T cell, inhibit] [hypothalamus, activate]
substrate—process	[dopamine, neurotransmission] [sodium, mitosis] [oxytocin, hydrolysis]	[food intake, associate] [neurotransmission, increase] [vasoconstriction, induce]

Table B: Example Seed Tuples for Filtering

For each relation class, a noun tuple filters for verbs and a noun+verb tuple filters for a wildcard subject-noun or object-noun.

	Total Relations Using Noun Queries			All Relations	(Noun+Verb Tuple)
Verb	**Substrates**	**Structures**	**Processes**	**Total Frequency**	**Most Occurred Relation Class**
increase	**254**	**86**	**96**	**2402**	substrate—substrate
decrease	**116**	31	57	**1574**	substrate—substrate
induce	**109**	**44**	**157**	**1488**	substrate—substrate
regulate	**65**	36	**66**	**1261**	substrate—process
treat	45	-	33	**1193**	substrate—process
inhibit	**62**	**35**	**79**	1184	substrate—process
associate	53	20	**101**	930	substrate—process
mediate	43	**57**	37	811	substrate—structure
stimulate	40	26	-	508	substrate—structure
correlate	30	-	33	471	substrate—process
express	27	**111**	29	940	substrate—structure
block	24	-	-	415	substrate—substrate
activate	22	30	-	331	substrate—structure
enhance	22	-	19	320	substrate—process
release	17	24	-	266	substrate—structure

Table C: Relational Noun-Filtered Verbs

Verbs showing the highest observed frequency out of 59,254 relations, having excluded those that occurred less than 10% relative to the highest frequency verb ("increase"). We sampled the composition of the verb used to describe each relation class and from there, eluded a semantic profile of each verb relative to our relation definitions. The above list of verbs (Table B - 14,094 relations total) were obtained by filtering with seed tuples as displayed in Table A. Using separate sets of nouns to query for verbs by category, the distribution of verb frequency for substrate nouns showed a linear relationship with the distribution of verb frequency for the entire set of all relations.

Table C: Relational Noun-Filtered Verbs

Verbs showing the highest observed frequency out of 59,254 relations, having excluded those that occurred less than 10% relative to the highest frequency verb (“increase”). We sampled the composition of the verb used to describe each relation class and from there, eluded a semantic profile of each verb relative to our relation definitions. The above list of verbs (Table B - 14,094 relations total) were obtained by filtering with seed tuples as displayed in Table A. Using separate sets of nouns to query for verbs by category, the distribution of verb frequency for substrate nouns showed a linear relationship with the distribution of verb frequency for the entire set of all relations.

Discussion

In the future, we would like to go beyond low level biochemical interactions and build on our system to handle higher level relations in biology. We would eventually like to perform data modelling at this level, using specialized sets of real-life parameters to extend knowledge discovery in our visualization to allow data clustering in a way that may reveal insights to high level processes in biological systems (i.e. cancer types, mental disease, intersections of disease conditions).

An added feature we wish to enhance that could help achieve this is a relation highlighting feature that allows a user to specify the degree of connection from an initial input (i.e. degree of 1 will highlight first degree connections to each initial object selected). This may help users gain understanding of targeted information on a particular object, process, or perhaps discover new patterns in the visualization.

In the beginning of the project, we used Gephi, an interactive visualization platform for graphs, to present our data. As we further developed our ideas, we realized that we could not solely rely on Gephi to show all the properties of the node and edges for our graph. Thus, we decided to build a more customizable visualization tool in the form of a web application in order to accommodate our needs. As a product of the web application, we were able to add features that didn't previously exist in the Gephi interface. Firstly, for each node of the graph, we added a descriptive side panel combining a short description obtained using the DBpedia API [9], a frequency histogram summarizing object verb occurrence, and hyperlinks to the original source of where the relations came from. Secondly, for each edge of the graph we've displayed comparative data on the two node types, summary data on verbs relating the two particular nodes in order of frequency, and again, hyperlinks to the original source.

The next stage in our research will comprise of more in-depth approaches in inference-based models for classifying trends in relations. As one of our top priorities, we intend to implement clustering machine learning methods (i.e. unsupervised learning) [12] for grouping common words and phrases in text mining and defining patterns in real-time as text streaming is occurring live. Such groups will allow determination of thematic elements, ontology extraction and text summarization.

Conclusion

We intend to run our software for another four weeks, gathering more precise statistics about its performance and accuracy; we also wish to increase our computing power, perhaps by allowing us to extract relations on multiple machines at once, using technology such as Apache Spark & Amazon Web Services.

References