Ion Channel ElectroPhysiology Ontology (ICEPO) – a case study of text mining assisted ontology development

Ravikumar Komandur Elayavilli, Ph.D.[1], Hongfang Liu, Ph.D.[1]
[1]Mayo Clinic, Rochester, Minnesota, USA

Abstract

Background
Computational modeling of biological cascades is of great interest to quantitative biologists. Biomedical text has been a rich source for quantitative information. Gathering quantitative parameters and values from biomedical text is one significant challenge in the early steps of computational modeling as it involves huge manual effort. While automatically extracting such quantitative information from bio-medical text may offer some relief, lack of ontological representation for a subdomain serves as impedance in normalizing textual extractions to a standard representation. This may render textual extractions less meaningful to the domain experts.

Methods
In this work, we propose a rule-based approach to automatically extract relations involving quantitative data from biomedical text describing ion channel electrophysiology. We further translated the quantitative assertions extracted through text mining to a formal representation that may help in constructing ontology for ion channel events using a rule based approach. We have developed Ion Channel ElectroPhysiology Ontology (ICEPO) by integrating the information represented in closely related ontologies such as, Cell Physiology Ontology (CPO), and Cardiac Electro Physiology Ontology (CPEO) and the knowledge provided by domain experts.

Results
The rule-based system achieved an overall F-measure of 68.93% in extracting the quantitative data assertions system on an independently annotated blind data set. We further made an initial attempt in formalizing the quantitative data assertions extracted from the biomedical text into a formal representation that offers potential to facilitate the integration of text mining into ontological workflow, a novel aspect of this study.

Conclusions
This work is a case study where we created a platform that provides formal interaction between ontology development and text mining. We have achieved partial success in extracting quantitative assertions from the biomedical text and formalizing them in ontological framework. ***Availability****: The ICEPO ontology is available for download at http://openbionlp.org/mutd/supplementarydata/ICEPO/ICEPO.owl*

Introduction

Driven by sudden surge in data, thanks to the recent spurt in the usage of high throughput technologies in research, scientists have shifted their focus from single gene studies to systems level approaches to understand biology. This has been potentiated by factors such as increased focus on studying the significance of quantitative measure in modeling biological pathways. Scientific literature has been a significant source of ever-growing repository of both qualitative and quantitative knowledge pertaining to biological pathways. On the other hand, biomedical ontologies have been popular in organizing biomedical knowledge from diverse resources including biomedical literature. However, it takes enormous time and manual effort in curating such knowledge resources. Text mining has the potential to bridge the gap between the knowledge resources and scientific literature.

Existing text mining studies are geared towards extracting the qualitative aspects of the biological events such as relationships between biological concepts and events. Recently there is a paradigm shift in biology from qualitative analysis to a more quantitative approach, as evident from emergence of the new discipline of quantitative systems biology. Obtaining kinetic parameters and their associated values directly from the literature for computational modeling is labor intensive, which involves huge manual effort and time. It has been a source of impedance in modeling cellular process. There have been considerable efforts to automatically extract quantitative data (e.g. units and their corresponding parameters) from the literature though largely in a limited context. Hakenberg et al., 2004 [1] proposed a support vector machine (SVM) based classifier to identify whether a full-text article contains kinetic data or not. However, their work did not include extracting the quantitative parameters (e.g. Kd, Vmax, IC50) and their corresponding values from the literature. On the other hand, KiPar [2], KID [3] and KIND [4] addressed the problem of kinetic data extraction from the literature in the context of enzymatic reactions and yeast metabolic networks. In a recent effort, Schomburg et al. 2013 [5] as part of BRENDA [6] database has extracted the enzymatic

kinetic parameters and its associated values through text mining for the complete PubMed database. There have been recent studies [7] to extract pharmacokinetic data in the context of drug-drug interactions from biomedical literature. Notable among them are the works of Wang et al., 2009 [8] and Wu et al., 2014 [9]. The authors besides developing an annotated corpus have developed a text mining to extract pharmacokinetic data from biomedical text and formalized the concepts into PK ontology. This system was further extended to extract pharmacokinetic data from full-text articles. This work serves as motivation of the current work for ion channel subdomain in biology.

Extracting quantitative information from the biological literature related to ion channel physiology is the main focus of the current work. Extracting quantitative data assertions from bio-medical text also offers certain natural language processing's challenges such as extracting information beyond clausal boundaries. While entity and event anaphora has been one kind of challenge, which been addressed extensively in earlier works [10-14], tailoring co-reference resolution to ion channel physiology sub-domain in addition to heuristic extra-clausal pairing of quantitative values, has been addressed in the current work. Besides, formalizing the electrophysiological concepts related to ion-channel physiology into a well-organized ontology is another aspect that is addressed in the current work. The biomedical domain has been leading the efforts in formalizing the intricacies of domain knowledge through the development of various resources such as semantic networks, terminologies, and ontologies [15]. In the past decade, there has been a serious effort to centralize and share the ontological resources. While, National Center for Biomedical Ontology (NCBO) [16] (http://bioontology.org) is an ideal platform for sharing biomedical ontologies it does not make any explicit recommendations for standards. The Open Biological and Biomedical ontologies (OBO) [17] foundry that provides a collaborative platform for developing ontologies in bio-medical domain do recommend certain standards to enable ontology more interoperable. Web-Protégé [18] is yet another recent effort in building ontologies in a collaborative framework. In parallel, there is an increase in recent efforts in building standards for representing quantitative biological knowledge such as Systems Biology Markup Language (SBML) [19], Systems Biology Graphical Notation (SBGN) [20] and Systems Biology Ontology (SBO) [21]. Hoehndorf et al., 2011 [22] has summarized the features of all the ontologies that exist in systems biology domain. There are few prior efforts to formalize the electrophysiological concepts of ion channel. CelO [23], SBO [21] and Gene ontology [24] have very preliminary representation of ion channel concepts in their ontologies. These resources do not cover the complete physiological landscape of the ion channels. Cardiac ElectroPhysiology ontology (CEPO) [25] and Cell physiology ontology (CPO) [26] partially attempts to represent the physiological processes related to ion channels. However, the relations defined by the concepts in these ontologies are far simpler than that exists in this domain. Hence, it is imperative to understand the electrophysiological concepts of ion channels and formalize them into a standard representation. Ion channel electrophysiology is centered on three important component namely, time, voltage and current. The relationships between the three are complex, which is not effectively reflected in the above-mentioned ontologies. Besides, the lag in the update of new concepts defined in the scientific literature has been a limiting factor in manual ontology development driven by domain experts. In this work, we attempt to overcome all the above-mentioned shortcoming by exploring a hybrid approach between manual and text-mining driven ontology development and curation.

In this study, we address three important issues:

1. Extract quantitative parameters and their associated values describing ion channel physiology from the biomedical literature using a rule based approach.
2. Provide a platform for interaction between text mining and ontology development by formalizing the quantitative assertions in ion channel electrophysiological events extracted from the literature into hierarchical relationships.
3. Define, create and integrate information from other ontology resources such as Cell Physiology Ontology (CPO) [26], Cardiac ElectroPhysiology Ontology (CEPO) [25] and Unit Ontology (UO) [27] that are relevant to ion channel electrophysiology.

Methods

Text Mining

Figure 1 outlines the system architecture of the text mining system that we have implemented in the current study. We developed a rule-based method approach [28] to extract relations between quantitative parameters, values, and their associated molecules and events from the literature related to ion channel electrophysiology subdomain. As a first step, we extract biological events and entities described in the biomedical text as described in Ravikumar et al., 2014 [10]. The following section briefly describes our approach for extracting the relations involving quantitative data.

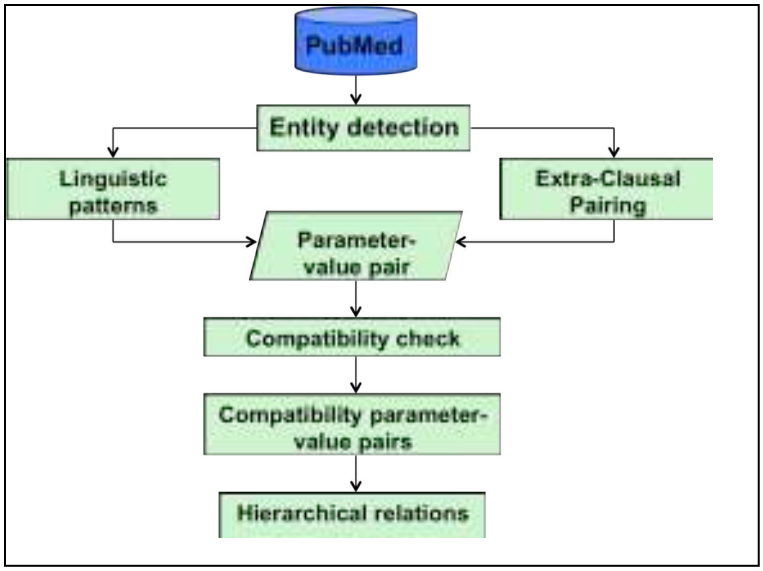

Figure 1. System architecture

Detection of quantitative parameters and values

We used dictionary lookup (e.g. conductance, membrane potential) and regular expressions (e.g., Kd, Ki) to identify electrophysiological parameters and detect their values (with units, e.g., 20 pS, -80 mV) as they occur in the text. While detection of units such as pS, mV, etc. is predominantly dictionary based, we use regular expressions to detect values along with units (e.g. [0-9]+ (pS|mV)). All the patterns for extracting relations involving quantitative data from biomedical abstracts are implemented in Perl. Consider the following sentences:

Example1: a) Measurement of the *conductance* of the *sodium channel* from current fluctuations at the *node of Ranvier* (Parameter=conductance, Protein=sodium channel, Tissue=node of Ranvier).

Example2: b) The *average gamma* from twelve measurements at *depolarizations* of *8-40 mV* was *7-9 +/- 0-9 pS* (S.E. of mean). (Parameter=*average gamma*, Parameter=*depolarizations*, Value=*8-40 mV* and Value=*7-9 +/- 0-9 pS*). The phrases that are italicized in the above two sentences were identified as entities and classified into respective categories as shown in paranthesis.

Extraction of relations involving quantitative data

The extraction of biological relations involving quantitative data includes extraction the relations i) within a single clause using patterns and ii) beyond clausal boundaries through compatible parameter-value pairing.

Pattern templates within a single clause - We use predefined template-filling model to extract relations involving quantitative data. Below, we briefly describe some of the template rules designed to extract kinetic parameters pertaining to channels and their associated values. The rules operate on a shallow parsed text where entities are already tagged and classified.

Pattern 1: **Parameter (PRP NP)* [be VP] [Value NP]**: the VP is headed by "be" verbs such as *is, was, are, and were*. PRP NP represents a prepositional phrase, which may occur zero or more times (as denoted by *). This pattern matches the sentence shown in Example2, to extract the relation between *average gamma* and *7-9 +/- 0-9 pS* as parameter and value respectively.

Pattern2: **[Parameter NP] around/of [Value NP]**: This pattern matches the sentence: "Noise power spectral densities were calculated in the ***frequency range* of *6-6-6757 Hz***" and pair frequency range with 6-6-6757 Hz. (Example3)

Pattern3: **[Parameter NP] = [Value NP]**: matches the sentence: "Increases in *sweat rate (DeltaSR)* were also significantly lower in grafted skin (***DeltaSR = 0.08 +/- 0.08 mg/cm/min***)" where the association between DeltaSR (Parameter) and 0.08 +/- 0.08 mg/cm/min (Value) are extracted. (Example4)

Pattern4: "**[Value Chemical NP]"** matches the sentence "External application of *150 nM* *tetrodotoxin (TTX)* and *10 mM* *tetraethylammonium (TEA)* ion." (Example5) and extracts <tetrodotoxin (TTX), 150nM> and <tetraethylammonium (TEA), 10mM> as chemical value pairs.

Table 1. Example of compatible parameter-unit filters for Parameter value association

S.No	Parameter	Compatible units
1	Conductance	pico Siemens, pS
2	Current	Amperes, pA, pA/pF
3	Chemical	nM, mM, micro M
4	Potential/Voltage	mV, milli volts
5	Hill, Open probability	Constant (without units)

Compatible pairing rules beyond clause boundaries - The patterns that we described in the previous section extract relations only within a single clause. For example, in the following sentence "The *single channel conductance* for *this channel* was approximately *20 pS* with *140 mM Na(+)* , *K(+)* , or *Cs(+)* in the patch pipettes **and** was approximately 13 pS with 100 mM Ca(2+) or Ba(2+) in the patch pipettes.", the parameter single channel conductance takes two values, one within the clause (20pS) and one beyond the clause (13 pS). While Pattern 1 associates the first value "*20pS*" with "*single channel conductance*", we use rules to pair "*single channel conductance*" with "*13 pS*" that occur beyond the co-ordination clause ("and"). The rules are further constrained by compatibility assessment between the pairs. Table 1 gives some of the compatible parameter-value units that are allowed in extra-clausal pairing. If more than one association is possible, then the closest compatible pair is selected. We also use compatibility rules to validate the associations extracted by the patterns described in an earlier section in order to filter incorrect associations.

Extraction of quantitative assertions as hierarchical relationships from biomedical text

Finally, we express the relations involving quantitative data extracted by the system in a way that can potentially lead to extraction of ontological relationships conveyed in a single abstract or article. Laurila et al., 2010 [29] has addressed this problem in a very limited context of extracting mutation impact from biomedical text. In this study, we summarize the relations across sentences from the biomedical abstracts into semantic assertions. For example, consider the two sentences from the abstract (PMID-10482751 shown in Figure 2). The semantic parser rule **"[Protein NP] [VP-PASS] [Parameter/Activity] of [Value]**" captures the assertive relation between the three entities "BK channels" (Protein) "conductance" (Parameter) and "223pS" (Unit-Value) in the first sentence. The system formalizes the extractions into three statements namely ("BK channels" hasProperty "Conductance") ("Conductance" hasUnit "pS") (Conductance hasValue "223"). The semantic constraints on the entity type in these rules help identify such assertions. In the second sentence while the parameter/activity ("*potassium permeability (PK)*") is associated with the value ("2.3 x 10(-13) cm(3) s(-1)") through simple rule **"[Parameter/Activity NP] (was/is/are/were)? [Value NP]**", the association of parameter (potassium permeability (PK)) with the "These channels" through extra clausal pairing. The anaphoric phrase "These channels" is resolved to "BK channels" based on the head noun "channels" and the semantic type of both entities. Hence, a clear relation between the protein ("*BK channels*"), parameter ("*potassium permeability (PK)")* and value ("*2.3 x 10(-13) cm(3) s(-1)").* Multiple assertions derived from the two sentences (shown in Figure 2) have a common node "*BK channels*". These assertions are subsequently rendered as hierarchical relation as shown in Figure 2.

Building Ion channel electrophysiology ontology (ICEPO)

The ion channel electrophysiology ontology is an outcome based on three approaches: 1) anecdotal and domain knowledge of the authors, 2) relations extracted from the biomedical text both abstracts and full-text articles, and 3) integration of vocabularies from other existing ontologies such as CEPO, CPO and Unit Ontology. We used Protégé Version 5.1 [30] to create the ontology. As a first step, we created top-level classes as instances of "*Thing*" shown in Table3. The descendants of these top-level classes contain concepts that have "is-a" relationship with their ancestors. Subsequently we defined the relationships between various class elements. Our approach to build ICEPO was a top down where we first outlined the most general nodes and subsequently added the descendant nodes. While the organization of the seed ontology consisting of high-level classes was based more on domain knowledge, some of the terms belonging to the leaf nodes were identified through text annotation and literature mining system. Moreover, the relations between the channels, its properties such as conductance, open probability and the technique used to characterize them were organized based on relations extracted by the text mining system. We followed the broad principles of OBO Foundry to better facilitate interoperability between other ontologies. We imported terms

from existing closely related ontologies namely CPO, CEPO, and Unit Ontology (UO). We used OntoFox [31] to extract selective classes and its instances of CPO, CEPO and UO, which we imported into ICEPO in Protégé.

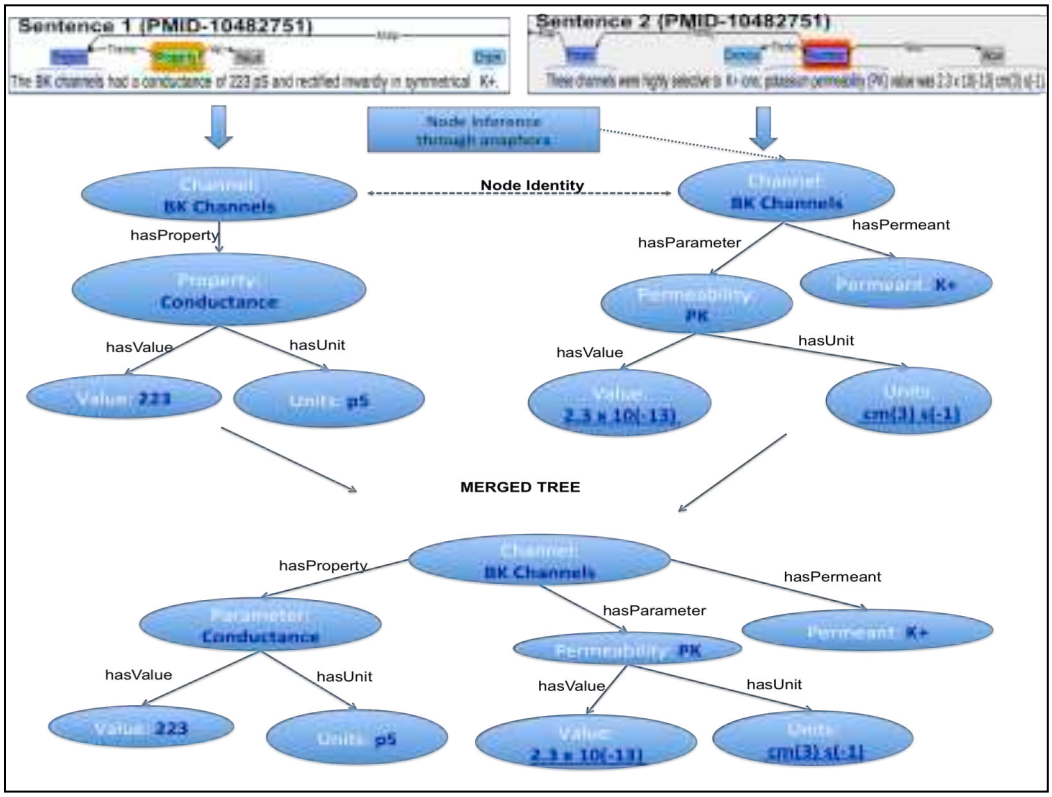

Figure 2. Mapping textual extraction to hierarchical representation

Results and Discussion

Text Mining

Data set

In order to the rules for entity and relation extraction, we manually annotated a development corpus for relations involving quantitative data, consisting of 180 biomedical abstracts and 5 journal articles related to ion channel electrophysiology. To evaluate the performance of our system, we used the CheQK corpus [32] as the blind test set, which consists of 105 Medline abstracts predominantly describing events related to channel proteins from the inward rectifying potassium channel (Kir) family. The annotation guidelines of the CheQK corpus were different from the one followed in the annotation of the development corpus the theme of annotation between them are common (relations involving quantitative data of ion channels). The development corpus contained 1,687 relations in total out of which 856 are relations involving quantitative data accounting for nearly 45% of the relations. The CheQK corpus consists of 1187 events in total out of which 755 are quantitative in nature. The CheQK corpus contains 5 different types of relations involving quantitative data namely, Reaction Parameters (ReactionP), Activity, Property, PhysProp and Comparison. We used the standard metrics of precision, recall and F-measure for the evaluation. While precision is a measure of accuracy (Total Correct/Total Extracted), recall is a measure of sensitivity of the system (Total correct/Total annotated). F-measure is harmonic mean of precision and recall.

Evaluation

Table 2 shows the performance of the relation extraction system against test corpus. In this study, we evaluated only the ability to extract the relations involving quantitative data.

The corpus had 154 ReactionP, 540 Activity, 44 Property, 4 PhysProp and 13 Comparison relations. The system extracted 687 relations involving quantitative data across all categories out of which only 497 were found to be correct leading to an overall precision, recall and F-measure to be 72.34%, 65.83% and 68.93% respectively. The

system achieved the highest performance (F-measure 71.60%) in the "Property" relation type and the lowest in PhysProp relations. This may also be partially due to very low number of "PhysProp" relations in the gold standard. The property is the simplest type of relations among the relations involving quantitative data, which may be due to the reason behind the high performance of the system. The performance in the "Activity" relations particularly the recall was very low despite they being very simple relations.

Table 2. Evaluation of relations involving quantitative data extraction on a blind data set

Relation Type	Total	Total Extracted (Total Correct)	Precision (%)	Recall (%)	F-Measure (%)
ReactionP	154	132 (93)	70.45	60.39	65.03
Activity	540	497 (364)	73.24	67.41	70.20
Property	44	37 (29)	78.38	65.91	71.60
PhysProp	4	3 (2)	66.67	50.00	57.14
Comparison	13	18 (9)	50.00	69.23	58.06
Total	755	687 (497)	72.34	65.83	68.93

Ion Channel Electro Physiology Ontology (ICEPO)

Ontology structure and content

Figure 3 gives the overall structure of ICEPO. The terms in ICEPO are distributed in 6 orthogonal classes described below.

We briefly discuss the classes and the relations between instances in our ontology. We have 19 simple relations in total, which are general to any other entity classes and certain relations that are very specific to ion channel events such as "hasPermeant", "hasGating" and "hasNoise". We also have certain relations such as "is inhibitor of" which are common to other sub-domains in biology.

1) Molecular entities: This class consists of three subclasses described below.

a. *Channel proteins* are broadly categorized based on the ions that flow through them and subsequently sub-categorized based on their gating. The terms in ICEPO has been integrated with CEPO. CEPO had "IonChannel" as an instance of Biomolecule, which include other categories such as "IonTransporter", Gene etc. However CEPO do not have "InterCellular channels" such as "Connexins" and neutral channels like "Aquaporins". Besides the Ion channels were not further categorized based on the type of the charge and the valence of the ions. We introduced a new sub-class "Channel_protein" under Biomolecule and made "IonChannel" of CPO as its child and "Neutral channel" and "Intercellular_channel" as its siblings. Under "IonChannel" we introduced new sub-classes to classify ion channels also based on the ion charge type and the valence of the ions it allows to permeate through them.
b. *Small molecules* – This category includes terms from category such as ions, small molecules that are used to block ion channels. PermeantIon listed as a subclass of Permeant is integrated from CEPO. However the individual ions are placed under new sub-classes namely Anion and "Cation" in ICEPO. However our list of inhibitors is very little and hence didn't consider integration with external standard resources such as ChEBI. In future, we plan to have additional terms of class drugs and inhibitors, in ICEPO through integration with ChEBI.
c. *Cell* – We have very limited category of cell types. Our Concept of 'Cell" is too preliminary and restricted to only four types of cells in the human body. We plan to extend the scope of concept "cell" through integration of well-known ontologies such as Cell Type Ontology (CTO) [33].

2) Parameter/Property – This class includes many sub-classes pertaining to properties/parameter of cell, properties/parameter of molecular entities (channel proteins), general properties common to both cell and molecular entities such as dimension, parameters pertaining to channel events such as inhibition constant, thermodynamic variables. Among these classes concentration and currents carried by individual channels (an instance of "Single channel current") are imported from Cell physiology ontology. Parameters often play a critical role while modeling ion channel electro-physiology. The various types of single channel current provide an ideal platform for integration with Cell physiology ontology. The "current by ion channel" class of CPO has been linked to respective classes under "Single channel current" class of ICEPO. The only difference being "ICEPO" attempts to categorize ion channels based on the ion that flows through the channel.

3) Unit - Terms related to units associated with parameters/properties of channels/molecular entities/cell are imported from Unit Ontology (UO).

4) Process - This class predominantly consists of terms involving processes such as gating, transport, state transitions etc. related to ion channel physiology, besides having terms related to experimental techniques such as patch clamping. In this category, all the experimental techniques related terms were imported from CEPO.

a. Gating – Gating refers to various physical states of ion channels during their activation, deactivation and inactivation.
b. State transitions – State transitions refers to the transition between two different states of ion channels. Typically in a two stage transition models we have two configurations of ion channels namely open and closed states.
c. Transport – This concept involve the different types of movement of ions and other Permeant molecules across cell membrane due to ion channels.
d. Techniques – It includes all experimental techniques employed in electrophysiological studies to characterize the function of ion channels. This category includes various sub-categories such as voltage-clamp, current clamp techniques including the configuration of the patch. All the terms described under this category are imported from CEPO.
e. Binding – Binding refers to a process of binding of a ligand, ion or small molecule to binding sites in ion channel to facilitate or block the activity of an ion channel.
f. Permeation – Process that allows transport of ions through the ion channels.
g. Activation is a process through which ion channels are activated, which may be due to the conformational change thereby allowing ions to pass through the channel.
h. Inactivation refers to the closing/inactivation of ion channels, i.e. the conformational change that prevents the passage of ions through them.
i. Reactivation – Process by which the ion channels returns to an activated state from the inactivation state.
j. Transport is the various mechanisms of transport of ions/molecules across cell membrane.

5) Noise – Class that contains terms that describe the random fluctuations while recording the signals of ion channel events. They fall into three broad categories: i) Channel configuration noise ii) Current noise and iii) Conductance noise

6) Mathematical representation - This class consists of various mathematical expressions that are used to model ion channel events. Mathematical expression is categorized as a separate class. Mathematical representation consists of two categories.

i) Equations – That are used to calculate parameters/properties of ion channels

a. Boltzmann equation is used to calculate the equilibrium potential of an ion species that permeates through channel membranes.
b. Chord conductance equation is used to calculate the cell membrane potential during the relative conductance due to the flow of all the ions across cell membrane.
c. *GHK flux equation* also known as "Goldman–Hodgkin–Katz flux equation" describes the flux due to flow of ions as a function of trans-membrane potential.
d. Goldman equation also known as "Goldman–Hodgkin–Katz voltage equation" determines the reversal potential due to flow of ions.
e. Langevin equation also known as "Schrodinger-Langevin equation" can be constructed for the case of ionic diffusion along potassium ion channels.
f. Nernst equation describes the balance between the two gradients namely "electrical" and "concentration" in electrically excitable cells.

ii) Mathematical models – Models that define the behavior of the function of ion channels. The major classes under this category

a. Morris–Lecar model describes the oscillatory behavior in neurons.
b. FitzHugh–Nagumo model also describe oscillatory behavior of electrically excitable cells.
c. Hindmarsh–Rose model describes the spike behavior in action potential in neurons.

Hybrid approach to building ontology

Our approach to building ontology for ion channel physiology is a hybrid one. The quantitative assertions that we extracted from the text are individual instances (e.g. “BK Channels, “Conductance”) of ontology. However the relationship between the instances represent a generic relationship between two nodes in a typical ontology. Figure 2 represents only very simple assertive relation and not a complex one. It is crucial to make abstractions from the individual instances to infer more generic relationships. Such inferences will transform textual extractions such as

("BK Channels" hasProperty "Conductance") to "("Potassium channel" hasProperty "Conductance") since "BK channel is a sub-type of "Potassium channel", which can be further generalized to ("Ion channel" hasProperty "Conductance"). This is achieved by taking advantage of the "is a" relationship outlined by the expert knowledge. While text-mining systems can be used to populate ontologies with instances (usually leaf nodes) and object relationships between them, the domain experts may focus in verifying the organization of ontology. Supplementing the efforts of domain experts with text mining will further drive inference by effectively combining the two approaches. This study is a humble beginning in that direction. We used clues from the quantitative data assertions during our ontology development stage, which helped significantly in defining object relationships during formal representation of ion channel electrophysiology relations. Complete integration of text-mining system into an ontological workflow is possible only when text-mining system extractions mature and their representation standards are capable enough to make complex layers of generalizations. However, we wish to say that knowledge of domain expert has played a dominant role and text mining has played only a limited role in the development of ICEPO ontology.

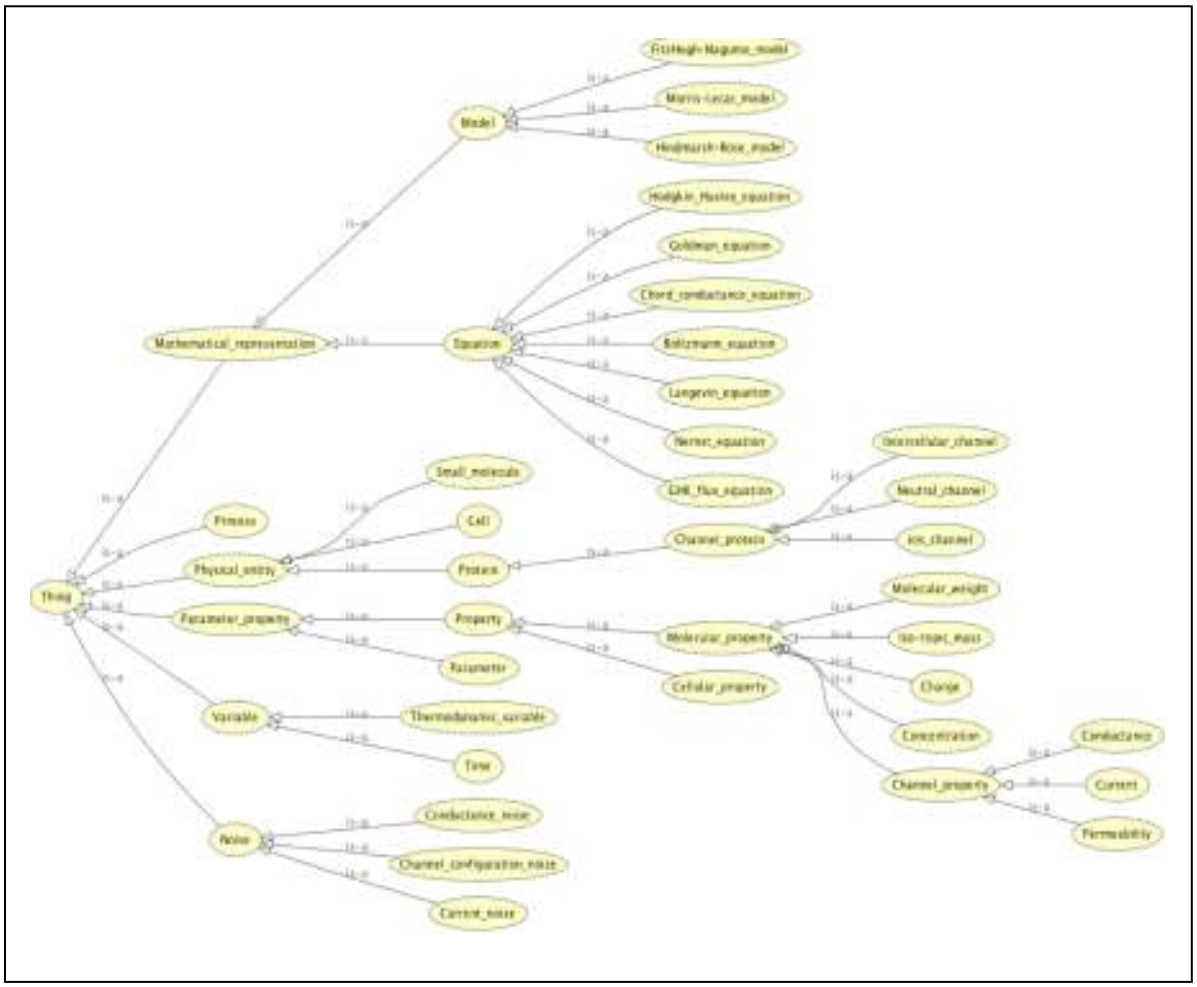

Figure 3. Graph of Ion channel ElectroPhysiology Ontology (ICEPO)

Conclusions

In this work, we focused on three important aspects 1) Extraction of quantitative information (parameters and their associated values) and formalized quantitative data assertions, which reflect our ICEPO ontological definitions from the literature on ion channel electrophysiology. 2) Formalizing all the concepts in ion channel electrophysiology into ontology (ICEPO) modeled as per the OWL definitions. 3) Integration of the ontological definitions in known existing ontologies such as CEPO and Unit Ontology with ICEPO.

The rule-based approach to extract parameter-value pairs both within the clause and beyond clausal boundaries has yielded the state of the art results in this domain. While there is no novel contribution to text mining in terms of methodology, tailoring the existing approaches to achieve good performance for quantitative assertion extraction for this sub-domain is a very important and needed one. The results point to the fact that while our pattern templates are extremely precise in extracting quantitative assertions within the clause, the rules for extra-clausal pairing and co-reference resolution play a significant role in boosting the recall. We have also improved upon the state of the art of text mining by formalizing the textual extractions into semantically meaningful assertions that may eventually help in formalizing them as ontological structures. We attempted to link the assertions from multiple into a hierarchical relation from a single biomedical abstract.

Formalization of electrophysiological concepts of ion channels in an ontological framework is one of the significant contributions of this work. The role of text mining system in the identification of certain leaf nodes and relationships between the classes further allow interaction of text mining with ontology development. The design of the basic framework of ICEPO ontology facilitated the smooth integration of other relevant ontological resources such as CPO, CEPO and UO. Integration of related ontologies in this sub-domain is another contribution to this work. We believe that the basic ontological framework that we proposed in this study can be generalized further to integrate other related ontologies and knowledge sources such as SBO and BioModels [34] and Reactome [35] databases.

Limitations and Future Work

One of the limitations of the current work is its focus on a narrow domain of ion channel physiology. From text mining perspective, we envisage the following as its limitation: 1) approach that we used to extract quantitative assertions from the bio-medical text is rule-based. 2) The blind data set that we used for evaluation in this study consists of only biomedical abstracts and not full-text articles. 3) Full-text articles are known to contain more quantitative data assertions than biomedical abstracts. 4) Another limitation of our text mining system is the scope of its extraction boundary to a single document. 5) We did not formally evaluate the utility of the quantitative assertions that we extracted from the text in the ontology development. The ontology that we have developed takes diversified scientific articles in this domain into consideration. The ICEPO ontology reflects synthesis of knowledge from multiple articles manually and other closely related ontologies. From ontology perspective, the structure of ICEPO may require further organization in order to facilitate its integration with other ontologies such as SBML, CellML and Gene Ontology.

We plan to address some of the limitations in the near future. Though, the scope of ICEPO is very limited to ion channel electrophysiology, care has been taken while designing ICEPO to ensure smooth integration with widely related bio-medical ontologies. We plan to integrate ICEPO with other related ontologies such as SBO, SBML, CellML and other generic ontologies such as Gene Ontology, ChEBI, Cell Ontology once the scope of ICEPO widens. We plan to extend the text mining solution for processing full-text articles.

Supplementary information: All the supplementary information related to the work is available at http://openbionlp.org/mutd/supplementarydata/ICEPO/

Acknowledgements

The authors acknowledge that the study was supported by two grants: National Science Foundation ABI:0845523 and National Library of Medicine R01LM009959 grants. We also thank the intramural support from Mayo Center of Individualized Medicine (CIM).

References

1. Hakenberg J, Schmeier S, Kowald A, Klipp E, Leser U: Finding kinetic parameters using text mining. *Omics: a journal of integrative biology* 2004, **8**(2):131-152.
2. Spasi I, Simeonidis E, Messiha HL, Paton NW, Kell DB: KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways. *Bioinformatics* 2009, **25**(11):1404.
3. Heinen S, Thielen B, Schomburg D: KID- an algorithm for fast and efficient text mining used to automatically generate a database containing kinetic information of enzymes. *BMC bioinformatics* 2010, **11**(1):375.
4. Tsay JJ, Wu BL, Hsieh CC: Automatic extraction of kinetic information from biochemical literatures. In*: 2009*. IEEE: 28-32.
5. Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A: BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic acids Research* 2012:gks1049.
6. Schomburg I, Chang A, Hofmann O, Ebeling C, Ehrentreich F, Schomburg D: BRENDA: a resource for enzyme data and metabolic information. *Trends in biochemical sciences* 2002, **27**(1):54-56.

7. Kolchinsky A, Lourenco A, Wu H-Y, Li L, Rocha LM: Extraction of Pharmacokinetic Evidence of Drug-drug Interactions from the Literature. *arXiv preprint arXiv:14120744* 2014.
8. Wang Z: Biomedical literature mining for pharmacokinetics numerical parameter collection. 2013.
9. Wu H-Y, Karnik S, Subhadarshini A, Wang Z, Philips S, Han X, Chiang C, Liu L, Boustani M, Rocha LM: An integrated pharmacokinetics ontology and corpus for text mining. *BMC bioinformatics* 2013, **14**(1):35.
10. Ravikumar KE, Wagholikar KB, Liu H: Towards pathway curation through literature mining-a case study using PharmGKB. In: *Pacific Symposium on Biocomputing: 2014*. 352-363.
11. Kim J-D, Nguyen N, Wang Y, Tsujii Ji, Takagi T, Yonezawa A: **T**he genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC bioinformatics* 2012, **13**(Suppl 11):S1.
12. D'Souza J, Ng V: Anaphora resolution in biomedical literature: a hybrid approach. In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine: 2012*. ACM: 113-122.
13. Lin Y-H, Liang T, Hsinehu T: Pronominal and Sortal Anaphora Resolution for Biomedical Literature. In: *Rocling: 2004*.
14. Choi M, Verspoor K, Zobel J: Evaluation of coreference resolution for biomedical text. In: *Proceedings of the SIGIR workshop on Medical Information Retrieval (MEDIR 2014): 2014*.
15. Freitas F, Schulz S, Moraes E: Survey of current terminologies and ontologies in biology and medicine. *RECIIS—Electronic Journal in Communication, Information and Innovation in Health* 2009, **3**(1):7-18.
16. BioPortal N: Welcome to the NCBO BioPortal. In.; 2010.
17. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 2007, **25**(11):1251-1255.
18. Tudorache T, Vendetti J, Noy NF: Web-Protege: A Lightweight OWL Ontology Editor for the Web. In: *Owled: 2008*.
19. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003, **19**(4):524-531.
20. Le Novere N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM: The systems biology graphical notation. *Nature biotechnology* 2009, **27**(8):735-741.
21. Juty N, le Novère N: Systems Biology Ontology. *Encyclopedia of Systems Biology* 2013:2063-2063.
22. Hoehndorf R, Dumontier M, Oellrich A, Rebholz-Schuhmann D, Schofield PN, Gkoutos GV: Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLoS One* 2011, **6**(7):e22006.
23. Matos EE, Campos F, Braga R, Palazzi D: CelOWS: An ontology based framework for the provision of semantic web services related to biological models. *Journal of biomedical informatics* 2010, **43**(1):125-136.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: Gene Ontology: tool for the unification of biology. *Nature genetics* 2000, **25**(1):25-29.
25. [http://bioportal.bioontology.org/ontologies/EP]
26. Cell physiology ontology [http://cybow.astem.or.jp/cpo/]
27. Gkoutos GV, Schofield PN, Hoehndorf R: The Units Ontology: a tool for integrating units of measurement in science. *Database* 2012, **2012**:bas033.
28. Ravikumar, K.E., Wagholikar, K.B. and Liu, H. (2014) "Automatic extraction of quantitative relations describing ion channel physiology from bio-medical literature" Proceedings of 17th Annual Bio-Ontologies workshop, July 11-12th, 2014, Boston, MA, USA.
29. Laurila JB, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker CJO: Algorithms and semantic infrastructure for mutation impact extraction and grounding. *BMC genomics* 2010, **11**(Suppl 4):S24.
30. Noy NF, Sintek M, Decker S, Crubézy M, Fergerson RW, Musen MA: Creating semantic web contents with protege-2000. *IEEE intelligent systems* 2001, **16**(2):60-71.
31. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y: OntoFox: web-based support for ontology reuse. *BMC research notes* 2010, **3**(1):175.
32. The CheQK Corpus [http://relagent.com/drafts/CheQK_v0.1.tgz]
33. Bard J, Rhee SY, Ashburner M: An ontology for cell types. *Genome biology* 2005, **6**(2):R21.
34. Chelliah V, Laibe C, Novère NL: BioModels database: a repository of mathematical models of biological processes. *Encyclopedia of Systems Biology* 2013:134-138.
35. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR: The Reactome pathway knowledgebase. *Nucleic acids research* 2014, **42**(D1):D472-D477.