UCSC's System for CLEF eHealth 2013 Task 1

Chunye Wang, Ramakrishna Akella

School of Engineering
University of California Santa Cruz
Santa Cruz, CA 95064, USA
{cwang,akella}@soe.ucsc.edu

Abstract. CLEF eHealth 2013 Task 1 requires participants to perform named entity recognition and normalization of disorder mentions from clinical reports, where two important questions need to be addressed: (a) discovering mentions of concepts that belong to the UMLS semantic group Disorders, and (b) mapping each mention to a SNOMED-CT concept represented by a Concept Unique Identifier (CUI). The disorder mentions here could be a single text span (e.g. "loss of consciousness") or multiple text spans (e.g. "right ventricular" ... "dilated"). The corresponding concepts are usually encoded in SNOMED-CT, but sometimes may not be formally defined. To tackle these challenges we designed a two-stage annotation system, where MetaMap serves as the first-stage annotator for the identification of phrases with potential interest, and a Rule-based annotator works on the second stage for fine-grained supplementation.

MetaMap has mature technical features and relatively good performance on medical text analysis. It is developed to link the text of medical documents to the knowledge embedded in UMLS Metathesaurus. Highly configurable on semantic types, it enables us to specify the output concepts of interest. However, MetaMap is incapable of mapping text to concepts undefined, which, unfortunately, is the case for a number of disorder mentions in the task. Thus, we proposed a rule-based approach as the second-stage annotator. The annotation rules are learned from errors MetaMap made on training data, and could successfully recognize those undefined concepts. We also proposed Normalization and Post-processing algorithms to normalize and prune the intermediate results for better matching.

The experiments on training data demonstrate the effectiveness of every system component. MetaMap fails on pinpoint identification, but has certain capability to parse and roughly recognize the phrases of interest. Once combined with normalization, it could attain 0.463 F-score on training data. Designed to correct false negative errors, the individual Rule-based annotator is able to identify up to 15% all true annotations. The entire system eventually achieves 0.68 F-score in Task 1a and 0.57 accuracy in Task 1b. From the final competition results, our system performs consistently on test data, and beats all other participating systems in the group with additional annotations.

Keywords: clinical notes, MetaMap, annotation rules, rule-based annotation, CUI, SNOMED-CT, UMLS, evaluation, NLP

adfa, p. 1, 2011.

1 Introduction to Task 1

Clinical reports, such as discharge summary, radiology reports, echocardiogram reports and electrocardiograph reports, are abundant in mentions of clinical conditions, anatomical sites, medications, and procedures, which is in stark contrast with the newswire domain where text is dominated by mentions of countries, locations and people. Many surface forms are representations of the same concept. Unlike the general domain, in healthcare area there are rich lexical and ontological resources that can be leveraged when building applications. The Unified Medical Language System[1,] (UMLS) represents over 130 lexicons/thesauri with terms from a variety of languages. The UMLS Metathesaurus integrates resources used worldwide in clinical care, public health, and epidemiology, including SNOMED-CT[2], ICD-9[3], and RxNORM[4]. In addition, the UMLS also provides a semantic network in which every concept in the Metathesaurus is represented by its Concept Unique Identifier (CUI) and is semantically typed [1].

Because the recognition and normalization of named entity mentions is a fundamental task, it becomes the focus of CLEF eHealth 2013 Task 1 [2]. Task 1 includes the identification of mentions of concepts that belong to the UMLS semantic group Disorders and the mapping from each mention to a unique UMLS/SNOMED-CT CUI. Here are a few examples:

1. The rhythm appears to be atrial fibrillation.
   "atrial fibrillation" is a mention of type Disorders with CUI C0004238. UMLS preferred term is "atrial fibrillation".
2. The left atrium is moderately dilated.
   "left atrium.... dilated" is a mention of type Disorders with CUI C0344720. UMLS preferred term is "left atrial dilatation".
3. 53 year old man s/p fall from ladder.
   "fall from ladder" is a mention of type Disorders with CUI C0337212. UMLS preferred term is "accidental fall from ladder".
4. The patient was admitted with low blood pressure.
   "low blood pressure" is a Finding in UMLS, and as such does not belong to the definition of the Disorder semantic group. In this case, however, because it does indeed describe a disorder, it should be annotated. The CUI is left empty as "CUI-less".

Example 1 above represents the easiest cases. Example 2 represents mentions that are disjoint. Example 3 is a synonym of the UMLS preferred term. Example 4 represents mentions that have no corresponding mapping concepts in UMLS.

---

1 https://uts.nlm.nih.gov/home.html
2 http://www.ihtsdo.org/snomed-ct/
3 http://www.who.int/classifications/icd/en/
4 http://www.nlm.nih.gov/research/umls/rxnorm/

The scope of current task is limited to clinical reports written in English language, with the normalization/mapping to SNOMED-CT CUIs in ULMS version 2011AA. Illustrated by the above examples, Task 1 requires us to solve two problems: (a) discovering the boundaries of disorder mentions, and (b) mapping each mention to a SNOMED-CT concept. The system output should contain both boundaries and CUIs information. In light of running the given evaluation code, every disorder annotation should follow the format below.

report name || annotation type || cui || char start || char end
00176-102920-ECHO_REPORT.txt||Disease_Disorder||C0031039||120||140

If the annotation contains disjoint spans (i.e., non-contiguous spans, such as in the sentence "Abdomen: no distention is noted." in which the single annotation for "abdominal distention, C0235698" encompasses the span 0-6 (abdomen) and 13-22 (distention)), additional character start and character end values of every following span will be appended to those of the first.

00176-102920-ECHO_REPORT.txt||Disease_Disorder||C0344720||430||441||456||463

2 System Pipeline and Approach

We designed a two-stage annotation system to tackle Task 1 (see Fig.1). At the first stage, MetaMap is employed to parse clinical reports and annotate disorder mentions. However, constrained by the accuracy of MetaMap and the special usage of some terms (e.g. Example 4), not all disorder mentions can be precisely recognized. Especially, all words/phrases that should be mapped to a "CUI-less" concept will not be annotated by MetaMap. Therefore, we proposed a second-stage rule-based annotation in our system to supplement MetaMap. Annotation rules are learned from two types of errors made by MetaMap on training data. Besides, for better matching, the intermediate results are normalized and pruned in Normalization and Post-processing steps. The functionality of system components will be detailed below individually.

2.1 MetaMap Annotation

We chose MetaMap as the first-stage annotator for three reasons. Firstly, MetaMap has mature technical features and relatively good performance on medical text analysis. It is developed to link the text of medical documents to the knowledge embedded in UMLS Metathesaurus. MetaMap employs a knowledge-intensive approach, NLP, and computational-linguistic techniques [9]. Its lexical/syntactic analysis functionalities, such as sentence boundary determination, POS tagging, acronym/abbreviation identification, shallow parsing and word sense disambiguation, cater to the needs of Task 1.

Secondly, because disorder mention is defined as span(s) of text that belongs to the Disorder semantic group, we need to limit the scope of annotation on text from that semantic group only. MetaMap is highly configurable on semantic types of concepts and thus enables us to specify the output of interest. To be specific, we restricted the annotation from one of the following ULMS semantic types:

- Congenital Abnormality
- Injury or Poisoning
- Disease or Syndrome
- Cell or Molecular Dysfunction
- Anatomical Abnormality
- Signs and Symptoms
- Acquired Abnormality
- Pathologic Function
- Mental or Behavioral Dysfunction
- Experimental Model of Disease
- Neoplastic Process

Thirdly, MetaMap has been broadly used in industry and academia. It can be treated as a benchmark and foundation to compare different algorithms and fulfill advanced analytics built over it.

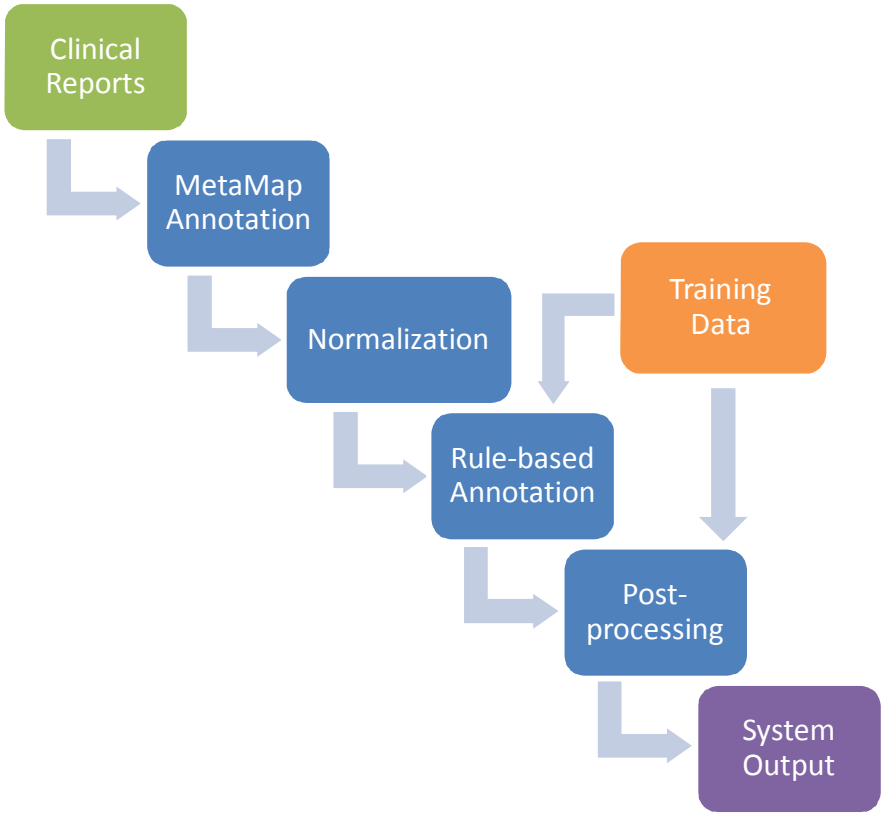

Fig. 1. System Pipeline

2.2 Normalization

The output of MetaMap is the phrase from shallow parsing with its corresponding concept in one of semantic types specified above. Table 1 shows three records extracted from MetaMap output. Record 1 is a perfect matching because the phrase string and concept string are identical. Then the boundaries of this disorder mention can be determined directly from the position of the phrase "Pericardial effusion" in the original report. Record 2 and 3, however, are cases that only part of the phrase string matches its concept string. In such scenario we need to normalize the phrase by virtue of the concept. Basically we only keep the words that appear in both phrase and concept. Thus, "left ventricular hypertrophy" and "SOB" will be the new annotation in Record 2 and 3, respectively.

Table 1. Examples of MetaMap Output

	Phrase from MetaMap	Concept in SNOMED	Semantic Type	CUI
1	Pericardial effusion	Pericardial effusion	Disease or Syndrome	C0031039
2	severe symmetric left ventricular hypertrophy	Left Ventricular Hypertrophy	Disease or Syndrome	C0149721
3	with increased SOB	SOB (Dyspnea)	Sign or Symptom	C0013404

2.3 Rule-based Annotation

MetaMap is capable of mapping text to existing concepts encoded in SNOMED-CT, but fails if corresponding concepts are not formally defined. Unfortunately, a portion of disorder mentions in Task 1 belong to the latter case, which enlightened us to propose a second-stage annotation for improved performance. We focused on the false negative errors, the true annotations missed by MetaMap, by comparing its output with the gold standards of training data, and then created corresponding rules to match the words/phrases of interest. To avoid overfitting and tune system performance, after applying a rule, we inspected the ratio of the size of its true annotations over the size of its false annotations, and set a threshold θ to control whether this rule should be included or not.

$$\frac{\#(true\ annotations\ by\ the\ rule)+1}{\#(false\ annotations\ by\ the\ rule)+1} \geq \theta \quad (1)$$

Basically the proposed annotation method is composed of three types of rules with regard to the style of required regular expressions, which are encoded to catch strings with certain patterns.

Single Span. The majority of disorder mentions are expressed in a single text span, such as Example 1, 3, 4 in Section 1. Thus the annotation rule can be written in regular expression simply using the text itself (see Table 2).

Table 2. Examples for Rules for Single Span

	Text Span	Concept	Regular Expression	CUI
1	cardiomegaly	Cardiomegaly	/cardiomegaly/i	CUI-less
2	free air	Pneumoperitoneum	/free air/i	C0032320
3	systolic murmur	Systolic murmur	/systolic murmur/i	CUI-less

Multiple Spans. A small number of disorder mentions contain two or more text spans, such as Example (2) in Section 1. Thus the annotation rule needs a generic expression to capture all variants, which is usually achieved by using metacharacter in regular expression (see Table 3).

Table 3. Examples of Rules for Multiple Spans

	Text Spans	Concept	Regular Expression	CUI
1	ascending aorta...dilated	Ascending aorta dilatation	/ascending aorta.*dilat /i	C0345049
2	tricuspid...regurgitation	Tricuspid valve regurgitation	/tricuspid.*regurgitation/i	C0040961
3	mitral...leaflets...thickened	Thickened mitral leaflet	/mitral.*leaflet.*thickened/i	C3164530

Acronyms and Abbreviations. Acronyms and abbreviations are used extensively in clinical notes. They are convenient shorthands in writing records, instructions, and prescriptions, and space-saving devices. Efforts have been made to standardize the form of them in some journals and books, but they generally vary from person to person. Learned from training data, frequent acronyms and abbreviations are linked to theirs CUIs by matching the entire word (see Table 4). We add the anchor meta-character "\b" in regular expression to match the word boundary, so that, for example, only the word "MR" will be matched by "/\bMR\b/", instead of the word "COMRADE".

Table 4. Examples of Rules for Acronyms and Abbreviations

	Acronym/Abbr.	Concept	Regular Expression	CUI
1	MR	Mitral valve regurgitation	/\bMR\b/	C0026266
2	JVD	Jugular venous distention	/\bJVD\b/	CUI-less
3	PNA	Pneumonia	/\bPNA\b/	C0032285

2.4 Post-processing

Similar to the normalization of the MetaMap output, phrases identified in Rule-based annotation step also need refinement. The normalization in post-processing removes stopwords, quantitative values, and descriptive words from annotations, such as "any", "severe" and "obvious". The final step is annotation pruning, which reduces

the false positive errors by filtering out phrases that match any rules on a blacklist. This blacklist is learned from training data set by analyzing the false positive annotations given by MetaMap.

3 Evaluation and Analysis

Task 1 provides a training data set and a withheld test data set. The training data contain 199 clinical reports with 5238 disorder mentions annotated, where rules are learned and parameters are tuned The test data contain another 100 reports with 4513 disorder annotations for evaluation purpose only. As mentioned earlier, Task 1 requires participants to solve two problems: (a) identifying the boundaries of disorder mentions and (b) mapping each mention to a SNOMED-CT concept. We will report our experimental results on these two subtasks separately, and in each task a strict evaluation and a relaxed evaluation are run individually. The strict evaluation requires the annotated text span to be identical to the reference standard span, while the relaxed evaluation only requires the annotated text span has overlap with reference standard span.

3.1 Experiments on Training Data

We conducted a series of experiments on training data to evaluate the effectiveness of the system components we proposed. Table 5 summarizes the performance of different system components and their combinations in two subtasks. Since the system is tuned and optimized for F-score under the strict evaluation standard, we will discuss and compare F-scores under this standard below, unless otherwise noted.

Table 5. Performance of Different System Components Combination

Components and Combinations		1	1+2	1+4	3	1+2+3	1+2+3 +4.1	1+2+3 +4.2	1+2+3 +4
Task 1a Strict	Precision	0.084	0.480	0.563	0.698	0.517	0.572	0.644	0.713
	Recall	0.076	0.447	0.384	0.151	0.586	0.623	0.612	0.650
	F-score	0.080	0.463	0.457	0.248	0.549	0.596	0.628	0.680
Task 1a Relaxed	Precision	0.703	0.704	0.872	0.904	0.739	0.761	0.86	0.895
	Recall	0.667	0.658	0.608	0.194	0.816	0.811	0.795	0.796
	F-score	0.684	0.680	0.716	0.319	0.776	0.785	0.826	0.843
Task 1b Strict	Accuracy	0.059	0.391	0.3	0.144	0.526	0.544	0.551	0.57
Task 1a Relaxed	Accuracy	0.782	0.874	0.781	0.956	0.897	0.873	0.9	0.876

System Components: 1 – MetaMap annotation, 2 – Normalization, 3 – Rule-based annotation, 4 – Post-processing including Normalization (4.1) and Pruning (4.2).

Comparing Comp. 1 and Comp. 1+2 in Table 5, we can see (1) MetaMap as a standalone annotator delivers quite poor results; (2) however, it has much better score in relaxed evaluation, and out of its raw output, near half of correct annotations can be obtained after normalization. This indicates that MetaMap fails on pinpoint identification, but has the capability to parse and roughly recognize the phrases of interest. Therefore, despite an unreliable tool to complete the work individually, MetaMap could serve as a reasonable platform for upper level algorithm development.

Designed to correct false negative errors, the Rule-based annotation alone (Comp. 3) is able to identify up to 15% of all true annotations. Working with Comp. 1+2, the combination attains a 0.549 F-score.

The post-processing step is also very important. Its normalization (4.1) and pruning (4.2) algorithms give 0.05 and 0.08 F-score lifts over Comp. 1+2+3, respectively. Unifying all components, the entire system eventually achieves 0.68 F-score in Task 1a and 0.57 accuracy in Task 1b.

3.2 Competition on Test Data

Participants are allowed to submit two runs of annotations on test data for each subtask. Systems using additional annotations (Group B) will be evaluated separately from systems without additional annotations (Group A). Our system is in Group B, since we employed MetaMap for the first-stage annotation. The competition organizers published the final results and team rankings for each group in each subtask online[5]. For easily reading and comparing, we compiled all teams together in Table 6-8 using the field "Group" for differentiation. The best result out of the two submissions is taken for each team. The ranking is based on F-score in Task 1a and accuracy in Task 1b.

From Table 6-8 we are glad to see that our system (UCSC) has consistent perfromance in every subtask on test data, and it outperforms all other participating systems in Group B in every subtask under either strict or relaxed evaluation standard. Even if we compare with all other participants, ignoring the group difference, our system successfully ranks 4th and 3rd in subtask 1a and 1b respectively. This achievement encourages the broad MetaMap users who contemplate a relatively high performance annotator built on top of MetaMap without spending much effort on NLP infrastructure.

Though the detail of the leading systems has not been published yet, we guess the gaps between our system and them are from three aspects: (1) probably MetaMap is not able to deliver annotations as precise and complete as those customized, advanced annotation systems; (2) with a portion of CUI-less concepts undefined, the vocabulary of SNOMED-CT is kind of limited for current task, resulting in a number of candidates unidentified by MetaMap; (3) the annotation rules learned from training data are difficult to capture certain features of the annotation, such as the sequential information among words, which may be supplemented by statistical learning algorithms. In future, we are going to advance our system along these directions.

---

[5] http://nicta.com.au/business/health/events/clefehealth_2013/results

4 Conclusion

In this paper we proposed a two-stage annotation system to solve CLEF eHealth 2013 Task 1, where MetaMap serves as the first-stage annotator for the identification of phrases with potential interest, and a Rule-based annotator works on the second stage for fine-grained supplementation. Learned from training data, the annotation rules are generated to eliminate two types of errors from MetaMap. The experiments on training data demonstrate the effectiveness of every system component, while the published competition results show that our system performs consistenly on test data and beats all other competing systems in the group using additional annotations.

Our results in this paper are applicable to healthcare text mining and disorder annotation from clinical reports. We anticipate that these results can be generalized further and that their use can be extended into many new domains such as network design diagnostics, semiconductor manufacturing, aerospace system operation, and automotive system design. This expectation is based on prior related work by us in the networks [3] [4], semiconductor [5], aerospace [6], automotive [7] contexts; we are exploring extensions to financial services. These results can also be adapted to other knowledge discoveries and information retrieval from clinical documents [8].

Acknowledgement

We wish to thank the organizers of the CLEF eHealth 2013 for preparing the datasets and organizing the shared tasks. Their work is supported by the Shared Annotated Resources (ShARe) project funded by the United States National Institutes of Health: R01GM090187.

References

1. O. Bodenreider and A. McCray. Exploring semantic groups through visual approaches. Journal of Biomedical Informatics, 2003. 36(2203): pp. 414-432.
2. Hanna Suominen, Sanna Salanterä Sumithra Velupillai, WendyW. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, BrettR. South, Danielle Mowery, Gareth J. F. Jones, Johannes Leveling,Liadh Kelly, Lorraine Goeuriot, David Martinez, and Guido Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013, Proceedings of CLEF 2013. To appear.
3. C. Wang, R. Akella, and S. Ramachandran. Hierarchical Service Analytics for Improving Productivity in an Enterprise Service Center, In Proc. of CIKM'10, pp. 1209-1218, 2010.
4. C. Wang, R. Akella, S. Ramachandran, and D. Hinnant. Knowledge Extraction and Reuse within "Smart" Service Centers, In Proc. of SRII'11, pp. 163-176, 2011.
5. W. Shindo, E. Wang, R. Akella, A. J. Strojwas. Effective Excursion Detection and Defect Source Identification Through In-line Defect Inspection and Classification, IEEE Trans. on Semiconductor Manufacturing, 12(1):3-10, 1999.
6. A. Srivastava, R. Akella, V. Diev, S. Kumaresan, D. McIntosh, M. Pontikakis, Z. Xu, and Y. Zhang. Enabling the Discovery of Recurring Anomalies in Aerospace System Problem

Reports using High-Dimensional Clustering Techniques, In Proc. of IEEE Aerospace Conference, 2006.
7. J. Voit, R. Akella, R. Kishore. Triggered Learning Process from Production to Product Development, In Proc. Of PICMET, 2003.
8. M. Daltayanni, C. Wang, and R. Akella. A Fast Interactive Search System for Healthcare Services, In Proc. of SRII'12, pp. 525-534, 2012.
9. A. Aronson and F. Lang. An overview of MetaMap: historical perspective and recent advances, Journal of American Medical Informatics Association, 2010, 17:229-236.

Table 6. Task 1a Results Using Strict Evaluation

Rank	Team	Precision	Recall	F-score	Group
1	UTHealth_CCB	0.800	0.706	0.750	A
2	NCBI	0.768	0.654	0.707	A
3	CLEAR	0.764	0.624	0.687	A
4	**UCSC**	0.732	0.621	0.672	B
5	Mayo	0.800	0.573	0.668	A
6	UCDCSI	0.745	0.587	0.656	A
7	CORAL	0.796	0.487	0.604	A
8	HealthLanguageLABS	0.686	0.539	0.604	A
9	LIMSI	0.814	0.473	0.598	A
10	AEHRC	0.613	0.566	0.589	A
11	RelAgent	0.651	0.494	0.562	B
12	Diganesan	0.614	0.505	0.554	A
13	steven_seeger	0.575	0.496	0.533	A
14	alamb	0.492	0.558	0.523	B
15	KPSCMI	0.494	0.512	0.503	A
16	THCIB	0.445	0.551	0.492	B
17	Rahul	0.397	0.465	0.428	B
18	ArvindWVU	0.230	0.318	0.267	A
19	SNUBME	0.191	0.137	0.160	A
20	FAYOLA	0.024	0.446	0.046	A

Table 7. Task 1a Results Using Relaxed Evaluation

Rank	Team	Precision	Recall	F-score	Group
1	UTHealth_CCB	0.925	0.827	0.873	A
2	NCBI	0.904	0.805	0.852	A
3	Mayo	0.939	0.766	0.844	A
4	CLEAR	0.929	0.759	0.836	A
5	AEHRC	0.886	0.785	0.833	A
6	UCDCSI	0.922	0.758	0.832	A
7	**UCSC**	0.883	0.742	0.806	B
8	ArvindWVU	0.788	0.814	0.801	A
9	Diganesan	0.885	0.731	0.801	A
10	HealthLanguageLABS	0.912	0.701	0.793	A

11	steven_seeger	0.848	0.741	0.791	A
12	alamb	0.740	0.840	0.787	B
13	RelAgent	0.901	0.686	0.779	B
14	Rahul	0.717	0.814	0.762	B
15	CORAL	0.942	0.601	0.734	A
16	THCIB	0.720	0.713	0.716	B
17	LIMSI	0.964	0.563	0.711	A
18	KPSCMI	0.680	0.687	0.684	A
19	SNUBME	0.381	0.271	0.317	A
20	FAYOLA	0.088	0.997	0.161	A

Table 8. Task 1b Results Using Strict or Relaxed Evaluation

Task 1b Strict Evaluation				**Task 1b Relaxed Evaluation**			
Rank		Accuracy	Group	Rank		Accuracy	Group
1	NCBI	0.584	A	1	AEHRC	0.939	A
2	Mayo	0.546	A	2	NCBI	0.89	A
3	**UCSC**	0.545	B	3	**UCSC**	0.879	B
4	UTHealth_CCB	0.510	A	4	Mayo	0.87	A
5	THCIB	0.470	B	5	KPSCMI	0.865	A
6	KPSCMI	0.443	A	6	THCIB	0.852	B
7	CLEAR	0.441	A	7	UTHealth_CCB	0.722	A
8	alamb	0.349	B	8	CLEAR	0.714	A
9	AEHRC	0.313	A	9	alamb	0.625	B
10	steven_seeger	0.309	A	10	steven_seeger	0.622	A
11	UCDCSI	0.303	A	11	AEHRC	0.553	A
12	Rahul	0.247	B	12	Rahul	0.531	B
13	Diganesan	0.242	A	13	UCDCSI	0.516	A
14	AEHRC	0.199	A	14	Diganesan	0.478	A
15	ArvindWVU	0.142	A	15	ArvindWVU	0.447	A
16	FAYOLA	0.113	A	16	FAYOLA	0.253	A
17	NCBI	0.584	A	17	AEHRC	0.939	A