GO for gene documents

Padmini Srinivasan and Xin Ying Qiu

Presented by Fernando Gutierrez

Goals of this work

- Gain better understanding of the GO annotation using SVM.
- Open issues in Go context:
  - Effect of hierarchical level in performance.
  - Negative examples overwhelm the positives in the training data.
  - Relaxed definition of classification correctness.

Background

- BioCreAtIve 2004 challenge
  - Performance of systems supporting GO annotations
- Approaches
  - SVM with synonymous terms.
  - Statistical methods, n-grams identification.
  - Hybrid method, sentence level classification.

Background:
Authors' approach

- Phase 1: document retrieval
  - Documents relevant to the gene.
- Phase 2: assigning codes
  - Which codes should assigned to each document.
- Phase 3: assigning the gene/gene product
  - Which code should be assigned to each gene in the document classified.

Background:
Authors' approach

Phase 1

- 5 retrieval ranking strategies.
- Locuslink summary and product information
  - Effective
- Consistent with other research
- In absence of summary
  - Manual designed generic query (Best)
    - Target: genetic domain
    - Gene name

Background:
Authors' approach, Phase 2

- Hierarchy structure of the codes
  - GO´s 3 hierarchies
- Link semantics:
  - Molecular functions: “is_a”
  - Biological process: “part_of” (1/5) and “is_a” (4/5)
  - Cellular component:“part_of” (1/2) and “is_a” (1/2)
- Document classification is not the final goal.

Methods:
GO

- GO structured way to annotate a gene
  - Molecular Function (MF)
    - *arbutin transporter activity*
    - *retinoic acid receptor binding*
  - Biological Process (BP)
    - *lipoprotein transport*
    - *phage assembly*
  - Cellular Components (CC)
    - *Nucleus*
    - *NADPH*

Methods:
Annotations

- Entries for Homo Sapiens with locus types:
  - *gene with protein product, function known or inferred*
- Entries selected
  - Have TAS or IDA, not both
- Total entries: 15,468
  - BP: 7,200 (1 to 579)
  - CC: 4,391 (1 to 789)
  - MF: 3,877 (1 to 333)

Methods:
Document representation

- Bag of words
  - All terms form a vector (except stop words)
- MEDLINE records
  - Title, abstract, RN, MeSH fields.
- Vector Space Model
  - TF*IDF
  - *atc*, *ltc*

Results and Discussion:
Code specific SVM classifier

- Distinct binary SVM for each code (class)
  - Document belongs to the class or not.
- Connection among the codes
  - Share dataset of documents.
- *k*-fold cross validation (*k*=5) (split)

Hierarchy	Recall	Precision	Fscore
MF	0.0419	0.0944	0.052
CC	0.0599	0.1461	0.0764
BP	0.0234	0.064	0.0398

Results and Discussion:
Hierarchy specific SVM score threshold

- Single Thresholding score for each hierarchy
  - Score assigned by SVM
  - Above the score are positive
- Threshold selected
  - Take a split and divide in 4 folds
  - Cross validation
  - Best threshold is selected

Results and Discussion:
Hierarchy specific SVM score threshold

			Training			Testing		
Hierarchy	Split	Threshold	Recall	Precision	Fscore	Recall	Precision	Fscore
MF	1	-0.84	0.5624	0.4136	0.4504	0.5992	0.4258	0.4684
MF	2	-0.86	0.5923	0.3835	0.4390	0.6775	0.4073	0.4769
MF	3	-0.86	0.5954	0.3734	0.4328	0.6817	0.3874	0.4684
MF	4	-0.84	0.5713	0.4046	0.449	0.6857	0.4487	0.5134
MF	5	-0.85	0.5921	0.4076	0.4541	0.6772	0.3945	0.4727
MF	Average	na	0.5827	0.3965	0.4451	0.6643	0.4128	0.48
CC	1	-0.82	0.4799	0.3185	0.3627	0.5301	0.3531	0.3986
CC	2	-0.82	0.4823	0.3214	0.3665	0.5359	0.3516	0.4006
CC	3	-0.86	0.5287	0.2976	0.3590	0.6553	0.3895	0.4571
CC	4	-0.85	0.5122	0.2997	0.3571	0.5703	0.2976	0.3715
CC	5	-0.85	0.5222	0.315	0.3714	0.599	0.29	0.3767
CC	Average	na	0.5051	0.3104	0.3633	0.5781	0.3364	0.4009
BP	1	-0.87	0.4304	0.2378	0.2847	0.4722	0.2585	0.3079
BP	2	-0.87	0.4377	0.2442	0.2908	0.5259	0.2713	0.3362
BP	3	-0.85	0.4019	0.2615	0.2948	0.4908	0.2884	0.3392
BP	4	-0.84	0.3706	0.2556	0.2794	0.4854	0.2966	0.3484
BP	5	-0.87	0.4519	0.2600	0.3069	0.4608	0.2220	0.2791
BP	Average	na	0.4185	0.2518	0.2913	0.4870	0.2674	0.3222

Results and Discussion:
LTC term weight

Hierarchy	Split	Threshold	Training Recall	Training Precision	Training Fscore	Testing Recall	Testing Precision	Testing Fscore
MF	1	-0.83	0.5971	0.4165	0.4640	0.6382	0.4162	0.4735
MF	2	-0.81	0.5732	0.4194	0.4614	0.6525	0.4537	0.5091
MF	3	-0.81	0.5687	0.4133	0.4551	0.6522	0.4311	0.4908
MF	4	-0.83	0.6099	0.4079	0.4634	0.6928	0.4416	0.5138
MF	5	-0.83	0.6107	0.4148	0.4672	0.6880	0.4036	0.4822
MF	Average	na	0.5919	0.4144	0.4622	0.6647	0.4292	0.4939
CC	1	-0.83	0.5426	0.3135	0.3772	0.5383	0.3248	0.3810
CC	2	-0.81	0.5075	0.3248	0.3765	0.5422	0.3437	0.3971
CC	3	-0.83	0.5194	0.3136	0.3685	0.6488	0.4075	0.4763
CC	4	-0.85	0.5606	0.3080	0.3779	0.6216	0.3161	0.3929
CC	5	-0.84	0.5689	0.3284	0.3956	0.6137	0.3136	0.3993
CC	Average	na	0.5398	0.3177	0.3791	0.5929	0.3411	0.4093
BP	1	-0.82	0.3773	0.2596	0.2881	0.4204	0.2894	0.3163
BP	2	-0.84	0.4124	0.2595	0.2965	0.4912	0.2898	0.3423
BP	3	-0.86	0.4405	0.2522	0.2983	0.5366	0.2737	0.3395
BP	4	-0.85	0.3935	0.2344	0.2713	0.5211	0.2856	0.3497
BP	5	-0.85	0.4472	0.2691	0.3131	0.4680	0.2466	0.3035
BP	Average	na	0.4142	0.2550	0.2935	0.4875	0.2770	0.3303

Results and Discussion:
Feature selection

- Document frequency: term not frequent, little class info

  Unique documents, 0.1% threshold.

- χ^2 statistical:
  - H_0 term's $\text{frequency}_{observed} = \text{frequency}_{expected}$
- *Z(t,c):* independence of *t*'s distribution

Hierarchy	Num Terms	FScore			
		None	Z	DF = 0.1%	CHI
MF	16	0.1211	0.0854	0.1445	0.3447
CC	12	0.1258	0.0677	0.1168	0.3079
BP	30	0.048	0.0366	0.0418	0.2333

Code specific SVM score threshold

- Before
  - Each hierarchy has a single threshold
- Specific threshold for each code
  - Average threshold is in a small range
  - Optimal threshold of each code
- Training and Test performance

Code specific SVM score threshold

Hierarchy	Split	Training FScore	Testing Recall	Testing Precision	Testing FScore
MF	1	0.6221	0.4499	0.3989	0.3852
MF	2	0.615	0.5364	0.4402	0.44351
MF	3	0.6128	0.5295	0.3892	0.4133
MF	4	0.6298	0.5793	0.4394	0.452
MF	5	0.6371	0.5467	0.4264	0.4451
MF	average	0.6234	0.5284	0.4188	0.4278
CC	1	0.5541	0.4679	0.3774	0.3842
CC	2	0.5052	0.5029	0.3435	0.3626
CC	3	0.5131	0.5632	0.3806	0.4239
CC	4	0.5554	0.5134	0.3273	0.3727
CC	5	0.5796	0.5148	0.3875	0.4201
CC	average	0.5415	0.5125	0.3632	0.3927
BP	1	0.4469	0.3994	0.2463	0.2554
BP	2	0.4472	0.4017	0.2727	0.2793
BP	3	0.4378	0.3951	0.2531	0.2589
BP	4	0.4248	0.4309	0.2654	0.2804
BP	5	0.4518	0.3710	0.2434	0.2543
BP	average	0.4417	0.3996	0.2562	0.2657

Analysis of results:
Recall versus precision

- Recall is always higher than the precision
  - Too much false positives
- Tighter constrains and filtering
  - Possible future research.

Analysis of results:
Hierarchical level and performance

Hierarchy	Level	# of Codes	Scores		
			Recall	Precision	FScore
MF	1	4	0.3176	0.1786	0.2205
MF	2	26	0.4846	0.2666	0.3176
MF	3	41	0.5261	0.3145	0.3695
MF	4	50	0.6780	0.4449	0.5066
MF	5	57	0.7799	0.4936	0.5732
MF	6	17	0.8937	0.5548	0.6505
MF	7	11	0.6961	0.3876	0.4728
MF	8	4	0.675	0.475	0.5233
MF	9	2	0.8	0.6	0.6667
CC	1	1	0.3171	0.2235	0.2537
CC	2	20	0.6476	0.4017	0.4675
CC	3	25	0.6062	0.349	0.4089
CC	4	26	0.5547	0.306	0.3741
CC	5	14	0.5502	0.2832	0.3622
CC	6	6	0.3955	0.2717	0.3135
CC	7	1	1	0.7917	0.8667
BP	1	3	0.1354	0.0481	0.0704
BP	2	10	0.3327	0.1767	0.2174
BP	3	34	0.5164	0.2517	0.3179
BP	4	54	0.4849	0.2563	0.3119
BP	5	49	0.4681	0.2516	0.3093
BP	6	52	0.4555	0.2734	0.3139
BP	7	51	0.5863	0.3251	0.3921
BP	8	21	0.4677	0.2834	0.3301
BP	9	8	0.4698	0.2840	0.3316

Analysis of results:
Number of positives for training & performance

More true positives, better results

Training size	# codes	MF-FScore	# codes	CC-FScore	# codes	BP-FScore
5	2	0.25	34	0.4067	128	0.2695
6–10	3	0.0833	25	0.3650	65	0.3875
11–15	9	0.4373	7	0.4528	22	0.3716
16–20	37	0.5645	4	0.4550	15	0.3306
21–25	39	0.544	3	0.4762	9	0.2588
26–30	31	0.5566	4	0.3687	4	0.3007
31–35	6	0.4663	3	0.5651	8	0.3566
36–40	7	0.5275	0	0	6	0.3579
41–45	10	0.4124	1	0.2009	5	0.3484
46–50	11	0.4276	1	0.2861	2	0.2553
51–75	18	0.3912	2	0.3430	12	0.3060
76–100	12	0.3936	1	0.2681	6	0.2726
101–125	5	0.4273	2	0.4089	0	0
126–150	4	0.4767	2	0.3226	0	0
151–last	20	0.3511	4	0.4586	1	0.2822

Analysis of results:
Correlations between level and number of positives for training

- Negative correlation
  - Between level and size in the case of MF and BP
- The CC hierarchy might need different classification method than MF and BP

Hierarchy	Level vs Size	Level vs FScore	Size vs FScore
MF	-0.2705*	0.3361*	-0.1146
CC	-0.0123	-0.1051	0.0904
BP	-0.2155*	0.1622*	-0.0191

Level specific thresholds

- Each level has a threshold for MF and BP
  - Level 2 and 3.
- No testing CC
  - No correlation between level and performance.

Level specific thresholds

Hierarchy	Split	Level	Original FScore	Threshold	Final FScore
MF	1	2	0.3299	-0.8	0.3665
MF	2	2	0.2782	-0.83	0.2973
MF	3	2	0.3298	-0.78	0.373
MF	4	2	0.3484	-0.81	0.373
MF	5	2	0.3016	-0.78	0.3263
MF	avg	2	0.3176	na	0.341 (+7.4%)
MF	1	3	0.3347	-0.87	0.3063
MF	2	3	0.3178	-0.84	0.3301
MF	3	3	0.4243	-0.88	0.3760
MF	4	3	0.4263	-0.87	0.3823
MF	5	3	0.3444	-0.86	0.3464
MF	avg	3	0.3695	na	0.3482 (-5.8%)
BP	1	2	0.2542	-0.87	0.2542
BP	2	2	0.2951	-0.89	0.2989
BP	3	2	0.2261	-0.89	0.2027
BP	4	2	0.1609	-0.87	0.2319
BP	5	2	0.1507	-0.88	0.1494
BP	avg	2	0.2174	na	0.2274 (+4.6%)
BP	1	3	0.2916	-0.86	0.3020
BP	2	3	0.3145	-0.83	0.3455
BP	3	3	0.3496	-0.82	0.3030
BP	4	3	0.3128	-0.83	0.3164
BP	5	3	0.3209	-0.83	0.3529
BP	avg	3	0.3179	na	0.324 (+1.9%)

Relaxing the correctness criteria

- Assume: document is assigned a GO code, also the ancestor GO code is assigned (implicitly).
- Ancestor
  - How up in the hierarchy: Ancestor_level
- Correctness
  - If the correct code or its ancestor is assigned, it is correct.

glucoside transport: carbohydrate transport (yes), *alpha-glucoside transport* (no).

Relaxing the correctness criteria

Hierarchy	ANC_LEVEL	Recall	Precision	FScore
MF	baseline	0.6643	0.4128	0.4800
MF	1	0.6643	0.419	0.4847
MF	2	0.6650	0.4229	0.4880
MF	3	0.6650	0.4243	0.4888
MF	4	0.6650	0.4245	0.4890
MF	5	0.6650	0.4245	0.4880
CC	baseline	0.5781	0.3364	0.4009
CC	1	0.5781	0.3471	0.4082
CC	2	0.5784	0.3509	0.4113
CC	3	0.5784	0.3536	0.4132
CC	4	0.5784	0.3540	0.4136
BP	baseline	0.4870	0.2674	0.3222
BP	1	0.4887	0.2724	0.3265
BP	2	0.4887	0.2746	0.3285
BP	3	0.4890	0.2773	0.3301
BP	4	0.4890	0.2776	0.3305
BP	5	0.4890	0.2778	0.3306

Relaxing the correctness criteria

Hierarchy	ANC_LEVEL	Recall	Precision	FScore
MF	baseline	0.6639	0.3076	0.4100
MF	1	0.6639	0.3195	0.4309
MF	2	0.6652	0.326	0.4370
MF	3	0.6652	0.3288	0.4396
MF	4	0.6652	0.3296	0.4403
MF	5	0.6652	0.3297	0.4404
CC	baseline	0.7442	0.3163	0.4432
CC	1	0.7442	0.3321	0.4586
CC	2	0.7458	0.3512	0.4769
CC	3	0.74583	0.3551	0.4804
CC	4	0.7458	0.357	0.4822
CC	5	0.7458	0.3572	0.4823
BP	baseline	0.5578	0.2286	0.3236
BP	1	0.5583	0.2360	0.3311
BP	2	0.5583	0.2432	0.3381
BP	3	0.5586	0.2466	0.3415
BP	4	0.5586	0.2482	0.3430
BP	5	0.5586	0.2485	0.3433

Code with less than five positive documents

- 239 codes with less than 5 associated documents
  - Not tried before
- Method of testing
  - Codes with more than 10 positive documents
  - Temporal sequence
  - 5-fold strategy
  - Test mode
    - First Five Test
    - Full Test

Code with less than five positive documents

Hierarchy	# +ves	Threshold	FullTest FScore	Threshold	FirstFiveTest FScore
MF	1	-0.942	0.141	-0.924	0.1682
MF	2	-0.892	0.1999	-0.9	0.2441
MF	3	-0.908	0.2583	-0.87	0.2825
MF	4	-0.906	0.2713	-0.86	0.3218
MF	GT4		0.4209		
BP	1	-0.942	0.0881	-0.95	0.1081
BP	2	-0.94	0.1440	-0.936	0.1791
BP	3	-0.904	0.1591	-0.894	0.1851
BP	4	-0.898	0.1931	-0.896	0.2251
BP	GT4		0.3480		
CC	1	-0.946	0.1439	-0.948	0.1791
CC	2	-0.916	0.1631	-0.896	0.1977
CC	3	-0.896	0.2012	-0.848	0.2067
CC	4	-0.872	0.2144	-0.844	0.2488
CC	GT4		0.3795		

Conclusion

- Thresholding at individual code level:
  - decreases performance.
- Performance by level and number of positives:
  - Better for MF and BP
  - CC different
- Relaxed evaluation criteria:
  - Improved

Conclusion:
Counter common intuition

- General codes in MF and BP are more difficult to classify.
- More positive in training data leads to a better performance.

Future Works

- Ensemble of classifiers:
  - Through hierarchy
- Explore other strategies for phase 3