Predicting ecological traits from fungal genomes

Joe Wan

Introduction

Fungi are a diverse group of organisms with key roles in natural ecosystems as well as in human health and industry. They have evolved diverse lifestyles, obtaining nutrients through **saprotrophy** (by decomposing dead matter), **biotrophy** (by harming living hosts), and **symbiotrophy** (by exchanging nutrients with living hosts).

Machine learning approaches integrating the large volume of available fungal genomes will improve understanding of **how genomic traits define the diverse ecological roles of fungi.** Additionally, as genome sequencing becomes increasingly cheap, genomic classification will serve as a useful tool for **predicting the ecology of poorly-understood species.**

This project will focus on one particular ecological trait: **ectomycorrhizal (ECM) symbiosis.** Estimated to have evolved independently in at least 16 lineages of fungi, this interaction involves highly specialized structures which allow **nutrient exchange** between fungi and plants. It is of critical importance to cycling of nutrients in land ecosystems. However, its genomic basis remains incompletely understood.

Ectomycorrhizal structures (the Hartig net) formed by the fungus *Amanita* on plant roots. Plant sugars are exchanged for phosphorus from the fungal partner.

1. Training data

FunGuild
ecological annotations

MycoCosm (JGI Joint Genome Institute, Department of Energy)
genome annotations

	ectomycorrhizal?	GO_0001	GO_0002	KEGG_0001	KEGG_0002	IPR_0001	KOG_0001	[...]
Boletus edulis	Y	10	0	51	1	2	0	
Tuber melanosporum	Y	0	1	0	1	5	0	
Amanita muscaria	Y	3	0	11	2	6	3	
Amanita thiersii	N	15	0	0	0	0	10	
Coprinellus micaceus	N	42	0	1	1	1	4	
Armillaria ostoyae	N	11	0	9	5	9	8	
[...]								

395 species (34 ectomycorrhizal)

(number of genes with each annotation present in the genome)

15792 features (possible annotations)

Example ectomycorrhizal species included in the training set: the porcini (*Boletus edulis*, at left) and the truffle (*Tuber melanosporum*, at right) are two edible fungi with major economic importance. Lineages leading to the two species diverged about 1.8 billion years ago, and ectomycorrhizal symbiosis is believed to have **evolved independently** in each lineage.

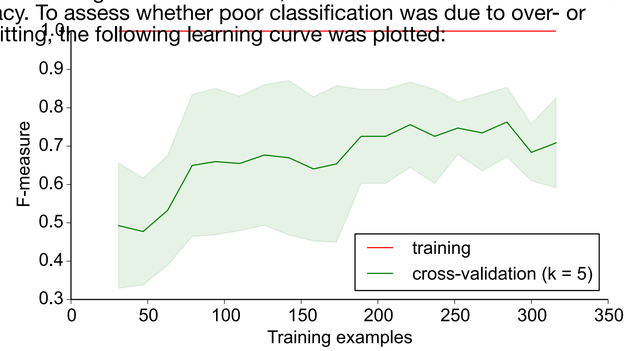
2. Linear SVM classifier

Using the Python scikit-learn library, a **linear SVM** was fit to the data. The **default value of *C* = 1** was chosen for the regularization parameter.

Five-fold cross validation was performed, giving the following average scorings:

accuracy: 0.94
precision: 0.70
recall: 0.70
F-measure: 0.73

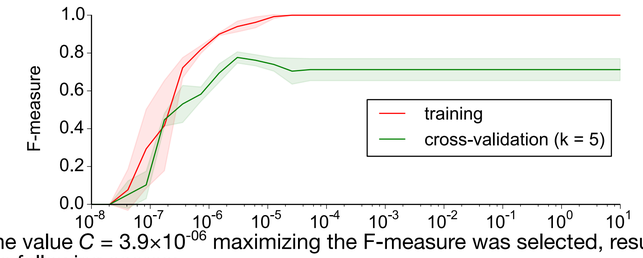
Since the training set is unbalanced, the F-measure is more informative than accuracy. To assess whether poor classification was due to over- or underfitting, the following learning curve was plotted:



Since training F-measure reaches 1.0 even with the smallest training example, the approach does not suffer from a bias error. Since the cross-validation error (our estimate for the generalization error) remains low with more training examples, the SVM appears to have high variance: it **overfits** the data.

3. Improving the linear SVM

To reduce overfitting, a smaller value of the parameter *C* can be chosen in order to favor a larger-margin classifier. To identify an appropriate value of *C*, training and cross-validation error was evaluated for different values of *C:*



The value $C = 3.9\times10^{-06}$ maximizing the F-measure was selected, resulting in the following scores:

accuracy: 0.97
precision = recall: 0.82
F-measure: 0.82

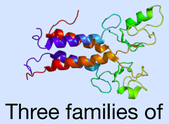
A new learning curve shows that overfitting no longer seems to be a major problem; instead, it appears that the most useful step would be to **obtain additional training examples** since the F-measure is still increasing with additional training data.

4. Interpretation of features

Using the optimized regularization parameter $C = 3.9\times10^{-06}$, a linear SVM was fitted to the entire dataset. Because a linear SVM was used, it was possible to interprets weights for each factor as a measure of influence on the classification. The following table lists the factors with greatest squared weight and whether each factor was associated with ectomycorrhizal (a check in the **ECM** column) or non-ectomycorrhizal genomes (no check). Natural language descriptions are given for each term.

Rank	ECM	Annotation Type	ID	Description
1		KEGG		pathway type: metabolic
2		KOG	KOG1550	extracellular protein SEL-1 and related proteins
3		KOG		KOG class: transcription
4		KOG		KOG class: general function prediction only
5	✓	KOG		KOG class: posttranslational modification, protein turnover, chaperones
6		IPR	IPR006597	Sel1-like repeat
7	✓	GO		GO term type: cellular component
8	✓	KOG	KOG1216	von Willebrand factor and related coagulation proteins
9		IPR	IPR002893	zinc finger, MYND-type
10	✓	GO	GO:0005622	intracellular
11		KOG		KOG group: poorly characterized
12	✓	IPR	IPR000719	protein kinase domain
13		IPR	IPR000210	BTB/POZ domain
14	✓	IPR	IPR001841	zinc finger, RING-type
15	✓	IPR	IPR001461	aspartic peptidase
16		IPR	IPR001878	zinc finger, CCHC-type
17		KEGG		pathway class: metabolism of complex lipids
18	✓	GO		GO term type: biological process
19	✓	IPR	IPR011009	protein kinases
20	✓	KOG		KOG group: cellular processes and signaling

Many top features corresponded to **very broad categories** (e.g. "KOG class"), indicating that common features may have been more useful for classification. Many features enriched in ectomycorrhizal genomes were related to **signaling** (e.g. "cellular processes and signaling", "post-translational modification"), which may reflect the importance of plant-fungal signalling in ECM symbiosis. Different **zinc finger domains** (possibly involved in transcriptional regulation) were associated with ECM versus non-ECM species, suggesting differences in regulatory processes may be predictive of ecology. Finally, many features were **poorly-characterized gene families** (e.g. "KOG group: poorly characterized") or families **not known from fungi** ("von Willebrand factor"), suggesting genes important for understanding fungal lifestyle adaptation remain poorly characterized and/or poorly annotated.



Three families of **zinc finger domains** (RING-like shown above) were among the highest-weighted features. Most zing finger domains are involved in DNA, RNA, or protein interactions.

Conclusions

— A **linear SVM** was developed to classify annotated fungal genomes into ectomycorrizal and non-ectomycorrhizal groups. Using 5-fold cross validation, the value of the parameter *C* was optimized to **reduce overfitting**, attaining precision and recall of 0.82.

— Factors with the highest weight in linear SVM revealed potential important **predictors** of ecological role. Many were **broad annotation categories**, reflecting the importance of common features. **Signaling and regulatory** processes appeared to be important in predicting role.

— In order to better understand the genetics of ectomycorrhizal symbiosis, **additional genomic sequencing** and **characterization of poorly-understood genes** will be crucial.