Gene Ontology Protein Function Prediction

Roman Eisner*, Alona Fyshe, Russell Greiner, Paul Lu, David Meeuwis, Brett Poulin, Duane Szafron, Chris Upton, David Wishart,
*eisner@cs.ualberta.ca

PENCE BIOINFORMATICS I

UNIVERSITY OF ALBERTA

www.cs.ualberta.ca/~bioinfo

aicml The Alberta Ingenuity Centre for Machine Learning

NSERC CRSNG

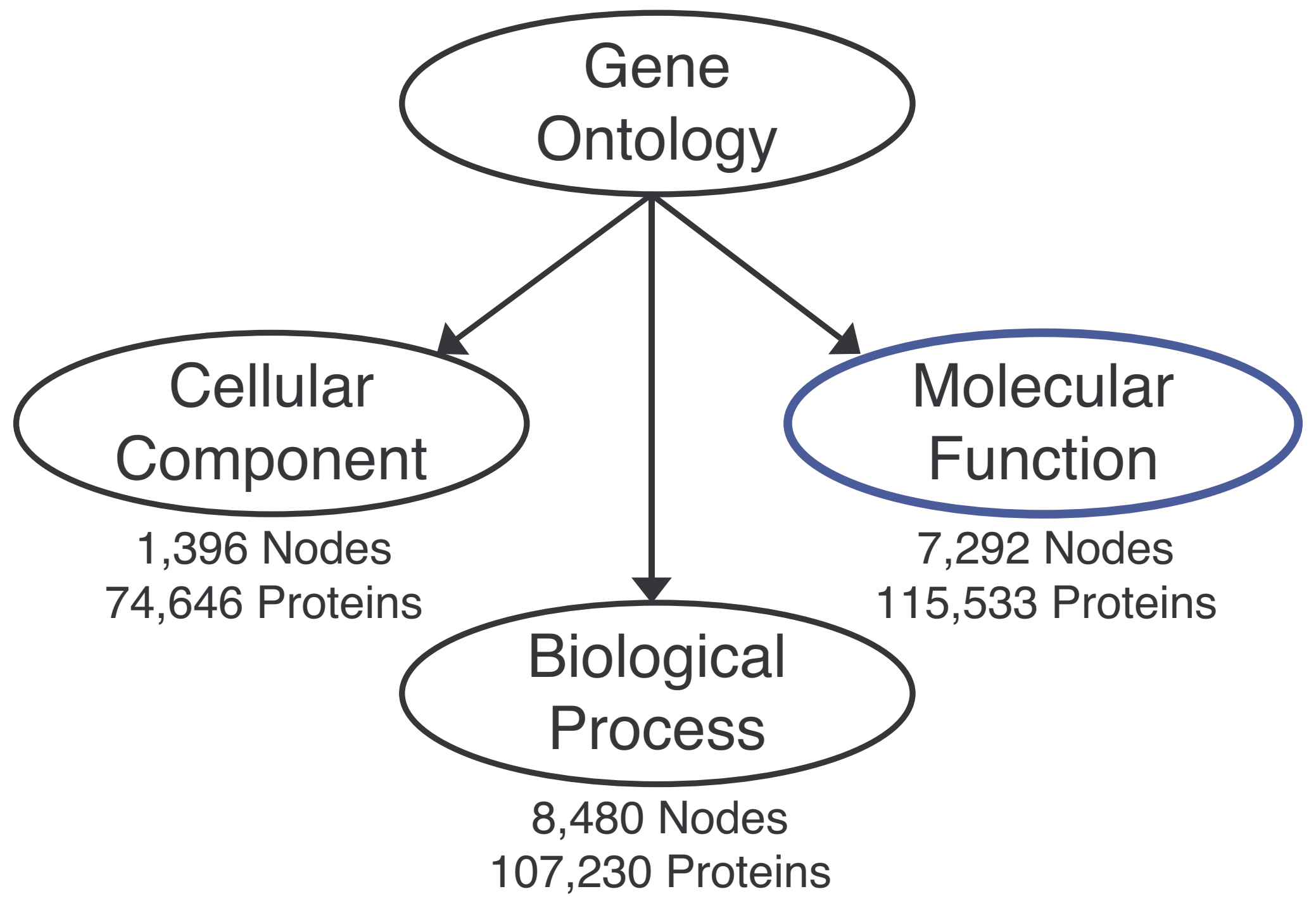
Sun microsystems

Introduction

Automating the extraction of useful information from dynamic databases, such as Swiss-PROT, is a daunting task, given that biological terminologies vary greatly between researchers. The Gene Ontology (GO) Consortium[1] has been very successfull in addressing this problem. We present our investigation of the *Molecular Function* aspect of GO, with respect to proteins found in the Swiss-PROT database, and our use of machine-learning algorithms to create a general function GO classifier.

What is Gene Ontology?

Gene Ontology is a controlled vocabulary of terms used to describe 3 aspects of proteins. These 3 aspects are independent networks of terms that describe the *Molecular Function*, *Biological Process*, and *Cellular Component* of gene products.

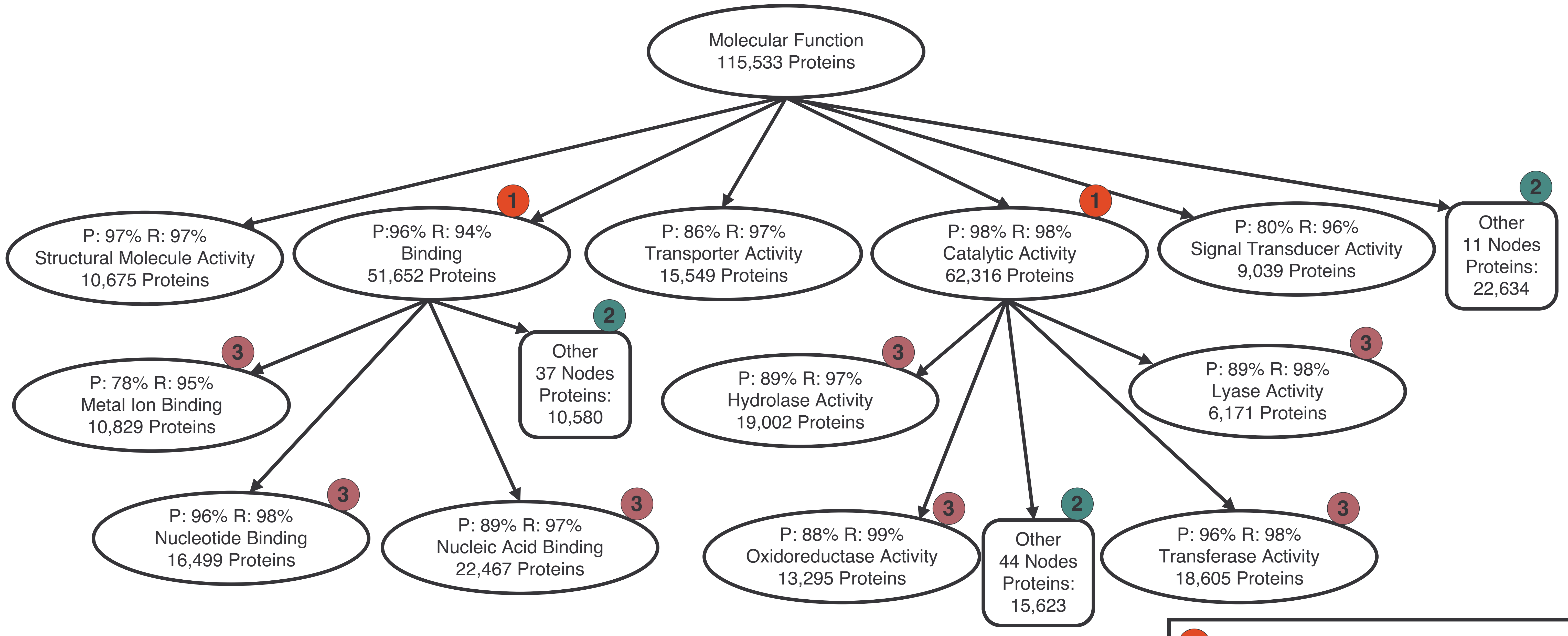


Gene Ontology

Cellular Component
1,396 Nodes
74,646 Proteins

Molecular Function
7,292 Nodes
115,533 Proteins

Biological Process
8,480 Nodes
107,230 Proteins



Molecular Function
115,533 Proteins

P: 97% R: 97%
Structural Molecule Activity
10,675 Proteins

1
P:96% R: 94%
Binding
51,652 Proteins

P: 86% R: 97%
Transporter Activity
15,549 Proteins

1
P: 98% R: 98%
Catalytic Activity
62,316 Proteins

P: 80% R: 96%
Signal Transducer Activity
9,039 Proteins

2
Other
11 Nodes
Proteins:
22,634

3
P: 78% R: 95%
Metal Ion Binding
10,829 Proteins

2
Other
37 Nodes
Proteins:
10,580

3
P: 96% R: 98%
Nucleotide Binding
16,499 Proteins

3
P: 89% R: 97%
Nucleic Acid Binding
22,467 Proteins

3
P: 89% R: 97%
Hydrolase Activity
19,002 Proteins

3
P: 89% R: 98%
Lyase Activity
6,171 Proteins

3
P: 88% R: 99%
Oxidoreductase Activity
13,295 Proteins

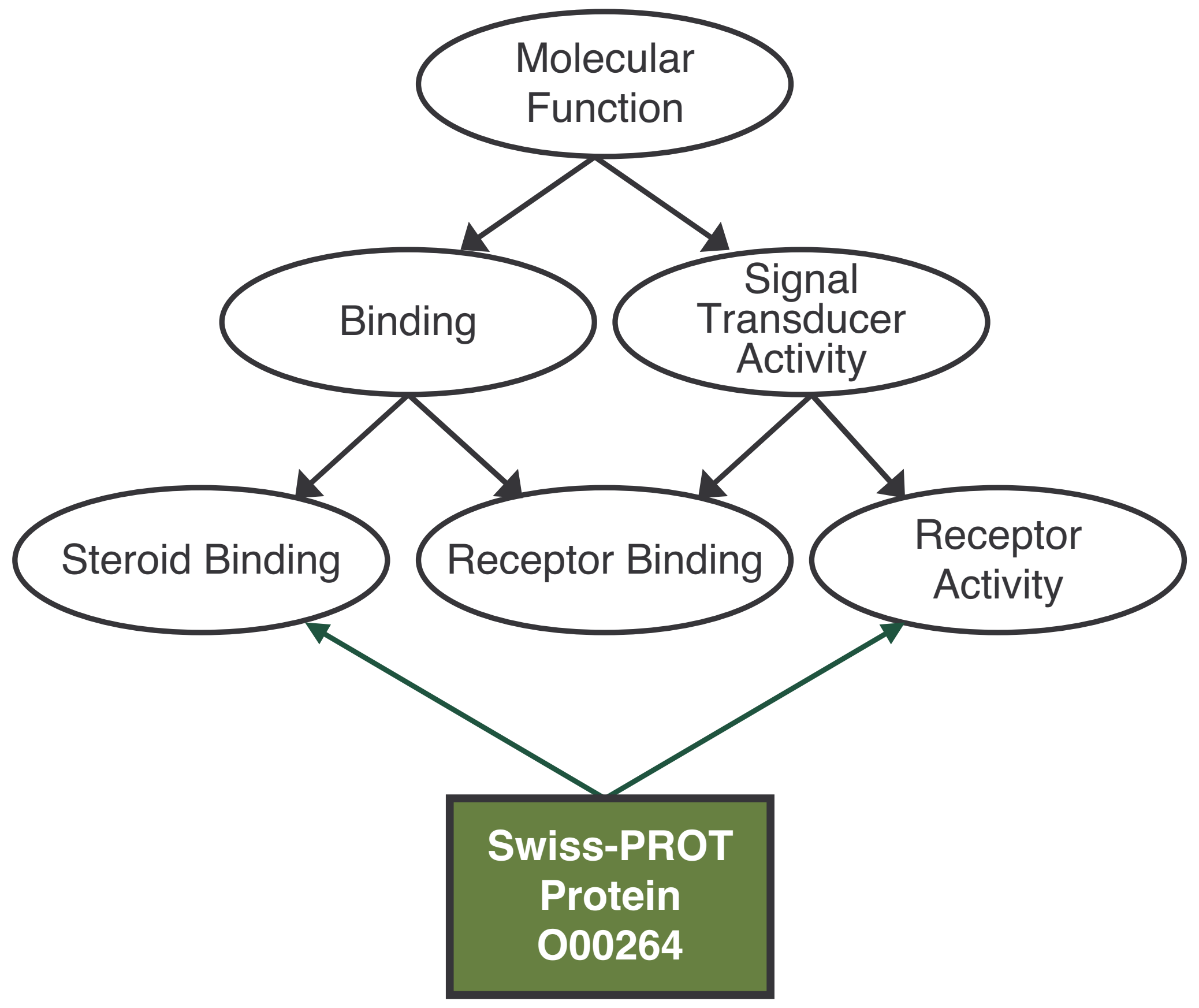
2
Other
44 Nodes
Proteins:
15,623

3
P: 96% R: 98%
Transferase Activity
18,605 Proteins

1 > 16,000 proteins: expand node
2 Nodes with < 6,000 proteins (not used for prediction)
3 Depth = 3: do not expand node
P: Precision of classifier
R: Recall of classifier

GO Structure

GO is a Directed-Acyclic Graph (DAG), which allows for greater representative power than simple hierarchies. Most proteins in the Swiss-PROT database have been annotated with their GO Node(s), and many of these map to more than one node in the Ontology.



Molecular Function

Binding

Signal Transducer Activity

Steroid Binding

Receptor Binding

Receptor Activity

Swiss-PROT Protein O00264

Methodology

The number of nodes in the *molecular function* network is very large. Consequently, have built a high-level function predictor on a subset of the classes. The nodes included in our pruned ontology are those that have more than 6,000 proteins, and were not deeper than the third layer in the *molecular function* network. Also, if a node has more than 16,000 proteins, we expand this node and apply the above criteria to the node's children. Those nodes that do not fit the criteria are put in the "other" class at that level and are discarded during the training of classifiers.

A Binary classifier (predicts "yes"/"no") was created for each of the 12 classifier nodes, using Naıve Bayes. All proteins that are mapped below a node *n* are considered to be contained within *n*, due to the nature of the ontology.

Feature Extraction

To use a Machine-Learning technique such as Naıve Bayes, each protein sequence must be represented by features. These results use the sequence annotation features found by homology (see the poster *Proteome Analyst: An Overview*), but excellent results were also found using sequence-based features as well (see the poster *Suffix Tree Methods for Protein Classification*). Results for each of the 12 classifiers using homology-based features are shown in the main diagram above.

References

1) The Gene Ontology Consortium
[http://www.geneontology.org/]