


JMB

Available online at www.sciencedirect.com

 ScienceDirect

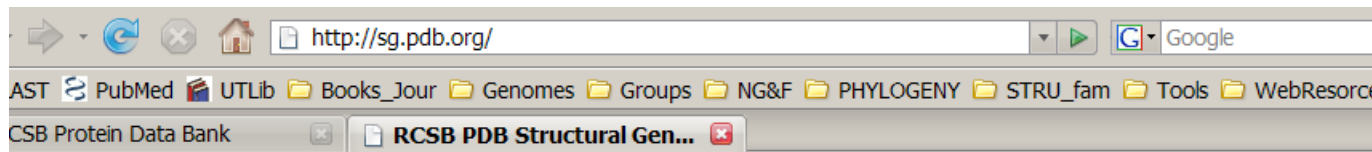


Towards Fully Automated Structure-based Function Prediction in Structural Genomics

James D. Watson¹  , Steve Sanderson², Alexandra Ezersky² Alexei Savchenko², Aled Edwards^{2,3}, Christine Orengo⁴ Andrzej Joachimiak⁵, Roman A. Laskowski¹ and Janet M. Thornton¹

3D

- Protein Structure Initiative (PSI)
- The Midwest Center for Structural Genomics (MCSG)
- Riken (Japan)
- SPiNE (Europe)
- Anglo-Canadian-Swedish SGC (Structural Genomics Consortium)



RCSB
PDB
PROTEIN DATA BANK

A MEMBER OF THE **PDB**

An Information Portal to Biological Macromolecular Structures

[SG Home](#) | [TargetDB Home](#) | [PepcDB Home](#) | [Functional Distributions Home](#)

RCSB PDB Structural Genomics Information Portal



Major aims of all centres

- High-throughput automation of protein production, structure determination and analysis
- Increased coverage of protein fold space and hence the number of protein sequences amenable to homology modelling methods
- Investigation of protein structure to elucidate function in health and disease
- Reduction of the cost of structure determination

30 September 2005, the **MCSG** had **over 5000 active targets** and a total of **319 structures** deposited in the PDB.

1/3 of these have **no** functional annotation

30 September 2005 there were over **1100 proteins** out of over **32,000 in the PDB** labelled as **unknown function**.

Current state of art; **42474 Structures**
 ~ 1500 unknown

pdbx_SG_project.full_name_of_center	
	Total Count (not null):
	0 967
RIKEN Structural Genomics/Proteomics Initiative	
Midwest Center for Structural Genomics	
Joint Center for Structural Genomics	
Northeast Structural Genomics Consortium	
Structural Genomics Consortium	
New York Structural Genomics Research Consortium	
TB Structural Genomics Consortium	
Center for Eukaryotic Structural Genomics	
Southeast Collaboratory for Structural Genomics	
Berkeley Structural Genomics Center	

Methods to infer a function

- Sequence based (similarity > 40%)
- Structure based methods:
 - analysis of global fold
 - identification of highly specific 3D cluster of functional residues
 - ligand binding

No single method will be successful in all cases, and there will be proteins for which no method is useful.

ProFunc server

Functional coverage of the MCSG dataset

(a)

Of the 282 non-redundant structures

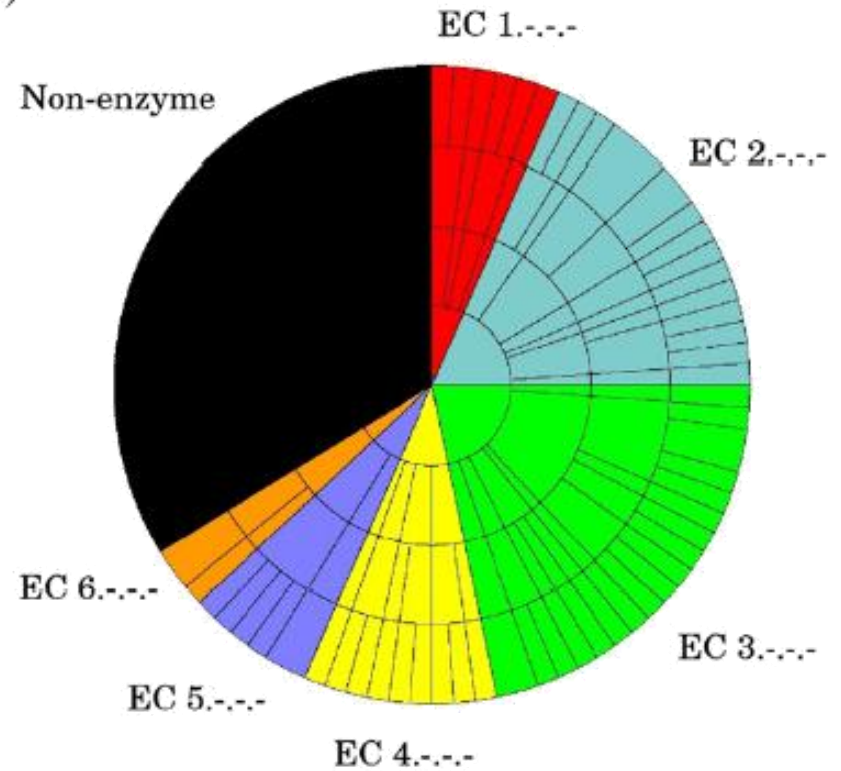
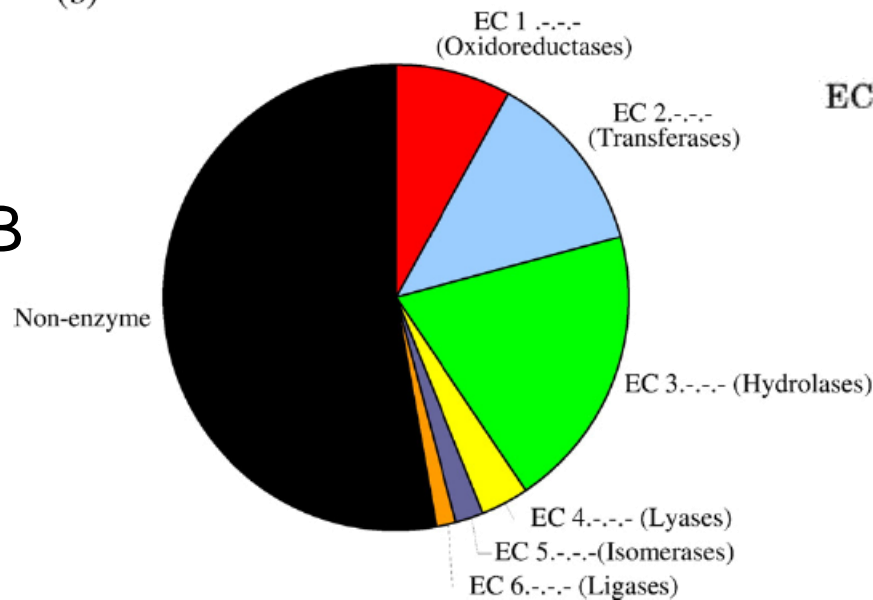
1/3 have known function →

21% have putative function

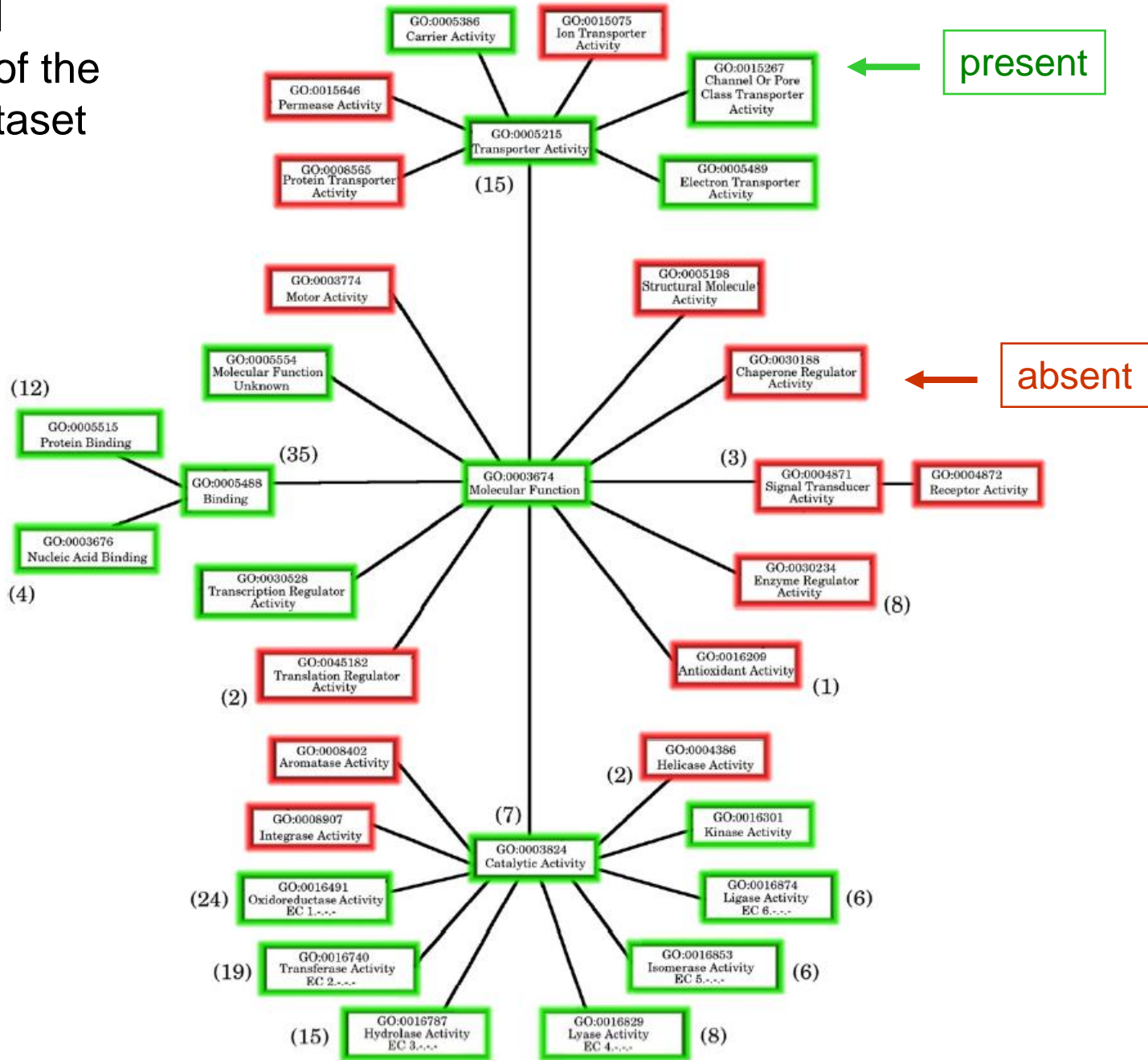
46% unknown function

(b)

PDB



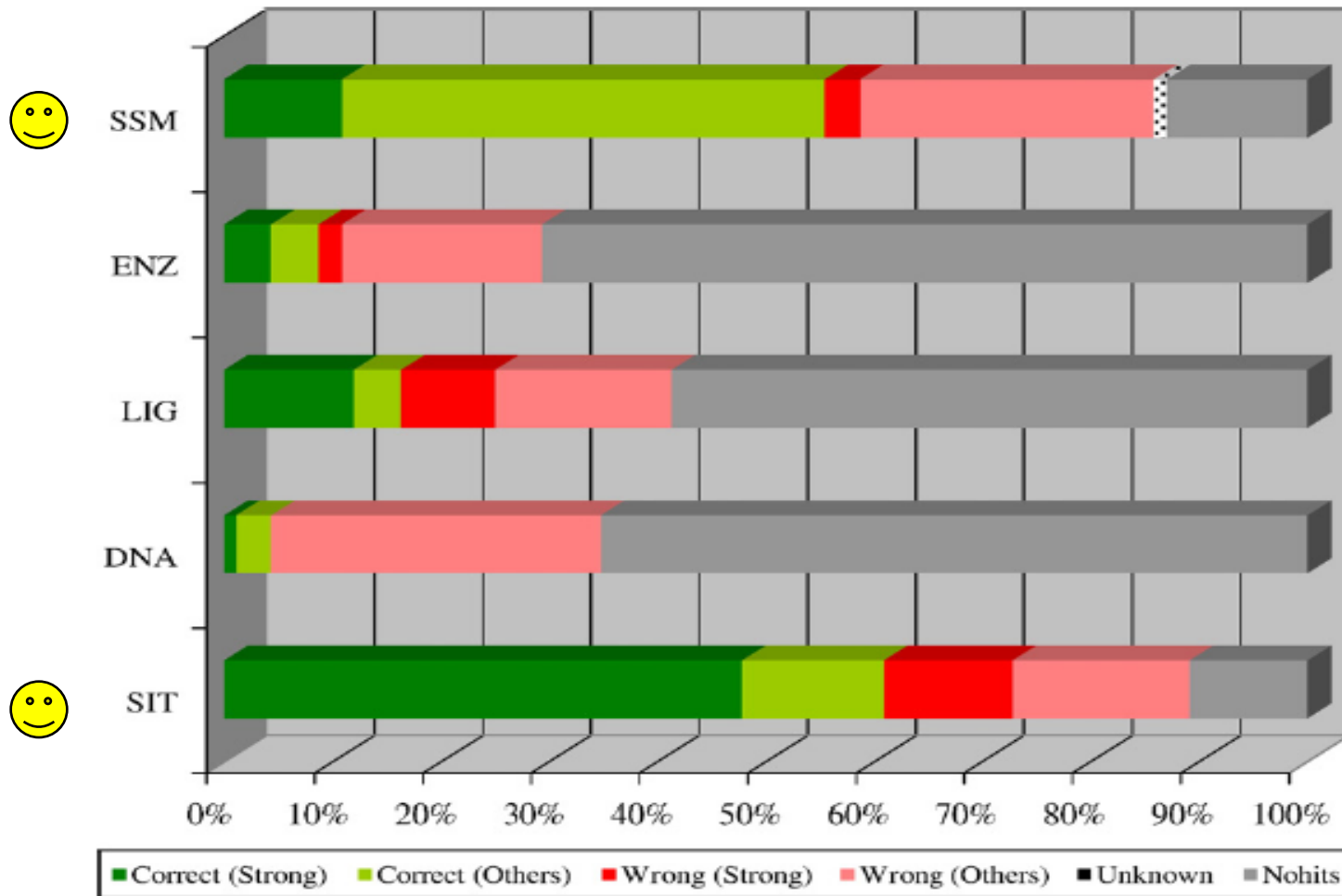
Functional coverage of the MCSG dataset



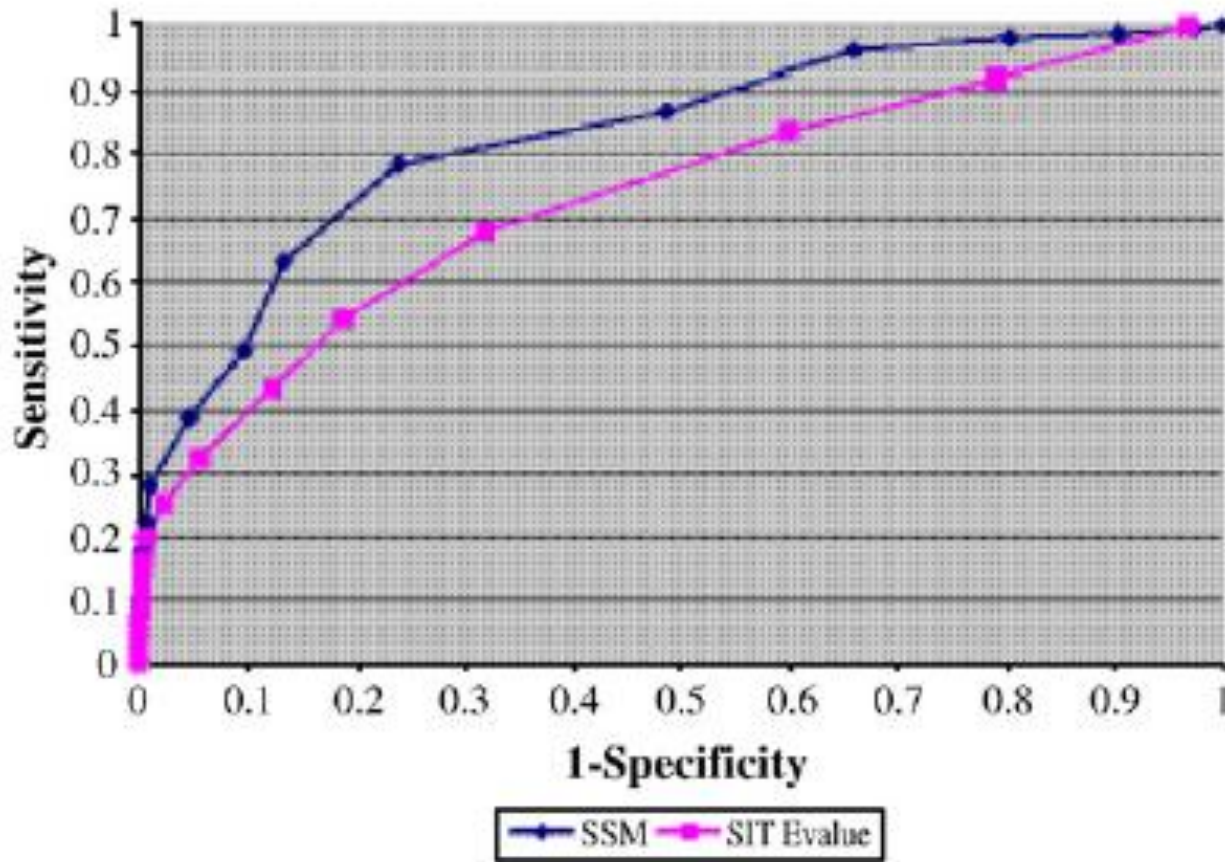
Results 1

Analysis:

- 92 proteins of known function
- The results have been backdated to the release date of the query by removing hits to structures released after that date
- That gives a picture of what the server would have suggested had it been available at the time



The SSM results show that in approximately 55% of cases the top fold match was able to provide the correct functional assignment (almost 20% of which are strongly predicted). The standard template methods provide some success but the most accurate structure-based method is the reverse template approach (SiteSeer SIT), which provides the correct function in 60% of the cases (of which over 75% are strongly predicted)



Area under curve:

0.5 – random

1 – ideal prediction

0.83 – SSM

0.70 – SIT

ROC (receiver operating characteristic) curves for SSM and SIT based on manual function assignment. The ROC curves are plotted for SSM results and for SiteSeer (reverse template) results. The cut-off used by SSM is the Z-score of the hit, whereas it is the E-value that is of interest in SiteSeer (reverse templates). **The ideal curve** would rise vertically from the origin and then horizontally out to the right, and would give an area under the curve of 1.

Overlapping

- In fact, in only **25 of the cases** did both methods return the same PDB file as their top hit.
- A further **25 cases** matched different PDB files but still obtained identical functional predictions.
- Of the remaining 32 cases, there were:
 - five where the reverse templates method found the correct match while SSM missed it
 - one case where SSM gave the correct answer and the reverse templates method was wrong

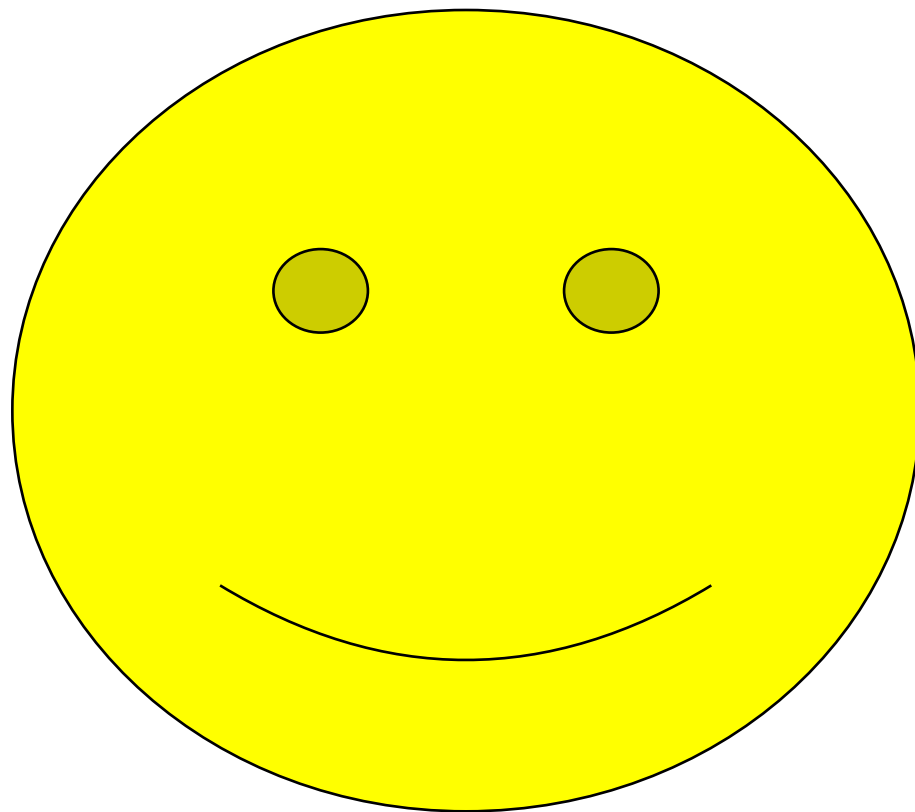
whether GO-slim terms can be used to assess the functional predictions in an automated way rather than requiring manual assessment ?

- the 77 proteins with GO annotation
- ProFunc results give a total of 207 structural matches:
 - 68 SSM fold match;
 - 74 reverse templates;
 - 8 enzyme templates;
 - 47 ligand templates;
 - 10 DNA templates

We tried both the generic GO-slims (31 terms) and our hand-curated molecular function (MF) GOslims (190 terms)

The cut-offs we tried were 25%, 50%, 75%, 100%, and a constrained 50% wherein a 100% match was required where the query protein has only two GO terms.

**Best results were obtained with a
75% cut-of on the MF-GO-slims**



Thanks !