UPPSALA
UNIVERSITET

Predicting protein function from its 3-dimensional structure

- Biological validation of protein function predictions

Anna Hennecke

Degree project in biology, 2005
Examensarbete i biologi 10 p, 2005
Biology Education Center and Linnaeus Center for Bioinformatics, Uppsala University
Supervisor: Torgeir R. Hvidsten

Summary

A classifier was developed by Hvidsten et al [6] that predicts the molecular function of a protein (described by Gene Ontology terms) from its 3-dimensional structure. I evaluated those predictions in order to find out whether they can assist the biological expert in finding proteins with missing Gene Ontology annotations and in the search for new protein functions. A method for systematic evaluation was developed. The method was applied on incorrect predictions (false positives) in a test set of partly annotated proteins. In average 6 % of the false positive predictions were evaluated to be actually true positives. Since the percentage of missing Gene Ontology annotations in the protein set is estimated to be higher, I can not state that false positive predictions are in general feasable to find missing annotations.

Many of the incorrectly predicted functions were still not complete nonsense predictions but had related functions to the actual protein function.
This suggests e.g. that substructures responsible for some functions (for example adenyl nucleotide binding and guanyl nucleotide binding) are structurally so similar that the classifier cannot distinguish between them.
In other cases proteins with an incorrect prediction for a certain function bind a substrate that was similar to the substrate bound by proteins that actually carry out this function. Those connections between function predicitions and actual functions might assist the biological expert in the search for new protein functions.

ABBREVIATIONS

Abbreviation	Explanation
AUC	Area under ROC Curve
EC	Enzyme Commision
FN	False Negative
FP	False Positives
GO	Gene Ontology
LCB	Linneaus Center for Bioinformatics, Uppsala University
MSD	Molecular Structure Database
mt	might be true
ND	Not Determined
PDB	Protein Database
rf	related function
t	true
TN	True Negative
TP	True Positive
ROC	Receiver Operator Characteristic
vc	vague connection

TABLE OF CONTENTS

2 Introduction

Proteins are life´s building blocks. In the tiniest intestinal bacteria, in plants, mice and men – in all living cells – proteins answer for both form and function. Naturally, research into proteins is therefore of greatest interest, especially for scientists wishing to know how things function at the molecular level.

20 amino acids with unique chemical properties are used to build a huge variety of amino acid chains that are then folded into 3-dimensional structures. The 3-dimensional fold of each type of protein permits it to carry out a specific molecular function, a molecular tool for the living cell to carry out functions from cutting DNA to transporting substances.

From the thousands of known proteins, the function is known for only a small number of them and just the structure or the amino acid sequence is known for others. Computational methods can help to determine the molecular functions and cellular role of proteins. The effort to accomplish high-throughput three-dimensional structure determination and analysis of biological macromolecules (like proteins) with computational methods is described by the term Structural Genomics. [5]

2.1 Gene Ontology

To make functional knowledge about proteins accessible for computational methods, a standardized vocabulary is needed to describe the proteins. The Gene Ontology Consortium provides a vocabulary, the **G**ene **O**ntology[1](GO) [1], for describing cellular

[1] Ontology (gr. ontos - to be, logos – word). In computer science, an ontology is the attempt to formulate an exhaustive hierarchical data structure containing all the relevant entities and their relationships. The computer science usage of the term ontology is derived from the much older usage of the term in a branch of metaphysics dealing with the nature of being.

roles of proteins. It is divided in three subontologies: One ontology contains terms for describing the function of a protein on molecular level, e.g. the GO terms *Protein kinase activity* or *Transferase activity* [2]. Each GO term comes with a precise definition of what function the term describes. E.g. Definition of *Protein kinase activity*: “Catalysis of the transfer of a phosphate group, usually from ATP, to a protein substrate”. The other two ontologies contain terms for describing the biological process a gene or gene product can be involved in and its location in the cell. To be able to describe the protein function on different detailed levels, the terms in the ontology are organized in a “parent-child-relationship” (Figure 1). A child term (more specialized term) can have man parent terms (less specialized terms).

[2] GO terms are written in italics in this report.

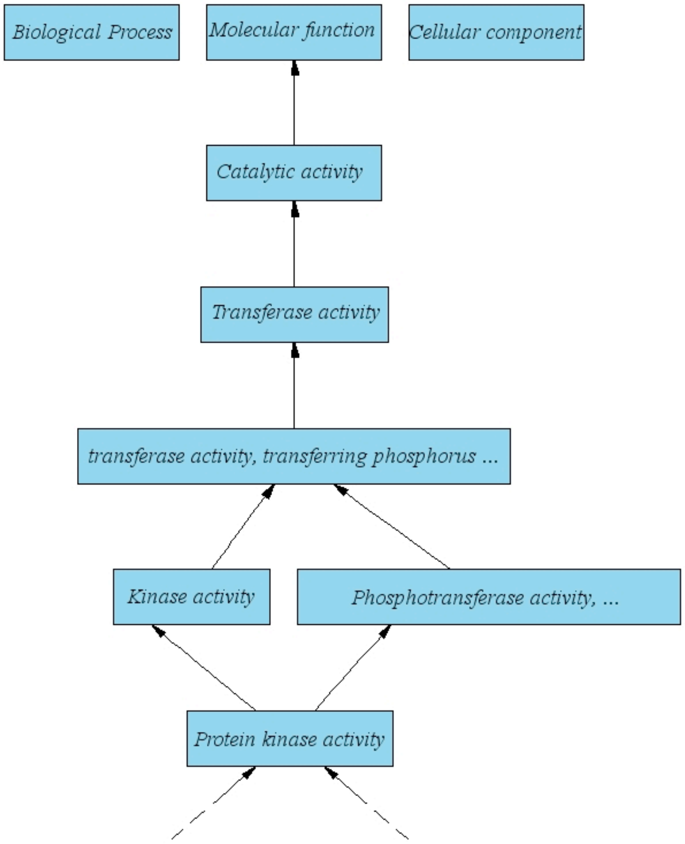

Figure 1 shows the organisation of the structured vocabulary provided by the GO consortium, examplified on the GO term *Protein kinase activity*. *Catalytic activity* is a parent term of the more specialized child term *Protein kinase activity.* They belong to the subontology describing molecular functions.

Proteins in databases, like the **M**olecular **S**tructure **D**atabase (MSD) [13], may be annotated with the Gene Ontology terms describing their function when the protein is characterised and knowledge about its function is available in the literature.

For example, the gene product cytochrome c can be described by the molecular function term *electron transporter activity*, the biological process terms *oxidative phosphorylation* and *induction of cell death*, and the cellular component terms *mitochondrial matrix* and *mitochondrial inner membrane*. Each GO term has an individual GO number, e.g. GO:0004672 *Protein kinase activity*. Through annotations the functional knowledge about proteins can be used for computational methods.

There are several ways a protein in a database can be annotated: Annotation by hand from curators in the GO consortium or from the scientists themselves who have characterised a protein or computer generated annotations. However, the GO annotation project is an ongoing process and it is known from experience that in publicly available databases a large number of annotations is still missing.

2.2 Databases

Below follows an overview of the databases that contain information about proteins. The focus lies on those used during the project. A good starting point to get information about a protein is Uniprot [2] which is crosslinked to various other databases. Some of those cross-references are also described here, with emphasis on what information can be found in those databases (catalytic activity, protein family, domains, literature links etc.)

- Uniprot [2]

The UniProt database consists of protein sequence entries. It is a container for protein sequence and function created by joining the information in the protein databases Swiss-Prot, TrEMBL and PIR. The joint information provide for each protein entry:

- core data consisting of sequence data,
- annotation data such as the description of protein function, catalytic activity, domains and sites, subunits, cofactors etc. and
- cross-references in form of pointers to information related to the protein entry and found in other data collections than Swiss-Prot. (e.g. the domains of the protein are named in form of pointers to Interpro and Pfam.)

• Enzyme (Enzyme nomenclature) [4]

Enzyme contains information concerning the nomenclature of enzymes. It describes each type of characterized enzyme for which an EC (**E**nzyme **C**ommission) number has been provided. It describes features that all enzymes of an EC class have in common. (Cofactors, Catalytic activity)

• Prosite [12]

Prosite is a database of protein families and domains. It contains profiles for protein domains and families grouped by similarities in their sequence. It aids the identification of newly sequenced proteins. Each of those profiles come with documentation providing background information on structure and function of these proteins.

• Pfam (**P**rotein **fam**ilies database of alignments and HMMs) [11]

Pfam is a collection of protein family alignments which were constructed semiautomatically using hidden Markov models (HMMs). Pfam families contain functional annotation and cross-references to other databases.

• Interpro (Integrated resource of Protein Families, Domains and Sites) [9]

Interpro is an integrated documentation resource for protein families, domains and sites. It combines several databases concerned with protein sequence classification like

UniProt, PROSITE, PRINTS, Pfam etc.

2.3 Strategies to predict the function of a protein

Predicting function from amino acid sequence and predicting function from protein structure are two different approaches to determine the molecular functions and cellular roles for proteins. Such automatic methods are primarily based on detecting proteins that have diverged evolutionarily from a common ancestor and then on inferring the function of the uncharacterized protein from its characterized homologues. Since structure is evolutionarily more highly conserved than sequence, and since only a few residues in functional sites need to be conserved for the function to remain stable during evolution, predicting function from structure often is more reliable than prediction from sequence.

Several strategies have been developed to assign a function or to suggest functional hypothesis for new structures. Martin et al. [8] used E.C. numbers as a measure of functional relationships and tried to identify the relationship between E.C. number and SCOP classification [7] or CATH [10] classification (both are databases that sort protein structure elements into classes, folds and families). Wallace et al. [14] assembled a database of 3-dimensional templates of active site residues. With this database, groups of residues in a query structure can be identified that are consistent with those in known active sites. In other approaches, the physical and chemical natures of proteins (electrostatic potential etc.) have been used to approach the analysis of a protein´s function.

The basic task in predicting a protein´s function with computational methods is to identify protein features like shape elements, electrostatic potentials, hydrophobicity patterns and sequence conservation, that are involved in the protein´s function. Feature-function- relationships can than be identified and used to infer the function of an unknown structure.

2.4 Computational methods for prediction

In order to answer the question which molecular function a protein with given features has, a classifier is needed that can assign the correct class (i.e. protein function) to the protein. A classifier is a mathematical function (=hypothesis) that was built by learning from examples. It should be able to predict the class of an object as good as possible when given the description of the object. The examples are called training set and are a set of already classified objects. The classifier is trained with cases from the trainings set and learns rules that connects features of the object (e.g. structure elements, hydrophobicity patterns etc. of a protein) to a certain class (e.g. a certain protein function).

When the classifier is constructed, the quality of its predictions need to be evaluated. One way to accomplish this is to apply the classifier on a test set. A test set contains objects that have not been used to train the classifier but the class for every object is known. If, for example, the test set contains proteins with different GO functions, the classifiers prediction can have four different qualities:

- **T**rue **P**ositive (TP): the classifier predicts the GO term correctly.
- **F**alse **P**ositive (FP): the classifier predicts a GO term where none exists.
- **F**alse **N**egative (FN): the classifier predicts no GO term where one exists.
- **T**rue **N**egative (TN): the classifier predicts correctly that no GO term exists.

Now interesting measures of a classifier´s quality can be calculated:

Specifity = TN/(FP+TN)

Sensitivity = TP/(TP+FN)

This measure is also called Coverage and it reflects how well members of the positive class are identified.

Precision = TP/(TP+FP)

Precision reflects how well members of the negative class are rejected.

A way to asses the prediction quality for a certain class (sometimes the classifier may predict a certain class better than another) is to calculate an AUC (**A**rea **u**nder the ROC[3] **C**urve) estimate value. With small trainings sets, this is done in k-fold cross-validation. In k-fold cross validation, the trainings set is split up in k subsets of equal size. k-1 of them are used to train the classifier again and one is used as a test set. This is done k times until every subset has been used once as test set. The number of FP, FN, TP and TN-predictions for each GO term are counted. The AUC takes the numbers of FP, FN, TP

[3] ROC (**R**eceiver **O**perator **C**haracteristics) is a graphical plot of the sensivity vs. 1- specifity

and TN for each GO term prediction into account and rates the classifiers performance in terms of numbers from 0.5 (no classification capability at all) to 1 (perfect classification).

2.5 The LCB classifier to predict function from structure

In the **L**inneaus **C**enter for **B**ioinformatics (LCB), Uppsala, a classifier was developed by Hvidsten et al. [6]. It predicts GO molecular functions[4] of a protein from the local descriptors of its 3-dimensional structure. Local descriptors are short 3-dimensional sequence fragments. A short summary of how the classifier was build by Hvidsten et al. follows below:

In the first step of building the classifier, a library of the most common (popular) local descriptors was created by browsing all available 3-D protein structures in the **P**rotein **D**ata**b**ase (PDB) [3] in an automatized manner. In a trainings set of characterized and annotated proteins, the structure of each protein was searched automatically for local descriptors. The local descriptors existing for a certain protein were then connected to the proteins GO molecular function by setting up IF- THEN rules.

IF descriptor_1 AND descriptor_9 THAN GO function_x

The large number of rules was reduced by computing the most relevant ones and used to build the classifier. The exact method will not be discussed here.

Then, the quality of the classifiers prediction performance was measured. This was done by calculating an AUC estimate value for each GO term in 10-fold cross validation.

[4] "GO molecular function" means all GO terms under the molecular function sub ontology.

The prediction method developed by Hvidsten et al. [6] predicts one or more GO functions for each protein. In the final step they tested the classifiers ability to identify the protein functions. It was applied on a test set of unseen proteins that were already partly annotated. Applied on this test set, the classifier could predict 50% of the new annotations (coverage equal to 50%) with four FP for each TP (precision equal to 20 %) for annotations with an AUC higher than 0.70[5]. For more details see Hvidsten et al. [6]

2.5 Aims

As mentioned, it is known from experience that in publicly available annotations, like in the test set, a large number of annotations is often missing. The aim of my work was to develop a method for systematic evaluation of the predictions and to use this method to evaluate the FP predictions [6] in the test set. In some cases this GO function just has not been annotated yet. That means the FP prediction is actually a TP prediction. I wanted to find out whether protein structure derived predictions could assist the biological expert in finding missing annotations [7] and in the search for new protein functions.

[5] An advanced classifier is currently under construction that uses more features than structure for predicition in order to improve those results.

[6] A FP exists when the classifier reports, incorrectly, that it has found a GO function where none exists in the test set.

[7] i. e. annotations for which information is present in databases or literature, but where this information has not been used to generate GO annotations.

3 Results

A table with a set of 404 proteins (referred to as test set) with their GO function annotations and the GO function predictions made by the LCB-classifier for those proteins was provided by T. Hvidsten, LCB, Uppsala University. Hvidsten compared the predictions automatically to existing annotations in test set and labeled them TP and FP (see materials and methods for more details about the protein set table). I selected 12 GO molecular function terms and evaluated all FP predictions for those GO functions. The predictions were sorted in the validation categories t, mt, rf or rf. Results are shown here. In order to get a feeling for how many annotations may be missing in the publicly available annotations and how feasible it might be to use false positive predictions to find them, the missing annotations for three GO terms were then quantified in the test set for three GO terms. Results are shown in the second part of this chapter.

3.1 Validation results for predictions of GO function

Table 1. Results of FP validation

GO term	AUC	*TP*	*FP*	*FN*	Validation Category				Number of “true FP” [e] in FP predictions
					t [a]	mt [b]	rf [c]	vc [d]	
GO:0004812 tRNA ligase activity	0,99	*3*	*7*	*1*	0			4	0
GO:0016160 Amylase activity	0,98	*1*	*4*	*0*	0		3		0
GO:0016886 ligase activity, forming phosphoric ester bonds	0,95	*4*	*19*	*2*	0			8	0
GO:0008235 metalloexopeptidase activity	0,93	*2*	*12*	*1*	1		1	7	1/12
GO:0042626 ATPase activity, coupled to transmembrane transport of substances	0,91	*2*	*7*	*2*	2	1			2/7
GO:0003755 peptidyl-prolyl cis-trans isomerase activity	0,90	*0*	*19*	*1*	0	1			0
GO:0015399 primary active transporter activity	0,90	*4*	*14*	*3*	3				3/14
GO:0004523 Ribonuclease H activity	0,84	*0*	*8*	*2*	0			1	0
GO:0030554 adenyl nucleotide binding	0,77	*25*	*42*	*17*	5	2	10	4	
GO:0004672 Protein kinase activity	0,77	*9*	*19*	*6*	0	1			0
GO:0003723 RNA binding	0,73	*6*	*44*	*16*	1 or more		7		1/44
GO:0003677 DNA binding	0,66	*36*	*70*	*24*	5 or more		5		5/70

[a] t- prediction is **t**rue

[b] mt- **m**ight be **t**rue. Not enough evidence to be sure.

[c] rf- **r**elated **f**unction. The predicted function is very close to the actual function.

[d] vc- **v**ague **c**onnection. The prediction has at least a (vague) connection to the actual function.

See materials and methods for detailed description of validation categories.

[e] A “true FP” is a FP prediction (according to test set table) which was found by my evaluation to be actually a TP prediction.

Table 1 shows the predicted GO molecular function, the AUC for this term and the number of proteins with a TP, FP or FN prediction in the test set. Numbers that are taken from test set table provided by T. Hvidsten are shown in italics. Note: this is the number of TP, FP, FN predictions BEFORE the FP predictions were validated. On the right side of the table, evaluation results of the FP predictions are shown. The percentage of FP that were found to be actually TP ("true FP" [8]) are given.

The number of FP predictions per GO term varied between 4 and 70 and between 0% and around 20 % (2 "true FP" in 7 FP and 3 "true FP" in 14 FP) of the FP predictions were "true FP". Out of 266 FP predicitions, 17 were evaluated as "true FP". This means that 17 missing annotations were found by using FP predictions. On the average 1 out of 16 FP predictions were true. 1 in 10 FP were true if DNA and RNA binding were not considered. These were only searched with search mode "b) search by protein name" (with this search mode less of the relevant existing information was found than when more intensive search literature and database search would were used).

The classifier predicted the 12 GO classes that were validated with an average precision

[8] They are called "true FP" to distinguish them from those predictions that were already scored TP in Hvidstens test set table. Example for a "true FP": the test set protein "Deoxynucleotide monophosphate kinase" had one publicly available GO function annotation: *transferase activity, transferring phosphorus-containing groups.* The LCB-classifier predicted two GO functions: *transferase activity, transferring phosphorus-containing groups* and *adenyl nucleotide binding*. The latter is labelled a FP prediction in the test set table. I validated this prediction and found that the protein actually does bind adenyl nucleotides. Therefore the FP prediction in the test set table was actually a TP prediction and is referred to as a "true FP" prediction. A GO annotation in MSD for *adenyl nucleotide binding* had not been generated yet.

of 25 % and a coverage of 55 %. Percentages were calculated from FP,TP, FN, TN numbers shown in Table 1. Considering the validation results that showed that 17 FP predictions were actually TP predictions ("true FP"), those numbers increase to 30% and 59% respectively.

It is interesting to note that the validation process revealed that even FP predictions that could not be evaluated to be true often were not complete nonsense predictions: 42 of the FP predictions were scored as being "related functions" or as having "vague connections" to the known protein function. Some of the FP predictions scored as 'related functions' or 'vague connections' are described in the following.

3.2 Related functions and vague connections

GO: tRNA ligase activity - 4 of 7 proteins with incorrectly predicted *tRNA ligase activiy* carry out an enzymatic function also involving *nucleic acid binding*. Those enzymatic functions were evaluated as vague connections. It seems that the classifier had difficulties in distinguishing between different types of enzymatic activities when the substrate is some kind of nucleic acid (DNA or RNA). In 6 of 7 cases, *tRNA ligase activity* was predicted together with *nucleic acid binding* (*DNA binding* or *RNA binding*).

GO: ligase activity, forming phosphoric ester bonds. This GO term is a parent term of *tRNA ligase activity*. 8 of 19 FP proteins with an incorrectly predicted ligase activity have the known function *GO: nucleic acid binding.*

GO: ligase activity, forming phosphoric ester bonds [9] in a protein comes always together with *GO: nucleic acid binding,* which therefore was evaluated a vague connection. The two substructures that carry out the functions always occur together in a protein. Presumably, *GO: nucleic acid binding* annotations are missing more often. I suggest that this might cause a difficulty for the classifier when setting up rules to predict the *GO function ligase activity, forming phosphoric ester bonds.*

GO: Metalloexopeptidase - 6 out of 12 proteins with a FP prediction had the known GO function*: metal ion binding. Metal ion binding* is a function that all proteins with GO function *metalloexopeptidase* have in common. In the FP group, 50 % of the proteins have this GO molecular function. This number is significantly different from the number of proteins with the function *metal ion binding* in the test set: Only 25 % (100 out of 404) of all proteins are annotated with *metal ion binding*.

Reason for this might be the same as suggested for *GO: ligase activity* – structural similarities in metalloexopeptidas-proteins and other metal ion binding proteins lie in the metal ion binding site. During the validation process the impression grew that annotation of binding activites like *GO: metal ion binding* lack more often than annotations of enzymatic activity. Because of this, in the course of building the classifier, incorrect rules connecting the metal ion binding site to an enzymatic activity might have been set up.

[9] It has two main child terms: *GO: DNA ligase activity and GO: RNA ligase activity*

However, the differences in the numbers might be due to missing annotations in the test set or to the fact that the FP group is too small (12 FP in the test set). Annotations in the test set need to be completed in order to evaluate this result properly.

GO: Ribonuclease H activity - In 7 of 8 cases of a FP prediction of this GO term, it was predicted together with *nucleic acid binding*. But the protein was annotated by this GO term only in one of those cases. The prediction is evaluated as a vague connection since *Ribonuclease H activity* always occurs together with *nucleic acid binding* in a protein.

GO: Adenyl nucleotide binding -10 out of 42 proteins with a false *adenyl nucleotide binding* predicition carry out related functions. The guanyl molecule and uracyl molecule have the same basic structure as adenyl and differ only in a few functional groups. This suggests that the 3-dimensional structure in binding sites for adenyl nucleotides of these proteins are structurally so similar to bindings sites for guanyl- uracil and NAD-binding sites that the classifier can not distinguish between them.

Binding of adenyl nucleotides in chemical compounds like NAD and NADH was also considered to be a related function. However, the molecular functions *GO: NAD/NADH/NADPH* are not closely related to *adenyl nucleotide binding* in the parent-child hierarchy of Gene Ontology.

GO: RNA binding and **GO: DNA binding** were only searched by protein name, with a more intensive search mode probably more FP would be validated as true.

Summarizing the observations described above I could see (for some GO terms) a tendency that proteins with the same incorrect (FP) GO function prediction often bind a substrate/cofactor that is identical to the substrate bound by proteins that actually carry out this GO function. For example, a significantly high number of proteins with incorrectly predicted *metalloexopeptidase activity* are annotated to *metal ion binding* and proteins with a known annotation by *metalloexopeptidase activity* are always also annotated by *metal ion binding.* The same observation was made for enzymatic *GO functions* (e.g. tRNA ligase activity, ligase activity, forming phosphoric ester bonds) that appear in a protein always together with the binding GO *function*: *nucleic acid binding*. Furthermore, FP predictions may be not nonsense predictions but may contain structural/functional information about the protein (e.g. adenyl- and uracil binding sites might have similar binding sites.

3.3 Missing annotations in test set of proteins

The fast search mode " b) search by protein name" was applied on all 404 test set proteins to get a picture of how many annotations are missing, and how many of those are found by the classifier. The real percentage of missing annotations would be expected to be higher since database and literature search would discover even more annotations.

Table 2 Missing annotation in test set

	AUC	Number of annotations in test set	Missing annotations	Percentage of missing annotations	Missing annotations found by classifier	FP	TP	TN+FN [f]
GO:0003677 DNA binding	0,6641	60	15 (or more)	23 %	5	70	36	298
RNA binding	0,7346	22	7 (or more)	24 %	1	44	6	354
GO:0004672 Protein kinase activity	0,7691	15	14 (or more)	48 %	0			

[f] The negatives (TN+FN) result from the calculation "all test set proteins" - FP - TP = TN + FN

Table 2 shows that 36 missing annotations were discovered in the test set by search mode b) search by protein name (see materials and methods). About a fourth of all “DNA binding” and “RNA binding” annotations were missing in the test set of proteins.

At least 50 % of all “protein kinase activity” annotations were found to be missing in the test set. None of them were found by the classifier. It can be assumed that the lack of annotations in the training set is similar. It would be useful to evaluate the effect of missing annotations in the training set. The cross validation AUC estimates were designed to measure prediction performance [6]; the AUC value of 0.76 for GO protein kinase activity is relatively low [10]. The extensive lack of annotation might cause difficulties during rule learning, resulting in a poorer prediction quality which is expressed in the AUC value.

The classifier was able to find 5 of the 15 “DNA binding” annotations missing in the test set and 1 of the 7 “RNA binding” annotations missing in the test set. I wanted to know if

[10] 1 is perfect classification capability, 0.5 is no classification capability at all. Hvidsten et al [6] showed that precision and coverage values of the LCB-classifier were best when they considered only predictions for GO terms with AUC´s over 0.7.

there were more proteins with a missing DNA binding annotation in the FP group than in the rest of the test set (i.e. FP group + TN group + FN group) . In that case, the FP predictions could assist the biological expert in finding missing annotations.

5 out of 70 (7 %) proteins with a FP prediction *DNA binding* are “true FP”, i.e. they miss this annotation. In contrast 15 of 368 (FP+FN+TN) proteins in the test set are FN and miss this annotation (4.1 %).

In other words: The biological expert would discover a protein that binds DNA in every 25th protein if he would be searching the whole test set and in every 14th protein if he would be searching only in the FP group instead. Therefore, the FP predictions may be a help for the biological expert to find proteins that miss annotation of the GO term *DNA binding*, because proteins with a missing binding annotation occur slightly more often in the FP group than in the whole test set. Still, in order to find all proteins with a missing annotations, the whole test set must be searched since the FP group lacks two third of the missing annotations.

For the GO function *RNA binding*, the number of proteins with a missing annotation in the FP group was not significantly different from the number of proteins with a missing annotation in the whole test set: every 44th protein (1 in 44 FP) in the false positive have a missing *RNA binding* annotation compared to every 57th protein (7 in 398) test set proteins) in the test set. For *Protein Kinase activity* none of the proteins that miss this annotations occurred in the FP group. Apparently, the feasibility of FP to find missing annotations varies from GO term to GO term and needs to be evaluated for each one separately. I can not say in general how useful FP predictions are to find missing annotations.

4 Discussion

4.1 Feasability of FP predictions to find missing annotations

On the average 6 % of the FP predictions were evaluated as “true FP”. The overall percentage of missing annotations in Gene Ontology probably is to be higher than 6% and I can not state it is generally feasable to find missing annotations from FP predictions. However, the number of “true FP” varied from GO term to GO term. While there was no missing annotation found for one half of the evaluated GO terms, the other half of the GO terms had several “true FP” in the FP group. Obviously, the FP groups for certain GO terms does contain a number of proteins missing that annotation. But since the percentage of missing GO terms in the test set is not known, I can not quantify the usefulness of using FP predictions to find missing annotations for each single GO term. Future work is needed to evaluate the FP predictions for all GO terms to be able to base the numbers on more results.

In order to try to estimate the potential for finding missing annotations from protein structure generated predictions, I quantified missing annotations for three GO terms (*DNA binding*, *RNA binding* and *Protein Kinase activity*) in the whole test set. I found that among proteins with a FP prediction *DNA binding* proteins that actually miss this annotation were present slightly more often, while FP predictions for *Protein kinase* *activity* were actually incorrect predictions so that it was not feasable to find missing annotations in this FP group.

4.2 Impact of structural similarities between proteins with different GO functions on performance of the classifier

A number of proteins in the FP groups carry out related functions or have at least a vague connection to the predicted function. The validation categories 'related function' and 'vague connection' were used to see if the classifier could be used to infer structural and functional knowledge about the proteins. In the FP group of GO *adenyl nucleotide binding*, 10 out of 42 FP proteins bind compounds that are structurally very similar to adenyl nucleotides (e.g. guanyl and uracil nucleotides). The 3-D structure of binding sites might be too similar for the classifier to distinguish. Learning and prediction at a higher GO level, i.e. taking a parent or grandparent term for learning and prediction, is therefore suggested to improve performance of the classifier(e.g. GO:0000166 *Nucleotide binding* is parent term of *adenyl nucleotide binding, uracil nucleotide binding* and *guanyl nucleotide binding)*.

It would be interesting to check automatically the relationship between predicted function and annotated function. This would show if learning and prediction at a higher GO level would maybe improve the prediction quality for some GO terms.

4.3 Impact of missing annotations in training set on performance of the classifier

An interesting observation was made in connection to “related functions” and “vague connections”. Proteins with a FP predictions of a certain GO function often bind a substrate that is identical or similar to the substrate bound by proteins that actually carry out this function. As indicated in section 5.3 it appeared that annotations for binding activities like *nucleic acid binding* and *metal ion binding* are missing to a greater extent than annotation of enzymatic activities in these proteins. It might be the case that those proteins in the trainings set have structural similarities in the binding site while the structures of the active site differ according to the enzymatic function. This might cause difficulties, since it could mean that incorrect rules connecting local descriptors of the metal binding site to an enzymatic activity might be set up in the course of building the classifier.

It would be interesting in the future to evaluate what effect missing annotations in the training set have on rule learning by the classifier, not only for the prediction of the GO term itself but also for other GO terms that often occur together in a protein (e.g. do missing *nucleic acid binding* annotations effect prediction quality of GO functions like *tRNAse activity*, that use nucleic acids as a substrate?) This could be done easily by erasing all annotations for one GO term in the training set and doing rule learning with this modified training set.

The effect of missing annotations on a GO term itself could be evaluated by creating various modified training sets, each one with a different number of missing annotations for this GO term. Changes in prediction performance could be measured by changes in the AUC value.

4.4 Future steps to faciliate validation GO function predictions

Future work on the search method itself could be to faciliate the search for validation evidence for predictions by automatical methods. Starting from the search method described in this report, a semiautomatic search method could be build up which could be based on the list of search criteria that was set up for each GO term, e.g. textmining like searching for the keyword “ATP” in the category “CATALYTIC ACTIVITY” of Uniprot. Keyword search in protein name and search for EC number can also be performed automatically based on the list of keywords.

5 Materials and methods

5.1 Method for systematic validation of predictions

FP predictions were validated for one GO term at a time for all concerned test set proteins instead of validating all false positive predictions for one protein at a time.

5.2 The test set

A table of 404 proteins (referred to as the test set) with their publicly available GO molecular function[11] annotations (annotations from **M**olecular **S**tructure **D**atabase (MSD) in february 2005) and the GO molecular function predictions made by the LCB-classifier for those proteins was provided by Torgeir Hvidsten, LCB, Uppsala University. Table 3

[11] “GO molecular function” here means all GO terms under the molecular function sub-ontology.

shows a fragment of the test set table. The proteins in the test set were not chosen according to certain criterias. It included all proteins that had been newly annotated by the GO consortium since the proteins for the training set were accquired from MSD and the classifier was built, i.e. a random selection of test set proteins. Between 0 and 18 predictions were made for each protein. The predictions were compared automatically by Hvidsten to the existing annotations and TP predictions were labelled with three exclamation marks. The proteins originate from a wide range of species from viruses to mammals and carry out several kinds of functions which are not more specified here.

Table 3: Fragment of the test set table provided by T. Hvidsten, LCB, Uppsala University.

PDB Entry ID	Protein name	GO molecular function annotations	GO molecular function predictions
1g72	Methanol dehydrogenase subunit 1, Methanol dehydrogenase subunit 2 [Precursor]	GO:0016616 *oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor* GO:0005509 *calcium ion binding*	GO:0005509 !!![12] GO:0016702 GO:0016747 GO:0016798 GO:0005125 GO:0004252 GO:0046914 GO:0008233 GO:0008083 GO:0004888
4crx	Recombinase cre	GO:0003677 *DNA binding*	GO:0003677 !!![13] GO:0030554 GO:0004601 GO:0003916 GO:0004672 GO:0015399 GO:0042623 GO:0042626 GO:0016818
1gde	389aa long hypothetical aspartate aminotransferase	GO:0008483 *transaminase activity*	GO:0008483 !!![14] GO:0016741 GO:0016747

5.3 Selection of GO terms

The prediction method developed by Hvidsten et al [6] set up rules to predict 72 different GO functions. I validated the FP predicitions for 12 of those. I chose those 12 GO terms

[12] The predictions were compared automatically to the existing annotations and true positive (TP) predictions were marked with three exclamation marks.

[13] See footnote 12

[14] See footnote 12

with the goal to have a selection of GO terms that included a range of different activities, from binding to enzymatic activities and AUC values ranging from 0.66 to 0.99.

5.4 Finding search criteria

Evidence for evaluation of FP predictions was found in the literature and in databases. The search was based on a list of search criteria. These search criteria were expressed as a list of keywords for each GO term and were inferred from the biological background knowledge associated with this GO term. This biological background knowledge was found from the GO term and its definition as well as by analysizing TP predictions in the test set and by analysizing proteins annotated to this GO term in Gene Ontology. (See Appendix A for lists of search criteria.) Depending on the GO term, different search strategies were effective to find evidence information for the concerned protein.

5.5 Search modes

Gene Ontology molecular function terms describe different types of protein activities, e.g. binding activity or enzymatic activity. Therefore, different search modes were appropriate for each GO function.

a) **Search by EC class:** For some enzymatic activities the GO term definition corresponds to an EC class. In those cases, the EC class was used as a search criteria as well. E.g. the GO function *tRNA ligase activity* is only performed by members of the EC class 6.1.1.- “Aminoacyl-tRNA synthetases” EC 6.1.1.-. Binding (e.g. ATP binding), on the other hand, occurs in many different enzyme classes. Therefore, this search mode was usually not appropriate for binding activities. In Appendix A appropriate GO terms and their EC classes are stated.

b) **Search by protein name:** The protein name in some cases reveals information on GO molecular function terms that can be used to describe its activity. E.g. protein names like Nuclease, Ribosomal Protein, Restriction enzyme hint to the GO term *DNA binding*. This fast search mode was useful when the EC class was not given. See list of "hinting" protein names for appropriate GO terms inferred from the GO terms biological background in Appendix A.

c) **Search by keywords in databases**: This was the most commonly used search mode. Validation started at Uniprot [2]. Varying amounts of information about function, catalytic activity etc. were found here. Cross-links to Pfam, Interpro, Enzyme provided additional information about domains in the searched protein, its protein family as well as its nomenclature class. Those databases as well as Entrez Gene were used to find hints to validate annotations for the given protein. Links to following databases were followed: **Enzyme [4]** contains for example information like the reaction catalyzed and cofactors for all proteins belonging to the same EC class as the searched protein. The "Comments" section of the Enzyme database was valuable sometimes for validation because it gives additional (in addition to Uniprot) information about e.g. special features for some members of this class. This was valuable because it gave information not only about the single protein for which a GO term was to be validated, but it gave information about all proteins in the same nomenclature class, which might be better characterized. **Prosite [4]** gave more extensive information about and description of the domains in the searched protein than Uniprot, e.g. information about function of other proteins sharing this domain. If this protein family in general or a member of this familiy was involved in the searched protein´s function, this could give useful hints when validating a predicted GO molecular function for individual proteins.

d) **Search by keyword in literature**: This most time consuming search mode was applied if search in databases didn´t provide information about the protein or if it suggested that the predicton might be true (e.g. predicted activity was found in a similar protein in other organisms or a related protein family etc.) but didn´t give enough evidence. Sources of information were literature links in Uniprot [6] as well as keyword search in Pubmed [15].

General strategies:

e) **Using existing annotations**: An existing annotation could hint to the existence of the searched annotation. E.g. an already existing annotation GO: *Protein folding* might hint that the FP prediction GO: *Peptidyl-prolyl-cis-trans isomerase* was true because protein folding often involves those isomerasing proteins. E.g., for a protein with annotated GO: *Kinase activity* the FP prediction GO: *adenyl nucleotide binding* must be true because all kinases use ATP (which is an adenyl nucleotide). See Appendix A for a list of “hinting” GO terms.

5.6 The validation categories

The idea during validation was to get a picture of what the protein does and than judge if the predicted GO term could be involved in that activity.

Using the search methods described above the FP predicitions were evaluated. The predicted GO terms were scored according to five different categories:

1) **True (t)**: The prediction was true. Source of evidence (database, literature) was given. The annotation for this protein is missing in Gene Ontology.

2) **Related function (rf):** The predicted function was very close to the actual function. This suggested that the binding or active sites in the protein are structurally very similar for predicted and actual function. Two related functions showed in some but not all cases close relation in the GO tree as well.

3) **Might be true (mt):** Literature and databases hinted to that annotation might be true, but further experiments will have to be carried out to be certain. A typical case would be that the predicted activity of this protein was found in another organism but not in the organism in the protein set.

4) **Vague connection (vc):** The prediction had at least some connection to the known protein function. There was presumably no structural similarity between those two functions but they often occur together in a protein. If for example the actual function was *RNA binding*, the prediction *tRNA ligase activity* has some connection to this because this activity involves RNA binding, in other words: a protein with an active site for *tRNA ligase activity* has always a binding site for RNA as well. Another example of a vague connection would be if actual and predicted function stated the same activity (e.g. hydrolase activity) but on a slightly different substrate. In this case, there might be a structural similarity between active sites in the protein for actual and predicted function.

5) **Not true (nt):** The prediction was not true and did not belong to one of the categories above.

6 References

1. Ashburner, M., Baöö. C. A., Blake, J. A. Botstein, D. Butler, H., Cherry, A., Hill, D.P., Issel-Tarver, L. kasarskis, A., Lewis, S. Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-9.
2. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS**.** UniProt: the Universal Protein knowledgedatabase. (2005) *Nucleic Acids Res*. 33
3. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov and Philip E. Bourne (2000) The Protein Data Bank. *Nucleic Acids Research* 28:235-242
4. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids* Res. 31:3784-8
5. [12] Goldsmith-Fischman S, Honig B. (2003) Structural genomics: computational methods for structure analysis. *Protein Sci.* 12:1813-21
6. Torgeir R. Hvidsten, A. Kryshtafovych, J. Komorowski and K. Fidelis. (2004)Predicting function from local substructures of proteins. *Submitted*
7. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP (2000) A structural classification of proteins database. *Nucleic Acids Res 28:257-9*

8. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM. (1998) Protein folds and functions. *Structure* 6:875-84

9. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley R, Courcelle E, Durbin R, Falquet L, Fleischmann W, Gouzy J, Griffith-Jones S, Haft D, Hermjakob H, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Orchard S, Pagni M, Peyruc D, Ponting CP, Servant F, Sigrist CJ; InterPro Consortium. (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform*. 3:225-35

10. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. (1997) CATH-a hierarchic classification of protein domain structures. *Structure 5:1093-108*

11. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26:320-2

12. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform.* 3:265-74

13. Velankar S, McNeil P, Mittard-Runte V, Suarez A, Barrell D, Apweiler R, Henrick K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids* Res. 33:D262-5

14. Wallace AC, Borkakoti N, Thornton JM. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci. 6:2308-23*

Appendix A

A.1. Description of GO term and search criteria

In this chapter, a short description of each GO term and its biological background is given. Based on its biological background, analysis of TP predictions and analysis of proteins annotated to a certain class in Gene Ontology, searching criteria were set up for each term. Searching criteria include a list of keywords for search in databases and literature and definition of “related functions” and “vague connections” are given. The lists are the base of the search and show therefore to which extend and to which limits databases are searched, they do not claim to be perfectly complete. Related functions and vague connections are mainly set up during evaluation process, only those are listed which were actually found in proteins with a false positive prediction.

A.1.1 GO:0030554 adenyl nucleotide binding (2804)

Gene Ontology definition: Interacting selectively with adenyl nucleotides, any compound consisting of adenosine esterified with (ortho)phosphate. [1]

In most cases (2762 of 2804 in Gene Ontology) adenyl nucleotide binding means ATP binding (child term). ATP hydrolysis is often used as energy source for active transport, synthesis, ligation and motility a well as Cotransmitter/Transmitter in some neurons.

The nucleic acid Adenin is component of other chemical compounds than adenyl nucleotides (NAD, cAMP etc.). Binding sites for those compounds might be similar to adenyl nucleotide binding sites and are therefore considered related functions or vaguely connected when adenine as a part of a DNA strain is recognized.

Table A.1. Search criteria for *adenyl nucleotide binding*

True	Related functions	Vage connection
Binding of compounds including adenyl nucleotides: ATP binding ADP binding AMP binding FAD binding **ATP binding enzymes** Kinase (GO:0016301 Kinase activity) ATPase Adenylate cyclase Restriction enzyme Type I and III **Acceptors for ATP as a neurotransmitter**	**Binding of guanyl nucleotides, their chemical structure is very close to adenyl nucleotides** Guanyl nucleotide binding (GO:0019001) GTP binding (sibling term) GDP binding (sibling term) GMP binding (sibling term) **Binding of adenin containing compounds other than nucleotides** NAD (Nicotinamide adenine dinucleotide) binding (GO:0051287) NADH binding (GO:0051288) NADPH binding (GO:0050661) cAMP binding **NAD binding enzymes** Dehydrogenases **Proteins involved in cAMP signaling pathways - cAMP synthesis and degradation** Adenylate cyclase (GO:0004016) cyclic-nucleotide phophodiesterase activity (GO:0004112)	- DNA binding if a specific A or G containing sequence is recognized.

This category is suitable for halfautomatical textmining involving Emzyme db. (e.g. get word "ATP" in catalytic activity.)

A.1.2 GO:0004812 tRNA ligase activity 0,9933

Gene Ontology definition: Catalysis of the formation of aminoacyl-tRNA from ATP, amino acid, and tRNA with the release of pyrophosphate and AMP. **[1]**

This very specific activity is only performed by members of the class "Aminoacyl-tRNA synthetases" EC 6.1.1.- . The EC class can therefore be used as a search criteria. As the first step in protein biosynthesis they activate amino acids by binding the carboxyl group of an aminoacid with ester linkage to Adenosin in the CCA sequence of a tRNA molecule using ATP as energysource. A protein has to bind at least those three components (ATP, aminoacid and the nucleic acid tRNA) and produce aminacyl-t-RNA to be classified as this.

The sibling term DNA ligase activity has a related function. It performes the ligase reaction with DNA instead of RNA. It builds the ester linkage between two nucleotides and not between a nucleotide and an aminoacid. This might be performed by a similar structure.

Nucleic acid binding in general and enzymatic activities with nucleic acid substrate are considered to have a vage connection to the GO term tRNA ligase activity.

Table A.2. Search criteria for *tRNA ligase activity*

True	Related functions	Vage connection
Members of **EC 6.1.1.- class -** Aminoacyl-tRNA synthetases tRNA AND ATP AND tRNA binding	DNA ligase activity (sibling term) (GO:0003909)	**Nucleic acid binding** DNA binding RNA binding tRNA binding **Enzymatic activities with nucleic acid substrate** Polymerase activity Translation factor activity Restrictionenzyme activity

A.1.3 GO:0016886 ligase activity, forming phosphoric ester bonds (584)

Gene Ontology definition: Catalysis of the ligation of two substances via a phosphoric ester bond with concomitant breakage of a diphosphate linkage, usually in a nucleoside triphosphate. [1]

582 proteins are annotated to GO term 0016886, the majority (530) have tRNA ligase activity, 39 DNA ligase activity. So for most of the proteins apply the same rules as for the above term G:004812 tRNA ligase activity.

A.1.4 GO:0003755 peptidyl-prolyl cis-trans isomerase activity (215)

Gene Ontology definition: Catalysis of the reaction: peptidyl-proline (omega=180) = peptidyl-proline (omega=0). [1]

Proteins excersing this GO activity are called Peptdidyl-prolyl Isomerases (PPIases). They are usually involved in protein folding and stabilization, which occurs mainly while forming the secondary and tertiary structure after translation. The cis-trans isomerization of peptidyl-proline bonds is thought to be one of the rate-limiting events in protein folding, hence isomerase (also called rotamase) activity is presumably involved in accelerating conformational transitions in folding intermediates.

If a FP protein is involved in protein folding (GO:0006457), this might hint the predicted "peptidyl-prolyl cis-trans isomerase activity" is true.

Bacterial rotamases are for example rotA, parvulin, trigger factor and SlyD. Also cyclophilins and FK 506 binding proteins have been shown to accererate folding of some proteins **[2]**

No related functions or vage connection terms are defined for this term.

Table A.4. Search criteria for *ligase activity, forming phosphoric ester bonds*

True	Related functions	Vage connection
Enzymes involved in Protein folding and stabilization GO:0006457 Protein folding PPIases Isomerase Rotamase rotA parvulin trigger factor SlyD Cyclophilin FK506 Proline	ND	ND

A.1.5 GO:0015399 primary active transporter activity (1048)

Gene Ontology definition: Catalysis of transport of a solute against a concentration gradient using a primary energy source. Primary energy sources known to be coupled to transport are chemical, electrical and solar sources**. [1]**

Primary active transporter activity is found in one main class of membrane proteins that shift specific molecules across the membrane: carrier proteins. Carrier proteins can be coupled to a source of energy to catalyze active transport. A broad range of protein families with very different structures have this transport activity: transport ATPase family – P-type and F-type ATPases (structurally different), ABC transporter (each member contains two highly conserved ATP binding cassettes).

All proteins that are transporters using a primary source of energy are classified “true” in contrast to a transporter working through “facilitated diffusion or chemiosmotic energy”. Furthermore, this GO term does not include “any molecular entity that serves as an electron accetor and electron donor in an electron transport system.

No related functions or vage connection terms are defined for this term.

Table A.5. Search criteria for *primary active transporter activity*

True	Related functions	Vage connection
Transporter activity **Carrier proteins** Transport ATPases: P-type transport ATPase family F-type ATPases also known as ATP synthases ABC transporter: MDR- multidrug resistance protein Molybdate-transporting ABC transporter (GO:0015412) **Child terms** cytochrome-c oxidase activity GO:0015451 : decarboxylation-driven active transporter activity GO:0015454 : light-driven active transporter activity GO:0015452 : methyl transfer-driven active transporter activity GO:0003957 : NAD(P)+ transhydrogenase (B-specific) activity GO:0008137 : NADH dehydrogenase (ubiquinone) activity GO:0015453 : oxidoreduction-driven active transporter activity GO:0015405 : P-P-bond-hydrolysis-driven transporter activity GO:0008121 : ubiquinol-cytochrome-c reductase activity	ND	ND

A.1.6 GO:0042626 ATPase activity, coupled to transmembrane movement of substances

Gene Ontology definition: Catalysis of the reaction: ATP + H_2O = ADP + phosphate to directly drive the transport of a substance across a membrane. [1]

No related functions or vage connection terms are defined for this term.

Table A.6. Search criteria for *ATPase activity, coupled to transmembrane movement of substances*

True	Related functions	Vage connection
Transporter activity **Carrier proteins** Transport ATPases: P-type transport ATPase family F-type ATPases also known as ATP synthases ABC transporter: MDR- multidrug resistance protein Molybdate-transporting ABC transporter (GO:0015412)	ND	ND

A.1.7 GO:0008235 metalloexopeptidase activity (191)

Gene Ontology definition: Catalysis of the hydrolysis of terminal peptide linkages in oligopeptides or polypeptides. [1] Enzymes of this class contain a chelated metal ion essential to their catalytic activity at their active sites.

The GO term corresponds to the EC class 3.4.17.22 Metallocarboxypeptidase.
This is a subclass of the big class of enzymes hydrolasing peptide bonds. (EC 3.4 Peptidases)
The mechanism used to cleave a peptide bond involves in all peptidases making an amino acid residue or a water molecule nucleophillic so that it can attack the peptide carbonyl group.

A FP prediction is only validated as “true”, if the protein belongs to this class.

Metalloexopeptidase activity requires metallion and protein binding and a hydrolase activity on carbon-nitrogen bonds. If any of those are present in the protein with a FP "metalloexopeptidase activity" prediction, this prediction is evaluated "vc", i.e. the actual function has a loose connection to the prediction. Even the uncle term "hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds" is evaluated "vc"

Table A.7. Search criteria for *metalloexopeptidase activity*

True	Related functions	Vage connection
Members of EC 3.4.17.- Metallocarboxypeptidases	Metalloendopeptidases	Protein binding Metalion binding **Enzymes with peptidase activity** Members of EC 3.4 Serine peptidase Threonine peptidase Cysteine peptidase Aspartic acid peptidase Glutamic acid peptidase **Enzymes with hydrolase activity, acting on carbon-nitrogen (but not peptide bonds)**

A.1.8 GO:0016160 Amylase activity (64)

Gene Ontology definition: Catalysis of the hydrolysis of amylose or an amylose derivative. [1]

Table A.8. Search criteria for a*mylase activity*

True	Related functions	Vage connection
Members of EC 3.2.1.1 Alpha Amylase **EC 3.2.1.2 Beta Amylase** **3.2.1.68 Isoamylase**	Hydrolase activity, acting on glycosyl bonds GO:0016798 (parent term)	Hydrolase activity GO: **Members of EC class 3**

A.1.9 GO:0003677 DNA binding (9675)

Gene Ontology definition: Interacting selectively with DNA (deoxyribonucleic acid). [1]

Table A.9. Search criteria for *DNA binding*

True	Related functions	Vage connection
DNA binding enzymes Nuclease Replication initian proteins Restriction enzymes Polymerase **Members of EC 3.1.11.- to EC 3.1.16 and EC3.1.21 to 3.1.22 and 3.1.25 to 3.1.27 Ribonucleases**	RNA binding **RNA binding enzymes** tRNA synthetase ribonucleoprotein Ribosomal protein	ND

Information sources

Validation of FP and search for missing annotations performed ONLY for keywords (from search criteria) in protein name and by EC class in Enzyme (http://au.expasy.org/enzyme/). Evidence check in Uniprot.

A.1.10 GO:0003723 RNA binding (2597)

Table A.10. Search criteria for *DNA binding*

True	Related functions	Vage connection
RNA binding enzymes tRNA synthetase ribonucleoprotein Ribosomal protein	DNA binding **DNA binding enzymes** Nuclease Replication initian proteins Restriction enzymes Polymerase	ND

A.1.11 GO:0004523 Ribonuclease H activity (89)

Gene Ontology definition: Catalysis of the endonucleolytic cleavage of RNA in RNA-DNA hybrids to 5'-phosphomonoesters. [1]

EC class 3.1.26.- contains different Ribonucleases. E.g. Ribonuclease M5, Ribonuclease P etc. The specific reaction of endonucleolytic cleavage to 5`phosphomonoester as described in the Gene Ontology definition is only performed by members of the EC class 3.1.26.4 called Ribonuclease H.

Table A.11. Search criteria for *Ribonuclease H activity*

True	Related functions	Vage connection
RNase H activity **Members of EC class 3.1.26.4**	Ribonucleases	**Nucleic acid binding** DNA binding RNA binding tRNA binding **Enzymatic activities with nucleic acid substrate** Polymerase activity Translation factor activity Restrictionenzyme activity

A.1.12 GO:0004672 Protein kinase activity (3674)

Gene Ontology definition: Catalysis of the transfer of a phosphate group, usually from ATP, to a protein substrate. [1]

Protein kinases are represented by EC class 2.A.-, EC 2.A.- to 2.7.4.- , 2.7.6.-

EC 2.7.9.-

Table A.12. Search criteria for *protein kinase activity*

True	Related functions	Vage connection
Kinases **Phosphotransferase** **Members of EC 2.A.- to 2.7.4.- , 2.7.6.-** **EC 2.7.9.-** Proteins involved in phosphorylation, ATP hydrolosis	ND	ND

Information sources

Search for missing annotations in testset performed ONLY for keywords (kinase) in protein name and by memberhip to EC class. For the search of true FP Uniprot was used in those cases where there was no EC number given.

Validation of FP with Uniprot Information.

A.2 GO terms and their search modes

Search mode a) Search by EC class

GO:0004523 Ribonuclease H activity

GO:0004672 Protein kinase activity

(GO:0016160 Amylase activity)

GO:0008235 metalloexopeptidase activity

GO:0016886 ligase activity, forming phosphoric ester bonds

GO:0004812 tRNA ligase activity

Search mode b) Search by protein name

GO:0004672 Protein kinase activity

GO:0003677 DNA binding

GO:0042626 ATPase activity, coupled to transmembrane movement of substances

GO:0003723 RNA binding

Search mode c) Search by keyword in databases

GO:0004672 Protein kinase activity (only when no EC number was given)

GO:0042626 ATPase activity, coupled to transmembrane movement of substances

GO:0030554 adenyl nucleotide binding

GO:0004812 tRNA ligase activity

GO:0003755 peptidyl-prolyl cis-trans isomerase activity

GO:0015399 primary active transporter activity

Search mode d) Search by keyword in literature

GO:0030554 adenyl nucleotide binding

and most of the other ones when more evidence was needed

Search mode e) Using existing annotations

GO:0030554 adenyl nucleotide binding

and others

Search for missing annotations

- only search mode a) and b)