GO Molecular Function Terms Are Predictive of Subcellular Localization

Z. Lu and L. Hunter

Pacific Symposium on Biocomputing 10:151-161(2005)

GO MOLECULAR FUNCTION TERMS ARE PREDICTIVE OF SUBCELLULAR LOCALIZATION

Z. LU AND L. HUNTER

Center for Computational Pharmacology
University of Colorado Health Sciences Centre
School of Medicine, Denver, CO

A protein's function is closely linked to its subcellular localization. Use of Gene Ontology (GO) molecular function terms to extend sequence-based subcellular localization prediction has been previously shown to improve predictive performance. Here, we explore directly the relationship between GO function annotations and localization information, identifying both highly predictive single terms, and terms with large information gain with respect to location. The results identify a number of predictive and informative GO terms with respect to subcellular location, particularly nucleus, extracellular space, membrane, mitochondrion, endoplasmic reticulum and Golgi. There are several clear examples illustrating why the addition of function information provides additional predictive power over sequence alone. Other interesting phenomena can also be seen in the results. Most predictive or informative terms are imperfect, and incorrect prediction may often call out significant biological phenomena. Finally, these results may be useful in the GO annotation process.

1. Introduction

High-throughput sequencing technology has heightened the need for automatic annotation of uncharacterized genes and gene products. As part of this annotation process, a number of systems have been developed that support automated prediction of subcellular localization of proteins. Most such methods are *sequence-based*, that is, they predict location based on features calculated from the amino acid sequence of a protein, such as degree of match to motifs constructed from short N-terminal signal peptides, or global amino acid composition. Recently, Chou & Cai[3] and Gardy et al.[7] have demonstrated that the addition of protein function information improves the performance of pure sequence-based predictions of protein subcellular localization. Chou & Cai reported the best performing predictive system, as measured on a carefully constructed gold standard, used a hybrid input containing both sequence patterns and protein annotations

based on the molecular function terms from the Gene Ontology (GO)[1].

Computational methods to predict subcellular localization are a crucial bioinformatic task[5]. Several broad classes of predictive information have been brought to bear on this task. One class of informative information is the global amino acid composition of the protein. For example, NNPSL[13] used neural networks with amino acid composition inputs, and SubLoc[8] took a similar approach using support vector machines as the induction method. A second class of predictive information is the presence of signal peptides, which are short sub-sequences of approximately 15 to 60 amino acids shown to play functional role in protein transport. Various computational methods have been used to characterize and recognize instances of these signals, and then use them to predict specific cell locations. An example of this approach is TargetP[6]. A third class of predictive information arises from protein homology, which as the breadth of annotation grows, becomes increasingly useful. The recent LOCkey[11] system offered an unusual twist on homology-based approaches, identifying putatively homologous proteins by sequence similarity search, using natural language processing technology to extract textual features from the annotations of the homologs, and using those features as inputs to an inductive classifier for prediction of location. Approaches that combine multiple classes of predictive information, such as Grady et al.'s and Chou & Cai's actually have a long history in this area, going back more than a decade to Nakai & Kanehisa's classic PSORT system[12].

Although the Chou & Cai results demonstrate the utility of inclusion of GO annotations in the set of information used for location prediction, the direct relationship between macromolecular function terms in the GO and subcellular location has, to our knowledge, not been previously explored. The main motivation for exploring this relationship is to understand the nature of the contribution that molecular function (and its existing annotation) makes to the prediction of subcellular localization; however, there are other aspects of this work that are also significant. As described below, there is a linkage in the GO annotation process between molecular function annotation and subcellular localization annotation; the assessment made in this study may be useful in improving that process. Also, most predictive or informative function terms are imperfectly discriminating; the minority (or mispredicted) localizations often call out significant biological phenomena.

2. Methods

2.1. *Source of annotations*

The GO provides controlled vocabularies in three broad categories: molecular function, biological process, and subcellular localization. Only protein annotations from the GO molecular function category are included in this analysis, and they are compared with two distinct subcellular localization gold standards. One localization dataset is derived from annotations from the SwissProt database[2], the other from a recent hand-curated set described in Nair & Rost[11].

We did *not* use the GO subcellular localization annotations for this work. The reason is that the GO location and function are annotated in an intentionally dependent manner. GOA curation guidelines specify that subcellular location annotations can be inferred by the curator directly from molecular function information when the curator cannot find any relevant evidence in the biological literature. The assessment of the true degree of correlation between molecular function and subcellular localization would be confounded by this practice if we used GOA's localization data. The use of the independently curated annotations for localization disentangles the annotation process from any actual biological relationship.

The GO biological process category was not assessed for relationship to location because it is intentionally designed to integrate sets of molecular functions. For example, the process of signal transduction inherently involves signals from the extracellular space interacting with receptors at the membrane, which in turn modify secondary messengers in the cytoplasm, which often end up causing changes in transcription occurring in the nucleus. While each of these activities is carried out by molecules which can reasonably be expected to have their own function annotation and location, the process itself cannot be said to have a subcellular location.

2.2. *Localization-specific GO Terms*

The GO molecular function terms and relationships used in this analysis were taken from the molecular function ontology flat file on the GO website[a] from May 05 2004, version 1.28. Protein annotations for both SwissProt generated localization (the SUBCELLULAR LOCALIZATION field in CC lines) and GO terms (GO field in DR lines) were taken from SwissProt

[a] `http://www.geneontology.org/`

release 42. Only GO terms that appeared in the molecular function flat file were used in the analysis.

Note that both of the above-mentioned fields are optional, and many proteins are missing one or the other field, or both. SwissProt release 42 includes 135,850 proteins; only 6,686 ($\sim$ 5%) have relevant annotations for both location and molecular function. Generally, the issue is lack of coverage of GO annotations; for example, 3,655 proteins are annotated as localized in the nucleus, but only 937 ($\sim$ 25%) of those have GO molecular function terms associated with them.

Certain subcellular localization is not well represented in SwissProt. For example, almost no GO molecular function terms are associated with proteins in chloroplast, which is an organelle typically found in green plants. Although GO annotation for plants is underway by several model organism database groups, particularly for *Arabidopsis thaliana* and the agriculturally important cereal grasses, these annotations are not yet very prevalent in SwissProt, where the priority for the ongoing GO annotation effort is on human proteins.

We took two different approaches to identifying molecular function terms that are associated with localization. The first approach calculated the proportion of occurrences of each molecular function term to the most frequent location for that term and, separately, for the term and all of its subtypes ("is-a" children in the GO DAG). Any term (or term and its subtypes) that is associated with one location over a threshold proportion (set here at 70%) is reported in Table 1.

In order to make this calculation efficient, this specific approach was taken. Define the set of locations of interest to be L $\equiv$ {nucleus, extracellular space, membrane, mitochondrion, endoplasmic reticulum and Golgi} and the set F to be all of the GO terms that are children of Molecular Function.

- Initialize four accumulators for each combination of location ($l \in L$) and function ($f \in F$): $acc_{f,l}, acc_{f,\sim l}, acc^{incl}_{f,l}, acc^{incl}_{f,\sim l}$ to zero.
- For each location $l \in L$, do
  - Let P_l be the set of proteins annotated to have location l. For each $p_l \in P_l$, do
    - Let F_{p_l} be the set of molecular function annotations associated with the protein p_l. For each $f_{p_l} \in F_{p_l}$, increment $acc_{l,f}$.
  - For each l and f such that $acc_{l,f} > 0$, let $P_{f,\sim l}$ be the set of

proteins annotated with function f but with a location other than l. For each $p_{f,\sim l} \in P_{f,\sim l}$, increment $acc_{f,\sim l}$

- Beginning from the leaves of the GO DAG and working up is-a links to the root node for the molecular function hierarchy, traverse each $f \in F$ and set $acc^{incl}_{f,l}$ to be the sum of $acc_{f,l}$ and the $acc_{f_{child},l}$ for each child f_{child} of function node f. Similarly for $acc^{incl}_{f,\sim l}$
- Use depth-first search starting from the root of the molecular function DAG to identify the highest-level terms whose ratios $acc_{f,l}/(acc_{f,l} + acc_{f,\sim l})$ or $acc^{incl}_{f,l}/(acc^{incl}_{f,l} + acc^{incl}_{f,\sim l})$ are greater than threshold

Caching of partial results is used to avoid redundant calculations. The approach avoids doing any work on the large portion of the proteins that are not annotated in a way that is relevant to the particular calculation.

2.3. *Identification of discriminative terms by information gain*

The informativeness of GO molecular function terms with respect to location can also be quantified by *information gain*, a measure of the amount of information (in bits) of that the knowledge of a feature (here, molecular function) contributes to knowledge of the class of the entity (here, location)[10]. An information gain of 1 means that knowledge of the molecular function is a perfect predictor of location, and an information gain of 0 means that knowledge of the molecular function provides no information regarding location.

In order to address concerns about the representativeness of the SwissProt dataset, we did this test on a completely independent set of annotations, that used by Nair & Rost[11]. This dataset classifies 1161 proteins into 10 different locations. These proteins are annotated with 207 different GO molecular functions. The information gain of the presence or absence of each molecular function term with respect to the distribution of locations was calculated. Information gain calculated in this way does not indicate *which* location is associated with a particular molecular function, or even that a function is associated with a single location. The implication of positive information gain is that "purity" of the locations associated with a set of proteins separated by the presence or absence of a particular function is higher than would be expected from a random division into the same size sets. This entropic measure has cleaner formal qualities, but is less directly useful, than the more ad hoc first method.

3. Results

The results of the first method, the ad hoc generation of localization-specific molecular function terms, found terms that either on their own or including annotations from their descendent terms provided $\geq$ 70% specificity for at least one of the 6 locations. Nineteen terms were over threshold for nuclear localization: 52 for membrane, 16 for extracellular, 2 for endoplasmic reticulum, 4 for mitochondrion, and 4 for Golgi. A sample of the highest-scoring terms is shown in Table 1. The complete set of results is available on a supplementary web site `http://compbio.uchsc.edu/Hunter_lab/Zhiyong/psb2005`.

Table 1. Selected highly discriminating terms (including children) for the six subcellular localization derived from the search for location-specific terms in SwissProt.

Location	Predictive GO Molecular Function terms
nucleus	GO:0003676 Nucleic acid binding
	GO:0008134 Transcription factor binding
	GO:0030528 Transcription regulator activity
membrane	GO:0004872 Receptor activity
	GO:0015267 Channel/pore class transporter activity
	GO:0008528 Peptide receptor activity, G-protein coupled
extracellular	GO:0005125 Cytokine activity
	GO:0030414 Protease inhibitor activity
	GO:0005201 Extracellular matrix structural constituent
mitochondria	GO:0015078 Hydrogen ion transporter activity
	GO:0004738 Pyruvate dehydrogenase activity
	GO:0003995 Acyl-CoA dehydrogenase activity
	GO:0015290 Electrochemical potential-driven transporter activity
E.R.	GO:0004497 Monooxygenase activity
	GO:0016747 Transferase activity, transferring groups other than amino-acyl groups
Golgi	GO:0016757 Transferase activity, transferring glycosyl groups
	GO:0015923 Mannosidase activity
	GO:0005384 Manganese ion transporter activity

The results from the information gain method were largely comparable with the location-specificity measure, even though the datasets were completely independent and the methods are quite different. 194 of the 207 terms had positive information gain, although only ten had information gains of .01 or greater, with the greatest gain at 0.047. These numbers are relatively low because while the presence of a function (such as DNA binding) may be quite specifically associated with a particular location, many other proteins without that function are likely to be in the same location,

driving down the information gain substantially. The top ten results are shown in Table 2, and the complete results are available from the supplementary web site (URL above).

Table 2. The ten highest location information gain GO molecular function terms.

Information Gain	GO Molecular Function terms
0.047	GO:0003677 DNA binding
0.024	GO:0005179 hormone activity
0.024	GO:0003676 nucleic acid binding
0.022	GO:0003700 transcription factor activity
0.016	GO:0008270 zinc ion binding
0.015	GO:0004129 cytochrome-c oxidase activity
0.015	GO:0003735 structural constituent of ribosome
0.013	GO:0008009 chemokine activity
0.011	GO:0008083 growth factor activity
0.010	GO:0016491 oxidoreductase activity

4. Discussion

4.1. *Comparability of the methods*

Most of the high information gain terms appear in the list of the location-specific terms, and some of the differences are methodological. For example, the high information gain set includes two terms which have a parent/child relationship ("GO:0003677 DNA binding" and "GO:0003676 Nucleic acid binding") while only the parent term would make it into the location-specific list generated by the ad hoc method. Other differences arise from use of different datasets: e.g., "GO:0008528 Peptide receptor activity" is a highly predictive function in the ad hoc method, but is not associated with any of the 1161 proteins in the Nair & Rost dataset used to calculate information gain.

However, some of the differences are more interesting. For example, two of the top information gain terms, "GO:0008200 Zinc ion binding" and "GO:0005179 Hormone activity" have relatively low location specificity scores. These terms strongly favor a biased subgroup of locations, rather than a single location (e.g. the molecular function "Hormone activity" is associated with proteins annotated to both extracellular and membrane locations, not exclusively one or the other).

4.2. *Why function is complementary to sequence in location prediction*

As described above, all published location prediction methods are at least in part sequence-based, but two recent methods[3],[7] that use functional information in addition to sequence outperform other methods. Using the set of proteins from our experiments that were labelled with highly predictive functions, we found several examples where sequence-only methods made inaccurate predictions. For example, MBL_DROME in Swiss-Prot is a protein associated with terminal differentiation of photoreceptor cells in Drosophila. It is annotated as a nuclear protein with GO function "GO:0003676 nucleic acid binding". However, it was not predicted as a nuclear sequence by predictNLS[4], a widely used server for identifying nuclear localization signals in sequence. Furthermore, TMHMM[9], a popular method for recognizing transmembrane helices in proteins, incorrectly predicted it as a membrane protein, since a putative transmembrane helix region was identified in the sequence. However, "nucleic acid binding" is highly predictive of nuclear localization, and was never observed (in our data) as a function of membrane proteins.

4.3. *Biological interpretations*

Several of the predictions (and prediction failures) have interesting biological interpretations. Focusing on the molecular function terms that are both predictive and occur in many annotations identifies useful biological knowledge implicit in the annotations. It may be possible to exploit these regularities in other applications, e.g. automated methods for constructing knowledge-bases. Here are several illustrative examples:

- Nucleus: The molecular functions that are predictive of nuclear localization are mostly related to nucleic acids. The most predictive term is "GO:0003676 Nucleic acid binding", which is also among the high information gain function terms. Other predictive terms include "GO:0030528 Transcription regulator activity", "GO:0008134 Transcription factor binding", and "GO:0004386 Helicase activity". However, none of these terms are associated solely with the nucleus; each annotates proteins that are localized elsewhere as well. For example, aside from the nucleus, "GO:0003700 Transcription factor activity", a child term of "GO:0030528 Transcription regulator activity" is commonly associated with proteins annotated as cytoplasmic. This duality reflects the fact that tran-

scription factors are often found in inactive form in the cytoplasm, and transported to the nucleus when activated. For example, consider the major signal transduction family Rel/NF-kappaB (NF-kB, KBF1_HUMAN in Swiss-Prot), which is involved in the control of a large number of normal cellular and organismal processes, such as apoptosis and inflammation. The interaction of NF-kB with (among other proteins) IkBa both inhibits its ability to bind to DNA and plays a role in maintaining its cytoplasmic localization. When a cell receives any of a multitude of extracellular signals, NF-kB dissociates from IkBa, is activated functionally and is transported to the nucleus. Many other proteins annotated with transcription factor activity are processed similarly, leading to the observed "imperfect predictiveness." However, an alternative view is that the annotations fail to capture the dynamism in the localization of these proteins.

- Membrane: More function terms are associated with membrane localization that with any other location. This is in part because, unlike in the case of the nucleus, the most abstract functions of membrane proteins (e.g. "GO:0005215 Transporter activity") are not specific to membrane, and more specific children of these terms must be used to achieve over-threshold predictiveness. For example, note that "GO:0005489 Electron transporter activity" is associated mainly with cytoplasmic proteins (although not enough so to be predictive at the 70% level), while "GO:0015267 Channel/pore class transporter activity" is predictive of membrane proteins.
- Mitochondrion: Mitochondrion carries out oxidative phosphorylation and produces most of the ATP in eukaryotic cells. We found 4 highly predictive terms for localization to this organelle, including two that are commonly characterized as specific to energy metabolism ("GO:0004738 Pyruvate dehydrogenase activity" and "GO:0003995 Acyl-CoA dehydrogenase activity"). The other two terms ("GO:0015078 Hydrogen ion transporter activity" and "GO:0015290 Electrochemical potential-driven transporter activity") describe functions not generally characterized in textbooks as specific to mitochondria. However, these are among the strongest associations found in this study ("GO:0015078 Hydrogen ion transporter activity" is associated with 81 mitochondrial proteins and no non-mitochondrial proteins).
- Endoplasmic Reticulum (ER): The difference between the biolog-

ical process carried out in a particular location and specificity of the functions involved in that process is illustrated in the results for the ER. The processes associated with the ER are primarily the synthesis of lipids and secretory proteins. However, the two molecular function terms that are predictive of ER localization, "GO:0004497 Monooxygenase activity" and "GO:0016747 Transferase activity, transferring groups other than amino-acyl groups", are not obviously representative of these processes.

4.4. *Implications for GO annotators*

As described in the introduction, GO provides an ontology of subcellular localization terms itself, and GO annotators sometimes infer such localizations on the basis of molecular function. Such inferences are assigned the evidence code IC (Inferred by curator), although it is not clear if all the localization annotations with IC evidence codes are done on the basis of molecular function alone.

We believe it might be possible to extend this kind of annotation on the basis of these results. For example, we did not find any localization annotations for the mitochondria with IC evidence codes. However, as noted above, the molecular function "GO:0015078 Hydrogen ion transporter activity" (among others) seems a very strong predictor for that localization. Several other instances also seem to provide evidence that could be used by curators to infer localizations.

4.5. *Conclusion*

The biological relationships among molecular function and subcellular localization are at least partially reflected in protein annotations. These parallel relationships can be demonstrated both in measures of information gain and in the development of effective ad hoc predictors of location. These results provide an explanation of why hybrid prediction methods perform better than sequence-based methods alone, but also suggest potential improvements that might be made in the annotation process, and illustrate important biological phenomena.

Acknowledgments

This work was supported by NIAAA grant 5U01 AA13524-03 (LH) and NIGMS graduate training grant T32-GM07635-25 (ZL).

References

1. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genet.*, 25(1):25–29, 2000.
2. B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.
3. K. Chou and Y. Cai. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem Biophys Res Commun.*, 311(3):743–747, 2003.
4. M. Cokol, R. Nair, and B. Rost. Finding nuclear localization signals. *EMBO Rep*, 1:411–415, 2000.
5. F. Eisenhaber and P. Bork. Wanted: subcellular localization of proteins based on sequence. *Trends in Cell Biology*, 8(4):169–170, 1998.
6. O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of Molecular Biology*, 300(5):1005–1016, 2000.
7. J. L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnády, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F. S. Brinkman. PSORT-B: Improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Research*, 31(13):3613–3617, 2003.
8. S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization. *Bioinformatics*, 17(9):721–728, 2001.
9. A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology*, 305:567–580, 2001.
10. T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
11. R. Nair and B. Rost. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, 18:S78–S86, 2002.
12. K. Nakai and M. Kanehisa. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14(4):897–911, 1992.
13. A. Reinhardt and T. Hubbard. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research*, 26(9):2230–2236, 1998.