

Sequence analysis

Improved enzyme annotation with EC-specific cutoffs using DETECT v2

Nirvana Nursimulu^{1,2,†}, Leon L. Xu^{1,†}, James D. Wasmuth³, Ivan Krukov³
and John Parkinson^{1,4,5,*}

¹Program in Molecular Medicine, The Hospital for Sick Children, 21–9709 PGCRL, 686 Bay Street, Toronto, ON M5G 0A4, Canada, ²Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada, ³Department of Ecosystem and Public Health, Faculty of Veterinary Medicine, University of Calgary, Calgary, AB T2N 4Z6, Canada, ⁴Department of Molecular Genetics and ⁵Department of Biochemistry, University of Toronto, Toronto, ON M5S 1A8, Canada

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on November 27, 2017; revised on April 13, 2018; editorial decision on April 27, 2018; accepted on May 1, 2018

Abstract

Summary: We present DETECT v2—an enzyme annotation tool which considers the effect of sequence diversity when assigning enzymatic function [as an Enzyme Commission (EC) number] to a protein sequence. In addition to capturing more enzyme classes than the previous version, we now provide EC-specific cutoffs that greatly increase precision and recall of assignments and show its performance in the context of pathways.

Availability and implementation: <https://github.com/ParkinsonLab/DETECT-v2>

Contact: john.parkinson@utoronto.ca

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Metabolic reconstruction and modeling provides a powerful approach to study metabolism. In addition to understanding relationships between metabolite availability and growth, modeling has proven effective at metabolic engineering and drug design (Bordbar *et al.*, 2014). Key to these studies is the availability of high quality enzyme annotations, typically inferred through sequence similarity searches (Altschul *et al.*, 1990; Hung and Parkinson, 2011) which yield high true positive rates (recall) at the expense of specificity; a consequence of not considering sequence diversity within and between enzyme classes (Hung *et al.*, 2010). To account for the impact of class-specific sequence diversity, we developed an original enzyme annotation tool, DETECT, hereafter referred to as DETECT v1 (Hung *et al.*, 2010). DETECT constructs positive and negative density profiles for each enzyme commission (EC) number (Bairoch, 2000), representing how sequences within a class align to one another and to other classes; an integrated likelihood score (ILS) is then attached to each EC assignment. Compared to BLAST,

DETECT provides substantially higher specificity predictions with higher true positive rate; ROC analyses yielded an empirical ILS cutoff of 0.2. Here, we present DETECT v2. In addition to covering more ECs (786 compared to 582 in v1) and protein sequences, we now provide EC-dependent cutoffs. We demonstrate that these improvements lead to higher precision and recall over existing tools and place our findings in the context of metabolic pathways, highlighting the ability of DETECT v2 to accurately reconstruct individual pathways.

2 Results and discussion

Following 5-fold cross-validation, we analyzed the precision-recall curve of each EC to characterize DETECT's performance. Encouragingly, 354 ECs can be predicted with 100% precision and recall, either irrespective of the cutoff [195 ECs; e.g.: 5.1.1.7 (diaminopimelate epimerase); Fig. 1A] or under the right cutoff [e.g.: 4.3.3.7 (4-hydroxy-tetrahydronicotinate synthase); Fig. 1A].

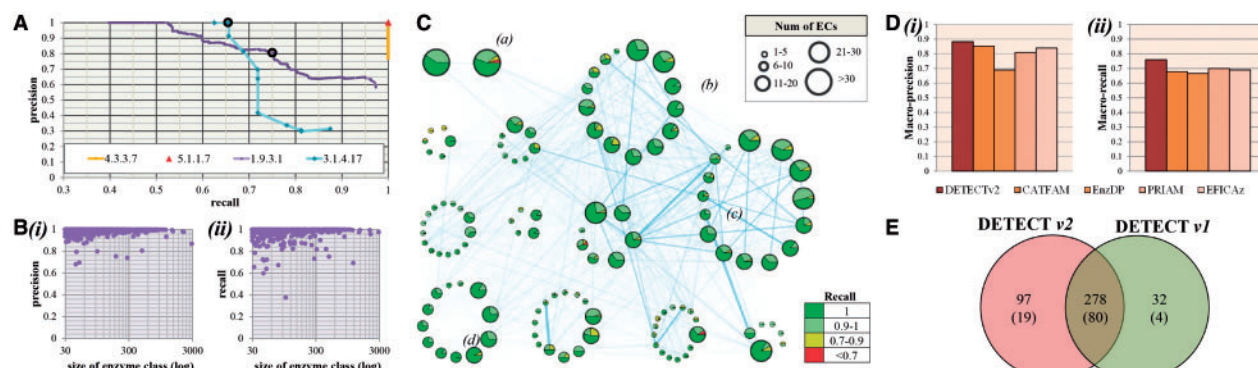


Fig. 1. (A) Example precision-recall curves. (B) Scatter plot showing F1-scheme's (i) precision and (ii) recall versus class size (log-scale). (C) Recall of ECs of various pathways, shown as nodes. Nodes, grouped by superpathway, are linked with heavier edges when sharing more compounds (as per KEGG). Node size captures the number of ECs predictable by DETECT in that pathway. Each node's pie chart shows the proportion of ECs predictable at different recalls under the F1-scheme. Annotated pathways are (a) purine metabolism, (b) lysine biosynthesis, (c) inositol phosphate metabolism and (d) 1-carbon pool by folate. (Details in Supplementary Material A.) (D) DETECT's performance compared with other tools in terms of (i) macro-precision and (ii) macro-recall. (E) Venn diagram showing number of ECs predicted in *C. elegans* by DETECT v2 and v1 and overlap with BRENDA indicated in brackets

However, at the cutoff of 0.2, of the 780 ECs with >80% precision, 709 have >80% recall (with 30 having <50% recall). At the lower cutoff of 0.01, of the 749 ECs with >80% recall, 700 attain >80% precision (with 22 having <50% precision). While this lower cutoff gives higher recall at the cost of precision, it is clear that annotation accuracy for individual EC classes would benefit from class-specific cutoffs. For example, for 1.9.3.1 (cytochrome-c oxidase) or 3.1.4.17 (3', 5'-cyclic-nucleotide phosphodiesterase), there is a compromise between precision and recall (Fig. 1A). To address this, for each EC, we define a cutoff which optimizes the F1-measure (Supplementary Material and Supplementary Table S4), yielding 772 ECs where >80% precision can be achieved at >80% recall.

Next, we investigated whether the number of protein sequences associated with an EC class influences precision or recall at the maximum F1-measure (Fig. 1B). 393 (50%) ECs predictable by DETECT have at most 115 protein sequences associated with them. However, for 321 of those, both precision and recall are >95%. More broadly, 737 ECs with both precision and recall >90% represent classes with 30 (minimum required by DETECT) to 2180 protein sequences, showing that the number of protein sequences in an EC class has little consequence on precision or recall at the reported cutoff. For exceptions such as 3.1.4.17, the lower reported recall is likely a consequence of high sequence diversity (Manganiello et al., 1995; Supplementary Material B); for 1.9.3.1, the lower precision reported is a direct result of compromising for a higher recall.

Moreover, we asked whether there are particular metabolic pathways [as defined by the KEGG database (Kanehisa et al., 2016)] that can be predicted with higher recall (Fig. 1C); we focus on recall due to DETECT's generally high precision [Fig. 1B(i)]. 105 of the 119 pathways covered by DETECT have $\geq 75\%$ of their enzymes predictable with $\geq 90\%$ recall under the F1-scheme. Remarkably, 71 pathways are completely predictable with $\geq 90\%$ recall under the F1-scheme; for example, the ECs in lysine biosynthesis, inositol phosphate metabolism and one carbon pool by folate (covering 12, 8 and 7 ECs, respectively) have >90% recall with >90% precision. 11 of these ECs (one being 5.1.1.7) are predicted with 100% precision and 100% recall. At the other end of the spectrum, 2 of the 47 ECs predictable in purine metabolism by DETECT have <70% recall (one being 3.1.4.17). However, the latter ECs represent exceptions rather than the norm given the range of recalls presented in Figure 1B(ii).

Following cross-validation, we compared the performance of DETECT v2 with that of other enzyme annotation tools (Supplementary Material A) on 2093 sequences annotated with (359) ECs in metabolic networks [from the BiGG database (King et al., 2016)]. While DETECT's macro-precision is comparable to the top-performers, its macro-recall is higher (Fig. 1D; Supplementary Material). Please refer to Supplementary Material for further comparisons, including against DETECT v1.

Last, we applied DETECT v2 and v1 to the proteome of *Caenorhabditis elegans* (WS235; Lee et al., 2017) and compared their EC predictions to annotations in the BRENDA database (Schomburg et al., 2017; Fig. 1E). We point out that the latter only provides high-confidence annotations reported in previous studies; EC predictions thus represent hypothetical functions to be validated biochemically. DETECT v2 makes 97 additional EC predictions (with 19 supported by BRENDA). We note that while 32 ECs were predicted by v1 alone, only eight were predictable by v2 (six of which would have been predicted at lower cutoffs). More importantly, the increased coverage of v2, together with the F1-scheme, give substantially increased performance. For example, DETECT v2 correctly identifies all three (as opposed to only 2) functions of gene D2085.1, essential for *de novo* pyrimidine synthesis (Franks et al., 2006), a direct consequence of the use of EC-specific cutoffs in v2. Further details of the performance enhancements of DETECT v2 are provided in Supplementary Material.

3 Conclusion

We present DETECT v2, an enzyme annotation tool tuned for high precision and high recall, that provides a powerful approach for the reconstruction of high confidence metabolic models.

Acknowledgement

The authors thank Drs Swapna Seshadri and Xuejian Xiong for helpful discussion.

Funding

This work was supported by grants from the Natural Sciences and Engineering Research Council (RGPIN-2014-06664) and the National Institutes of Health (R21AI126466). N.N. was supported by a SickKids

RestraComp scholarship. Computing resources were provided by the SciNet HPC Consortium; SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund—Research Excellence and the University of Toronto.

Conflict of Interest: none declared.

References

- Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
- Bordbar,A. *et al.* (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.*, **15**, 107–120.
- Franks,D.M. *et al.* (2006) *C. elegans* pharyngeal morphogenesis requires both de novo synthesis of pyrimidines and synthesis of heparan sulfate proteoglycans. *Dev. Biol.*, **296**, 409–420.
- Hung,S.S. and Parkinson,J. (2011) Post-genomics resources and tools for studying apicomplexan metabolism. *Trends Parasitol.*, **27**, 131–140.
- Hung,S.S. *et al.* (2010) DETECT—a density estimation tool for enzyme classification and its application to *Plasmodium falciparum*. *Bioinformatics*, **26**, 1690–1698.
- Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
- King,Z.A. *et al.* (2016) BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.*, **44**, D515–D522.
- Lee,R.Y.N. *et al.* (2017) WormBase 2017: molting into a new stage. *Nucleic Acids Res.*, **46**, D869–D874.
- Manganiello,V.C. *et al.* (1995) Diversity in cyclic nucleotide phosphodiesterase isoenzyme families. *Arch. Biochem. Biophys.*, **322**, 1–13.
- Schomburg,I. *et al.* (2017) The BRENDA enzyme information system—From a database to an expert system. *J. Biotechnol.*, **261**, 194–206.