Efficient use of accessibility in microRNA target prediction

Ray Marín and Jiří Vaníček*

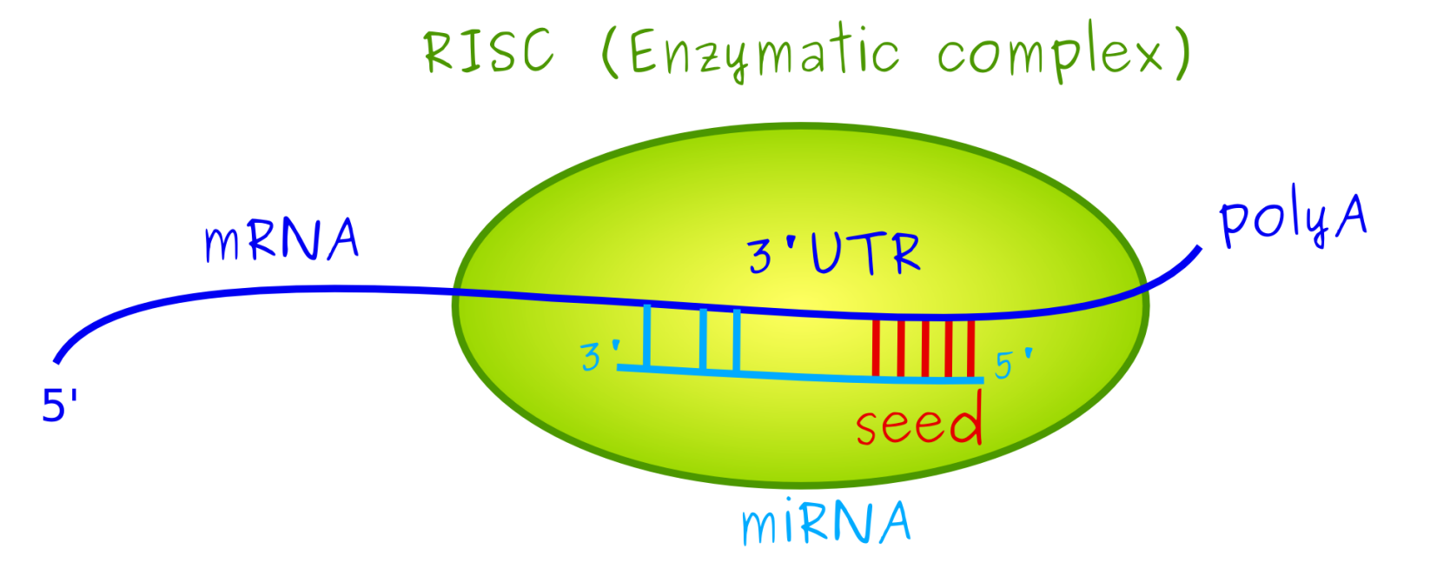
Laboratory of Theoretical Physical Chemistry, Institut des Sciences et Ingénierie Chimiques,
École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

*jiri.vanicek@epfl.ch

Abstract

Considering accessibility of the 3'UTR is believed to increase the precision of microRNA target predictions. We show that, contrary to common belief, ranking by the hybridization energy or by the sum of the opening and hybridization energies, used in currently available algorithms, is not the most efficient way to rank predictions. Instead, we describe an algorithm which also considers only the accessible binding sites but which ranks predictions according to over-representation of the accessible seed matches [1]. When compared with experimentally validated and refuted targets in the fruit fly and human, our algorithm shows a remarkable improvement in precision while significantly reducing the computational cost in comparison with other free energy based methods. In the human genome, our algorithm has at least twice higher precision than other methods with their default parameters. In the fruit fly, we find five times more validated targets among the top five hundred predictions than other methods with their default parameters. Furthermore, using a common statistical framework we demonstrate explicitly the advantages of using the canonical ensemble instead of using the minimum free energy structure alone. We also find that "naïve" global folding sometimes outperforms the local folding approach. The proposed method also allows combining the accessibility filter with a conservation filter using multiple sequence alignments [2]. Predictions in the human genome show that the combined filter increases precision more than either filter alone. It is shown that some conserved but non-functional sites can only be rejected by means of an accessibility cutoff.

PACMIT Method



- Complementarity to seed (e.g. positions 2-8)
- Filter by accessibility (and/or conservation)
- The model assumes that functional miRNA-target pairs arose by coevolution, therefore, complementary sites in real targets should correspond to overrepresented 7-mers.

Single hypothesis P value (P_{SH})

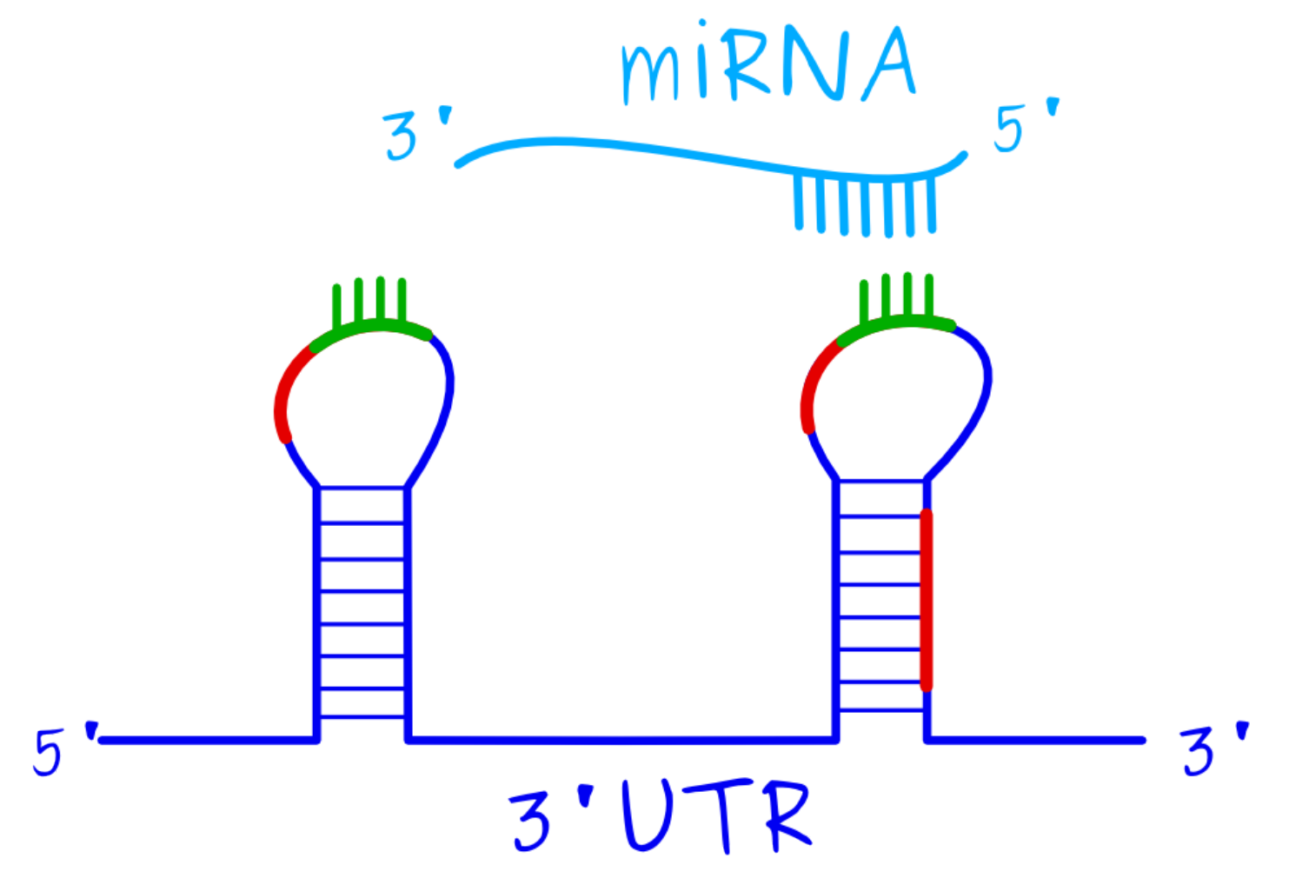
gives an approximate probability that a given oligomer (or n-mer), complementary to the miRNA seed, is found by chance at least c times in the corresponding 3'UTR.

$$P_{SH} = \sum_{i=c}^{l-n+1} \binom{l-n+1}{i} p^i (1-p)^{l-n+1-i},$$

- l = length of the 3'UTR
- n = number of nucleotides in the seed (e.g. 7)
- p = probability to find the given n-mer by chance.

Partial accessibility

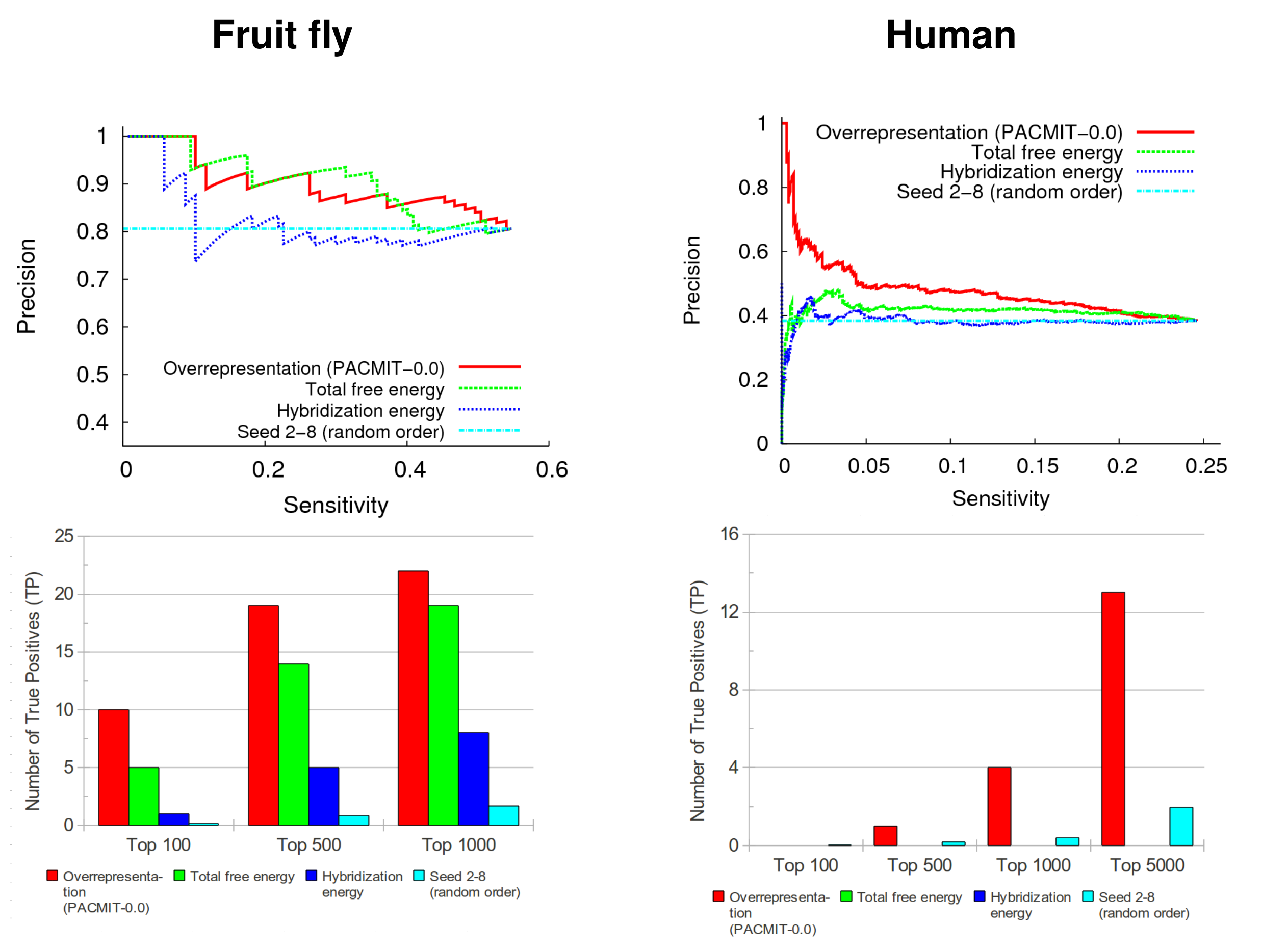
GGGGCACGGGGGGGGAAGAAGGCCAAAACGTGCCCC
GGGGCACGGGGGGGGAAGAAGGCCAAAACGTGCCCC
GGGGCACGGGGGGGGAAGAAGGCCAAAACGTGCCCC
GGGGCACGGGGGGGGAAGAAGGCCAAAACGTGCCCC



- Minimum Free Energy Structure (MFE) using RNAfold [3].
- Canonical Ensemble of Secondary Structures using RNAplfold [4].
  - **Local folding:** fold all subsequences of 80 nt. Restrain distance between paired bases to 40 nt.
  - **Global folding:** the full sequence is folded.
  - In the two cases the probability that 4 consecutive nucleotides in the seed region are unpaired is used to select the partially accessible sites complementary to the seed (c_{access}) among all the partially accessible sites in the 3'UTR (t_{access}).

$$P_{SH} = \sum_{i=c_{\text{access}}}^{t_{\text{access}}} \binom{t_{\text{access}}}{i} p^i (1-p)^{t_{\text{access}}}$$

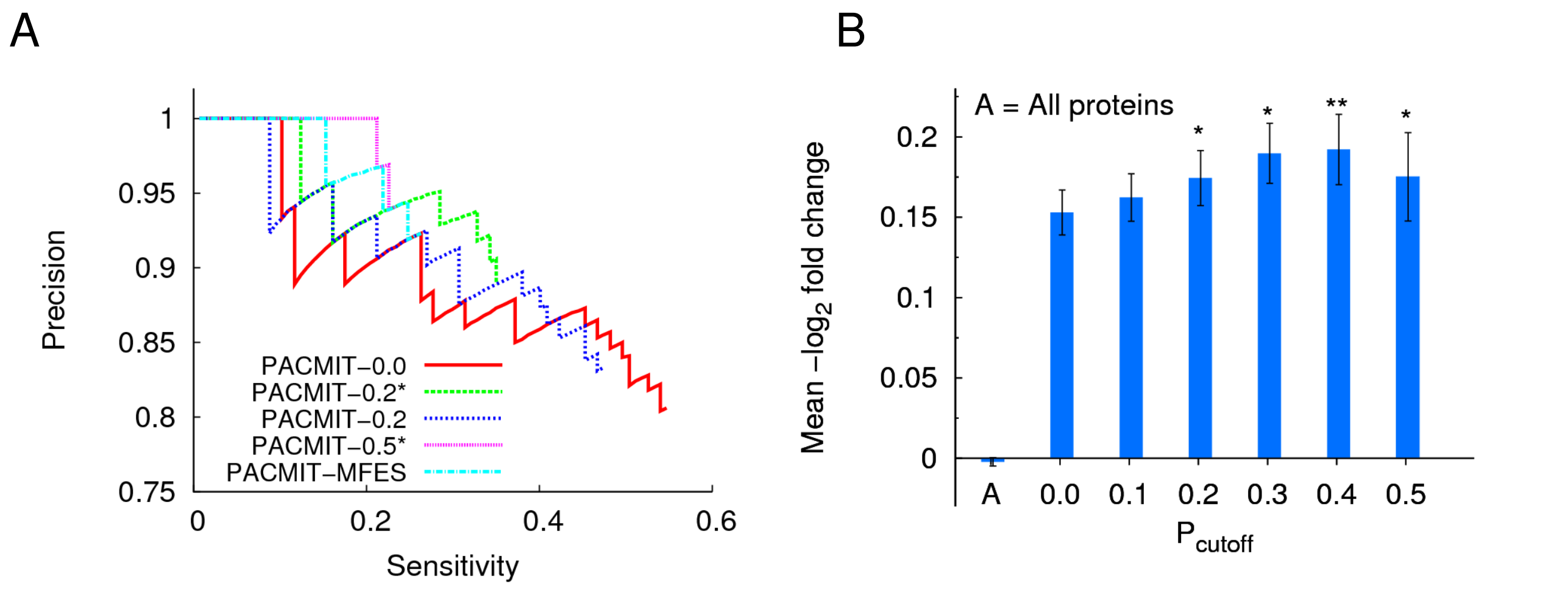
Overrepresentation better ranking criterion than free energies


Fruit fly

Human

- **Fruit fly:** 137 functional, and 83 non-functional miRNA-3'UTRs pairs [5,6].
- **Human:** 2,406 functional, and 13,400 non-functional miRNA-3'UTRs pairs [7,8].
- **Hybridization energy**: assumes that the binding site is completely free to interact with the miRNA.
- **Total free energy**: Takes into account both the hybridization energy and the energy required to make the binding site accessible.

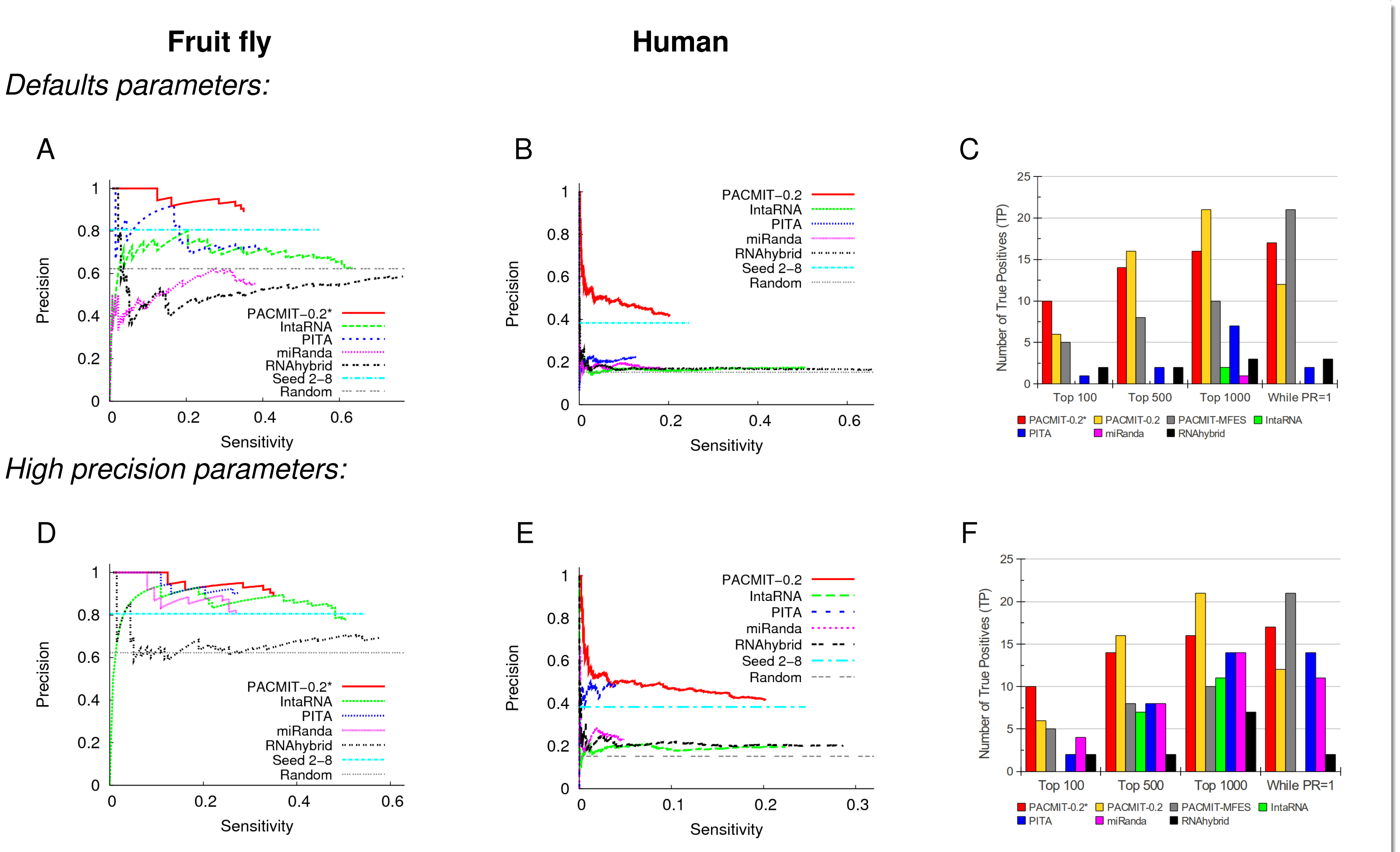
The accessibility filter improves the performance of the method



(A) The accessibility filter helps to improve the precision.

(B) Targets predicted using the accessibility filter in Human are in average more repressed than those predicted without considering this feature.

PACMIT better than free energy based methods


Fruit fly

Human

Defaults parameters:

High precision parameters:

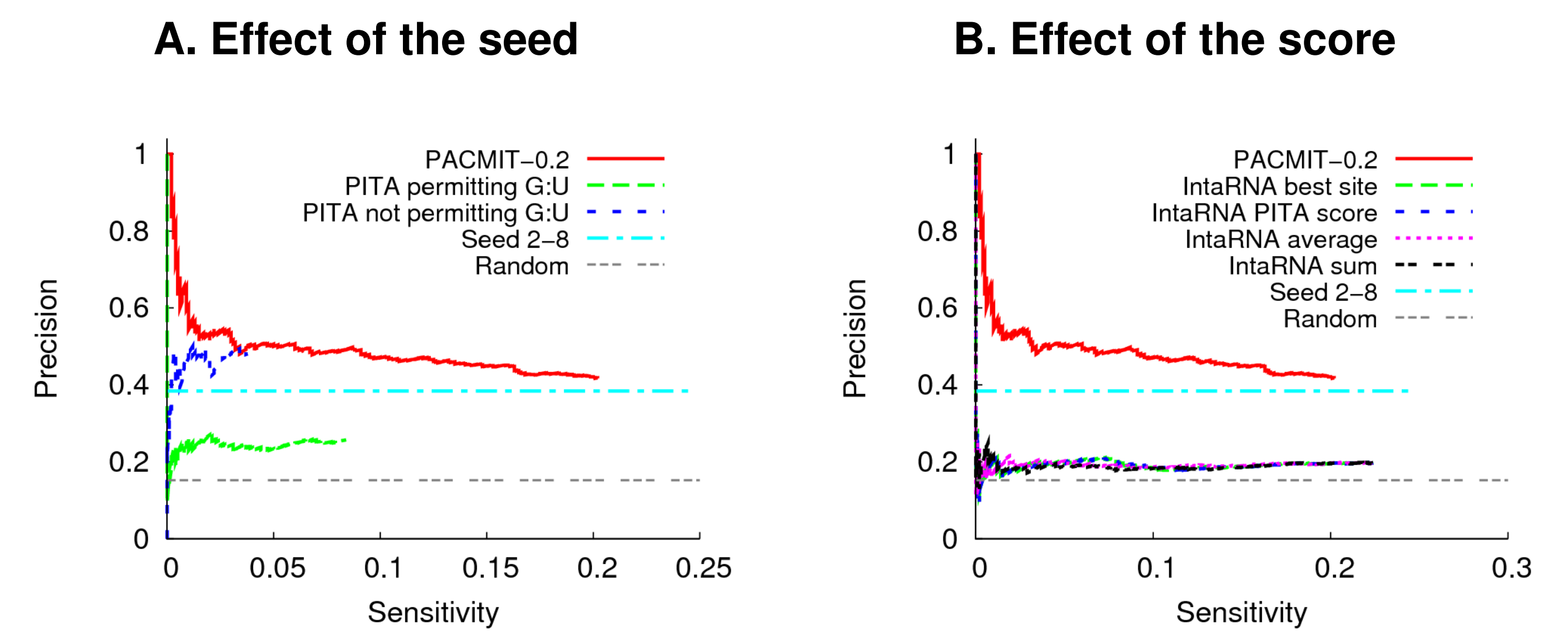
- With the default parameters most free energy based methods perform worse than the simple seed 2-8 method (panels A and B).
- Only PITA with strict parameters can reach a precision similar to that of PACMIT in the fruit fly and in the human (panels D and E).
- The top predictions of PACMIT contain at least twice more true positives than the predictions of the other methods (panels C and F).
- PACMIT is faster than the other methods (table).

CPU time consumption (fruit fly)

Method	time (hours)	Accessibility
IntaRNA	50	yes
PITA	15	yes
RNAhybrid	3	no
miRanda	0.33	no
PACMIT	0.17 (0.66[a])	yes

[a] Only the first calculation for the same genome

Energy based scores are not enough. The seed restriction helps


A. Effect of the seed

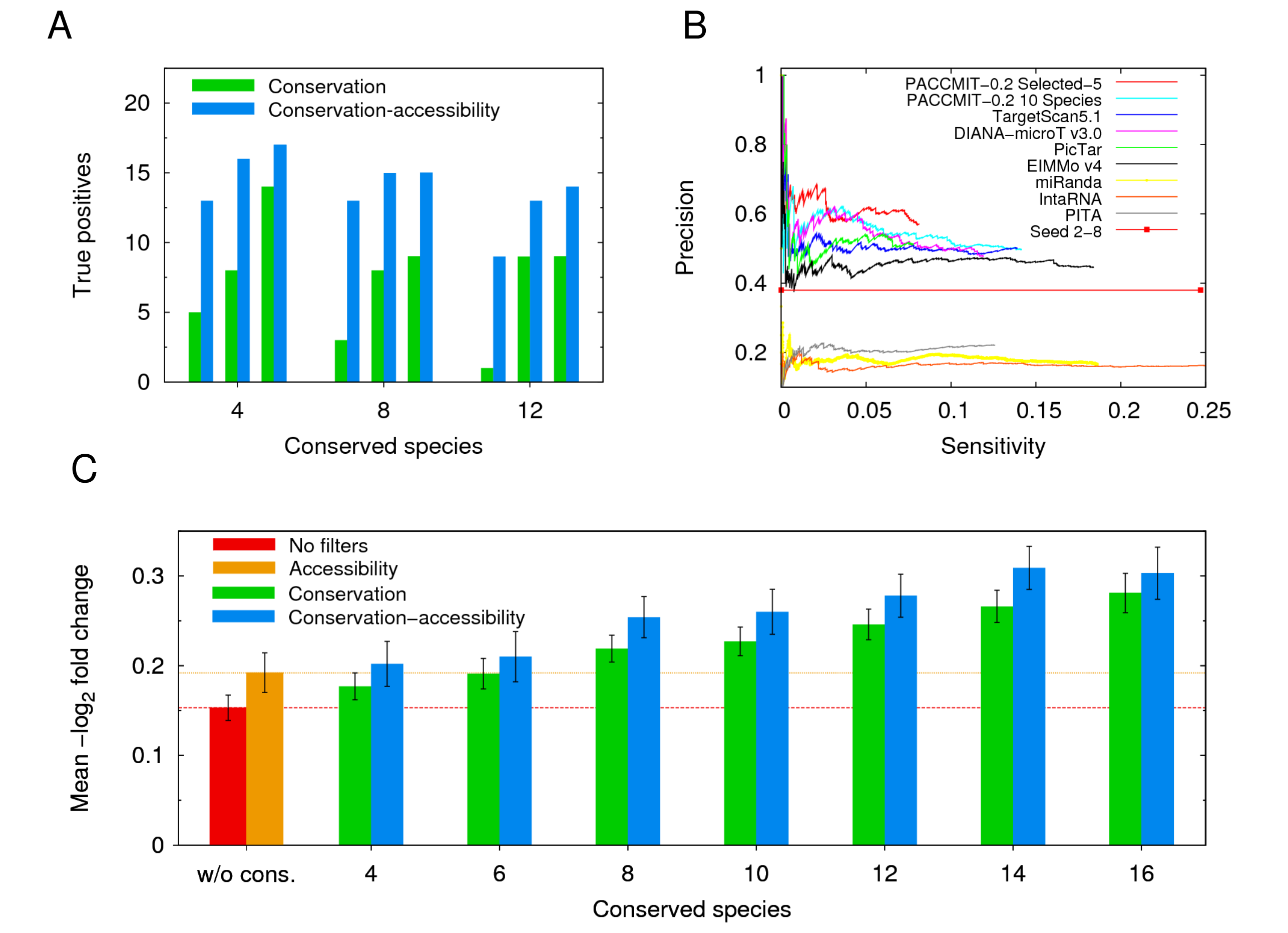
B. Effect of the score

- **Best site**: lowest free energy found for a miRNA-3'UTR pair.
- **Average & Sum**: average and sum of the free energies for all sites (respectively) .
- **PITA score:** $=-\log\sum_i e^{\Delta\Delta G_i}$ where $\Delta\Delta G$ is the total free energy for site i.

(A) Requiring seed matches without GU pairs works better .

(B) Using different scores based in the free energies has no effect in the performance of the method.

Combining conservation and accessibility filters



(A) The accessibility filter rejects false sites not discarded by the conservation filter.

(B) Comparison with widely used methods shows that ours (PACCMIT) has the highest precision.

(C) Targets predicted using the combined filter show in average the highest repression.

In general the quality of the predictions increases according the use of the following filters: accessibility $<$ conservation $<$ conservation-accessibility.

References

1. Marin, R.M. and Vanicek, J. *Nucleic Acids Res.*, **2011**, *39*, 19.
2. Murphy, E., Vanicek, J., Robins, H., Shenk, T., Levine, A. J. *Proc. Nat. Acad. Sci. USA*, **2008**, *105*, 5453.
3. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., Schuster, P. *Monatsh. Chem.*, **1994**, *125*, 167.
4. Bernhart, S. H., Hofacker, I. L., Stadler, P. F. *Bioinformatics*, **2006**, *22*, 614.
5. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., Segal, E.*Nat. Genet.*, **2007**, *39*, 1278.
6. F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao and T. Li, *Nucl. Acids. Res.*, **2009**, *37*, D105.
7. Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R. Rajewsky, *Nature*, **2008**, *455*, 58.
8. Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M., Hatzigeorgiou, A. G. *Bioinformatics*, **2009**, *25*, 3049.
9. Robins, H. and Press, W. H. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 15557.
10. Robins, H., Li, Y. and Padgett, R. W. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 4006.