Cancer Gene/Module Identification

Cesim Erten
Antalya Bilim University

[Joint work with R. Ahmad, I. Bali, E. Hoxha, H. Kazan]

Informal Definition

- Given mutations data (TCGA etc.):

  Find driver genes/modules in cancer

- Two problems:
  - Candidate Genes
  - Candidate Modules
- Problem with Gene Ranking:

  Mutations at different loci could lead to the same disease

Driver Module Identification

- De novo Methods:
  - Rely only on genetic data
  - [Miller et al. 11; Vandin et al. 11, Leiserson et al. 13, Liu et al. 17]
  - Algorithmically, heavy submatrix type problems
  - Disadvantage:
    - Solution space is large
    - Prefiltering based on frequencies may miss rare mutations

Driver Module Identification

- Knowledge-based Methods:
  - Interaction networks as well as genetic data
  - Coverage oriented methods:
    - Heat diffusion of mutation frequencies
    - Algorithmically, heavy subgraph type problems
    - Hotnet, Hotnet2, HierHotnet …
  - Coverage+Mutual Exclusion:
    - Simultaneous mutations of genes in shared pathways not frequent
    - Algorithmically, greedy seed-and-extend type heuristics
    - MEMo, BeWith, MEMCover ...

Driver Module Identification

- Formally, given S_i and PPI network G,

$$MEX(M) = \frac{|\bigcup_{\forall g_i \in M} S_i|}{\sum_{\forall g_i \in M} |S_i|} \qquad MS(P) = \sum_{\forall M_q \in P} RS(M_q) \times MEX(M_q).$$

$$CO(M) = \frac{|\bigcup_{\forall g_i \in M} S_i|}{|\bigcup_{\forall g_i \in V} S_i|}. \qquad CS(P) = \sum_{\forall M_q \in P} \frac{1 - RS(M_q)}{\sum_{\forall M_t \in P} 1 - RS(M_t)} \times CO(M_q),$$

- Find P, maximizing $DMSS(P) = MS(P) \times CS(P)$ such that
  - Subgraph induced by each M is connected
  - Min module size and total size below given thresholds
- Computationally intractable

Algorithm for Module Identification

- Generate node-weighted edge-weighted graph
  - reflecting coverage and mutual exclusion

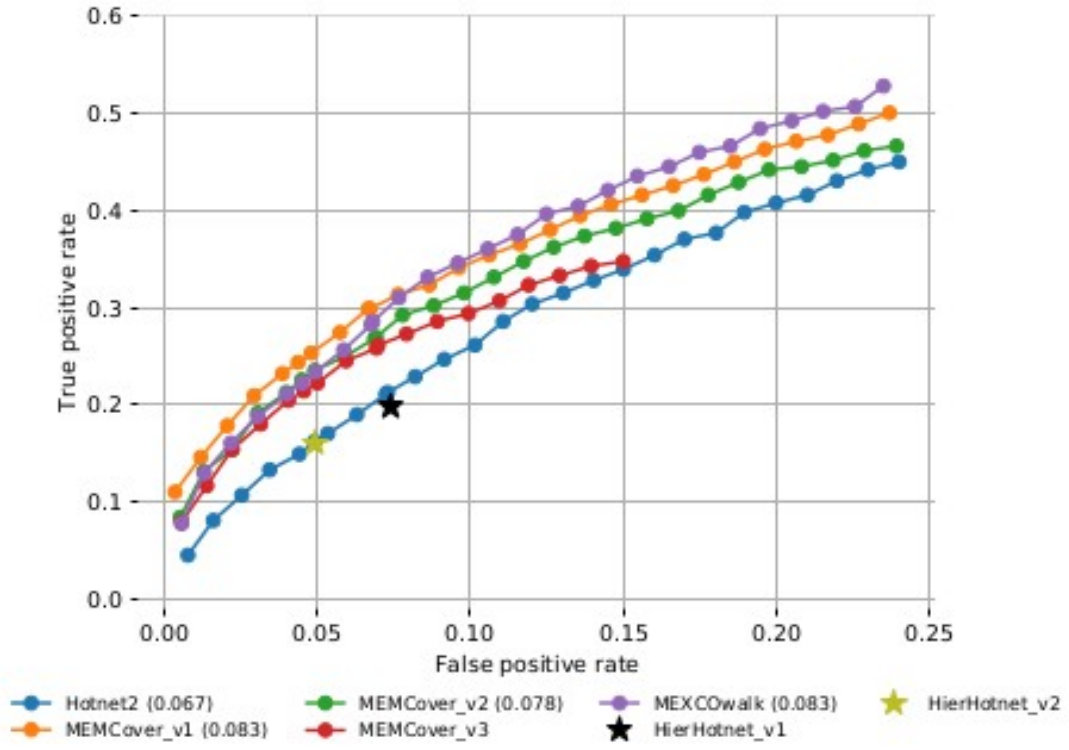
$$w(g_i, g_j) = MEX_n(g_i, g_j) \times CO(\{g_i\}) \times CO(\{g_j\}).$$

- Edge-weighted random-walk with restart
- Initial modules:
  - strongly connected components
  - similar to other heat-diffusion methods
- Split-and-extend large modules

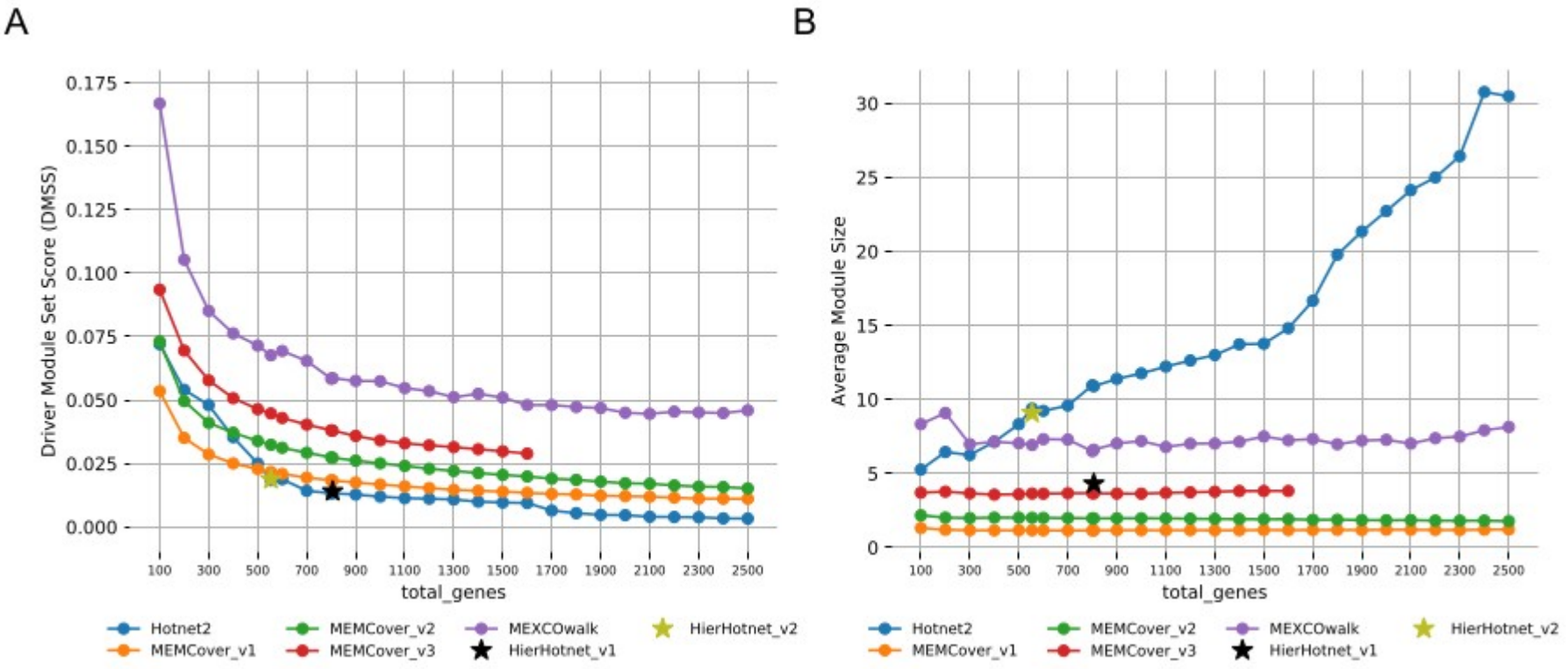
Evaluation Metrics

- Systematic evaluations of previous work:
  - Static evaluations based on reference sets (COSMIC etc.)
- Module-specific systematic evaluations missing:
  - Carefully defined optimization scores as in DMSS
  - Cancer type/subtype specificity score
  - Classification (normal vs tumor) accuracy score

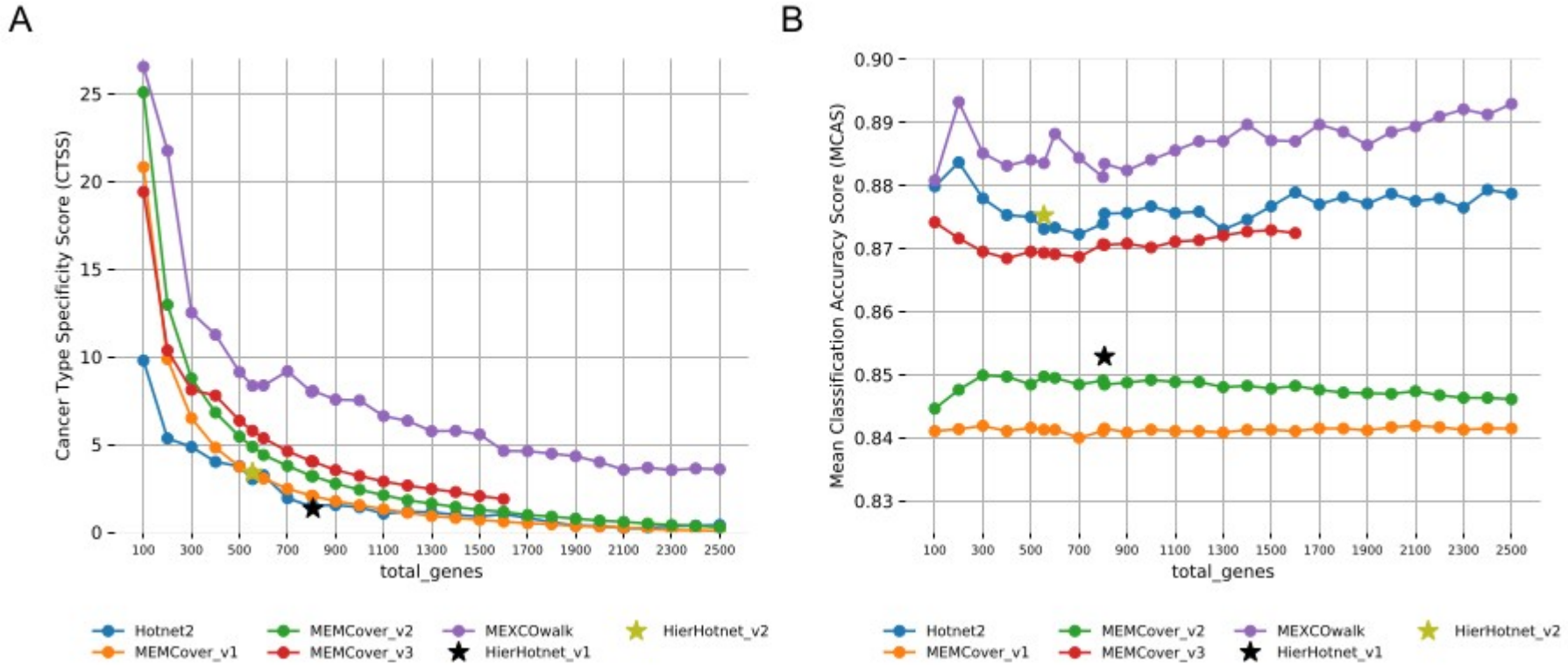
Static Evaluations



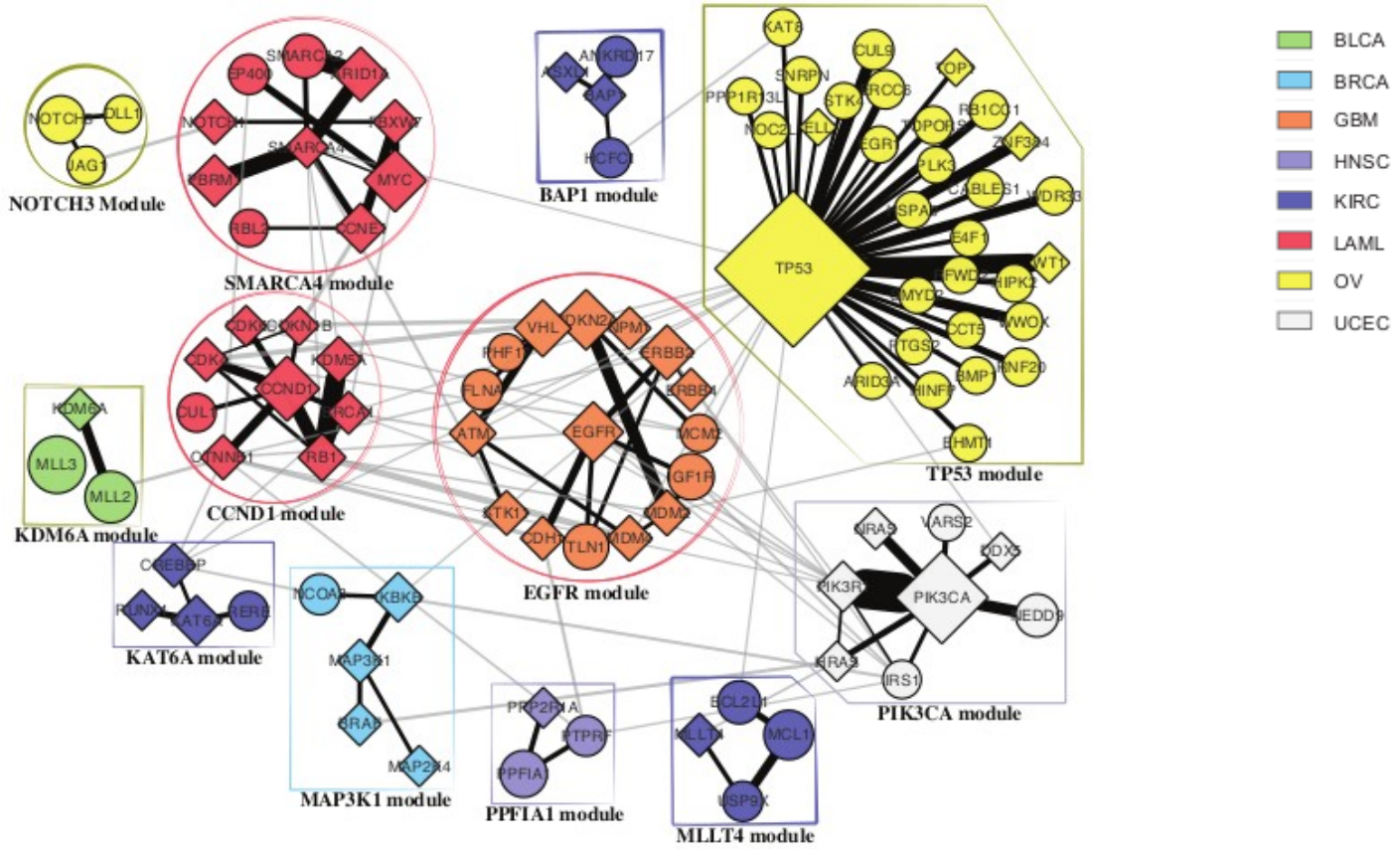
DMSS and Average Module Size



Type Specificity and Classification



Modules on Pancancer Data



Open problems

- Module-oriented
  - optimization problem definitions
  - evaluation criteria
- Especially for the overlapping modules case
- Computational complexity results on sparse graphs

THANK YOU