The inputs to the sequential forward floating search (SFFS) algorithm are the top *k*th (as determined by apparent accuracy) link community (LC) sets. Generally, I set k to be 50.

While the link communities were defined by utilizing the protein-protein interaction (PPI) data, it is assumed that there is still useful information content within the PPI data which has not been captured by the initial link communities. That is, the biomolecular networks defined by the LC sets should be refined to improve biomarker fidelity.

Main ideas of the SFFS algorithm:

Therefore, the sequential forward floating search seeks to adapt each LC set by iteratively adding and removing genes. The pool of available genes to add on an iteration is the set of genes which share an edge (putative interaction) in the global PPI graph with at least one of the genes already in the seed set. (I have also explored using ‘n hops’ (n=2,3) away in the graph, in the event that the addition pool is too limited – generally, this does not improve performance.)

Furthermore, on each iteration, following adding the ‘best’ gene (as determined by apparent accuracy), SFFS begins conditional exclusion of genes already within the set. This is in hopes of constructing the smallest set which still provides the best accuracy. Imagine a set of genes {*a,b,c,d*} – to which is added gene *e*. Gene *e* provides all of the information (in the sense of acting as a distinguishing biomarker for phenotypes x versus y) already provided by gene *b*, and then some. Thus, SFFS would be able to safely remove gene *b*. If instead the best gene to remove would be that just added on this iteration, gene *e*, then SFFS removes no genes and goes on to the next iteration.

The conditional exclusion step proceeds as long as possible in a given iteration, and thus on a single SFFS iteration, exactly one gene may be added, but zero or more genes may be removed.

In the following examples, it is interesting to note that (according to SFFS at least, and for these representative samples) the initial LC sets contain many extraneous genes. Generally, the first iteration of SFFS clears many of these out to a most important ‘basis set’, and then sets out on a more balanced give-and-take.

P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters, Volume 15, Issue 11, November 1994, Pages 1119-1125, ISSN 0167-8655, DOI: 10.1016/0167-8655(94)90127-9.

**SFFS in pseudocode:**

Initialize PPI graph, PPI

For each LC set, X with |X| = k

Set θ[2,…,k] = apparent accuracy of LC set X

(*Iterate)*

While apparent accuracy of set X improves at least η,

(*Inclusion)*

Define addition pool as gene set A, such that ∀a∈A, ∃x∈X. ∃(a,x)∈ PPI

∀a∈A, compute the apparent accuracy of the new set Xa = X ∪ {a}

Let X’ be the set among these with the highest apparent accuracy, θ

Let θi+1← θ, where |X| = i and |X’| = i+1

If |X’| below minimum (usually 3), do not remove any elements and begin a new iteration

(*Conditional exclusion)*

Else, ∀x∈X’, compute the apparent accuracy of the new set X’x = X’ \ {x}

Let X” be the set among these with the highest apparent accuracy

If X” = X (removed x = added a), begin a new iteration with X ← X” (keep a in the set),

Otherwise,

(*Continuation of conditional exclusion)*

∀x∈X”, compute the apparent accuracy of the new set X”x = X” \ {x}

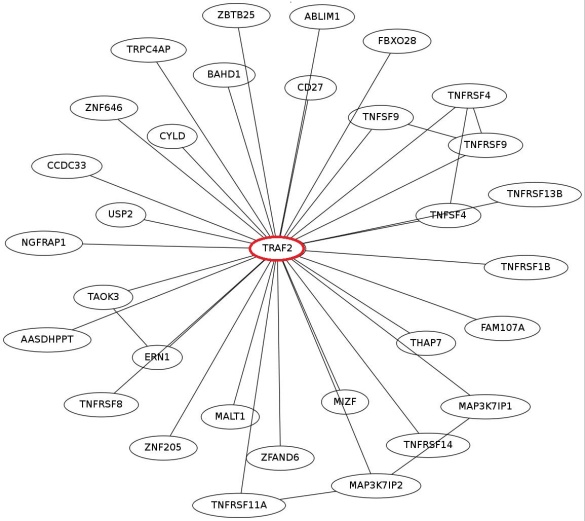
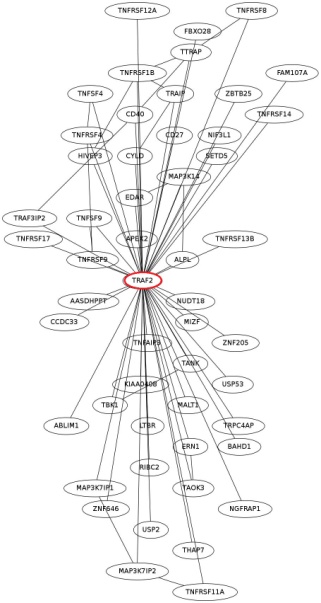
Let X”’, |X”’| = i, be the set among these with the highest apparent accuracy, θ

If θ is greater than the previously observed apparent accuracy for any set of size i, θi,

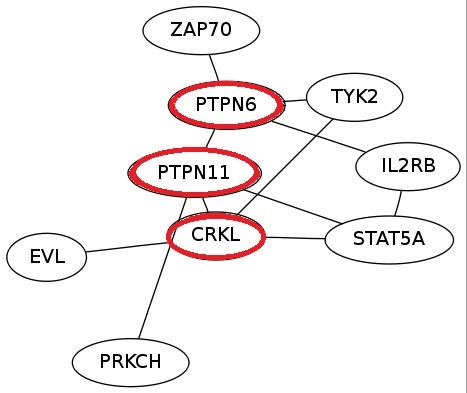
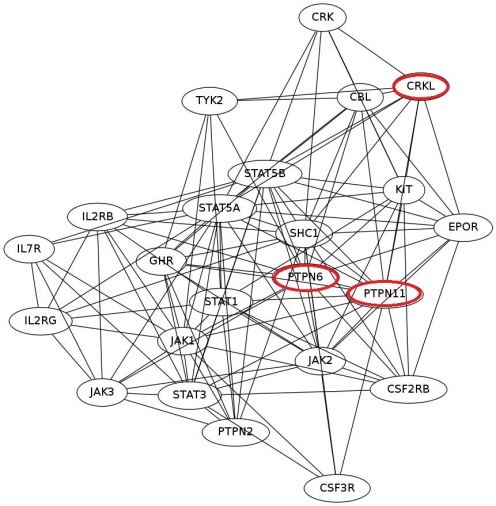
continue conditional exclusion with X” ← X”’, θi ← θ, unless |X”’| < 3

Else, begin a new iteration with X ← X”

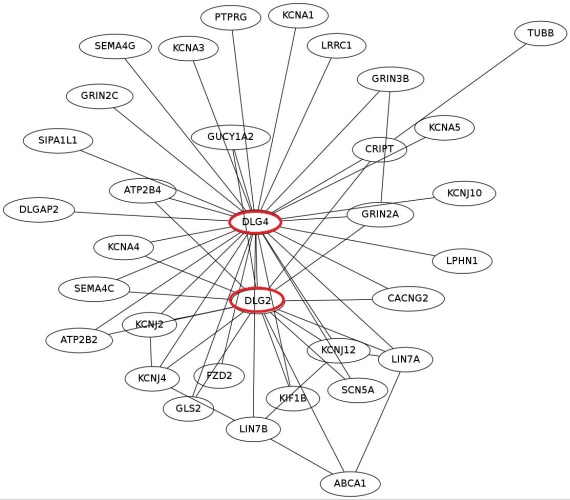
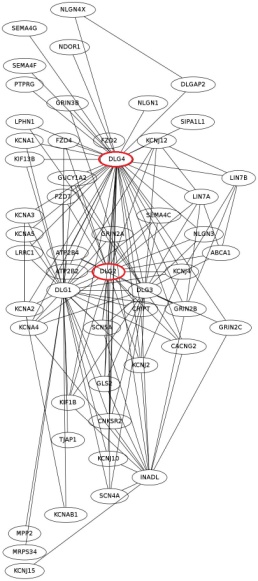
PPI induced graphs of three seed (LC) networks and corresponding SFFS networks (lung cancer dataset):



Seed network 1 SFFS network 1



Seed network 2 SFFS network 2



Seed network 3 SFFS network 3

The complete network evolution for each step of SFFS:

Network 1:

+PCBD1-TTRAP-TBK1-HIVEP3-KIAA0408-NIF3L1-RIBC2-SETD5-TANK-EDAR- ALPL- APEX2-MAP3K14-NUDT18-LTBR-CD40-PCBD1-TNFAIP3- TNFRSF12A- TNFRSF17-TRAF3IP2-TRAIP-USP53

+ZFAND6 (terminated)

Network 2:

+IL21R-JAK3-CRK-CSF3R-EPOR-GHR-IL21R-IL2RG-JAK1-JAK2-KIT-PTPN2- SHC1-STAT1-STAT3-STAT5B

+EVL

+PIK3CG-IL7R-PIK3CG

+PRKCH

+EIF5B

+PRLR-CSF2RB-CBL-PRLR

+ZAP70 (terminated)

Network 3:

+LRP2-SCN4A-SEMA4F-CNKSR2-DLG1-DLG3-FZD4-FZD7-GRIN2B-INADL- KCNA2-KIF13B-LRP2-NDOR1-NLGN1-NLGN3-NLGN4X

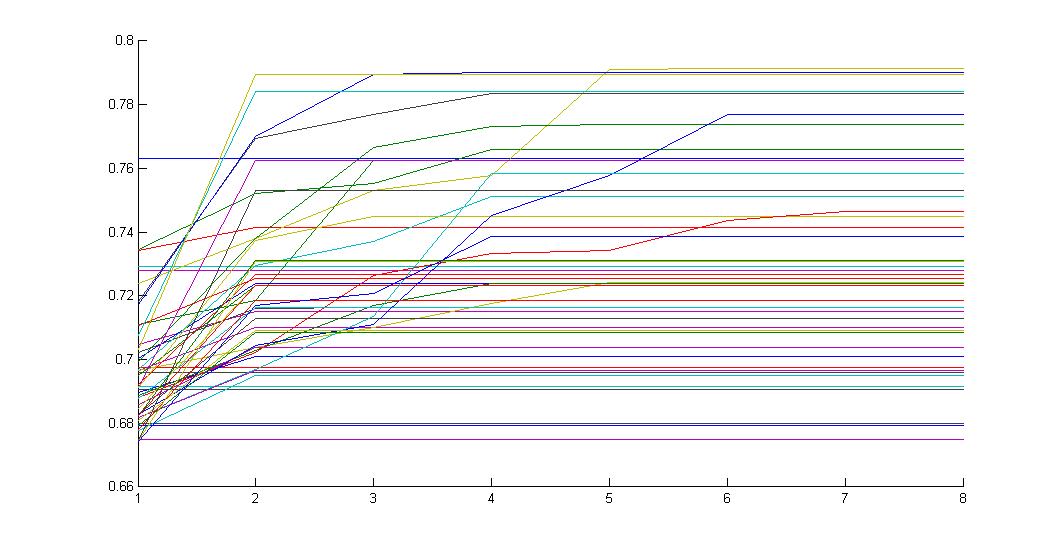
+TUBB (terminated)

Apparent accuracy for candidate networks (50) in Lung Cancer over course of SFFS

Apparent accuracy

Vs

Iteration



Each plot is one of 50 different networks; the search was halted for most of them by a lack of improvement (eta cutoff) after only a few iterations. In general, the high-accuracy SFFS networks have one prevalent paradigm in terms of structure: the seed network is reduced in size to one or a few central ‘hub’ genes which provide the base from which further interactions are drawn. Apart from the hub, the induced interaction graph is sparse.