Biomarker Candidate Generation using Small Sets of Samples

Preliminary Results on PHD Data Analysis

Mark-Oliver Stehr

SRI International

PHD Data and Analysis Objective

Characteristics of PHD Data:

- High-dimensional data (> 10000 genes)
- Nontrivial temporal dimension (non-equidistant time series with many gaps)
- Baseline measurements (usually more than one, but not always)
- Large amount of noise (baseline measurements differ, spikes)
- Small sample size (19-20 subjects)
- Classification into symtomatic/asymtomatic not necessarily objective
- Unbalanced samples classes

Objective:

- Prediction of eventual symptoms based on measurements at early or middle time points of the time series
- Synthesis of predictors with low error rates that: have the potential to generalize and can be easily interpreted and hence can be confirmed/refuted based on biological expertise or follow-up studies
- Development of a general method to synthesize such candidate predictors
- Method should generalize beyond gene expression data
- Method should enable incorporation of external/background knowledge

SRI International

Approach and Rationale

Challenges

- Sample size is not sufficient to afford luxury of independent test set
- Training set of much less than 20 does not make much sense
- Noisy high-dimensional data makes easy to find patterns for almost everything

➔ Learning must be highly constrained to obtain predictors that can generalize
(cf. regularization, minimum description length principle)

Overview of Approach

- Select subset of genes as relevant for the analysis to focus on most promising candidates (reduces computational complexity and inherently noisy genes)
- Reduce noise by filtering/averaging/abstraction techniques
- Reduce degree of overfitting by avoiding model/optimization parameters
- One level of leave-one-out crossvalidation to estimate error rate of the method
- Avoid second level of crossvalidation for parameter optimization
- Perform randomization test to determine p-value of the estimated error rate
- Use all samples (all available information) to construct best predictor

SRI International

Preprocessing & Noise Estimation

Time Series Preprocessing:

- Gaps in the time series are filled with a replica if available.
- If two measurements are available for a given time point we use the average.

Definition of Baseline:

The baseline interval is defined as the initial interval up to the time of infection.

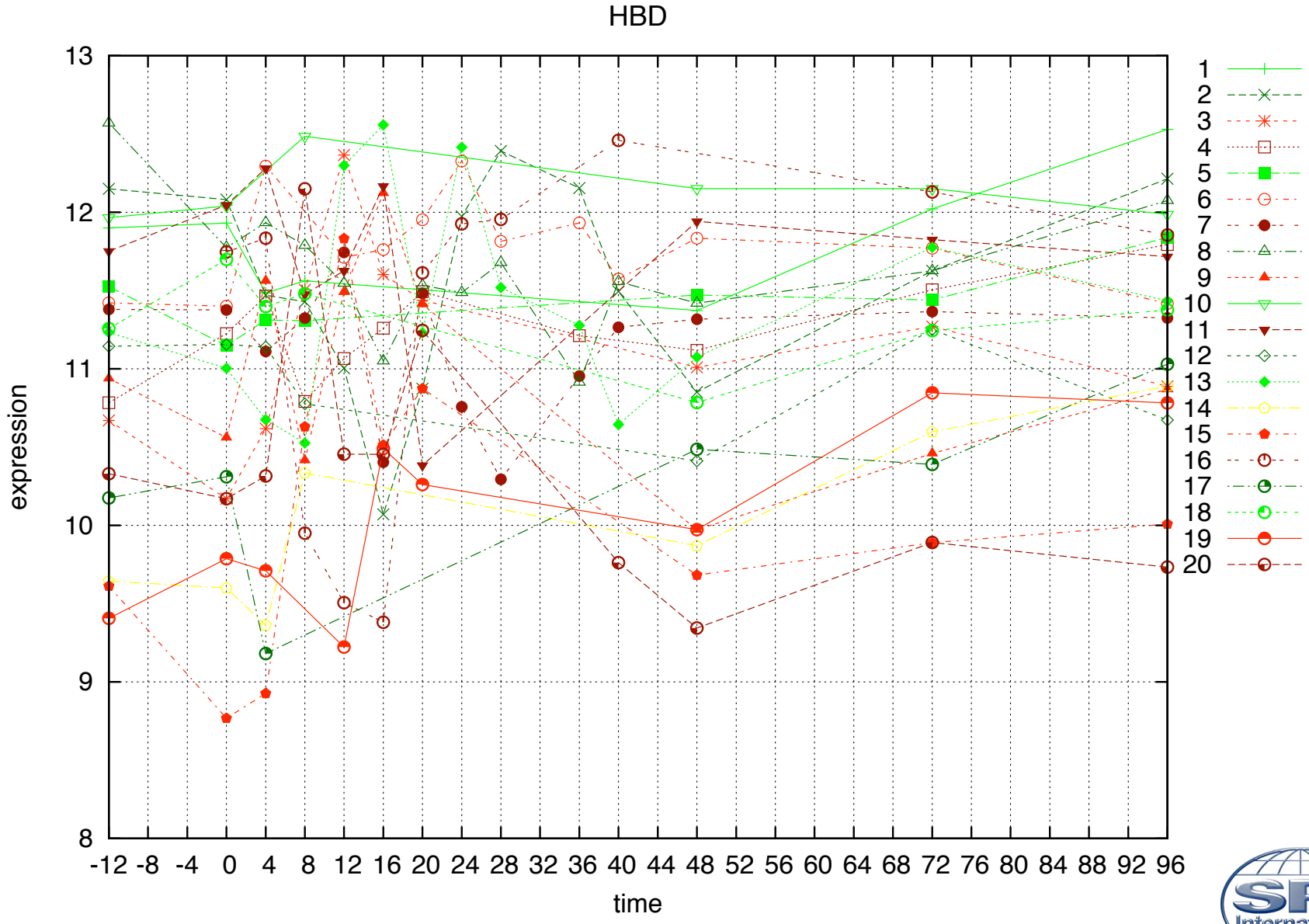
Noise Estimation:

- For a given subject and gene,
  the range (max-min) of the baseline measurements
  can be used as an estimate of (subject/gene-specific) noise.
- However:
  there a not more than two and sometimes only one baseline measurement available in the study, making the estimate inaccurate or impossible.
- Hence:
  we estimate, instead, the gene-specific noise by computing
  the average range over all subjects for a given gene.

Other methods for noise estimation, e.g. use of standard deviation, have been investigated as well.

SRI International

Sample Time Series after Preprocessing



Prescreeing Genes

The prescreening stage reduces the set of candidate genes from which biomarkers are selected in later stages:

- Rank genes according to a quality metric, e.g. range-to-noise ratio, where range is defined over the entire time series (min-max), and noise is estimated as explained before.

- Define a gene as relevant for the subject is it ranks among the top, e.g. 10%.
- Compute such a ranking independently for each subject.
- Then define relevant genes as the union of all genes relevant for some subject.

More generally:

a gene needs to be relevant for a certain fraction/number of subjects (e.g. 1/3) before it is considered relevant. In other words, relevance needs to be somewhat consistent across subjects.

This reduces the number of genes considered relevant due to noise or measurement errors, but might also miss some interesting genes.

SRI International

Time Series Normalization

For each relevant gene and subject,
we perform the following two operations on the times series:

- Baseline alignment:

  Subtract the baseline from all values of the times series, where baseline is defined as the average value in the baseline interval.

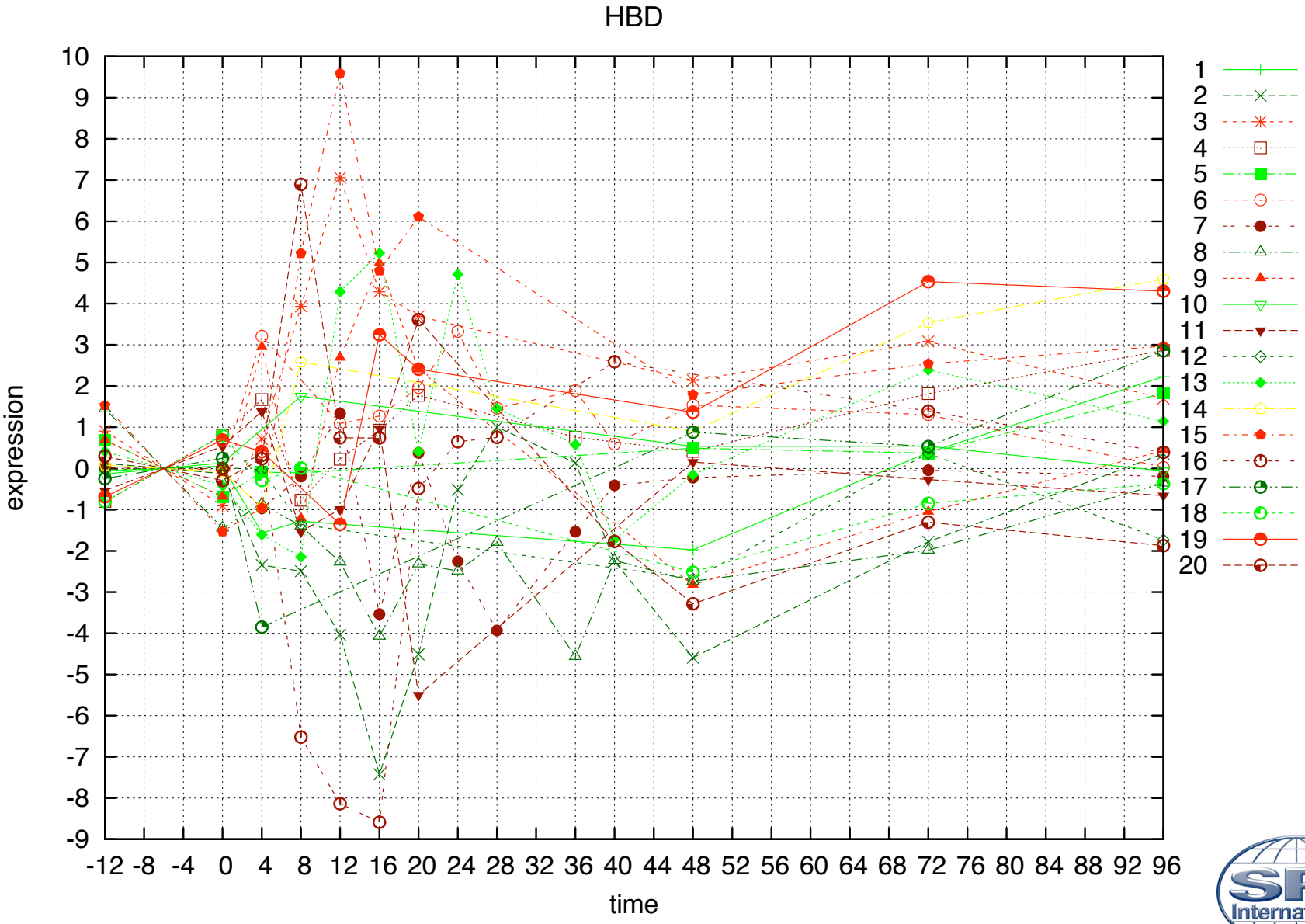
- Rescaling (noise adjustment):

  All values of the time series are divided by the noise of the underlying gene.

Note:
The baseline average and also the gene-specific noise can be obtained from any large population of subjects (possibly different from the subjects of the study) so that measurements before the infection may not be necessary. However, the impact on the quality of the predictor remains to be investigated.

SRI International

Sample Time Series after Normalization



Time Series Temporal Abstraction

The temporal abstraction defines the basis of our predictors.

Furthermore, it improves robustness and enables us to deal with gaps and reduce noise (e.g. spikes) in the original time series data.

Similar to the baseline interval, we perform an interval-based abstraction of the remaining time points:

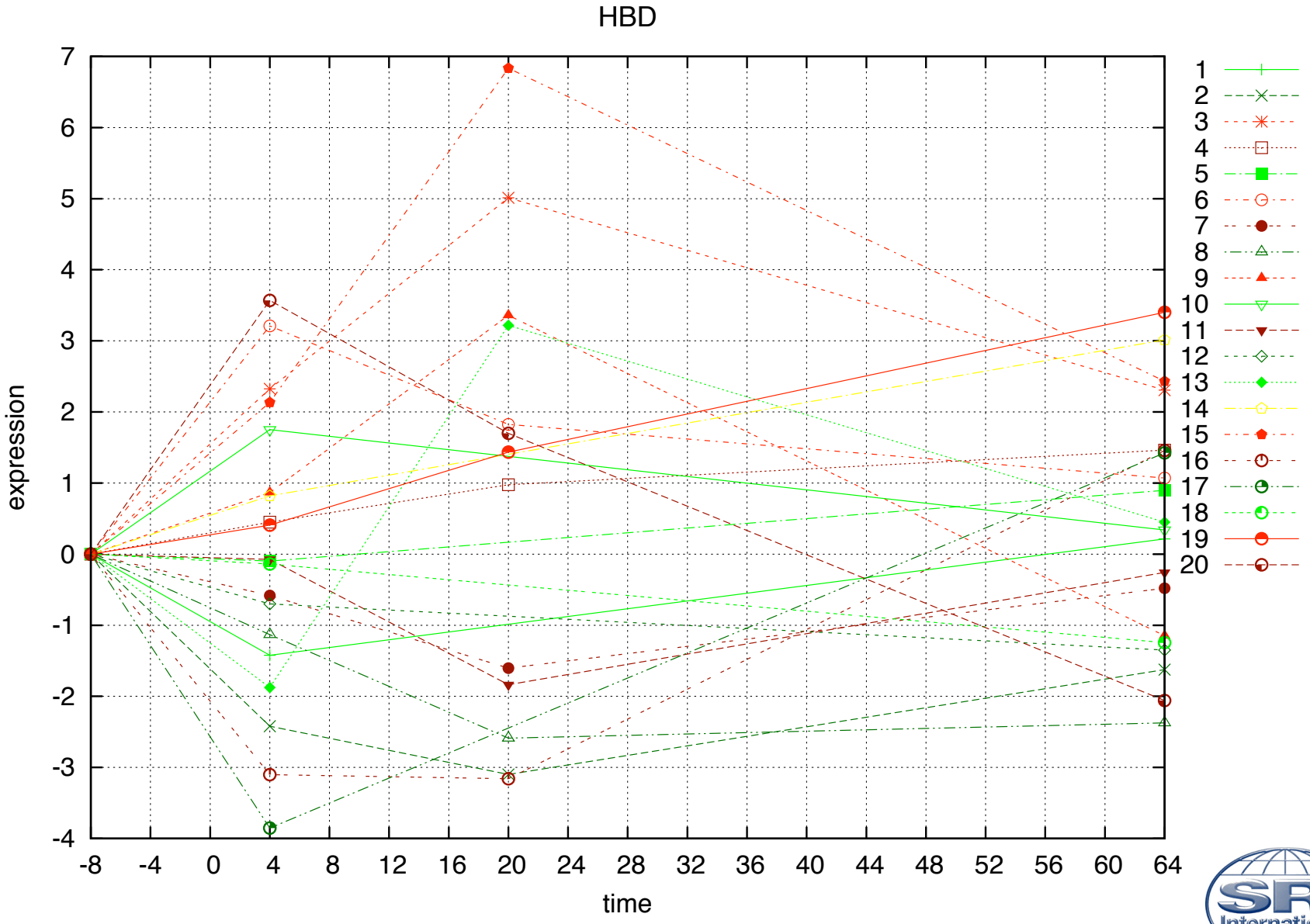
- We introduce three additional intervals, namely the early, middle, and late interval, to cover the remaining time points.
- Each interval will be replaced by a single measurement defined as the average of all values in the interval.

Note:

Clearly, to be effective each interval should cover more than one measurement.

We currently use non-overlapping intervals for the temporal abstraction, but this is not a strict requirement.

Sample Time Series after Abstraction



Single Gene Predictors

- Predictor genes are selected from relevant genes only
- With each gene we associate a single gene predictor
- Predictor is a ternary logic threshold predictor that can give three answers:
  -1 (negative) if the gene expression is below 1.0
  0 (undecided) if gene expression is between -1.0 … 1.0
  +1 (positive) if gene expression is higher than 1.0

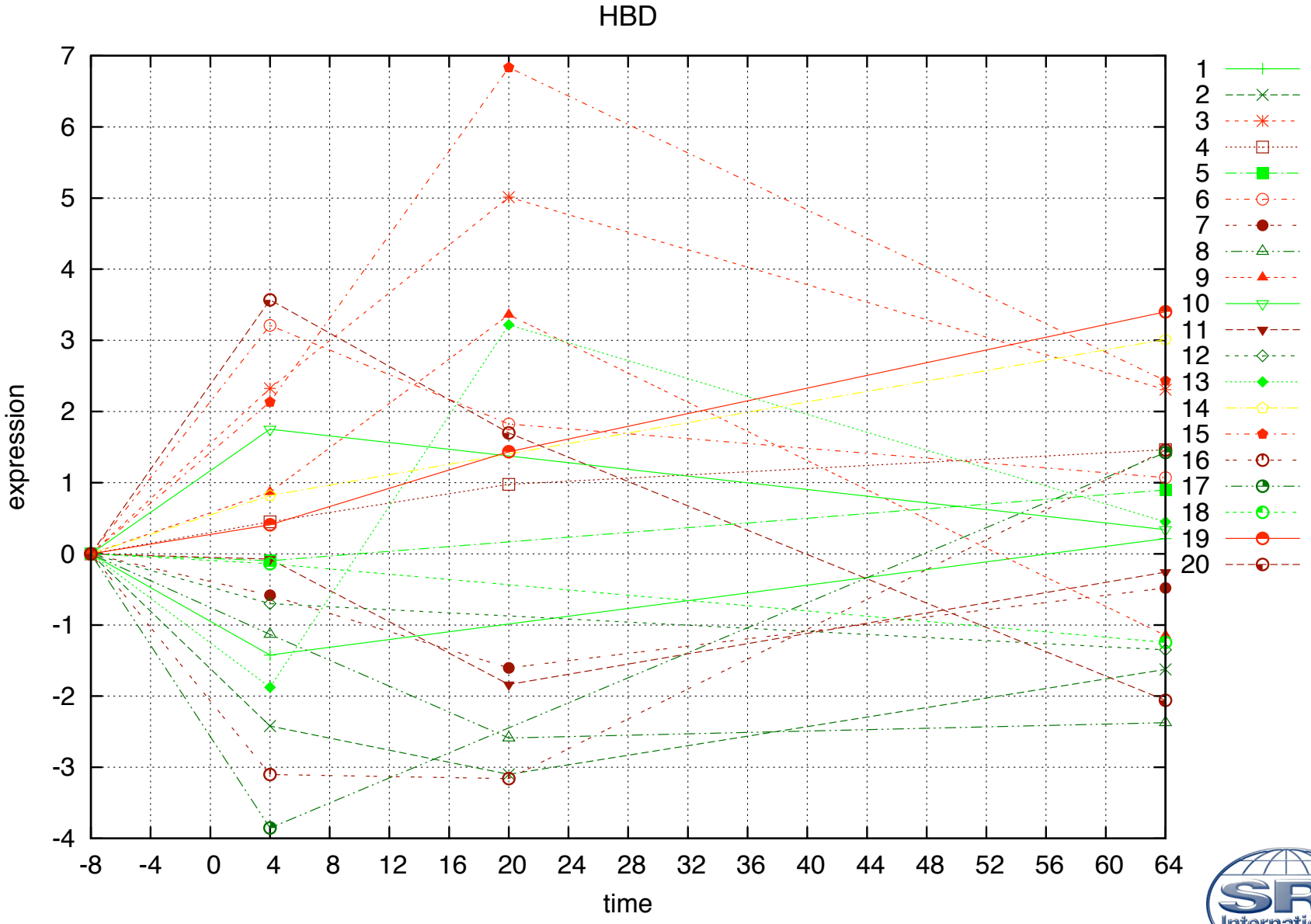
Note: The threshold 1.0 is fixed and corresponds to the noise-level in the normalized time series. It is not a gene-specific model parameter that is subject to optimization.

Definitions:

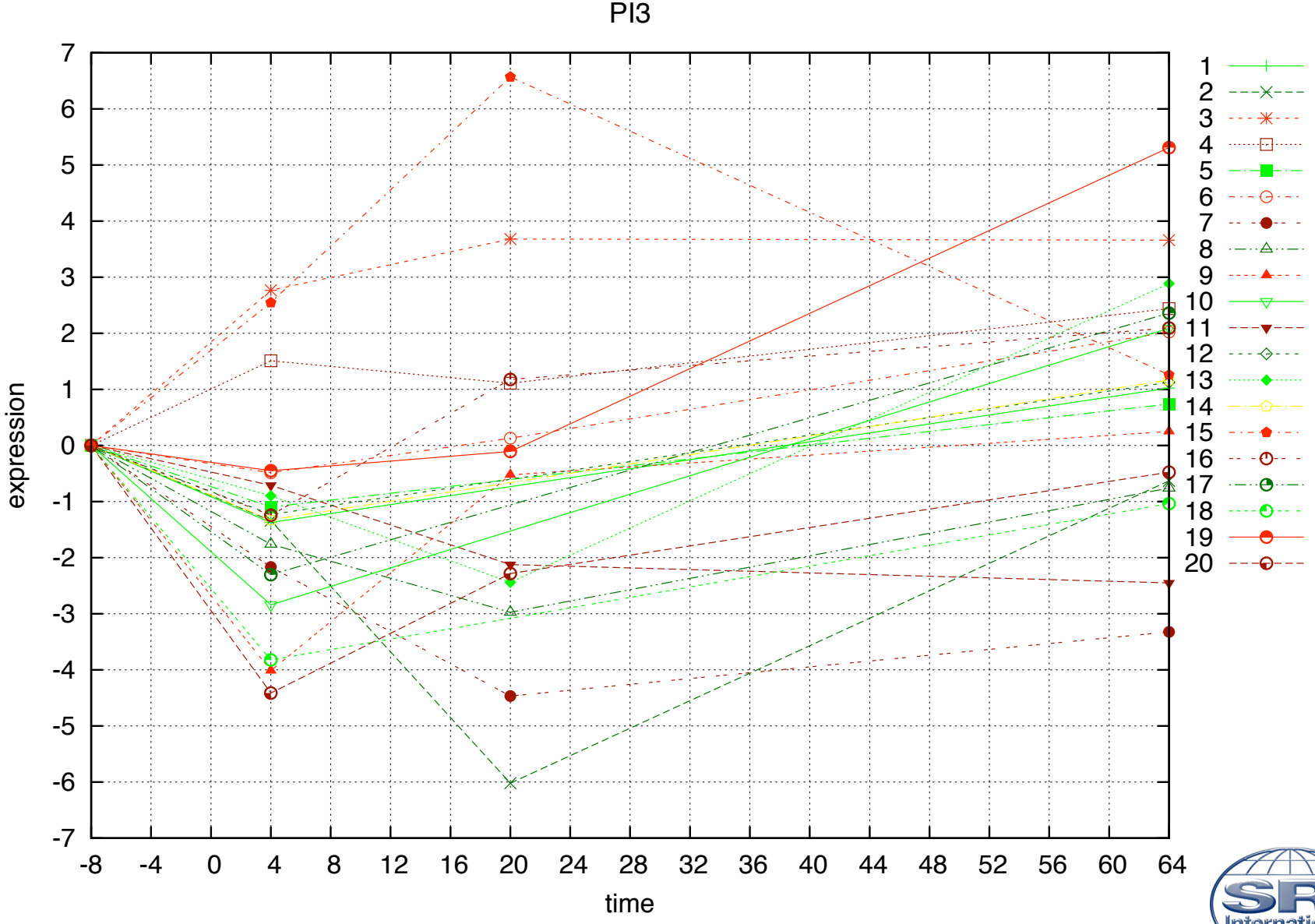
- The prediction rate (PR) is defined as the fraction of classified subjects, i.e. the predictor gives a positive or negative answer.
- The error rate (ER) is simply the fraction of missclassified subjects based on the positive/negative answers (undecided is never considered an error).
- To compare predictors with different capabilities, we also use a normalized error rate NER = PR * ER + (1-PR) * 0.5 (error rate of a corresponding binary predictor that is forced to toss a coin if the ternary predictor is undecided).

SRI International

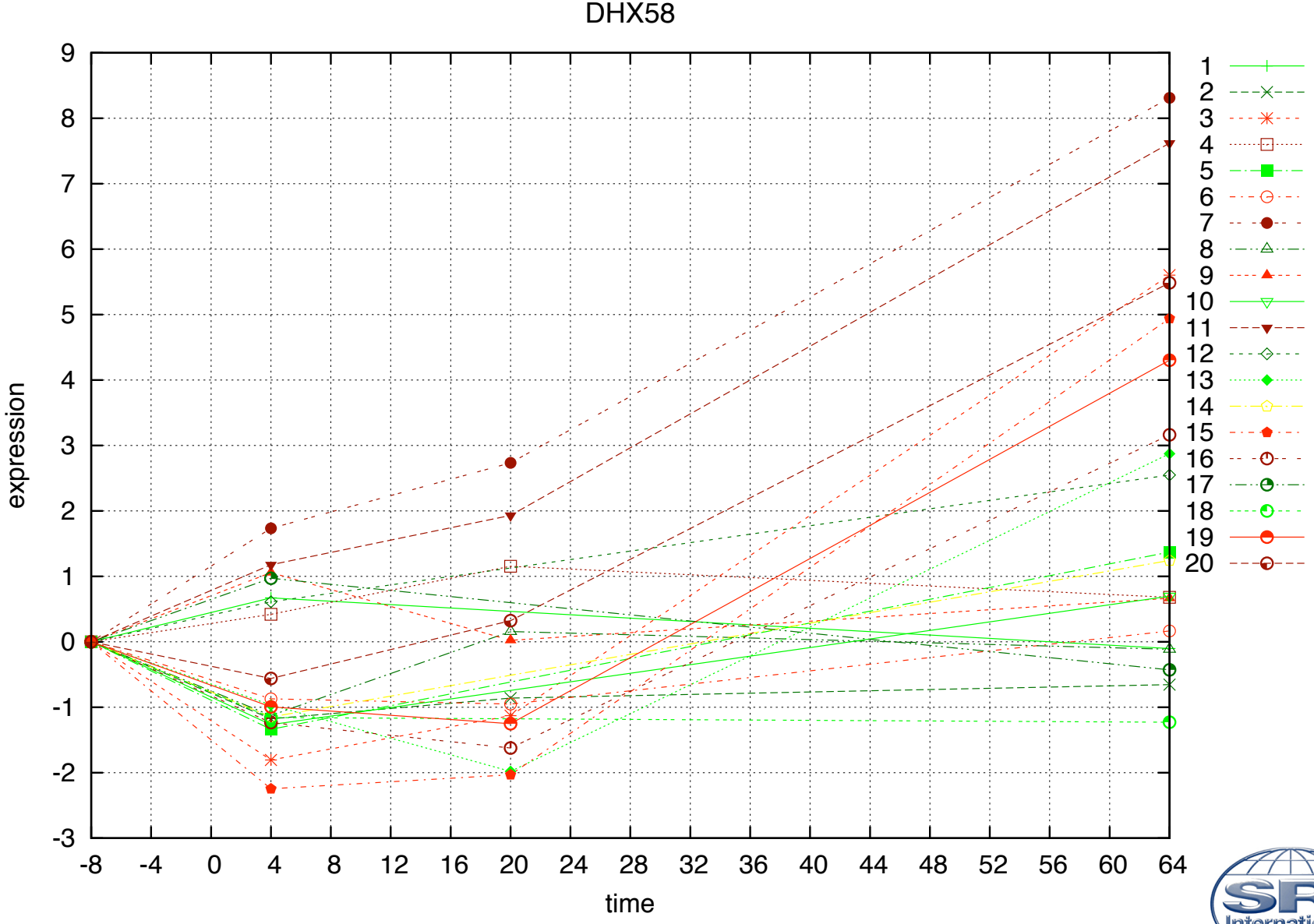
Rank 1 - Early Predictor (Positive)



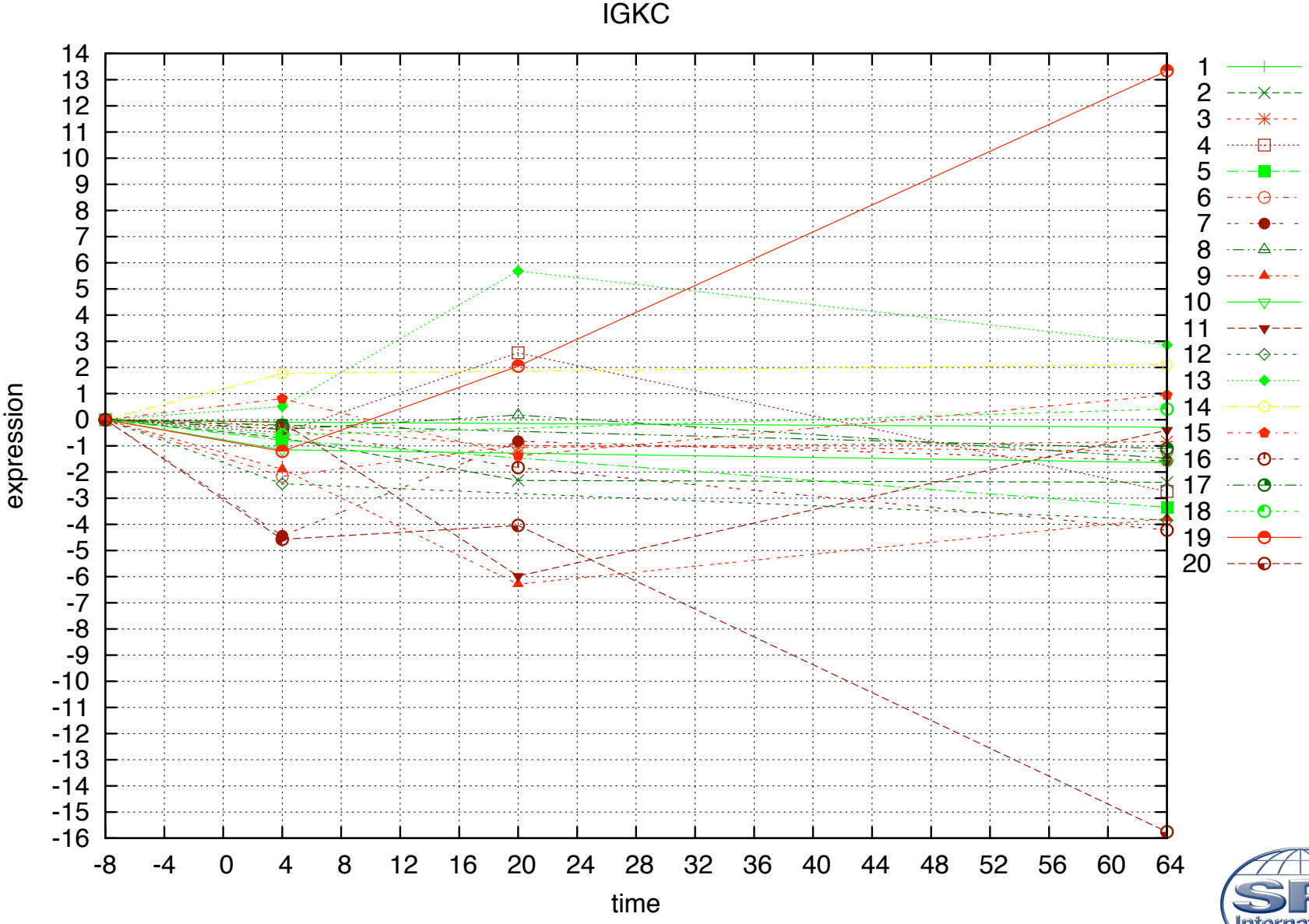
Rank 2 - Early Predictor (Positive)



Rank 3 - Early Predictor (Positive)



Rank 4 - Early Predictor (Negative)



Voting-Based Predictor Composition

A set of single gene predictors can be composed by majority voting:
The answer of the composite predictor is simply
the sum of the answers of its single gene predictors.
This yields again a ternary predictor with three possible answers:
> 0 (positive), 0 (undecided), and < 0 (negative).

Predictor composition is highly constrained to reduce overfitting,
i.e. not every set of genes can be composed.
All single gene predictors are ranked according to their normalized error rate.
As composite predictors we only consider compositions of the top n single gene predictors for all n up to some bound (e.g. 100).

Note that in this approach each composite predictor is an extension of the previous one (starting with the best single gene predictor).
Additional heuristic constraints can be imposed to increase the independence of the component predictors, e.g. add a single gene predictor only if

- if coverage, i.e. set of classified subjects, increases.
- if it is non-correlated (in a sense to be defined) to the previous ones.

SRI International

Predicting Best Predictor Complexity

- Note that the only model parameter in the predictor synthesis is n, the number of predictors to compose.
- The standard way to estimate n is to use cross-validation, but that would introduce another level of cross-validation for the purpose of parameter optimization and the size of the training set would further shrink.
- Hence, we simply estimate n based on the training error rate.
- One simple heuristic uses the first local minimum, that is point where adding another component predictor does not improve the training error rate.
- The underlying rationale is that if the training error rate does not improve it its unlikely that the real error rate would improve.

Dimensionality Reduction

Remaining problems:

- high-dimensionality of the data set (danger of overfitting)
- sensitivity to individual gene noise (non-robustness)

Natural idea:

- move to a higher level of abstraction
- instead of a single gene use a set of genes (a cluster in some sense)
- then apply the predictor synthesis at this abstract level.

Clusters should contain functionally related genes with similar expression profiles.

- A rapidly growing amount of public data is available to derive such clusters in a fully automatic fashion, e.g. from gene ontology or sequence data bases, microarray data repositories
- However, our analysis shows that already the PHD data alone leads to a meaningful notion of cluster if we exploit the temporal dimension (without subject classification information)

Correlation Analysis

Definition of correlation:

- For each time series we compute the difference time series, i.e. each value is replaced by the change relative to the previous available time point.
- For each subject and pair of relevant genes, we compute the Pearson correlation between the gene's corresponding difference times series.
- For each subject s we define a binary correlation relation $\sim_s$ such that genes g $\sim_s$ g' are correlated iff the the absolute value of the Pearson correlation is larger than some threshold (e.g. 0.8)

Note:

- We restrict the computation to relevant genes for computational feasibility.
- Time series across subjects are not compared in this approach.
- Negative correlations are as important as positive correlations.

Definition of consistent correlation:

- Now we define ~ such that g ~ g' iff g $\sim_s$ g' for all subjects s
- More generally, g ~ g' iff g $\sim_s$ g' for a number/fraction of all subjects s (e.g. 1/3).

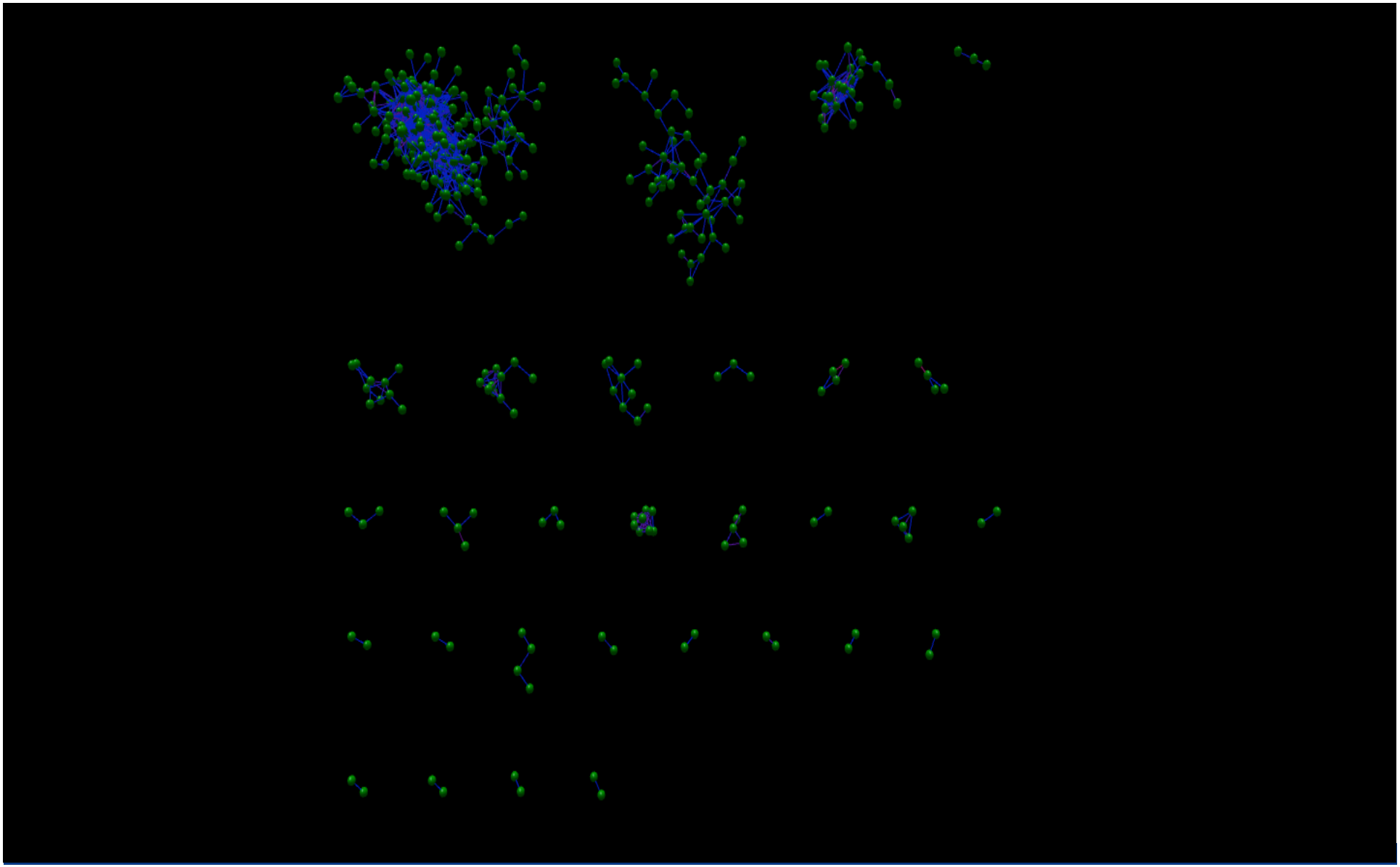
A similar approach can be used to define a consistent non-correlation relation.

SRI International

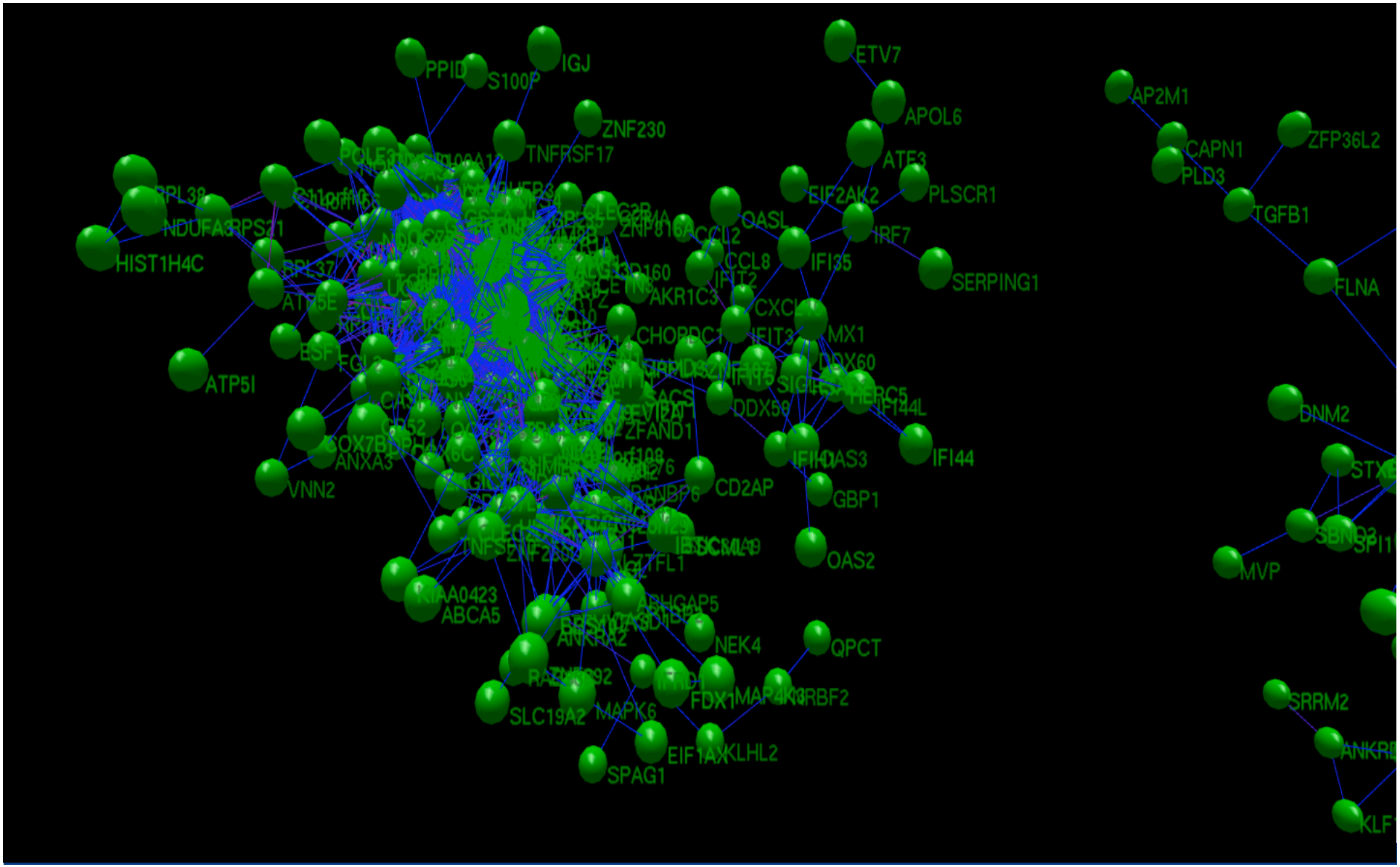
Graph-Based Clustering

- Uses consistent correlation graph (but no underlying correlation coefficients).
  - Advantage: The graph abstraction is less prone to small changes (although can significantly depend on the correlation threshold).
- We define the top-level clusters of the graph as its connected components.
  - Unfortunately, this naïve clustering based on transitive closure is almost always too coarse-grained, ignoring the fine-structure of each cluster.
- Hence, working directly with the similarity relation, each cluster will be further decomposed into a hierarchy of potentially overlapping subclusters.
- The key notion is that of a ken, defined as a maximal clique, i.e. in out context, a set of genes that are pairwise consistently correlated.
  - Unfortunately, already the clique problem is NP-complete and there can be exponentially many (maximal) cliques (impossible to enumerate).
  - However, using a logical encoding and a state-of-the-art SAT-solver (more precisely MaxSAT) it is feasible to find at least one ken in a graph with hundreds of nodes.
- Instead of enumerating all kens, we interleave search and abstraction:
  - search for an arbitrary (non-trivial) ken using the SAT-solver
  - perform a graph abstraction by collapsing the clique into a single node.
  - These steps are repeated till the graph collapses to a single node.

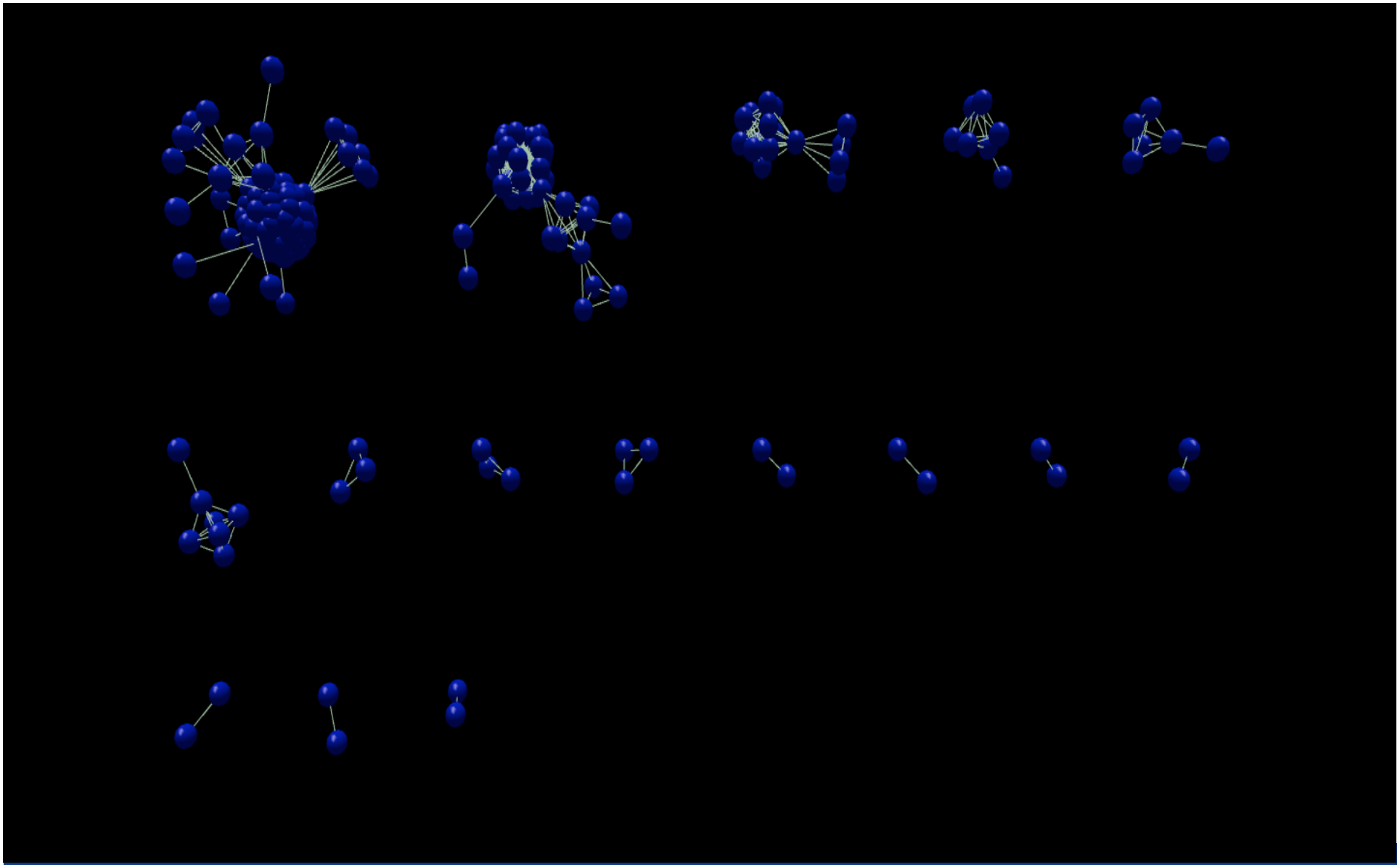
Sample Consistent Correlation Graph



Sample Consistent Correlation Graph



Abstract Ken Graph



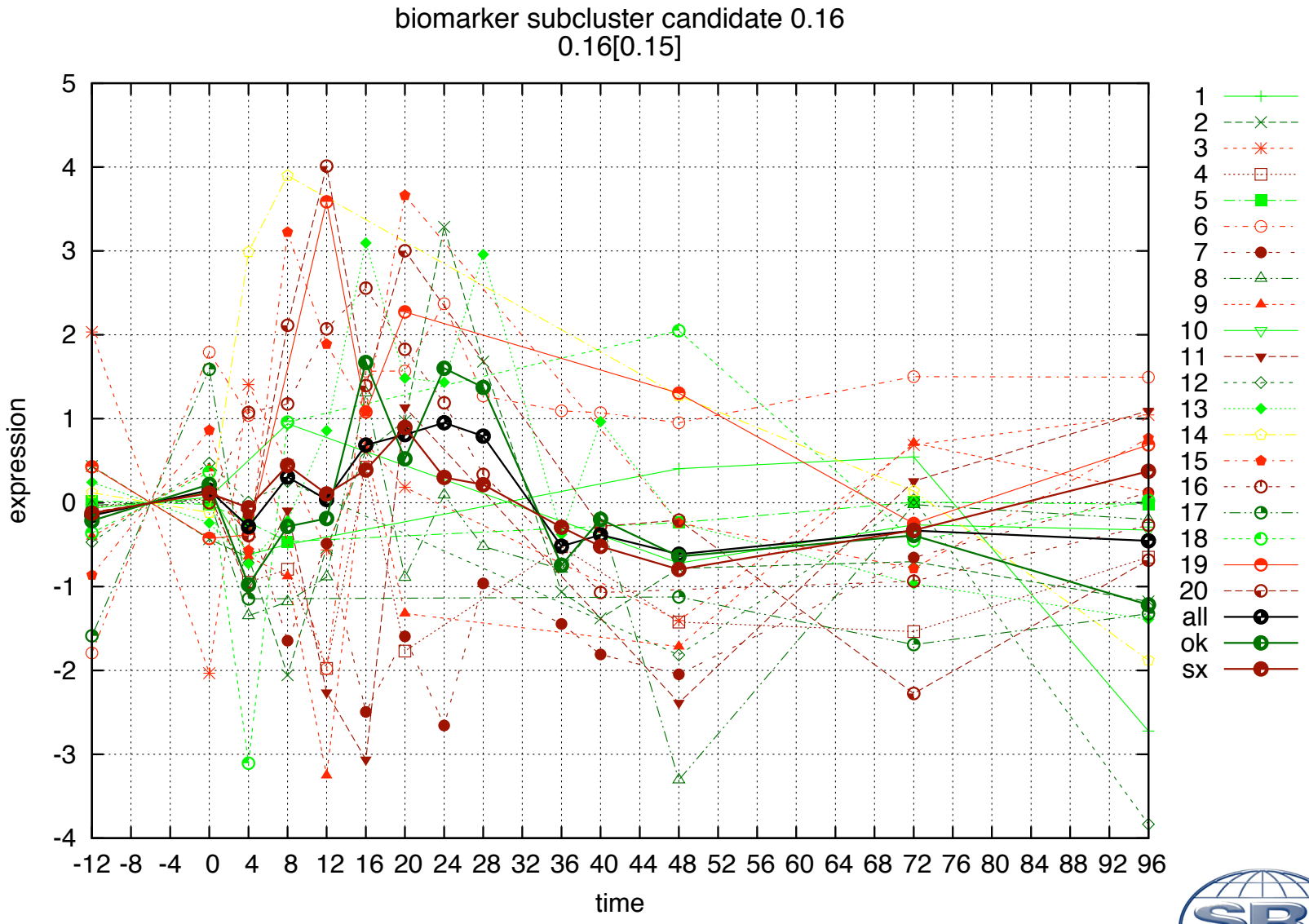
Gene Cluster Predictors

- For each (sub)cluster of genes we compute the cluster time series as the average of the individual time series of its constituents
- Since averaging reduces the noise level is reduced, the resulting time series is rescaled by the noise estimate for the (sub)cluster.
- Temporal abstraction of the (sub)cluster time series is performed as before.
- Gene cluster predictors use (sub)clusters of relevant genes instead of single relevant genes. We again evaluate and rank all gene cluster predictors.
- As before, a composite predictor is obtained by a voting-based composition of the top n gene cluster predictors.
- To improve independence, we require that (sub)clusters of component predictors are non-overlapping. A stronger notion of non-correlation at the cluster level still needs to be developed.

This is ongoing work …

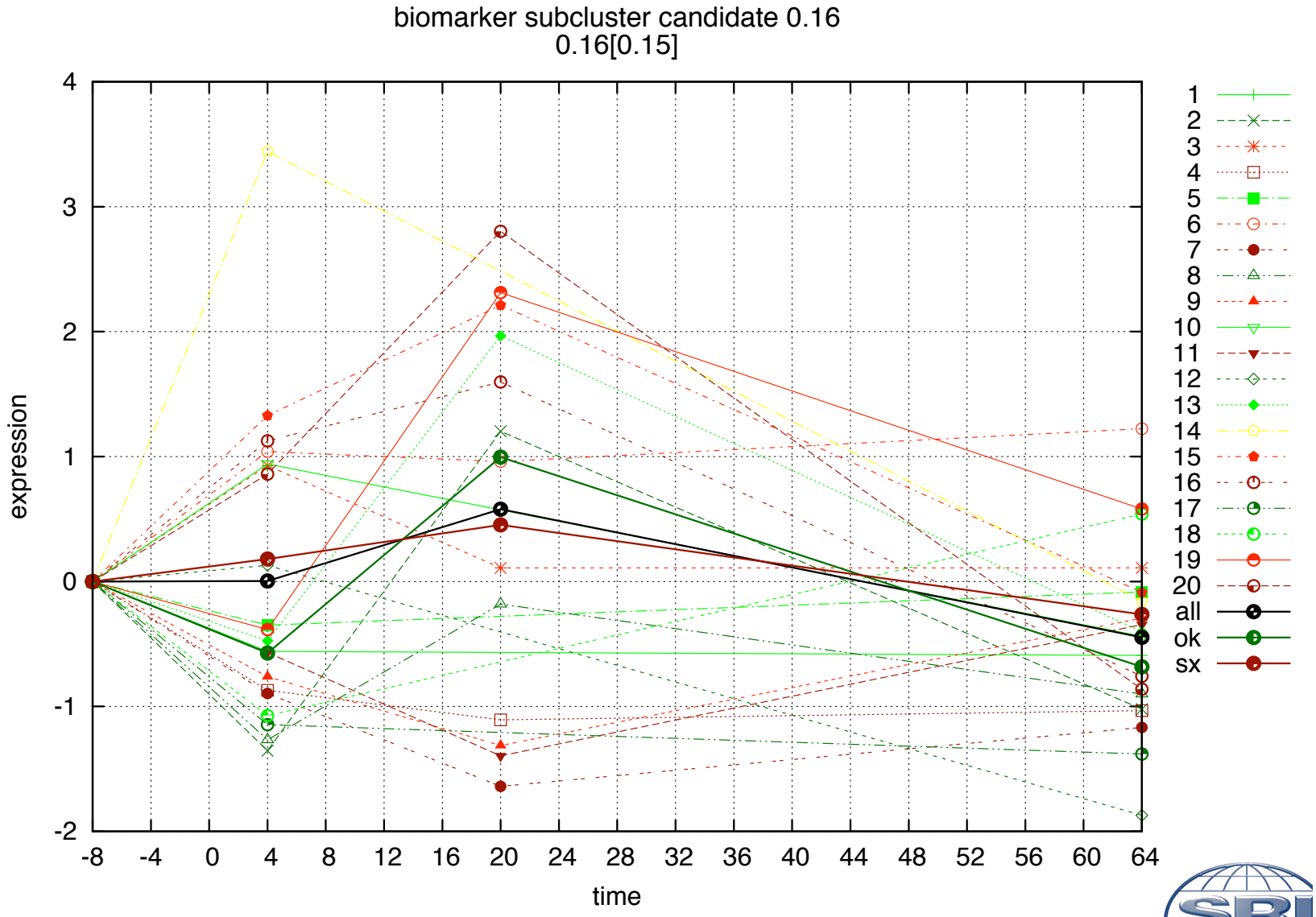
Sample Gene Subcluster

CNPY2
EIF3M
CEBPG
CETN2
PTGES3
DBF4
DSTN
PARK7
CSTA
SEC11A
HIBCH
C6orf66
METTL5
CHMP4A
DNAJC15
SLC25A5
H2AFZ
ANXA2
HMGN1
ILF2
NDUFB8
NEDD8
CLEC4A
POP5
POMP
MRPS17
UFC1
PDHB
ATP5J
PPIB
GLT8D1
PSMD4
PTGER2
RFC4
RPL36AL
RPS6
MRPL9
SNRPD2
TPD52
TTC1
C2orf47
TMEM14B
DPM1
GADD45GIP1
MINPP1
RBM39



Rank 1 – Early Predictor (Positive)

CNPY2
EIF3M
CEBPG
CETN2
PTGES3
DBF4
DSTN
PARK7
CSTA
SEC11A
HIBCH
C6orf66
METTL5
CHMP4A
DNAJC15
SLC25A5
H2AFZ
ANXA2
HMGN1
ILF2
NDUFB8
NEDD8
CLEC4A
POP5
POMP
MRPS17
UFC1
PDHB
ATP5J
PPIB
GLT8D1
PSMD4
PTGER2
RFC4
RPL36AL
RPS6
MRPL9
SNRPD2
TPD52
TTC1
C2orf47
TMEM14B
DPM1
GADD45GIP1
MINPP1
RBM39



Sample Gene from Subcluster