Incorporation of reactivity data in RNA secondary structure determination using Pareto optimization

Cédric Saule

Faculty of Technology
Bielefeld University

saule_cedric@yahoo.fr

February 13, 2017

Decision making under independent objectives

Love over gold? Mind over matter?

How to combine multiple objectives?

Two approaches to work with multiple objectives

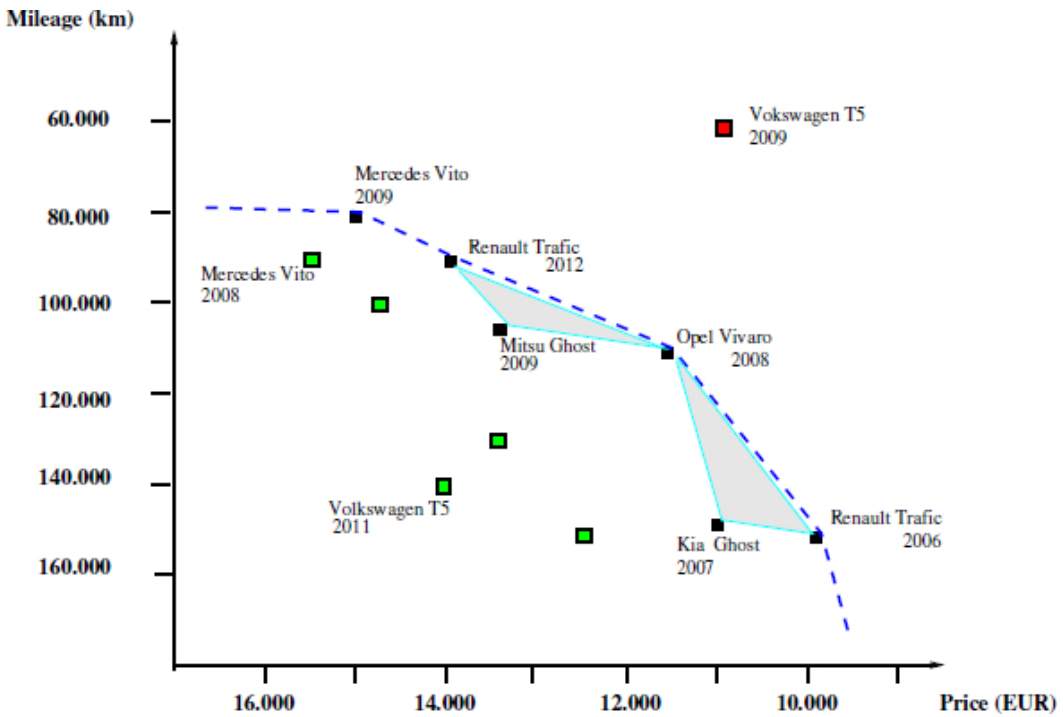
- Amalgamation of multiple objectives into a single one
- Pareto optimization (which keeps the objectives separate)

For discussion: Focus on (multiple = TWO) objectives

Vilfredo Pareto, Italian economist, 1848-1923

Buying a used car

Objectives: mileage, cost (diagram), ..., age, outfit, colour ...



Uninteresting offers: dominated data points

Interesting offers: Pareto optimal data points

Sequence comparison

Various objectives:

- global similarity (max)
- number of matches (max)
- number of gaps (min)
- number of exons/introns (prior knowledge)
- longest perfect match (max)
- number of 'jumps' (min) in jumping alignments

RNA structure

- MFE versus MEA score (min/max)
- Sequence similarity versus structure conservation (Sankoff problem)
- Sequence similarity versus covariance (RNAalifold)
- Chemical probing data used with MFE or MEA folding

Note: Multiple objectives for the same search space may be correlated, anti-correlated, independent, ... Handled by

RNA structure

- MFE versus MEA score (min/max)
- Sequence similarity versus structure conservation (Sankoff problem)
- Sequence similarity versus covariance (RNAalifold)
- Chemical probing data used with MFE or MEA folding

Note: Multiple objectives for the same search space may be correlated, anti-correlated, independent, ... Handled by pseudo-scores, bonuses, penalties, scaling parameters, ...

Pareto domination

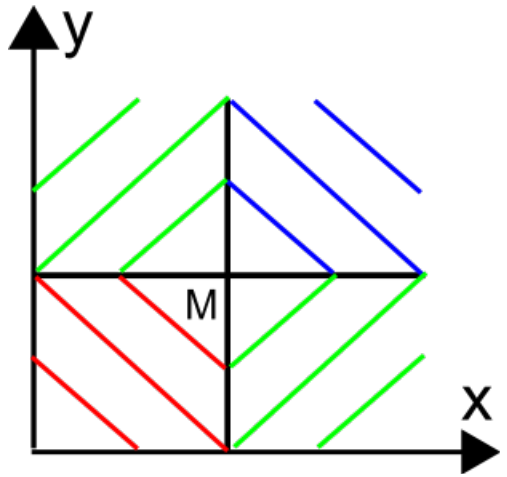
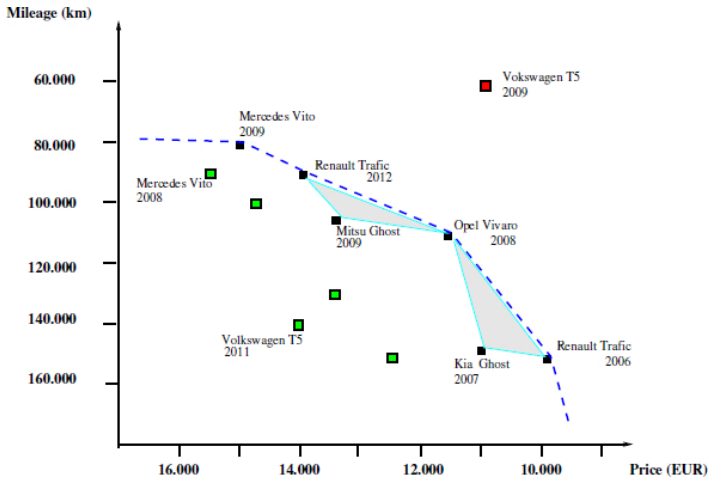


Figure : The point M defines dominant, dominated or co-dominant points areas.

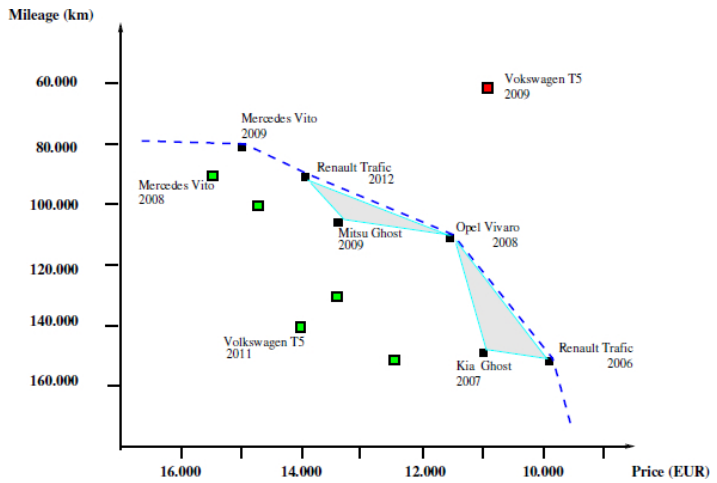
- In red, dominated area.
- In blue, dominating area.
- In green, co-dominating areas.

Pareto front properties



- Total order on each dimension.
- Pareto product is defined over a partial order.
- Sorted points: Increasing order on the first dimension, decreasing on the second.

Pareto front properties



- Maximum on each dimension are part of the front.
- Optimal solutions of criterias linear combinations are on the convex hull.
- We call other solutions **ghosts solutions**.

Past results

In dynamic programming, there is no asymptotic or constant factor overhead of Pareto optimization, compared to computing a similar amount of information with classical means.
With Bellman's GAP, Pareto optimization comes for free!

Virtues of Pareto optimization:

- obtain all interesting trade-offs (but then, we must choose by other means)
- carefully evaluate competing scoring models
- a small Pareto front indicates a well-posed optimization problem
- explore behavior of weighted additive combinations
- easy to code thanks to Bellman's GAP!

Past results

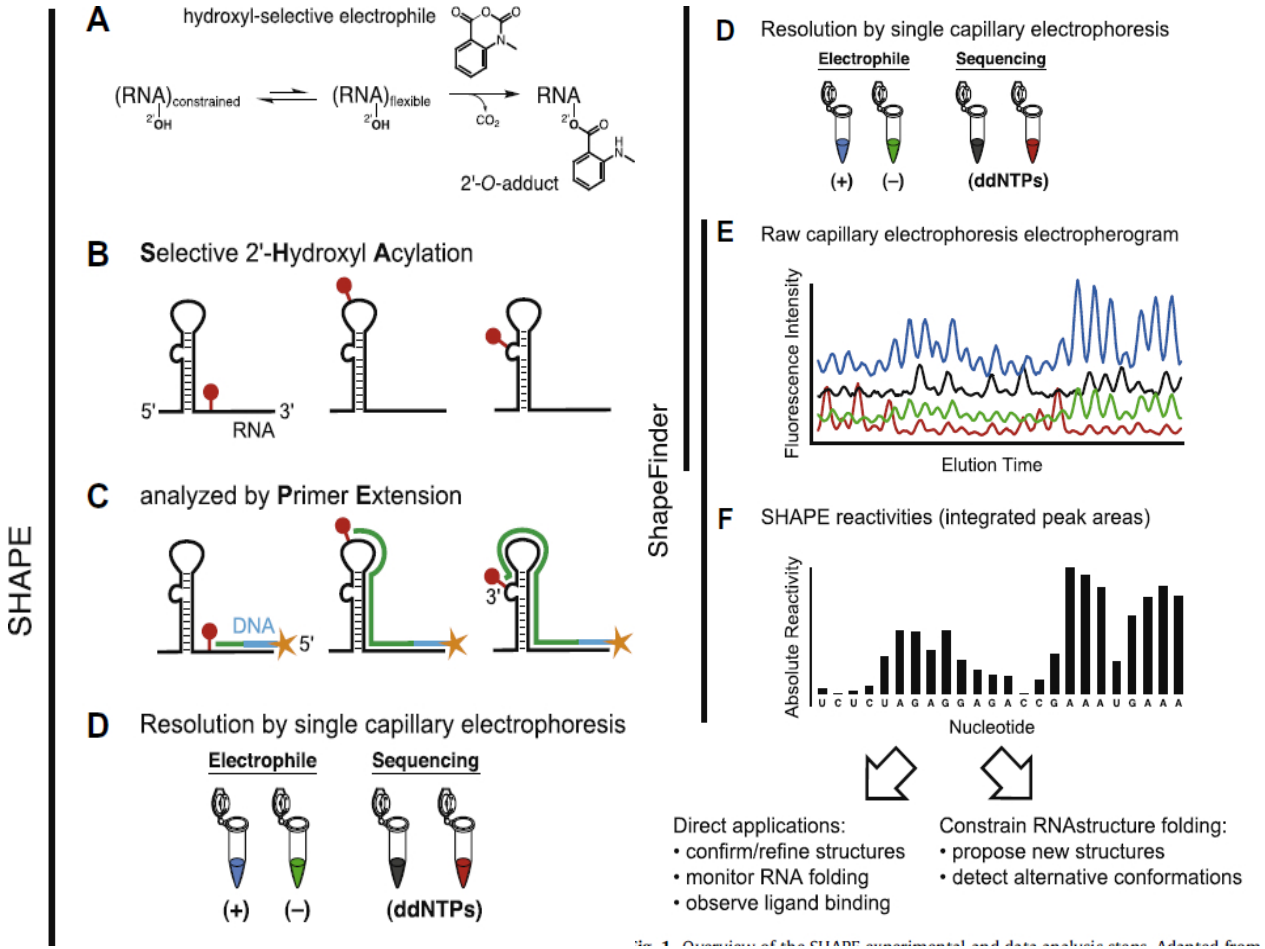
- Saule and Giegerich, Algorithms for molecular biology, 2015
- Gatter, Giegerich and Saule, 2016

Integration of probing data in RNA structure prediction

Experimental probing data are traditionally integrated as pseudo-scores in RNA folding
Rough idea: Pairing bases with high probing accessibility incurs a (positive) energy penalty

Original work: David Mathews with RNAstructure

Principles of RNA probing



Probing scores

$$score = ln(\frac{\#\text{reads with primers}}{\#\text{reads negative control}})$$

Various normalization processes, slope and intercept parameters changing over publications.
By definition, this score should not to be negative.

Probing techniques considered

Probe	Nucleotides targeted	Other
SHAPE	A, C, G, U	Unpaired bases
DMS	A, C	Unpaired bases
CMCT	G, U	Unpaired bases, near GU or stem begin

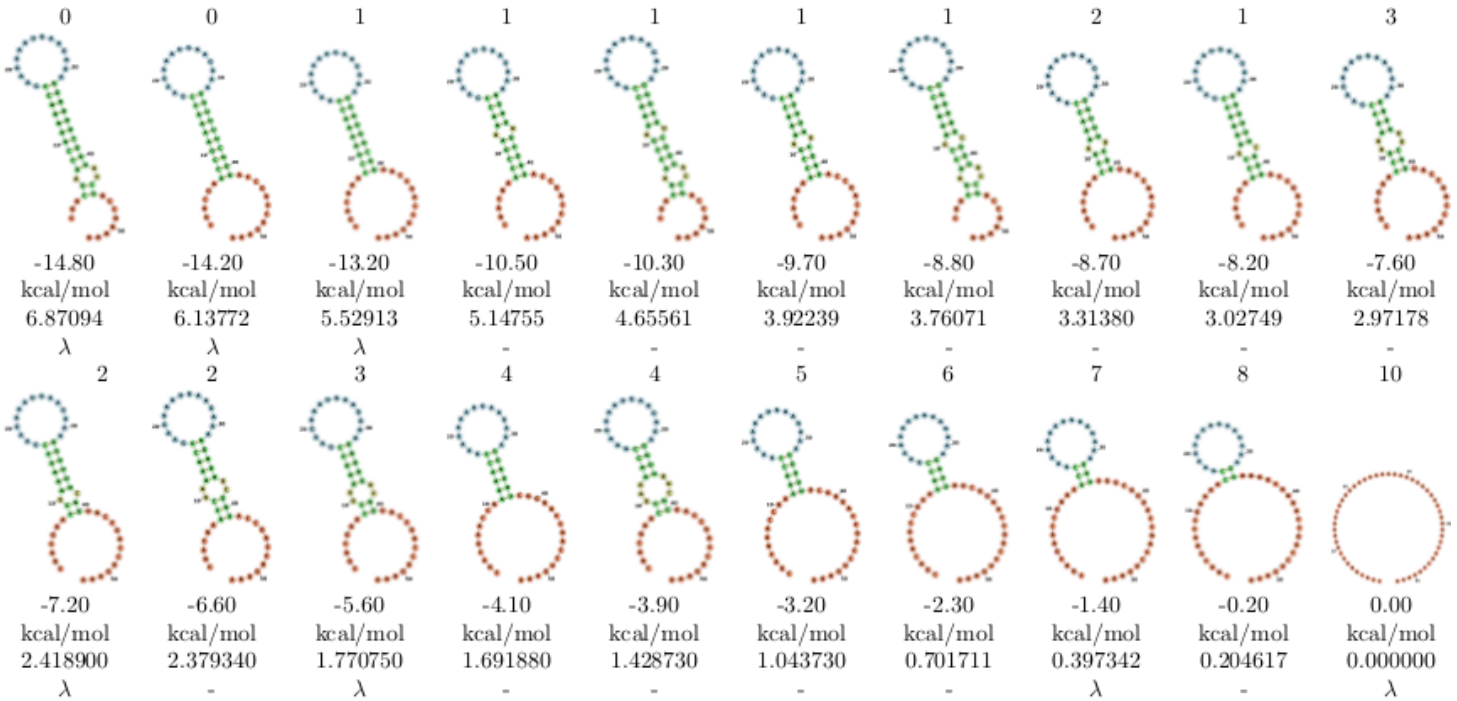
Experiments

Questions asked:

- Does chemical probing help at all with structure prediction?
- How about using probing data alone, possibly several methods in combination?
- Where is the best answer in the Pareto front?
- Are there good ghost solutions?

Observations drawn from ca. 175 sequences with probing data and reference structure. SHAPE (142), DMS (18), CMCT (15)
12 sequences have several types of probing data.
We use probing data in the same way as RNAstructure

MDLoop Pareto front



Probing helps(1)

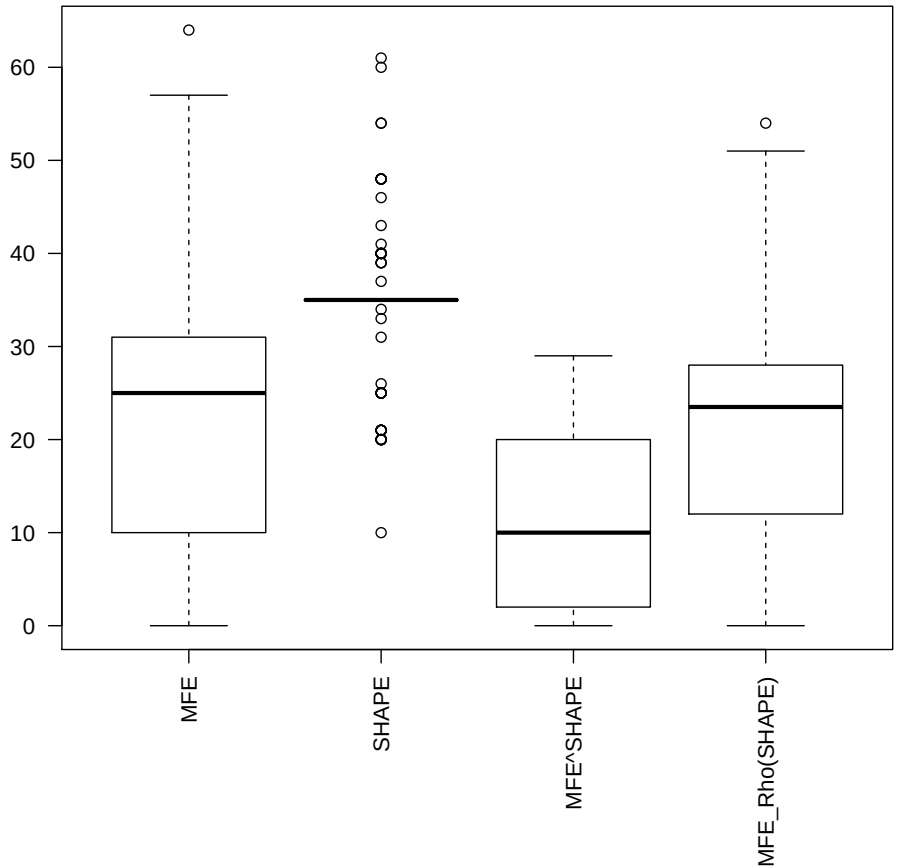
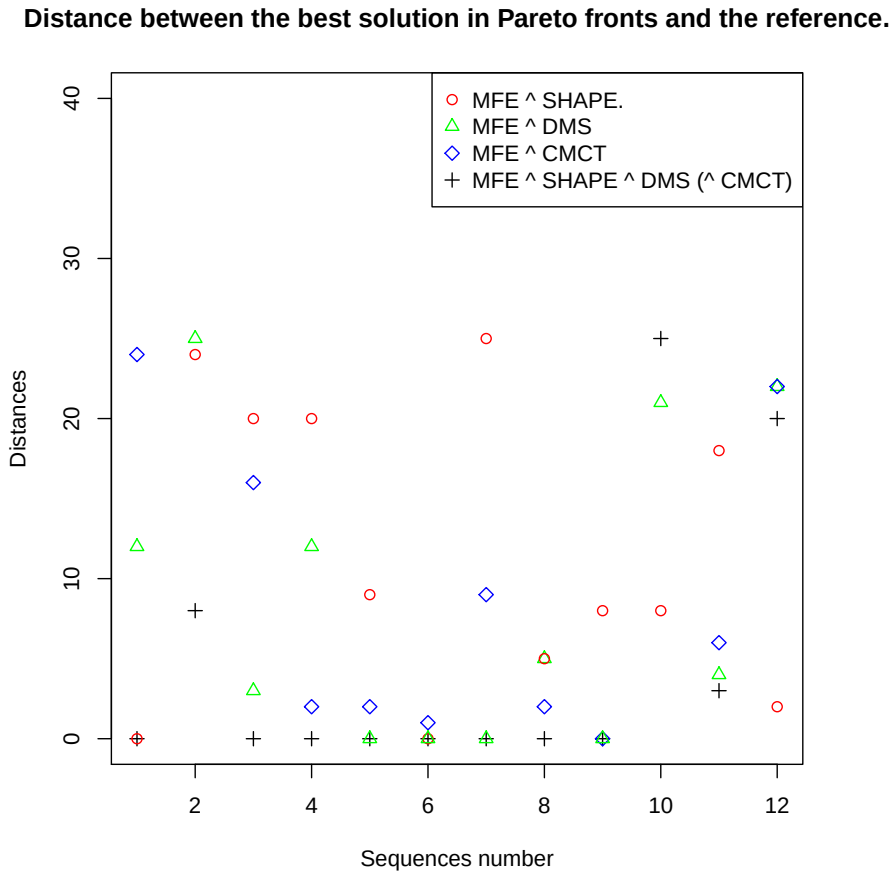
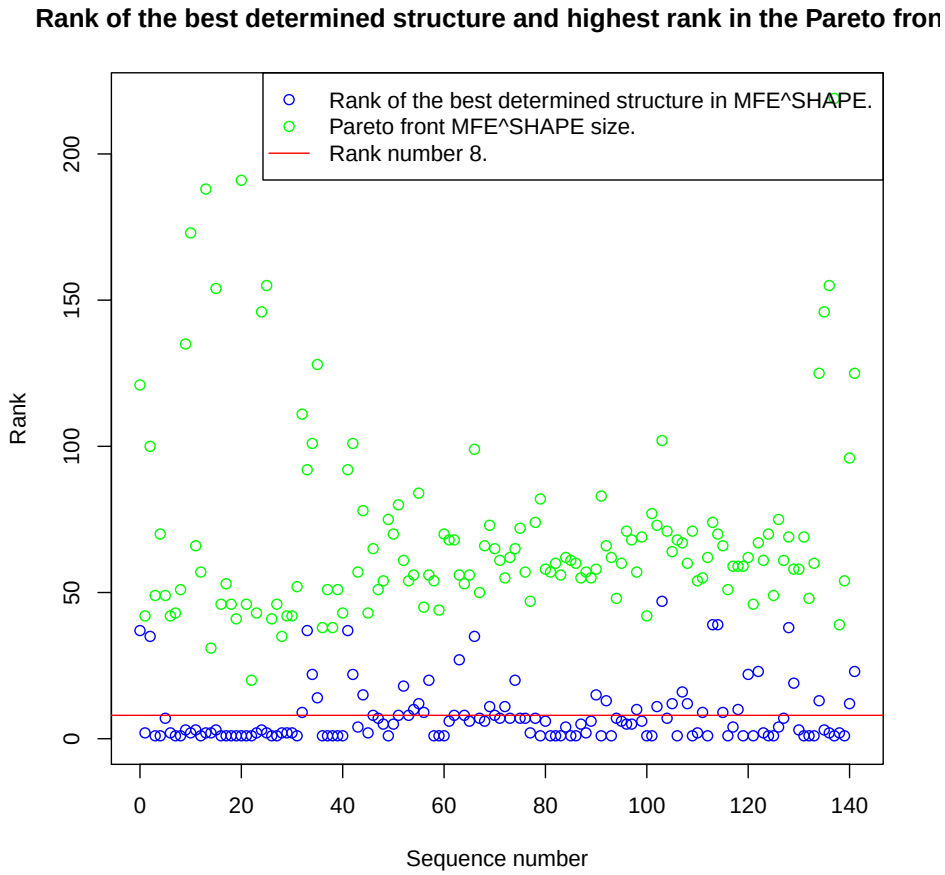


Figure : Boxplot of the distances between the best prediction and the reference.

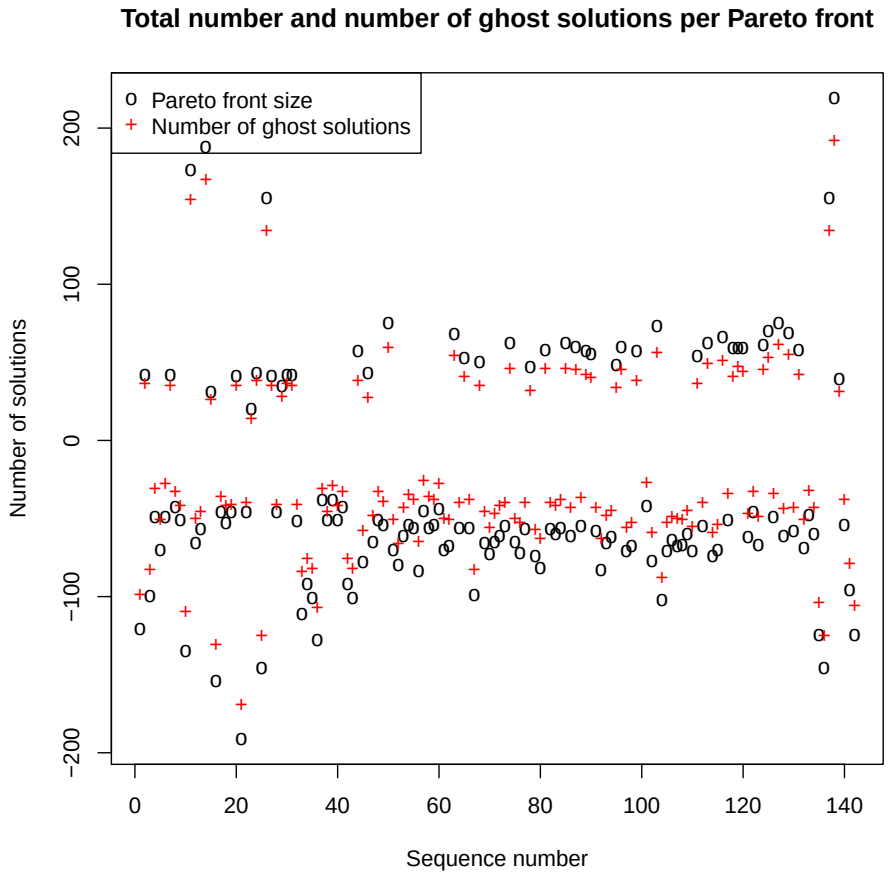
Probing helps(2)



Picking the best solution



Good ghosts everywhere



Lesson and advice

- Use Bellman's GAP!
- Re-think scoring scheme when dissecting $A + \lambda B$
- Avoid plain correlation or anti-correlation when choosing A and B
- Ongoing ...

The end

Have fun with Pareto optimization !!

The game has just begun.

Thanks for your attention.

Co-authors: R. Giegerich, S.Janssen and T. Gatter.
Thanks also to T. Schnattinger and H.A. Kestler of Ulm University.

Pareto front size indicates RNA family relationship

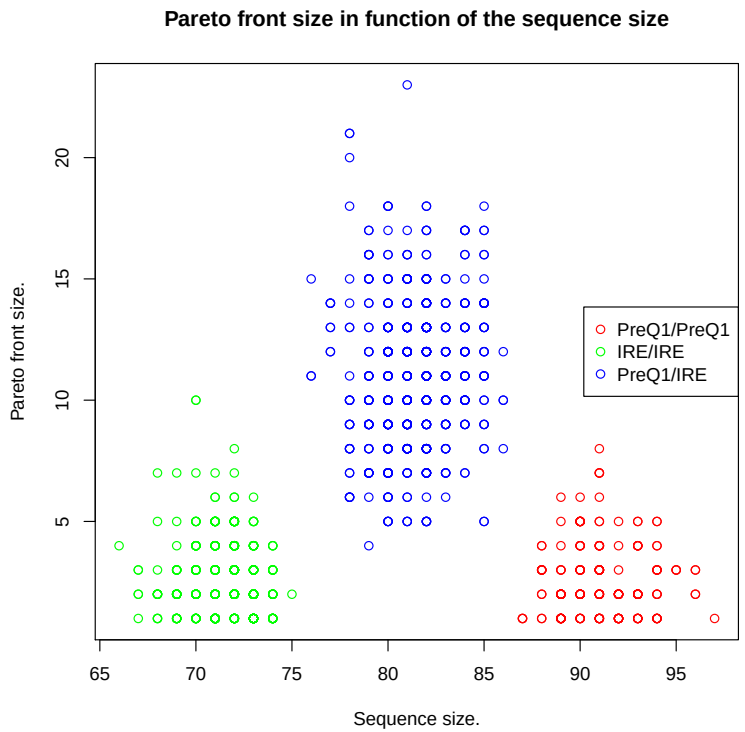


Figure : Empirical Pareto front size of $\mathcal{G}_{Sankoff}(BPP\hat{}SIM, x)$. All pairwise comparisons for RNAs from families IRE and PreQ1.