Probes Design

By Daniel Cazalis
For Dr. Giri Bio-informatics Course

Content

- Description of the problem
- Biological importance
- Articles studied
- Algorithms
- Random test results

A General Problem Description

1) Given a set of elements and a set of characteristics present in some of them find a minimum subset of characteristics that differentiate all elements.(MCPS)
2) Find a subset of size n that differentiate a maximum number of elements.(MDPS)

Informal Example

We have 8 cars, some of them are red, some are black, some have 4 doors other only 2, some are American some foreign cars,some are luxury cars, some front wheel drive, some 6 cylinders some 8 and so on.

Ideally 3 characteristics well chosen will specify all 8 cars…

Table for the Cars Example

Car Number	Color	4 Doors	American	Front Wheel	6 Cilender
1	Black	Yes	Yes	Yes	Yes
2	Black	Yes	No	Yes	Yes
3	Black	Yes	Yes	Yes	No
4	Black	Yes	No	Yes	No
5	Red	No	No	Yes	Yes
6	Red	No	No	No	Yes
7	Red	No	Yes	Yes	No
8	Red	No	No	Yes	No

Design Probe Problem

- *Minimum Cost Probe Set (MCPS)*
- *Instance: a set C of clones and a set P of probes*
- *Feasible solution: a subset S of P such that all clones are distinguish .*
- *Measure: /S /to be minimized.*

Design Probe Problem

- *Maximum Distinguish Probe Set (MDPS)*
- *Instance: a set C of clones and a set P of probes and an integer k.*
- *Feasible solution: a subset SP with /S /= k.*
- *Measure: Maximize some measure of distinguish.*

Measures of Distinguish

- Entropy: $\sum -|S_i|/|P| * Log_2(|S_i|/|P|)$

  Where S is a cluster and P are all the elements
- Number of distinguish Pairs
- Number of Clusters
- Size of the largest cluster

Biological importance of the problem

- The method of Hybridization of short synthetic oligonucleotide probes to clones DNA sequences has become a powerful tool in gene sequence analysis.
- Microorganism are of great importance for science, with little rDNA clone sequencing taxonomic units can be built with significant reduction of cost an effort.
- One of the biggest challenge of this technique is the selection of the oligonucleotide probe set.

Articles studied

- **Information theoretical probe selection for hybridization experiments**: Ralf Herwing, Armin O Schmitt, Matthias Steinfath…
- **Probe selection algorithms with application in the analysis of microbial communities**: James Boreman, Marek Chrobak, Gianluca Della Vedova, Andres Figueroa and Tao Jiang.

Algorithms Implemented

- Greedy
- Pivot Last
- Pivot Greedy
- Affinity
- Genetic Probes
- Random Pivot
- Simulated Annealing

Greedy

1. Select a probe that partition the set best
2. Select a probe the together with the former ones partition the set better (depending on the measure).
3. If number of probes selected less than k go to 2.

Pivot Last

1. Apply Greedy to the problem.
2. (Pivot). For every selected probe and for every not selected probe interchange and measure.
3. If measure improves change selection and go to 2.

Pivot Greedy

1. Select a probe that partition the set best
2. Select a probe that together with the former ones partition the set better.
3. For every selected probe and for every non selected probe interchange and measure.
4. If improvement goto 3
5. If number of probes less than k go to 2

Affinity

1. Build a Matrix of the affinity for every pair of probes. From a 1,-1 matrix probesXclones.
2. Select the 2 probes with the best affinity.
3. Select a probe that together with the former ones generates a better sum over the affinity matrix until size=k.
4. Equivalent to maximize $x' \cdot Q \cdot x, \{0,1\}^n$

Random Pivot

1. Select a random set of probes of size=k
2. Do a random pivot evaluate.
3. If improvement change the set
4. if no improvement for more than k*(size of set of probes) exit else go to 2.

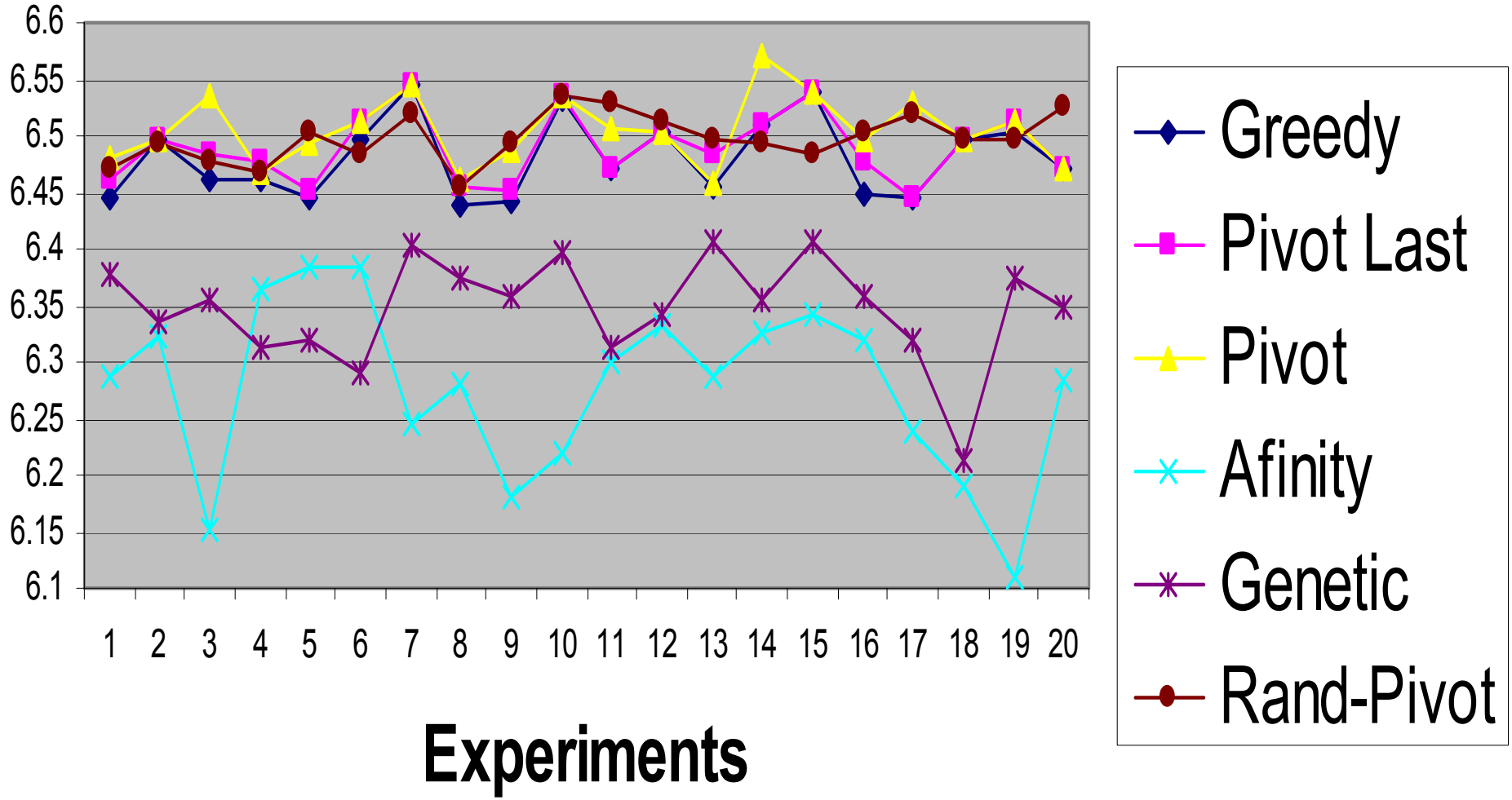
Simulated annealing

1. Select a random set of probes of size=k
2. Set initial temperature
3. Do a random pivot(set)=>newset.
4. Change to the new pivot with probabilty
5. Min(1,exp((m(newset)-m(set))/t)
6. Diminish temperature by a factor.

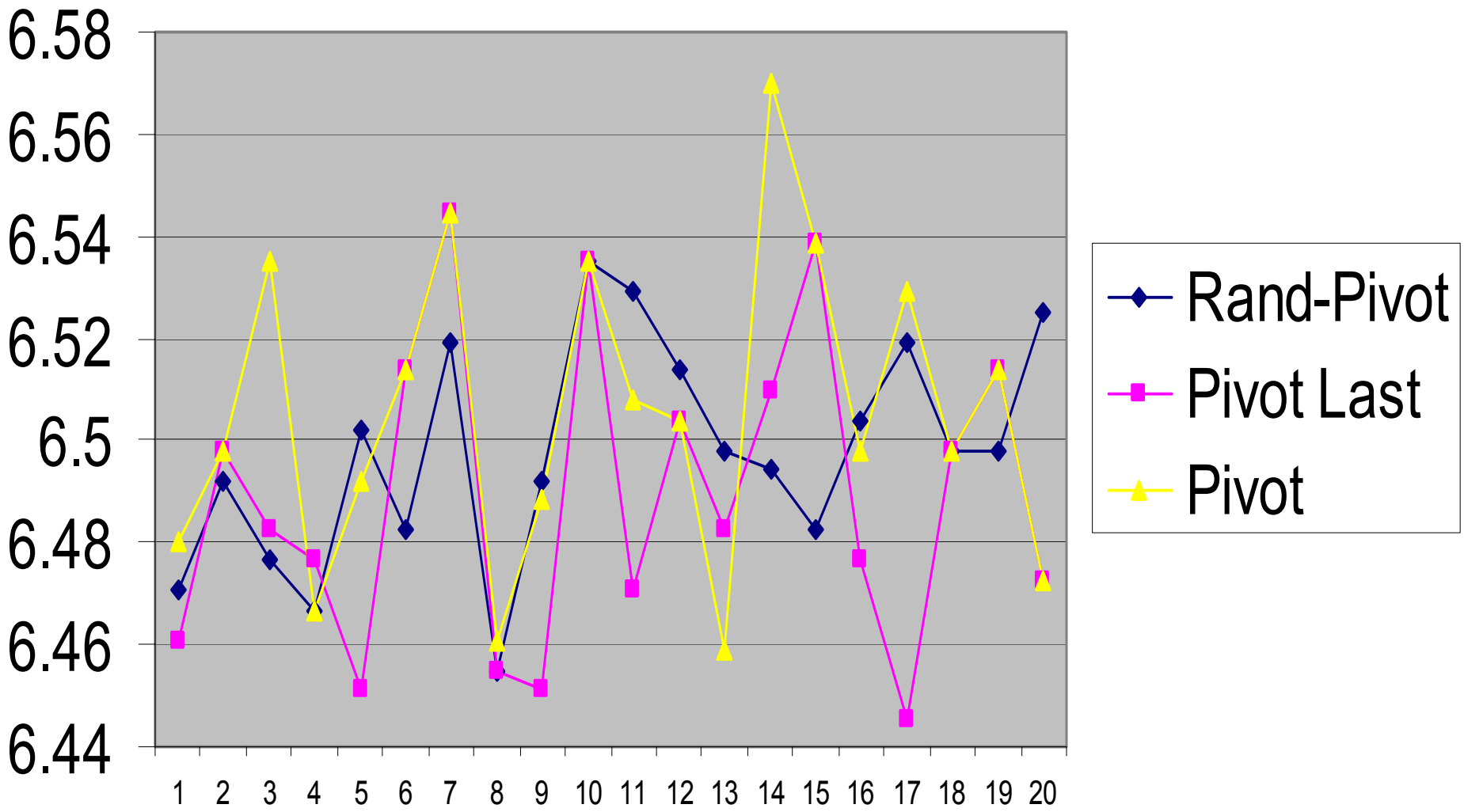
Entropy Data

	Greedy	Pivot Last	Pivot	Afinity	Genetic	Rand-Pivot
1	6.44516	6.46078	6.48024	6.28684	6.37676	6.47051
2	6.49793	6.49793	6.49793	6.32399	6.33578	6.49203
3	6.46078	6.48231	6.53508	6.15005	6.35524	6.47641
4	6.46078	6.47641	6.46668	6.36497	6.31219	6.46668
5	6.44516	6.45106	6.49203	6.38266	6.31809	6.50176
6	6.49793	6.51356	6.51356	6.38266	6.28891	6.48231
7	6.54481	6.54481	6.54481	6.2438	6.40211	6.51945
8	6.43926	6.45489	6.46078	6.27918	6.37293	6.45489
9	6.44309	6.45106	6.4882	6.1813	6.35731	6.49203
10	6.53125	6.53508	6.53508	6.21844	6.39828	6.53508
11	6.47051	6.47051	6.50766	6.29864	6.31426	6.52918
12	6.50383	6.50383	6.50383	6.33372	6.34168	6.51356
13	6.45489	6.48231	6.45872	6.28684	6.40801	6.49793
14	6.50973	6.50973	6.57016	6.32606	6.35524	6.4941
15	6.53891	6.53891	6.53891	6.34168	6.40723	6.48231
16	6.44899	6.47641	6.49793	6.32016	6.35907	6.50383
17	6.44516	6.44516	6.52918	6.23997	6.32016	6.51945
18	6.49793	6.49793	6.49793	6.19102	6.21255	6.49793
19	6.50383	6.51356	6.51356	6.11083	6.37293	6.49793
20	6.47258	6.47258	6.47258	6.28477	6.34934	6.52535

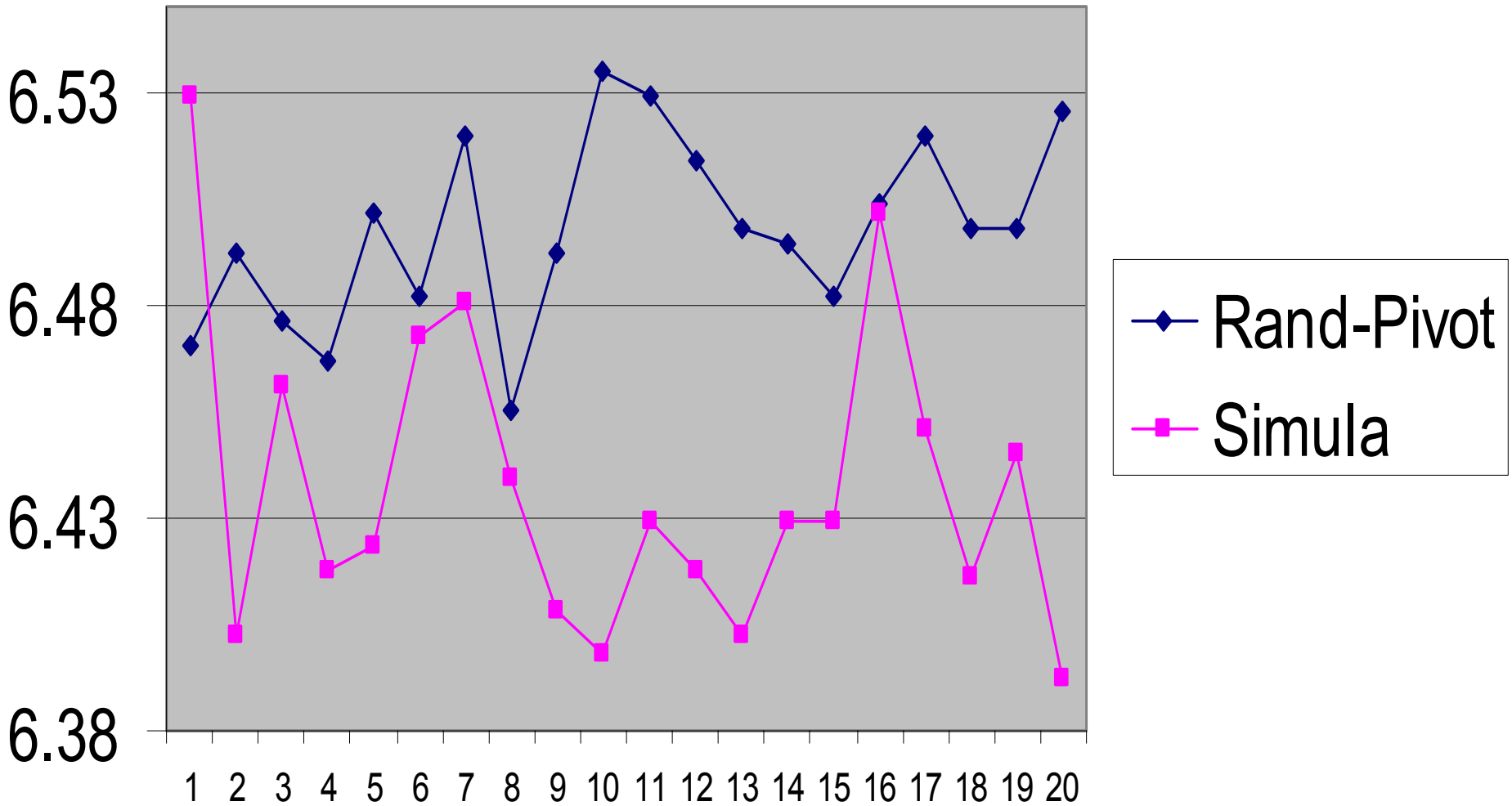
Entropy


6.6
6.55
6.5
6.45
6.4
6.35
6.3
6.25
6.2
6.15
6.1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Experiments
Greedy
Pivot Last
Pivot
Afinity
Genetic
Rand-Pivot

Entropy 3 best



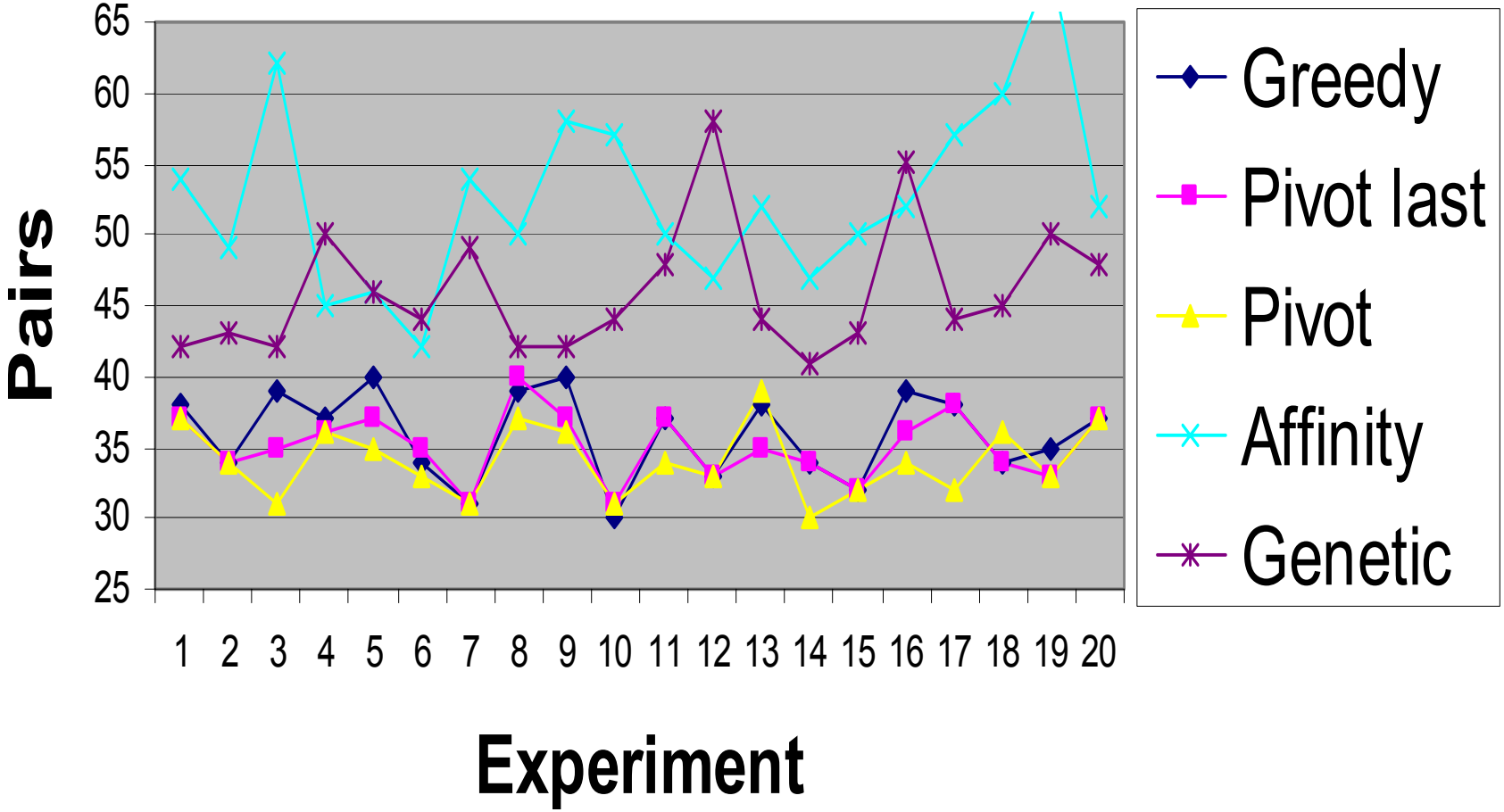
Entropy


6.53
6.48
6.43
6.38
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Rand-Pivot
Simula

Undistinguished Pairs Data

	Greedy	Pivot last	Pivot	Affinity	Genetic
1	38	37	37	54	42
2	34	34	34	49	43
3	39	35	31	62	42
4	37	36	36	45	50
5	40	37	35	46	46
6	34	35	33	42	44
7	31	31	31	54	49
8	39	40	37	50	42
9	40	37	36	58	42
10	30	31	31	57	44
11	37	37	34	50	48
12	33	33	33	47	58
13	38	35	39	52	44
14	34	34	30	47	41
15	32	32	32	50	43
16	39	36	34	52	55
17	38	38	32	57	44
18	34	34	36	60	45
19	35	33	33	69	50
20	37	37	37	52	48

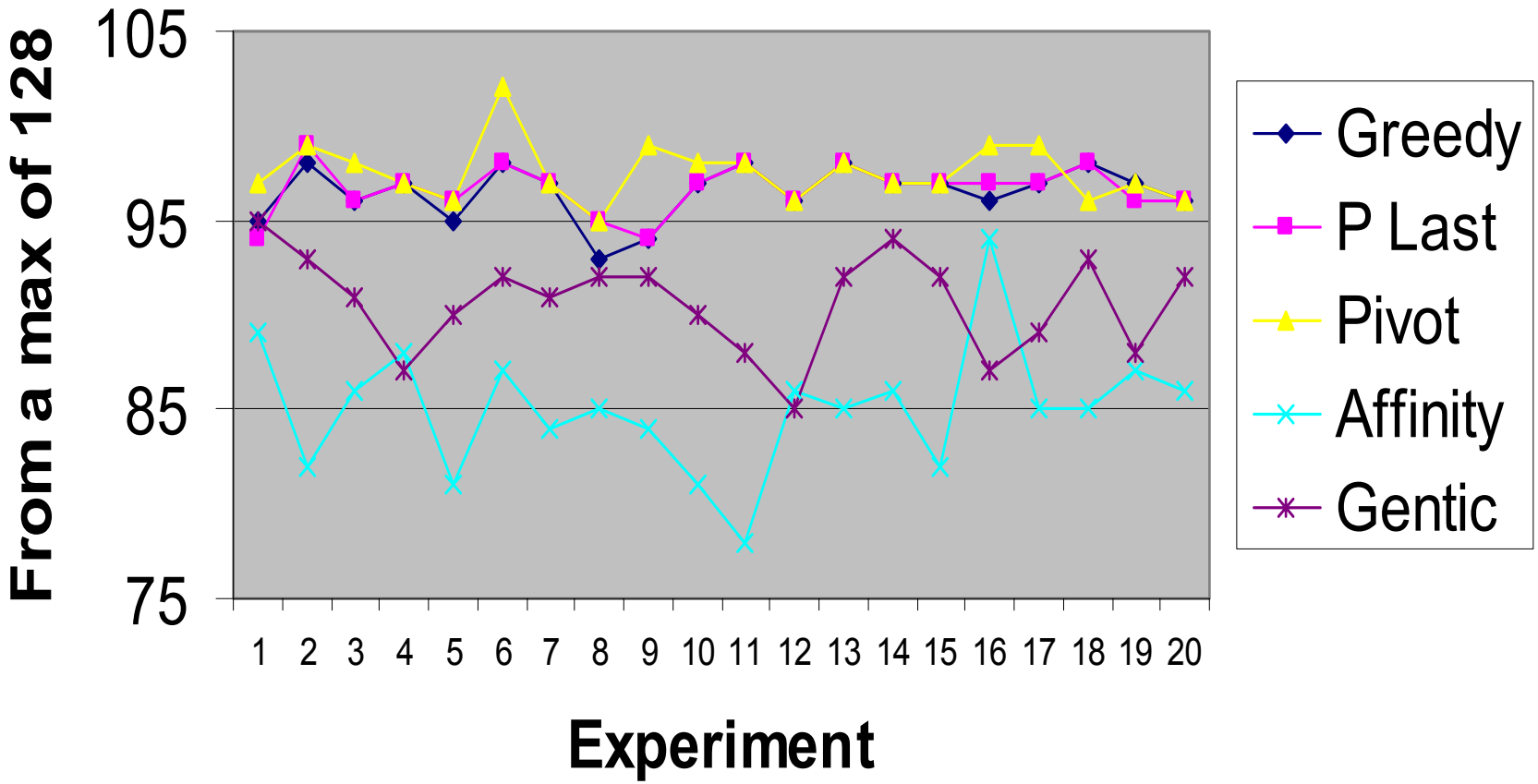
Undistiguished Pairs



Number of Clusters Data

	Greedy	P Last	Pivot	Affinity	Gentic
1	95	94	97	89	95
2	98	99	99	82	93
3	96	96	98	86	91
4	97	97	97	88	87
5	95	96	96	81	90
6	98	98	102	87	92
7	97	97	97	84	91
8	93	95	95	85	92
9	94	94	99	84	92
10	97	97	98	81	90
11	98	98	98	78	88
12	96	96	96	86	85
13	98	98	98	85	92
14	97	97	97	86	94
15	97	97	97	82	92
16	96	97	99	94	87
17	97	97	99	85	89
18	98	98	96	85	93
19	97	96	97	87	88
20	96	96	96	86	92

Number of Clusters



Real Data

- We are working on the development of some perl programs that will allow us to get and transform real data problems to our format.
- The selection of a working probe set from a random generated one is part of the work need to be done.
- We have run some of the algorithm for small real data problems but the result are still inconclusive