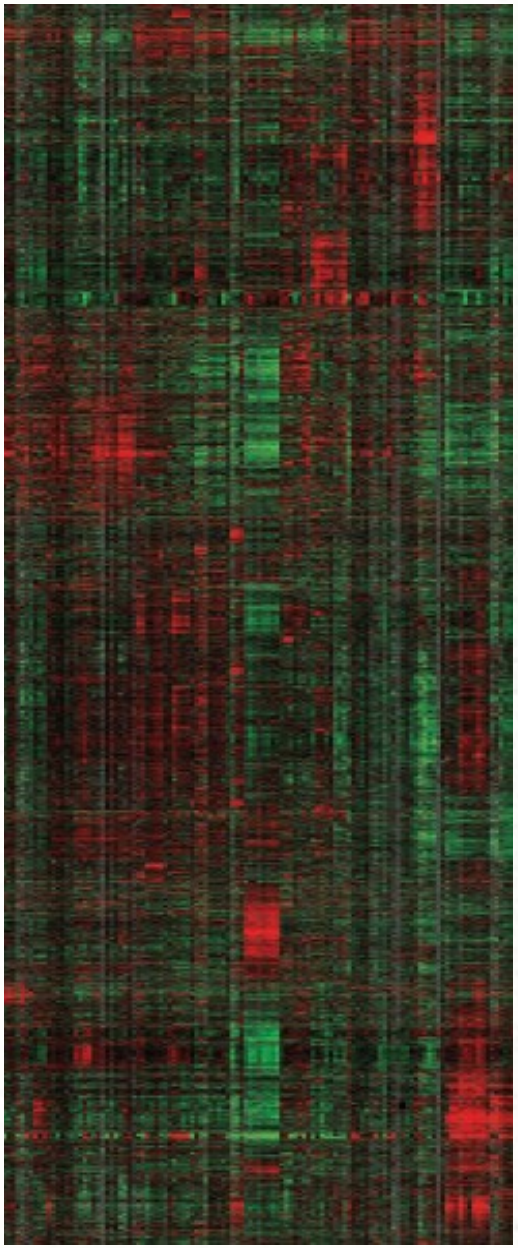


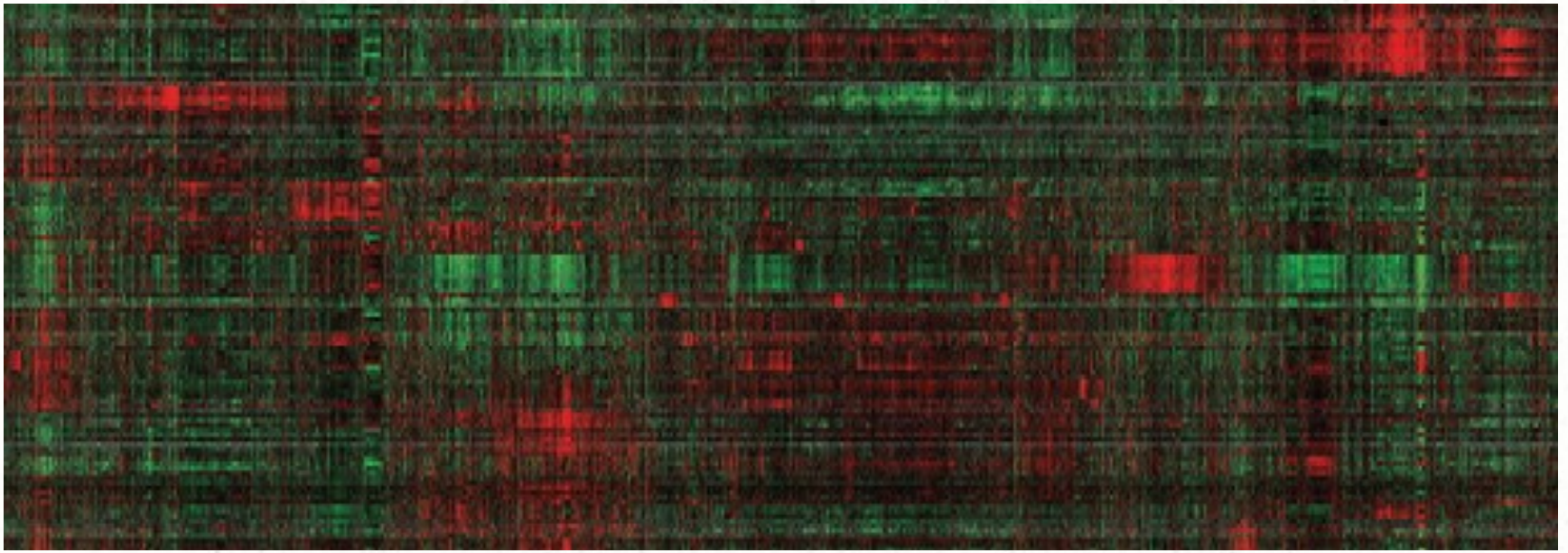
CUDA-Accelerated Data-Mining for Heteromeric Transcription Factor Complexes

Edward A. Salinas, Independent Researcher
Dr. Amitava Karmaker, Univ. WI, Stout,
Dr. Stephen S. Kwek, Medio Systems, Acknowledged &
Dr. Michael Allan, US Patent Office, Acknowledged

Agenda/Overview

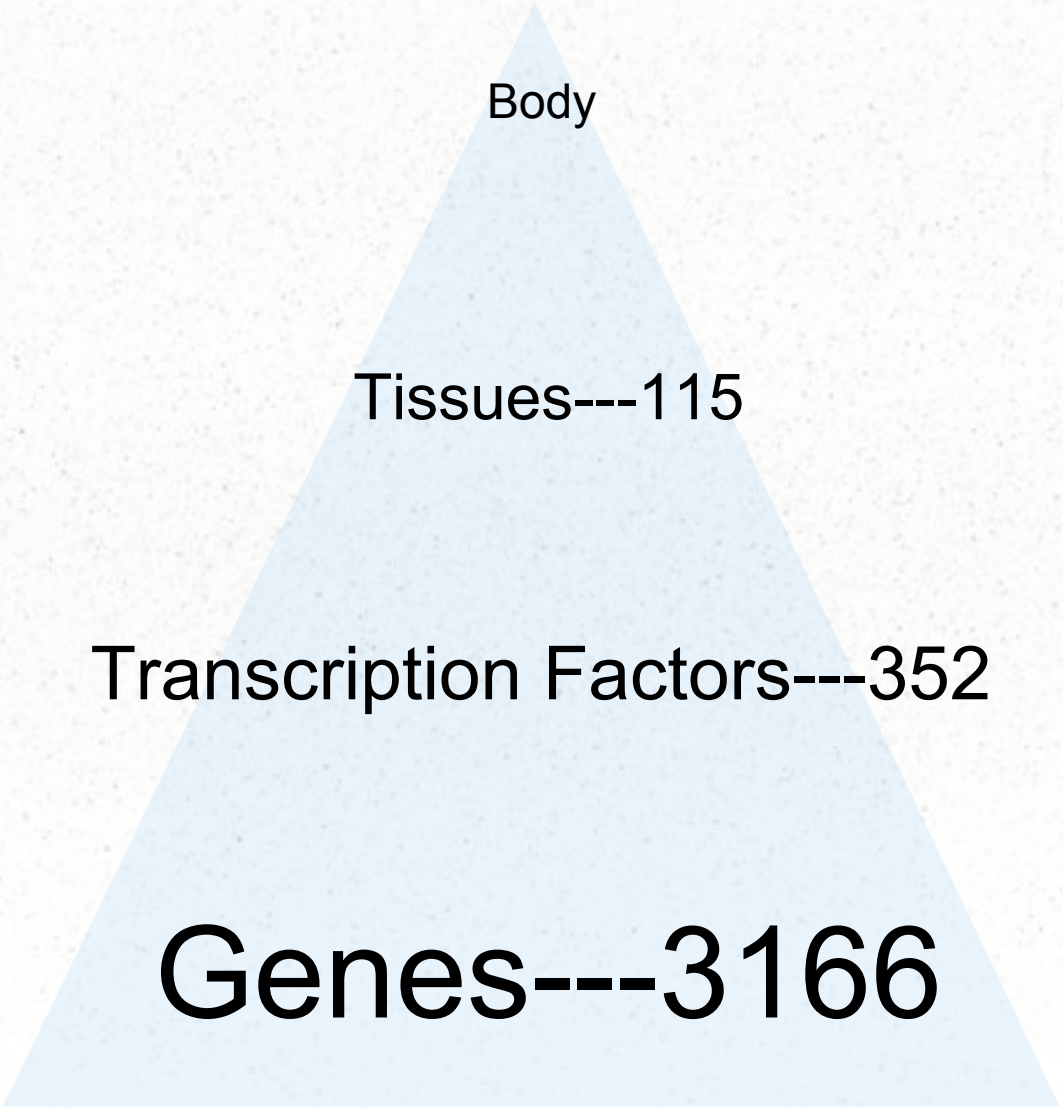
1. Project Overview
2. Technique overview
3. exploration/validation analysis with cFos/cJun (AP-1)
4. “k=4” CPU-based approach (execution-time!)
5. “k=4” GPU-based approach (CUDA)
6. Execution & SpeedUp Discussion
7. Conclusion/Discussion

TF Expression Profile Data Mining May Facilitate ID of TF Complexes



MEANS: Select data, compute hypothetical expression profiles & coefficients, analyze, rank, and find hypothetical polymers.

Our algorithm calls for expression analyses with Gene and TF data.



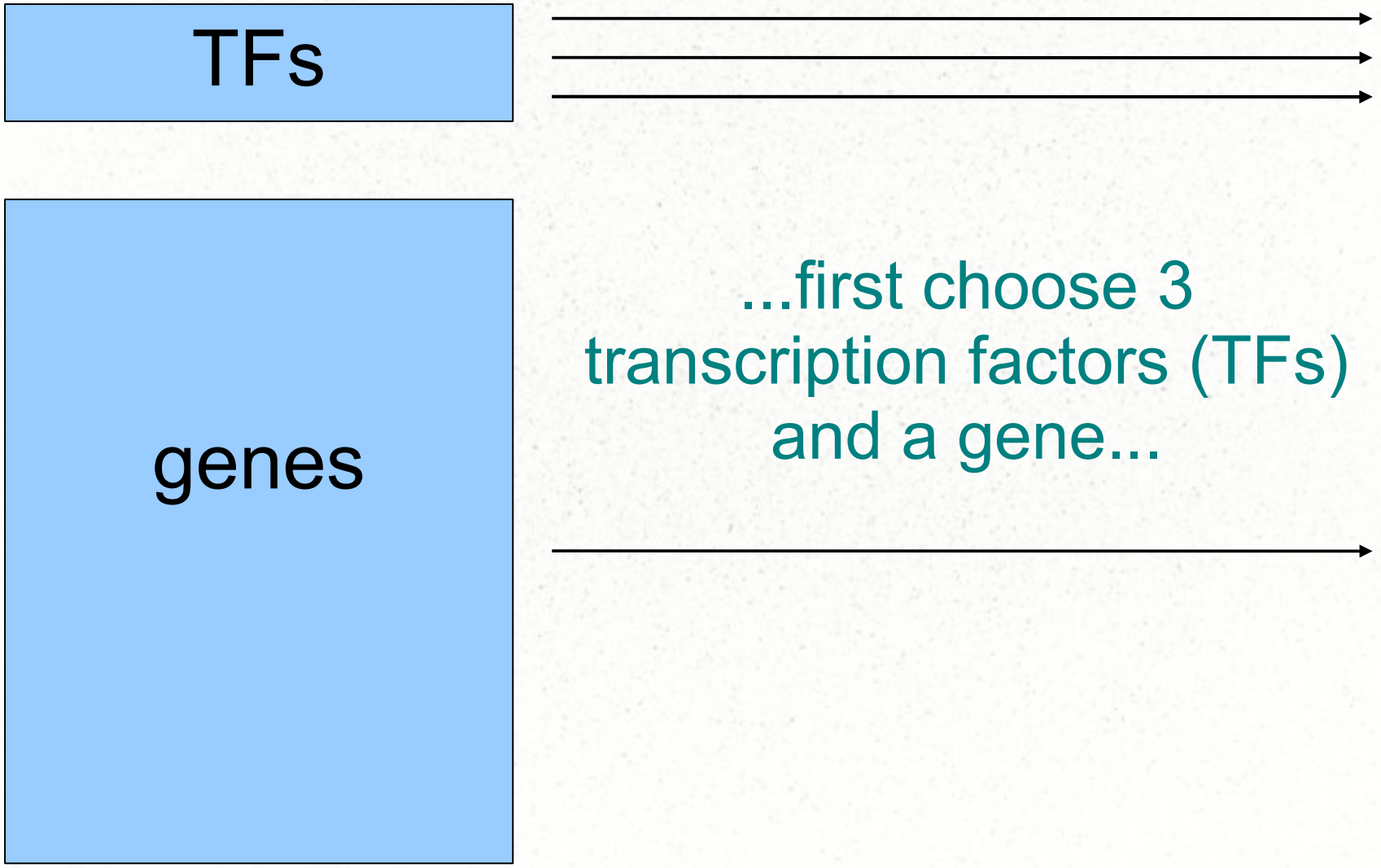
For a given gene (or Transcription Factor(TF)), the data consist of a row of expression values across the tissues. Correlation coefficients are computed between two rows (profiles) of values (a gene and a TF)

	HEART	BRAIN	LYMPH NODE	OVARY
HMGA1	**0.1**	**-0.3**	**0.7**	**0.9**
CFOS	**-0.2**	**.4**	**-.6**	**.2**

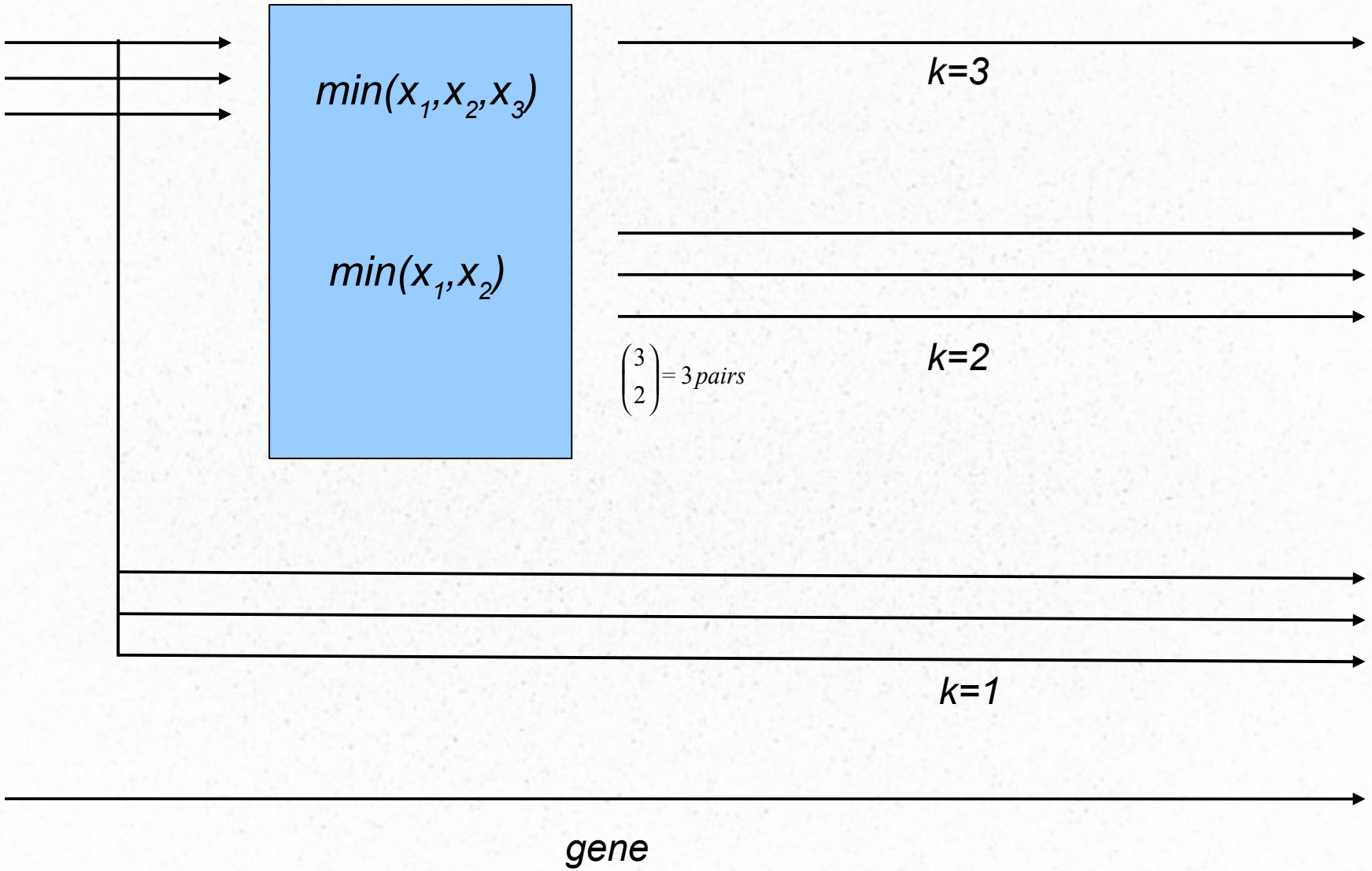
For a given set of TF profiles, a hypothetical TF composite profile is estimated by taking minima across tissues. Motivation: *limiting reactant*

	HEART	BRAIN	LYMPH NODE	OVARY
TF1	0.1	-0.3	0.7	-0.9
TF2	-0.2	0.4	-0.6	0.2
$TF(12)_H$	-0.2	-0.3	-0.6	-0.9

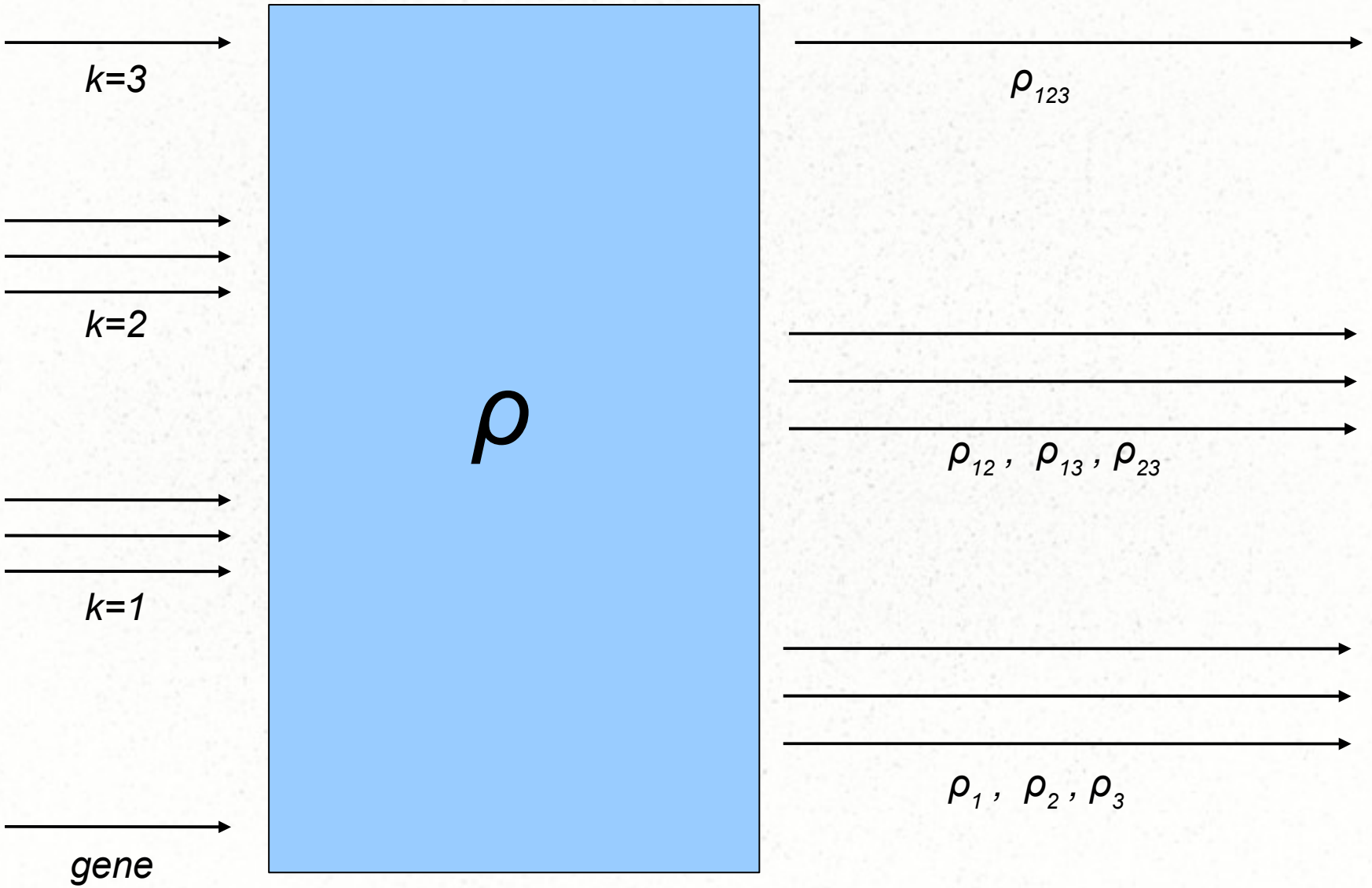
To carry out a k=3 analysis.....



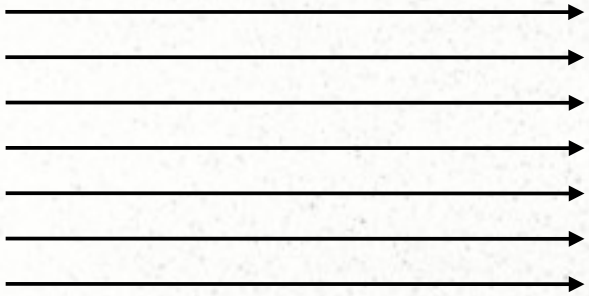
...second, compute hypothetical dimer/trimer profiles...



...third, compute coefficients between the gene profile and each of the 1+3+3=7 TF profiles....



...and finally compare the resulting coefficients to compute an absolute improvement score.



$$\min_{y \neq k_{\max}} \left(| \rho_{k_{\max}} - \rho_{y} | \right)$$

Repeat for all possible pairs of genes and combinations of selections of 3 TFs if you want to complete a full k=3 analysis!

We used a pair of TFs that are known to dimerize to explore the validity of the algorithm: CFOS & CJUN.

	g	tf1	tf2	c1	c2	cc	i
1	VARS2	FOSL1	JUN	0.43135	-0.318	0.0735	0.3578
2	RNU3IP2	FOSL1	JUN	0.50372	-0.228	0.1559	0.3479
3	ZFX	FOSL1	JUN	-0.4036	0.369	-0.0594	0.3442
4	AP2S1	FOSL1	JUN	0.46311	-0.212	0.1232	0.3355
5	LRP6	FOSL1	JUN	-0.3653	0.375	-0.0468	0.3185
6	MAP4K5	FOSL1	JUN	-0.3528	0.318	-0.0351	0.3177
7	LOC56902	FOSL1	JUN	0.41646	-0.225	0.0893	0.3141
8	EGR1	FOSL1	JUN	-0.0400	0.614	0.3041	0.3101
9	HMGA1	FOSL1	JUN	0.40596	-0.250	0.0549	0.3046
10	TAPBP	FOSL1	JUN	0.38650	-0.223	0.0802	0.3035

$$P(X \geq 2) = P = 8.4 * 10^{(-5)}$$

2 target genes (“true positives” out of 4 total) found in top 10 of 3166 genes

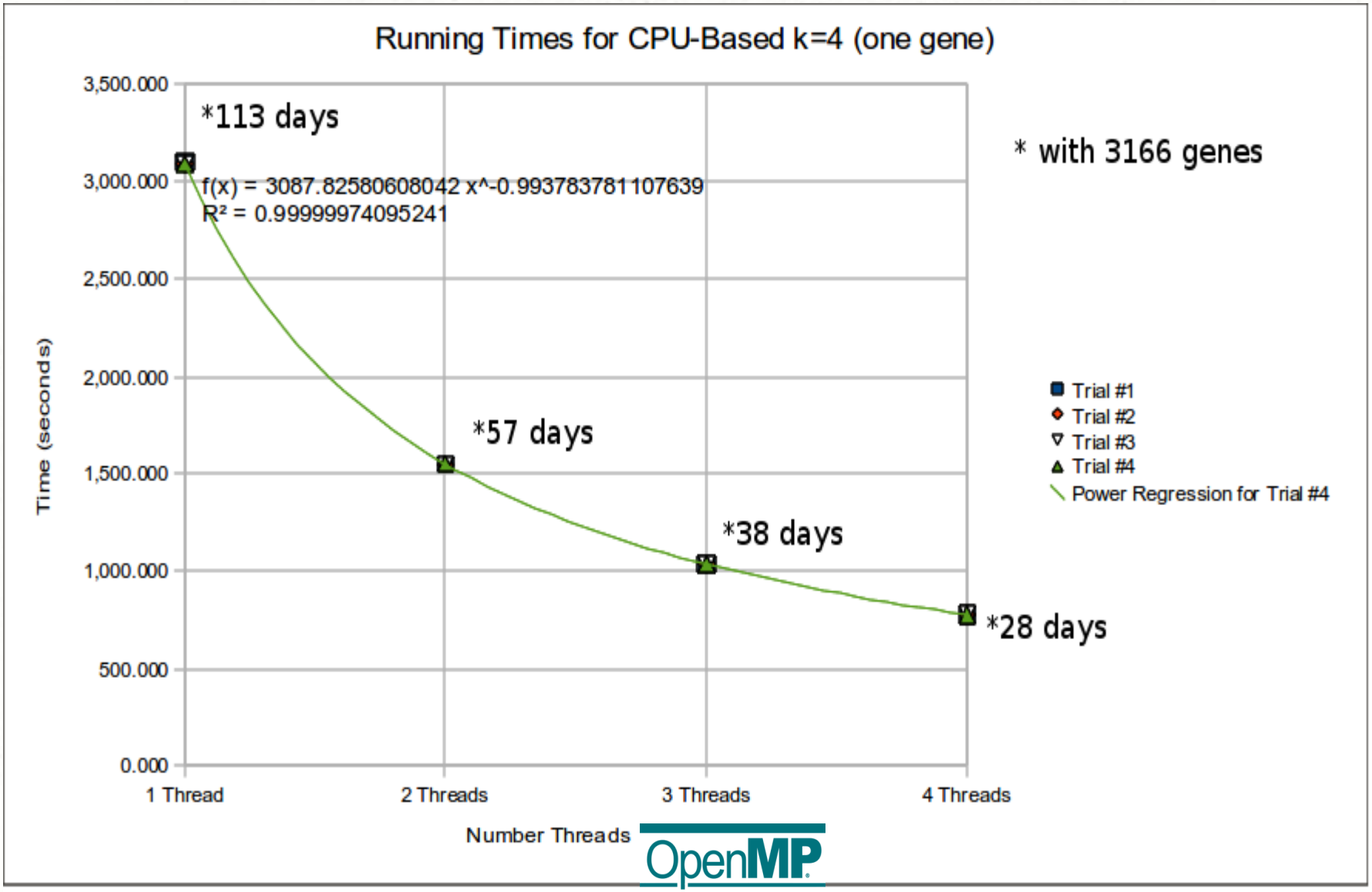
There's a *low* probability of a result at least as extreme happening by random chance!

Reject H_0 at $\alpha = 1\%$

Searching for higher-order polymers means “deeper data-mining”...

k	Number Coefficients to Compute
1	1,114,432
2	196,380,648
3	23,014,375,848
4	2,013,884,773,648
5	140,578,442,425,168

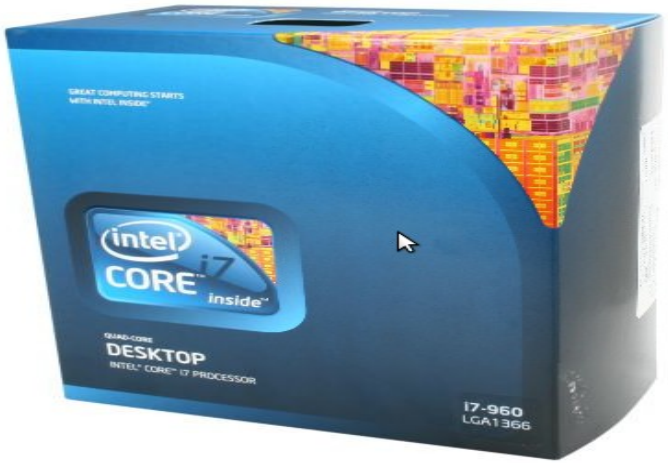
...a full k=4 analysis would take a long time.....



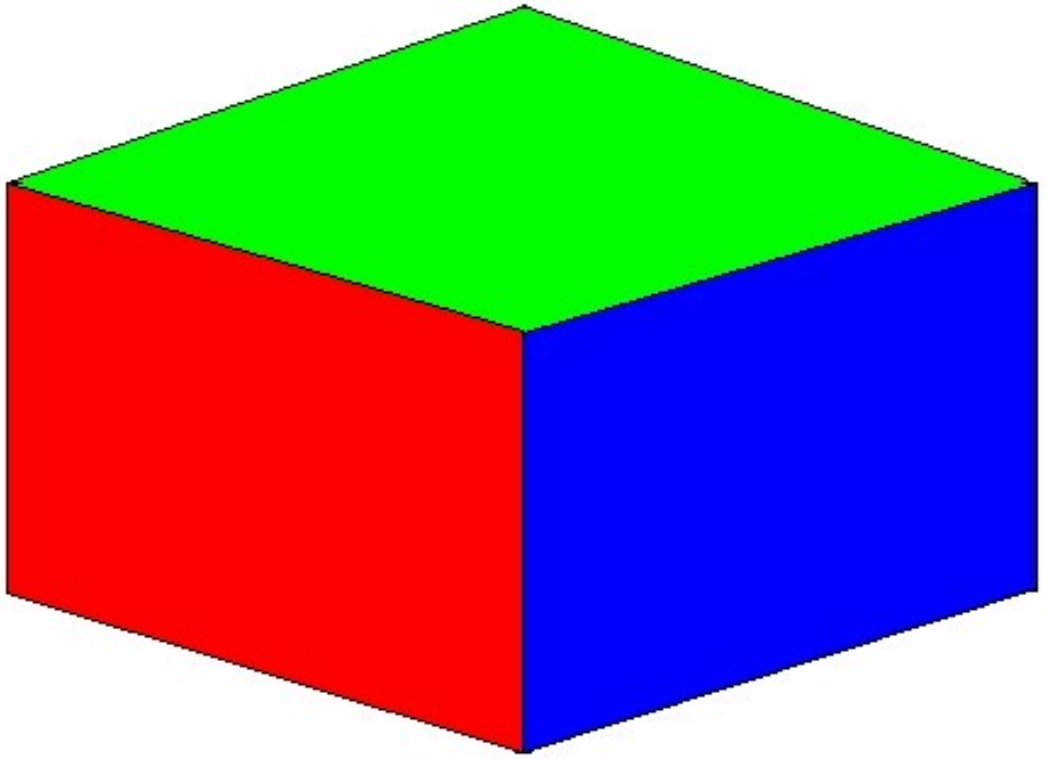
...so we decided to apply a heterogeneous set of parallel computing technologies to speed things up!

GTX 590

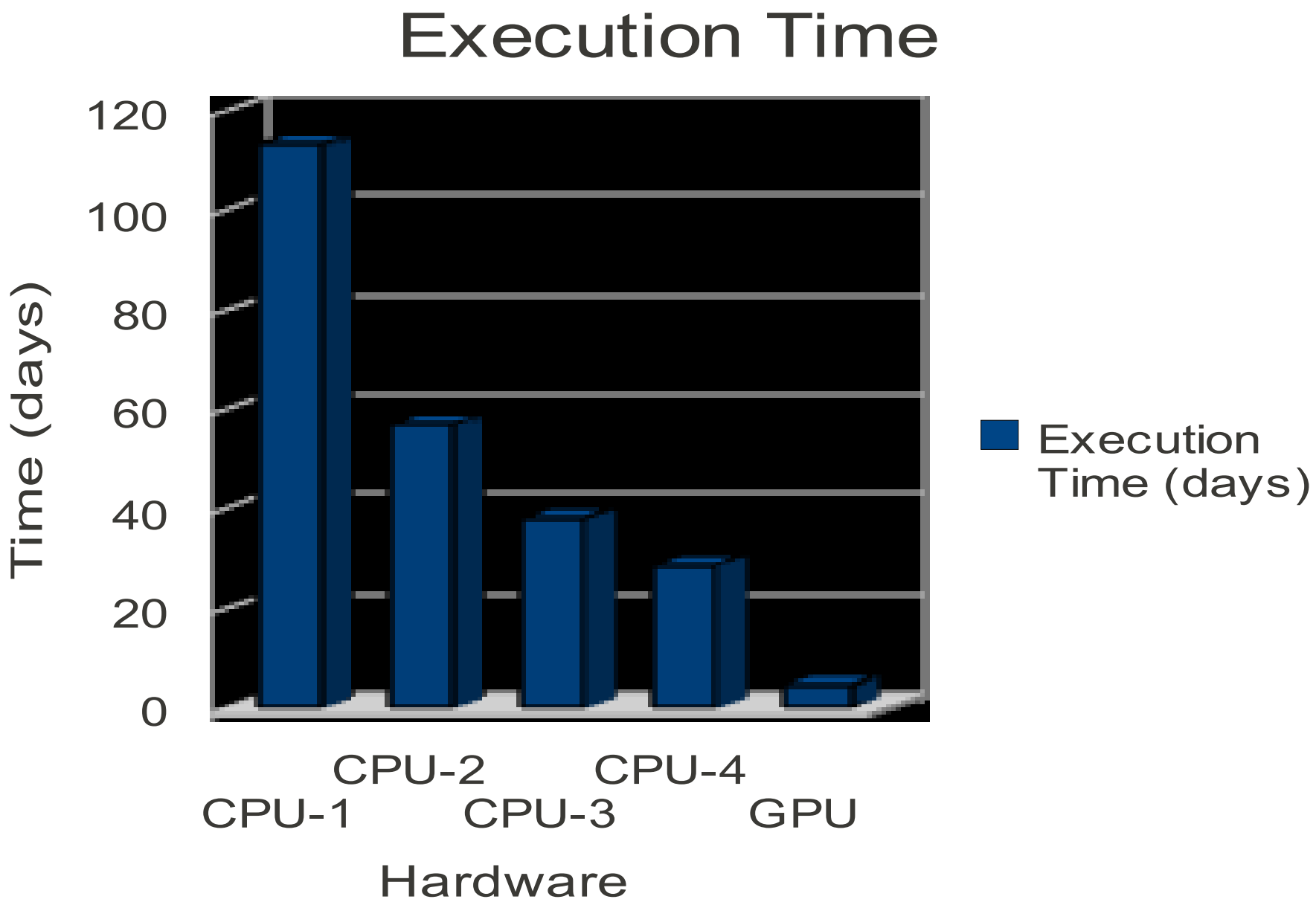
OpenMP



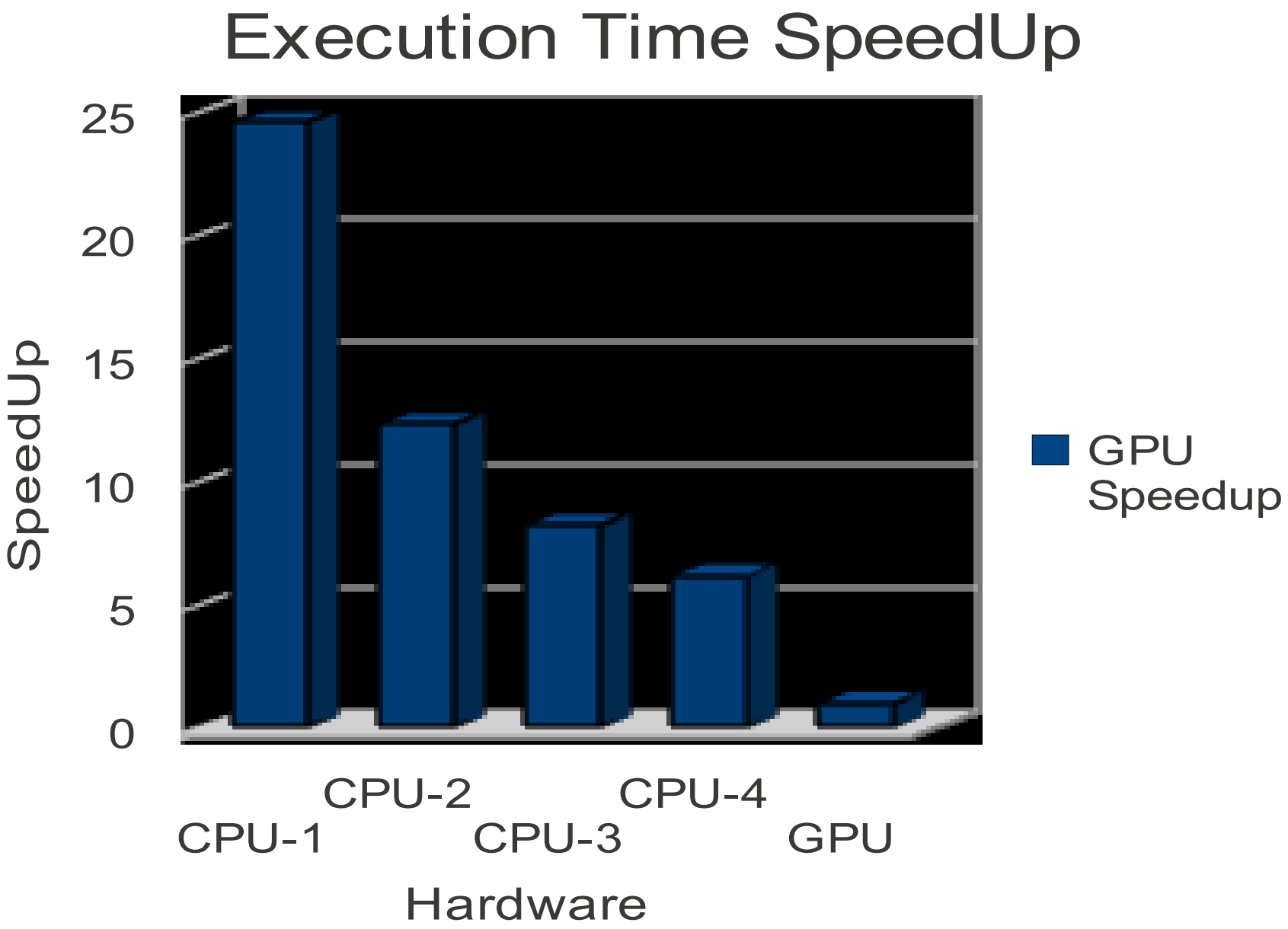
Our (C/C++ & OpenMP) CUDA-based implementation included an exec. conf. of a 35x35x35 grid of 10x10x10 blocks for a total of 350x350x350=42,875,000 threads that compute correlation coefficients per kernel call!



The CUDA-based implementation ran faster,....



... gave notable SpeedUp,...



...and these results for the k=4 analysis.

AI	NG	NT1	NT2	NT3	NT4
0.733009	SERPINA6	EPC1	PLAGL1	WT1	ZNF10
0.728841	FGB	IRF1	MGA	PAPOLA	SNAPC3
0.727331	FGB	PRKAR1A	TWISTNB	ZNF155	ZNF83
0.721789	FGB	EPC1	HMGB2	ITGB3BP	SP110
0.718291	FGB	EPC1	ITGB3BP	PAPOLA	SP110
0.709569	AFP	BCL6	ID4	SIAH2	ZNF212
0.704491	C4BPB	BCL6	PPP2R1B	PPP2R1B	PPP2R1B
0.70428	FGB	E2F5	HMGB2	MGA	SP110
0.701733	FGB	HMGB2	MGA	SP110	ZNF83
0.701466	FGB	TWISTNB	ZNF155	ZNF198	ZNF83

Notes on genes

- FGB: fibrinogen beta chain (role in blood clotting)
- EPC1: part of HAT (histone acetyltransferase)
- ZNF10: transcriptional repressor
- PLAGL1: zinc-finger protein ; possibly a tumor-supressor
- WT1 : TF with zinc-finger motifs ; associated with Wilm's (kidney) tumors
- BCL6: TF zinc-finger ; maybe involved with pathogenesis of diffuse large-cell lymphoma (DLCL)

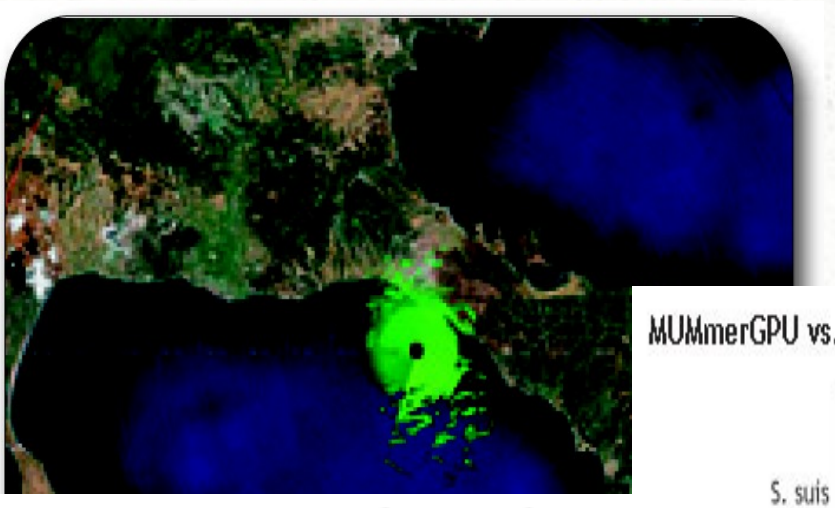
Some concluding thoughts.

- More efficient use of the GPUs running k=4 might give running times closer to ≈*2.3* days (instead of *4.6*)

- Polymeric TFs with degrees of homopolymerization (homopolymer)

- For certain computing problems, CUDA is valuable tool!

Example DOD/BioInformatics GPU Applications



Because radar relies heavily on FFTs, it is the kind of computationally intensive application for which GPGPU technology is ideally suited.

SOAP3/GPU

华大基因 BGI

Future possible projects and/or directions in this thread of research

- Deployment/implementation in “the cloud” – develop locally, but run in production in the cloud
- “CUDAfy” the computation and comparison of absolute improvement scores (currently done by CPU)

Future possible projects and/or directions in this thread of research, cont'd…

- Apply MPI and/or OpenACC? ; a more heterogeneous set of HPC-style technologies (MPI+OMP+GPU)?

- MAAS? ("micro-array analysis as a service") www.minemyarray.org?

THANK YOU!
QUESTIONS?