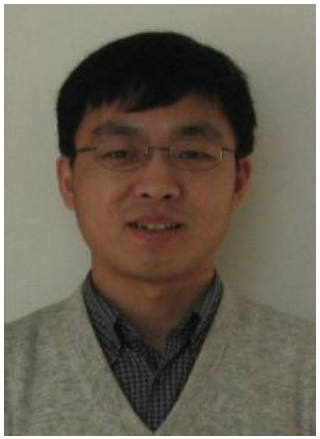
University of New Brunswick, April 11, 2013

Homology Search

Ming Li

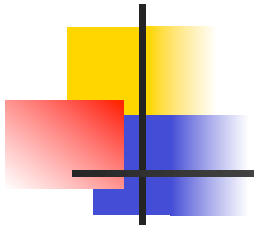
Canada Research Chair in Bioinformatics
School of Computer Science
University of Waterloo

Main Coauthors



Bin Ma

John Tromp

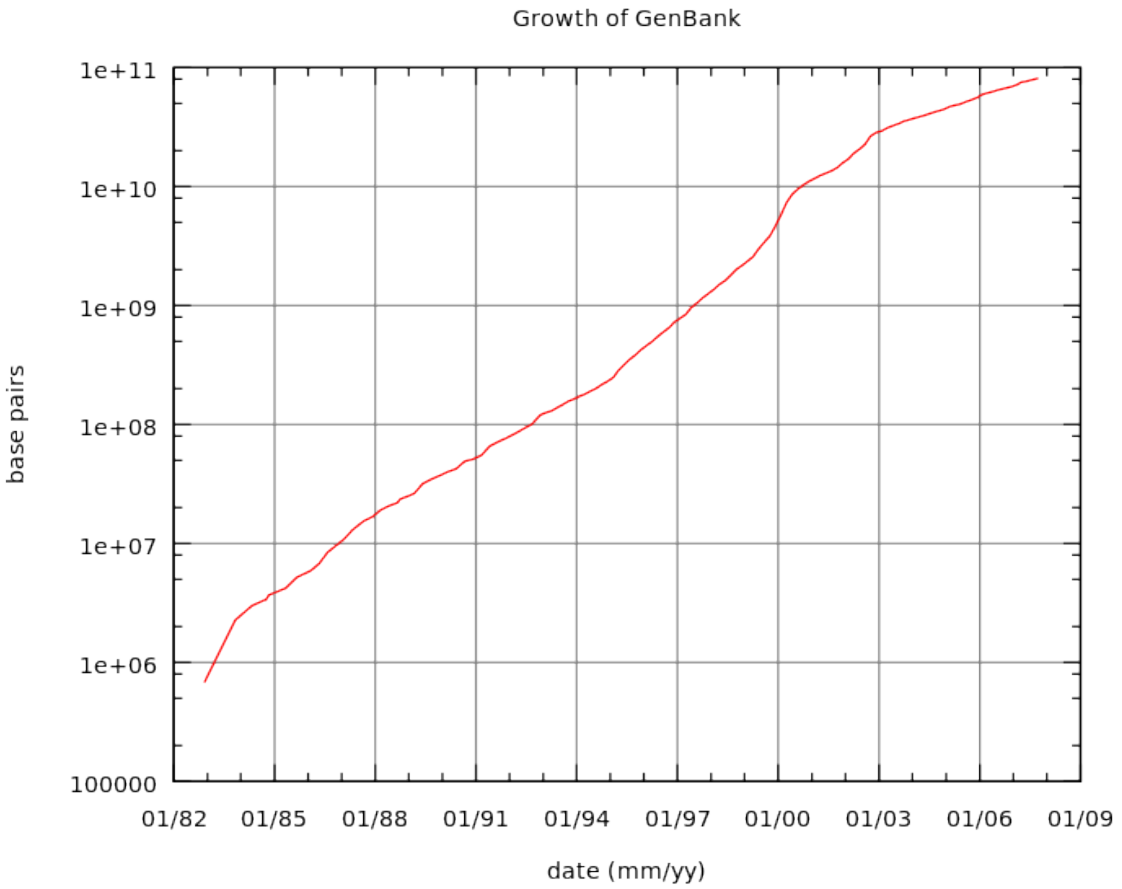


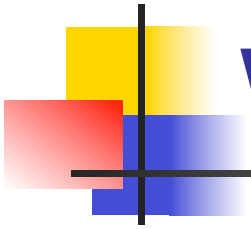
Research is not about complicated math, large engineering systems, but it is about ideas, simple ideas.

I will present one simple idea

Background

- GenBank doubles every 18 months
- From 100,000 distinct organisms
- 2013 Feb, over 150 billion bases
- $1000-one day-genome sequencing



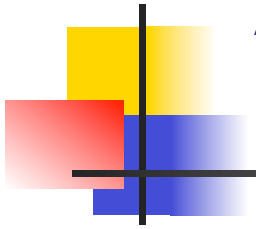


What is homology search

- Given two DNA sequences, find all "similar regions".

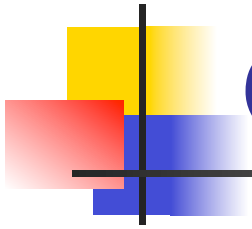
- Specifically, let's fix "edit distance"
  - match=1,
  - mismatch=-1,
  - gapopen=-5, gapext=-1

A comparison



- Homology search
  - Upper bound: 5 billion people x 3 billion basepairs + millions of species x billion bases
  - Data size: 150 GB
  - Query frequency: NCBI BLAST -- 150,000/day
  - Query type: approximate match.

- Internet search:
  - Upper bound: 5 billion people x homepage size
  - Data size, 2006: 840TB
  - Google, 2011: 4.7 billion queries/day
  - Query type: exact keyword match --- easy to do



Old Homology Search

- Dynamic programming (1970-1980)
  - Too slow: Human vs mouse genomes: 10^4 CPU-years
- BLAST, FASTA heuristics (1980-1990)
  - Trading sensitivity for speed
  - Yet, still not fast enough -- Human vs mouse genomes: 19 CPU-years (2001).

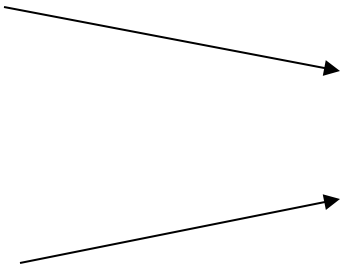
Our Goal

Old paradigm

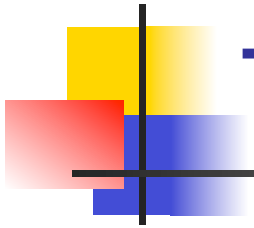
We want

Dynamic Program.
Sensitive, but slow

BLAST: Fast,
but low sensitivity

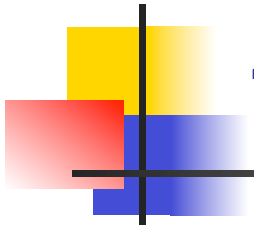


100% sensitivity
and faster



Talk Outline

1. Optimal spaced seeds
2. Multiple seeds
3. A new application
4. Open questions



1. Optimal Spaced Seeds

BLAST Algorithm: location, location …

- Find seeded matches of eleven base pairs, represented as 1111111111.
- Extend each match to right and left, until the scores drop, to form an alignment.
- Report all local alignments.

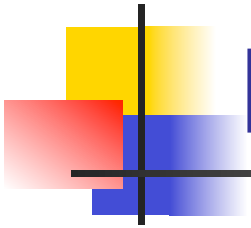
Example:

0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 0 1 1 1 1 0

AGCGATGTCAGGCGCCCGTATTTCCGTA

| | | | | | x | | | | | | | | | | | | | | | | | | | | |

TCGGATCTCACGCGCCCGGCTTACCGTG



BLAST Dilemma:

- Speed & sensitivity have contradictory requirement for seed length:
  - increasing seed size speeds up, but loses sensitivity;
  - decreasing seed size gains sensitivity, but loses speed.
- How do we increase sensitivity & speed simultaneously? Many have tried: suffix tree, better programming …

The Idea: Optimal Spaced Seed

BLAST seed was:

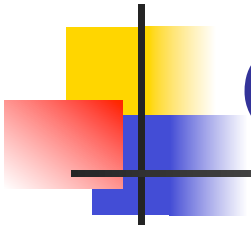
11111111111

And this:

11111*11*11*11

Optimizing gives: 111*1**1*1**11*111

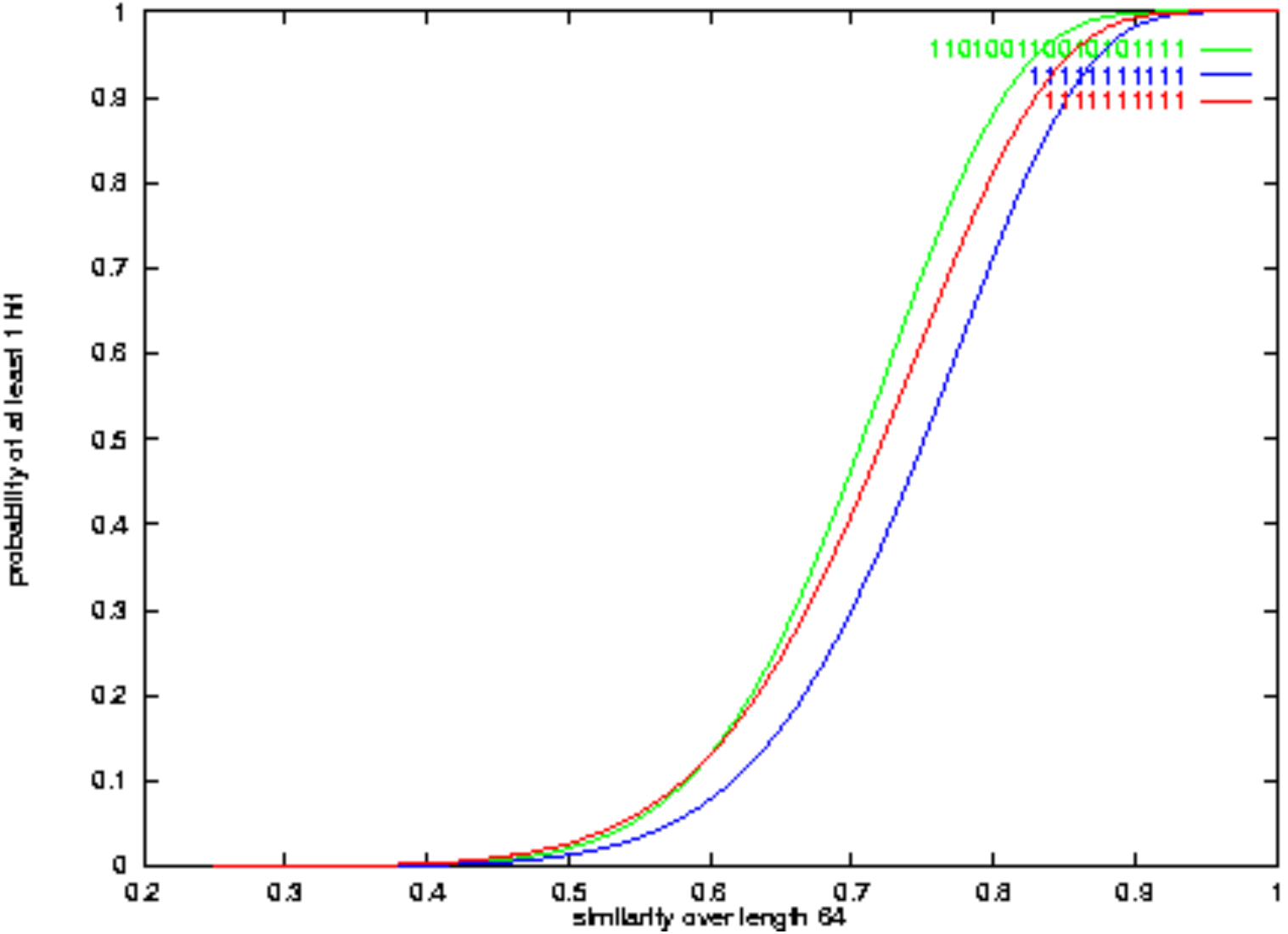
- 1 means a required match
- * means "don't care" position



Optimal Spaced Seed

- Spaced Seed: nonconsecutive matches and optimize match positions.
- BLAST seed 11111111111 is the worst seed
- Spaced seed: 111*1**1*1**11*111 is optimal
  - 1 means a required match
  - * means "don't care" position
- This seemingly simple change makes a huge difference: significantly increases hit to homologous region while reducing bad hits.

Sensitivity: PH weight 11 seed vs BLAST 11 & 10

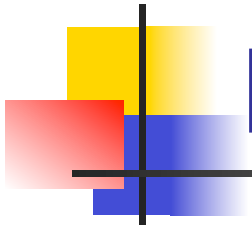


Formalize

- Given i.i.d. sequence (homology region) with Pr(1)=p and Pr(0)=1-p for each bit:

```
1100111011101101011101101011111011101
                          111*1**1*1**11*111
```

- Which seed is more likely to hit this region:
  - BLAST seed: 11111111111
  - Spaced seed: 111*1**1*1**11*111



Expect Less, Get More

- Lemma: The expected number of hits of a weight W length M seed model within a length L region with homology level p is

  $$(L-M+1)p^W$$

  Proof. $E(\#hits) = \sum_{i=1 \ldots L-M+1} p^W$ ■

- Example: In a region of length 64 with p=0.7
  - Pr(11111111111 hits)=0.3
    E(# of hits by 11111111111)=1.07
  - Pr(111*1**1*1**11*111 hits)=0.466
    E(# of hits by 111*1**1*1**11*111)=0.93

Why Is Spaced Seed Better?

A wrong, but intuitive, proof: seed s, interval I, similarity p

E(#hits) = Pr(s hits) E(#hits | s hits)

Thus:

Pr(s hits) = Lp^w / E(#hits | s hits)

For optimized spaced seed, E(#hits | s hits)

111*1**1*1**11*111	Non overlap	Prob
111*1**1*1**11*111	6	p^6
111*1**1*1**11*111	6	p^6
111*1**1*1**11*111	6	p^6
111*1**1*1**11*111	7	p^7

.....

- For spaced seed: the divisor is $1+p^6+p^6+p^6+p^7+$ ...
- For BLAST seed: the divisor is bigger: $1+ p + p^2 + p^3 +$ ...

Computing Spaced Seeds by DP

(Keich, Li, Ma, Tromp, *Discrete Appl. Math*)

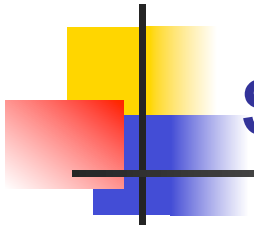
Let $f(i,b)$ be the probability that seed s hits the length i prefix of R that ends with b.

$$f(i,b) = \begin{cases} 1, & \text{if } s = b \\ (1-p)f(i-1,0b') + pf(i-1,1b') & \text{o.w.} \end{cases}$$

where b' is b deleting the last bit. Thus,

$$\text{Prob(s hitting R)} = \Sigma_{|b|=M} \, Prob(b) \, f(L-M,b)$$

Complexity of finding the optimal spaced seeds



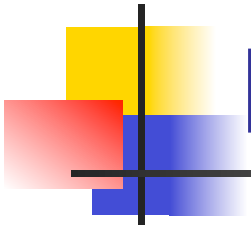
Theorem 1 [Ma-Li]. Given a seed, it is NP-hard to find its sensitivity, even in a uniform region.

Theorem 2 [Ma-Li]. The sensitivity (including very small sensitivities) of a given seed can be efficiently approximated with high probability.

Open: Determine the complexity of finding an optimal spaced seed.

Theorem 4 [Buhler-Keich-Sun, Ma-Li] The asymptotic hit probability is computable in exponential time in seed length, independent of homologous region length.

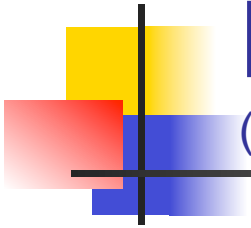
Theorem 5 [L. Zhang] If the length of a spaced seed is not too long, then it strictly outperforms consecutive seed, in asymptotic hit probability.



Related Literature

- Prior work. Random or multiple spaced q-grams were used in the following work:
  - FLASH by Califano & Rigoutsos
  - Multiple filtration by Pevzner & Waterman
  - LSH of Buhler
  - Praparata et al on probe design

- Many extensions to HMM seeds, vector seeds, variable length seeds … Spaced seeds bibliography http://www.lifl.fr/~noe/spaced_seeds.html



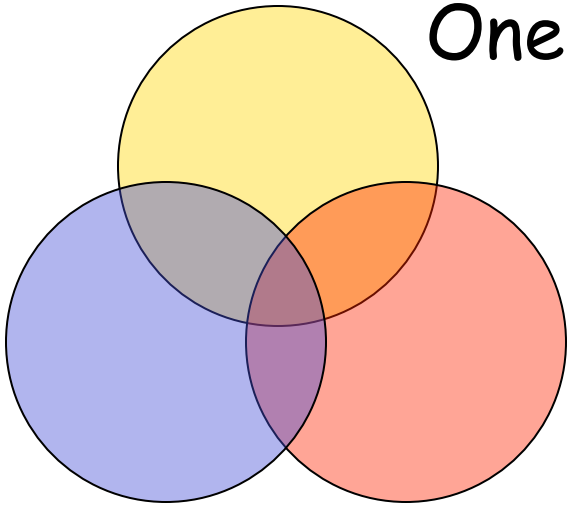
PatternHunter

(Ma, Tromp, Li: *Bioinformatics*, 18:3, 2002, 440-445)

- PH used optimal spaced seeds
- Written in Java.
- Used in Mouse Genome Consortium (*Nature*, Dec. 5, 2002), as well as in hundreds of institutions & industry.
- Optimal spaced seeds today are used in almost all homology search software, including BLAST, serving tens of thousands of queries daily.

2. Multiple Seeds: Full Sensitivity

Space of homologous regions



One seed

Two seeds

Three seeds

PatttternHunter II:

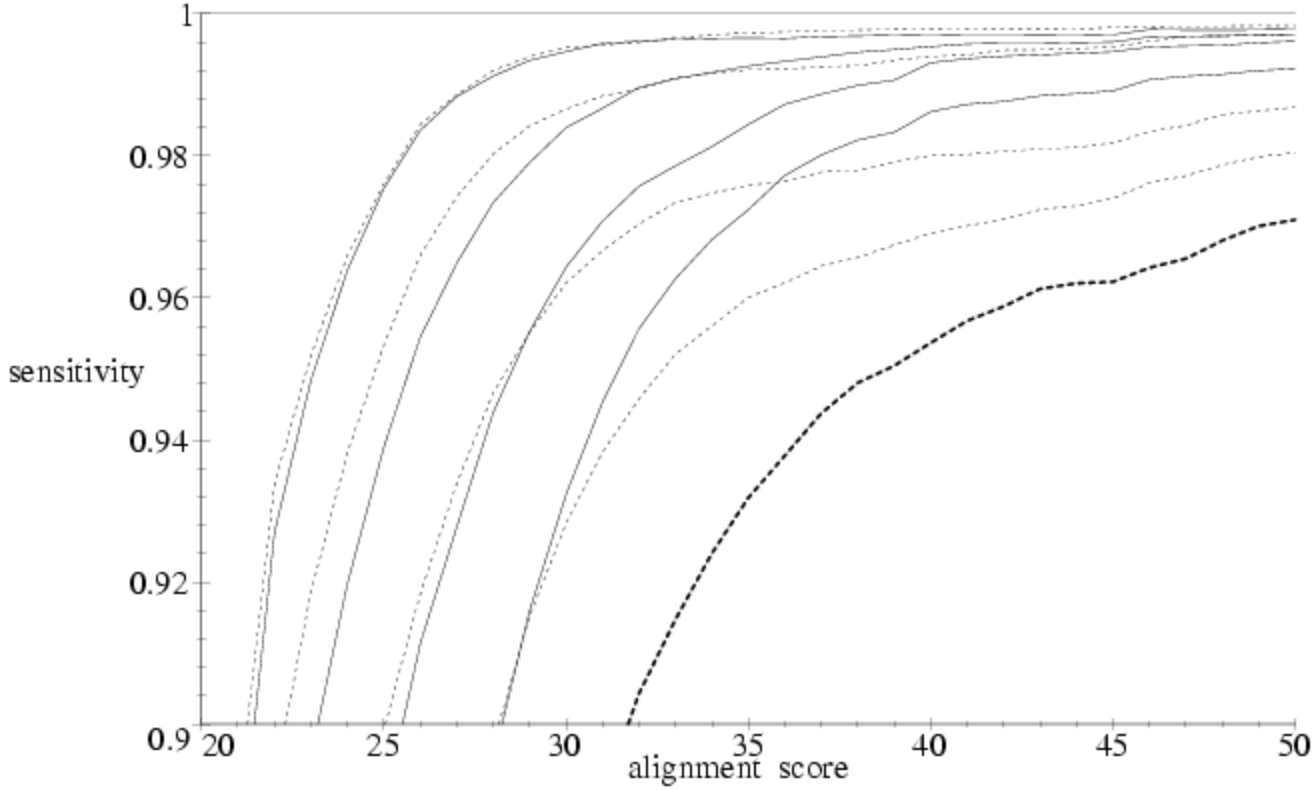
-- Fast search at full sensitivity

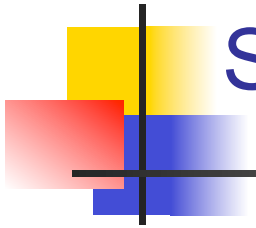
(Li, Ma, Kisman, Tromp, *J. Bioinfo Comput. Biol*. 2004)

- The biggest problem for BLAST was low sensitivity. Massive parallel machines are built to do S-W exhaustive dynamic programming.
- Spaced seeds give PH a *unique* opportunity of using several optimal seeds to achieve optimal sensitivity, this was not possible by BLAST technology.
- Using multiple optimal seeds. PH II approaches Smith-Waterman sensitivity & 3000 times faster.

Sensitivity Comparison with Smith-Waterman (at 100%)

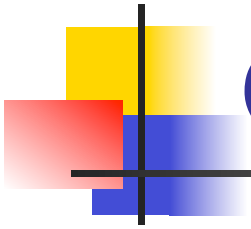
The thick dashed curve is the sensitivity of BLAST, seed weight 11. From low to high, the solid curves are the sensitivity of PH II using 1, 2, 4, 8 weight 11 coding region seeds, and the thin dashed curves are the sensitivity 1, 2, 4, 8 weight 11 general purpose seeds, resp.





Speed Comparison with Smith-Waterman

- Experiment: 29715 mouse EST, 4407 human EST.
- Smith-Waterman (SSearch): 20 CPU-days.
- PatternHunter II with 4 seeds: 475 CPU-seconds. 3638 times faster than Smith-Waterman dynamic programming at the same sensitivity.



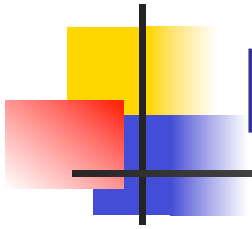
One example.

- DOTM Project has one million EST's for the Brassica napus genome.
  - They initially depended on TimeLogic special hardware to do exhaustive Smith-Waterman alignment, needing 800 days.
  - At > 99% sensitivity, Patternhunter II can finish the job in 40 days on one PC.

3. Trend Prediction

Zou-Deng-Li: Detecting Market Trends by Ignoring It, Some Days, 2010

- 4.6 billion dollars are traded at NYSE daily.
- Buy low, sell high.
- Essentially, a "buy" indicator must be:
  - Sensitive when the market rises
  - Insensitive otherwise.



Background

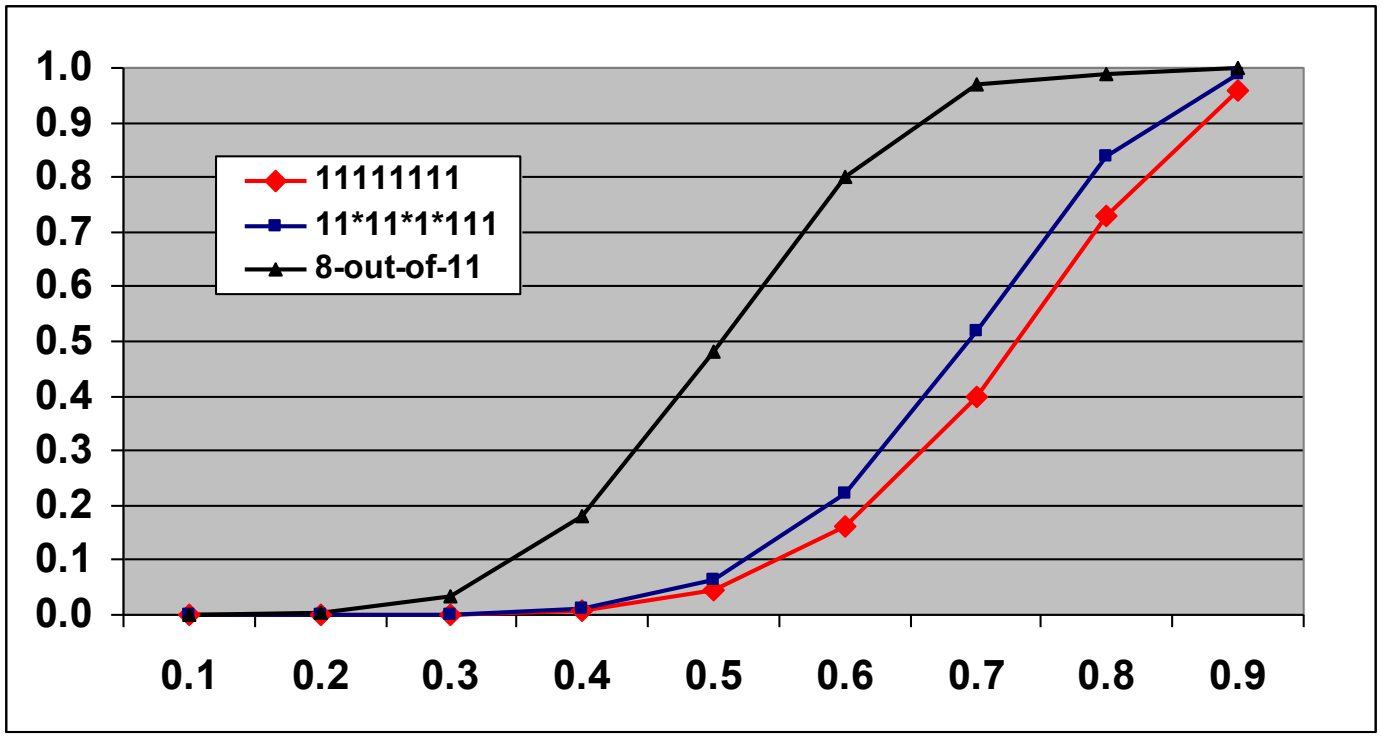
- Hundreds of market indicators are used:
  - Common sense: if the past k days are going up, then the market is moving up.
  - Moving average over the last k days. When the average curve and the (plain) price curve intersect, buy/sell.
  - Special patterns: a wedge, triangle, etc.
  - Volume
  - Hundreds used in automated trading systems.

Problem Formalization

- The market movement is modeled as a 0-1 sequence, one bit per day, with 0 meaning market going down, and 1 up.
- S(n,p) is an n day iid sequence where each bit has probability p being 1 and 1-p being 0. If p>0.5, it is an up market
- $I_k=1^k$ is an indicator that the past k days are 1's.
  - I_8 has sensitivity 0.397 in S(30,0.7), too conservative
  - I_8 has false positive rate 0.0043 in S(100, 0.3). Good
- I_i^j is an indicator that there are i 1's in last j days.
  - I_8^{11} has high sensitivity 0.96 in S(30,0.7)
  - But it is too aggressive at 0.139 false positive rate in S(100, 0.3).
- Spaced seeds 1111*1*1111 and 11*11111*11 combine to
  - have sensitivity 0.49 in S(30,0.7)
  - False positive rate 0.0032 in S(100, 0.3).
- Consider a betting game: A player bets a number k. He wins k dollars for a correct prediction and o.w. loses k dollars. We say an indicator A is better than B, A>B, if A bets after B and it always wins more and loses less than B does.

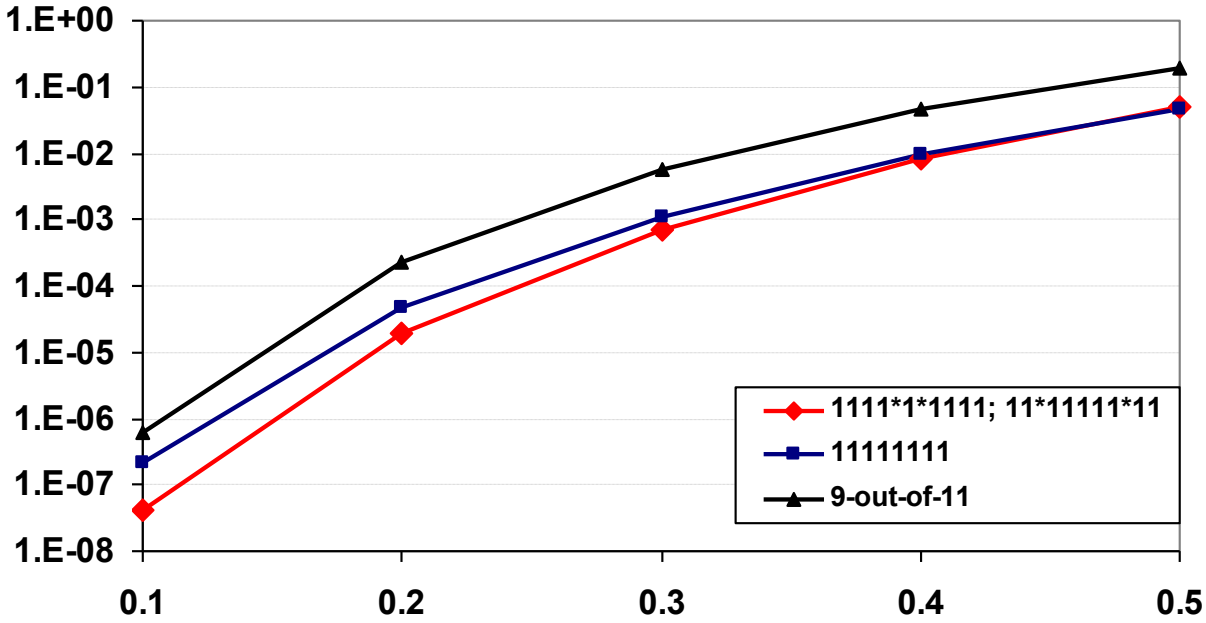
Sleeping on Tuesdays and Fridays

- Spaced seeds are beautiful indicators: they are sensitive when we need them to be and not sensitive when we do not want them to be.

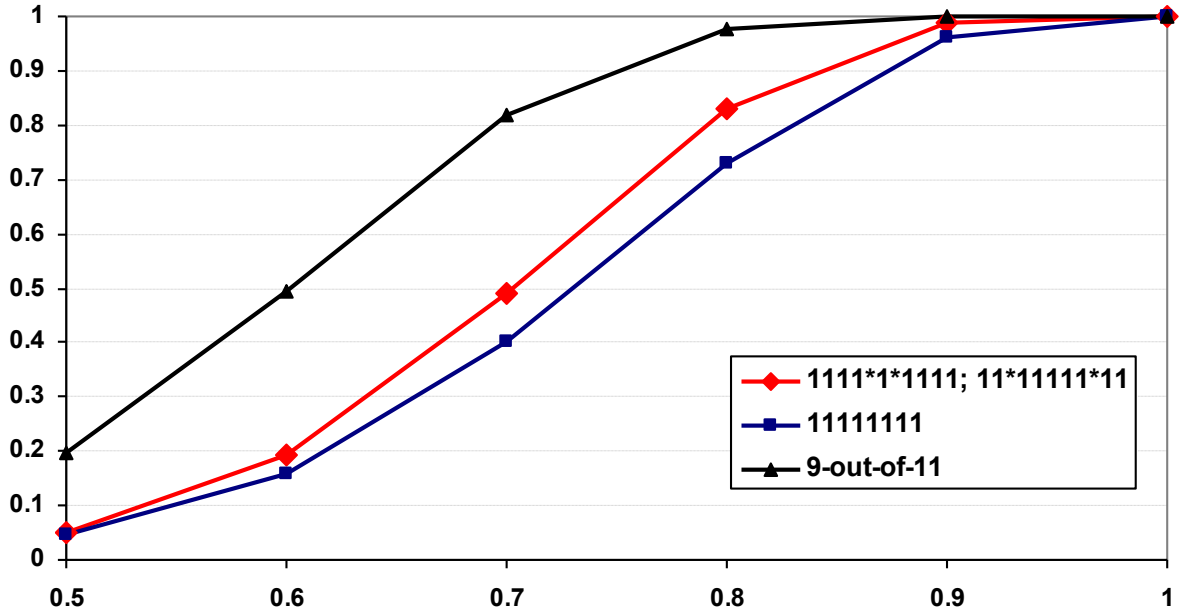


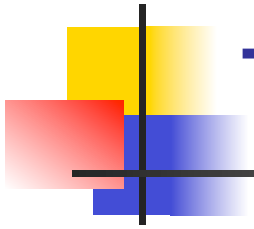
11*11*1*111 always beats I_8^{11} if it bets 4 dollars for each dollar I_8^{11} bets. It is $>I_8$ too.

Two spaced seeds



Observe two spaced Seeds curve vs I_8, the spaced seeds are always more sensitive in p>0.5 region, and less sensitive when p<0.5



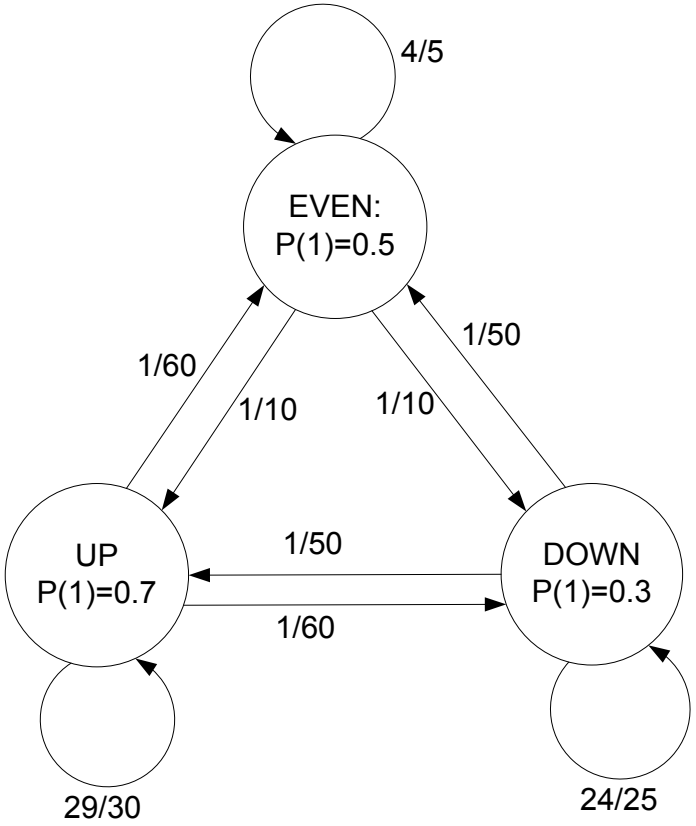


Two experiments

- We performed two trading experiments
  - One artificial
  - One on real data (S&P 500, Nasdaq indices)

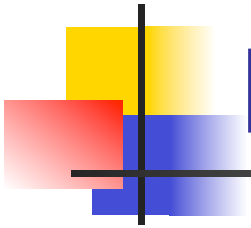
Experiment 1: Artificial data

- This simple HMM generates a very artificial simple model
- 5000 days (bits), start at $100, average over 250 simulations.
- Indicators: I_7, I_7^{11}, 5 spaced seeds.
- Trading strategy: if there is a hit, buy, and sell 5 days later.
- Reward is: #(1)-#(0) in that 5 days times the betting ratio



Results of Experiment 1.

	R	#Hits	Final MTM	#Bankrupcies
I_7=1111111	$30	12	$679	16
I_7^{11}	$15	47	$916	14
5 Spaced seeds	$25	26	$984	13



Experiment 2

- Historical data of S&P 500, from Oct 20, 1982 to Feb. 14, 2005 and NASDAQ, from Jan 2, '85 to Jan 3, 2005 were downloaded from Yahoo.com.
- Each strategy starts with $10,000 USD. If an indicator matches, use all the money to buy/sell.

Trading Details	Trading Indicators				
	12 Month MA crossover	I_7^9	I_7= 1111111	1 Optimal Seed 111*11*11	2 Optimal Seeds 111*1*111, 11*1111*1
Initial Investment	10,000	10,000	10,000	10,000	10,000
S&P 500	20-Oct-82: 139.23 to 14-Feb-05: 1206.4				
Mark-to-Market	68,923	29,384	32,343	74,689	80,582
# Trades	43	51	3	8	10
# Trades with Profit	12	25	2	7	8
# Trades with Loss	31	26	1	1	2
Avg Gain per $1,000 per trade	29.4	18.8	309.3	210.0	178.8
NASDAQ	2-Jan-85: 353.20; 3-Jan-05: 2152.15				
Mark-to-Market	88,436	104,208	110475	111,105	144,496
#Trades	41	73	32	18	22
# Trades with Profit	15	40	19	13	17
# Trades with Loss	26	33	13	5	5
Avg Gain per $1,000 per trade	51.3	24.7	68.6	125.2	126.6

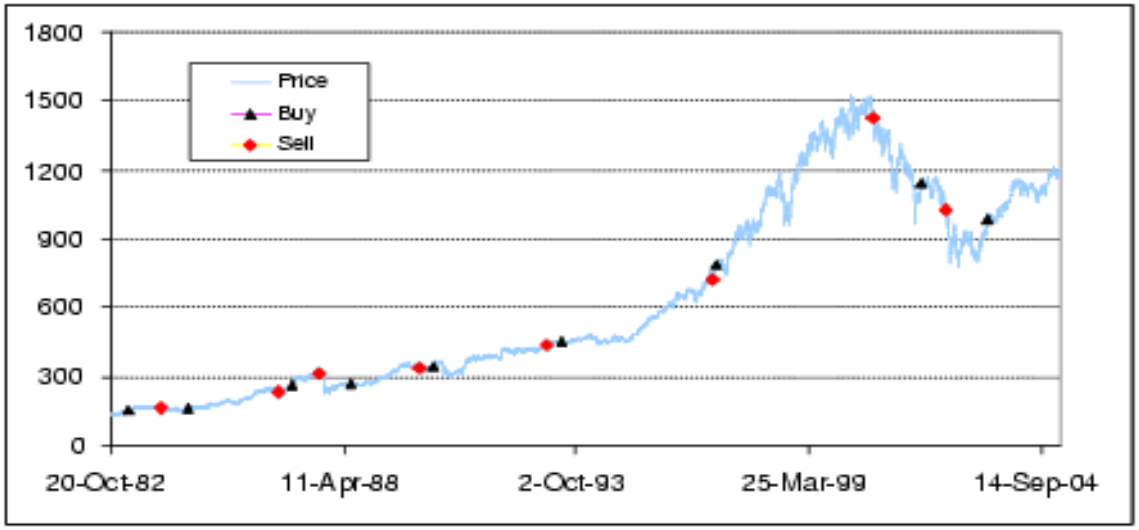
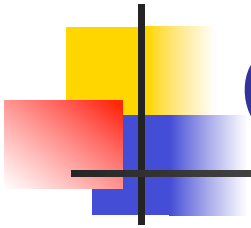


Fig. 5. The buy-sell points of 111*11*11 on S&P 500 index.



Conclusion

Simple ideas are often the better ones.

Open Question:

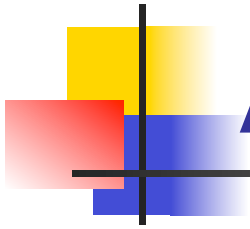
1. Complexity of finding an optimal seed, in a uniform region. Note $L=\{(1^L, 1^W, S_{opt})\}$ is not NP-hard, as it is sparse in a uniform distribution. Note, for arbitrary distribution, it is NP-hard.
2. Alternating seeds
3. Extend our work for financial market.
4. Can the spaced seeds be applied to other areas?

An idea and open question

- The optimal spaced seed has the least self correlation.
- Idea: can we further improve this by using different (or alternating) spaced seeds as we scan through the sequences?

```
111*1**1*1**11*111
 1*111**1*11*11**11
  111*1**1*1**11*111
   1*111**1*11*11**11
        ...
```

- Open Question: Prove this is no good?



Acknowledgement

- PH is joint work with Bin Ma and John Tromp
- PH II is joint work with Ma, Kisman, and Tromp
- Some joint theoretical work with Ma, Keich, Tromp, Xu, Brown, Zhang.
- Financial market prediction: J. Zou, X. Deng

- Financial support: NSERC, Killam Fellowship, Steacie Fellowship, CRC chair program, Bioinformatics Solutions Inc. MOST 863 Project.