Strong motifs are easy to find

dan brown

cheriton school of computer science

university of waterloo

browndg@uwaterloo.ca

joint work with brona brejova, ian harrower, and tomas vinar

At CPM 2005…

Our paper: simple PTAS for motif finding
requires very slow runtimes to guarantee good approximation ratio

As much as $\Omega(\ell(nm)^{1/\epsilon^2})$ runtime, when *n* = # of sequences, *m* = sequence length, *l* = motif length, for worst-case motifs we presented.

Ming Li asked Ian Harrower, "What about average case motifs?"

CPM 06: good motifs are easy to find.

Strong motifs (for many definitions of strong) can be **approximated** efficiently.

Can also be found **exactly** in reasonably efficient runtimes.

Instead of requiring $\Omega(\ell(nm)^{1/\epsilon^2})$ runtime to guarantee 1+ε approximation ratios, we only need runtime with a logarithm in exponent.

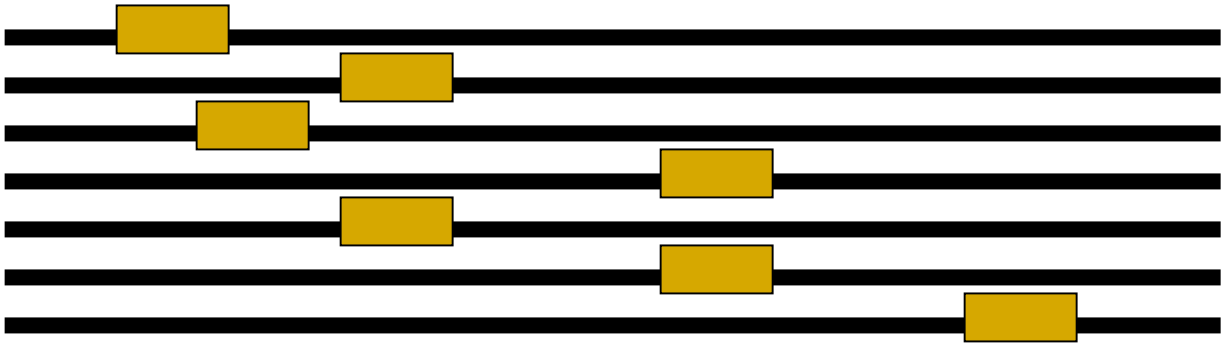
All interesting motifs are strong.
Motif-finding is easier than thought.

Motif finding:
abstraction of real problem

The real problem:

Given a collection of genes, all “controlled” by same protein, find out how that control works.

Simplest form: in DNA “before” each gene, a common element present. Find that element.



An abstraction

Consensus-Pattern:

- Given n sequences, s_i, each of length m, over finite alphabet Σ, and a parameter l.
- Find: Sequence x of length l (motif centre), and subsequence x_i of length l from each sequence s_i with minimum Hamming distance from x.
- Minimize sum of Hamming distances between all x_i and x, the motif centre.

Example of Consensus-Pattern

Sequences:

AGATTACATAGCATATGGGACATAGGATT

ACATAGGTATAGAGAAAAAGCCCCAGATA

GTATTTTACAACGGGAGAATTTTCAAGAT

TTAGTATATTTTTTAACAACGTTAGTATTA

GTATTACCTAGTAGGGACACCCCATATTA

ATATTAGGATTCATATGGATACCATATGA

Motif centre： ACATAG

A family of heuristics

- Suggest a set *X* of motif centre choices
- Look for best match in each sequence to each member of *X*.
- Return member of *X* with best total score.

Look-up phase takes O(*nml*) time.

If *X* is polynomial in size, heuristic is poly-time.

(Note: the reason this problem isn't trivial is that the motif centre may be a subsequence of none of the input sequences!)

A provably good set X.

A PTAS for this problem:

- Look at **all** sets of r subsequences of length l from the input set.
- X = {consensus sequences of r-element sets}. [Break ties arbitrarily]

For constant r, a poly-time algorithm:
$O((nm)^{(r+1)}l)$, because $|X| = O((nm)^r)$

Theorem [Li, Ma, Wang]:

An O($1 + 1/\sqrt{r}$)-approximation algorithm

Is this a useful theorem?

In some sense, yes:

- Sampling-based algorithms work well for Consensus-Pattern. Maybe this is why?

In a very real sense, **no**.

- To obtain 1+ε-approximation algorithm, need sample size $r = \Omega(1/\epsilon^2)$ which gives an algorithm with $\Omega(\ell(nm)^{1/\epsilon^2})$ runtime.
- That's pretty terrible.

Things get worse…

We showed [Brejova, B, Harrower, Lopez-Ortiz, Vinar 05]:

For any value of *r*, there exists an instance of the problem for which the approximation ratio of a very close relative of the sampling PTAS is $\Omega(1 + 1/\sqrt{r})$.

But people actually use sampling-based algorithms, even simpler than the PTAS! Huh?

A first definition for strong motifs

For simplicity, assume binary alphabet.

Motif instances: best matches to consensus.

Columns of the motif: the *i*th position of each of the *m* motif instances.

Consistently strong motif: every column of the motif is at least .5+ε fraction 0's, for some nonnegative constant ε.

Our bad instances had strength $1/2 + 1/\sqrt{r}$

A simplification and some probability

Can restrict attention to seeing what happens when the input are only the *l*-length motifs.

(The problem is trivial then, but the PTAS may not pick the right motif centre.)

First moment principle:
if we analyze the performance of a random sample, *some* sample does at least that well.

We're enumerating *all* samples.

Sampling with replacement

Consider a 1-column strong motif, with *pn* zeros and (1-*p*)*n* ones. (*p* well above 0.5)

Random sampling *with replacement*:
like flipping a biased coin with heads probability *p* *r* times.
Does it come up heads more than tails?

Weak motifs and strong ones

In our CPM 05 paper. we showed that for weak motifs (where $p = 1/2 + 1/\sqrt{r}$), the probability of more tails than heads is at least a constant.

With **strong** motifs, probability of more tails than heads at most $(4p(1\text{-}p))^{r/2} = \alpha^r$

Converges to zero **exponentially fast** as a function of *r*.

How to show that?

This uses the Hoeffding bound

- Like a Chernoff bound, but overall probability of the bad event is based on the probability of each individual bad event occurring.

What does this give us?

- The column has probability at most α^r of being guessed wrong.

Is that enough to get us a good theorem?

An ok theorem, but not great.

Suppose *m* columns, all with exactly *pn* zeros.

Cost of getting column right: (1-*p*)*n* (# of ones)

Cost of getting it wrong: *pn* (# of zeros).

Expected cost: at most $\alpha^r pn$ +(1- α^r)(1-*p*)*n*.

Expected approximation ratio: at most 1+α^r(p/(1-p)).

Converges to 1 exponentially in *r* (for weak motifs, it converges like $1 + 1/\sqrt{r}$)

What's not great about that?

Consistently strong: **at least** *p* fraction of zeros in each column.

What if more than *p*?

Lower probability of getting the column wrong, but approximation ratio gets very bad.

Must trade off two probabilities. Is that ok?

Yes: for every *p*, there *exists* an *r* such that $1+\alpha^r(p/(1-p))$ is decreasing once the sample size is *r*.

Full statement of theorem

- For consistently strong motifs of minimum strength *p*, there exists an *r* such that as sample size grows past *r*, approximation ratio converges to 1 exponentially fast.

Another fun theorem: if the expected number of mistaken columns is less than 1, then the PTAS **will** find the optimum.

Details in the paper.

What about random motifs?

Random motif of fixed content: a *p* fraction of the entries in the motif are zeros and a 1-*p* fraction are ones.

Score of the optimum is *not* (1-*p*)*nl*, though!

Some columns may have more ones than zeros!

Also, we might get a bad instance of the problem, with lots of columns very close to 50% zeros and 50% ones.

A slight modification to the PTAS

Allow only one sample from each sequence.

Hoeffding bound still applies here. Shown using machinery by Panconesi and Srinivasan on applying Chernoff-style bounds to non-independent samples.

(Their paper should also be more well known.)


Put this together

Bad instances: at least $\alpha^{r/2}$ columns with more ones than zeros.

They are exponentially rare.

Good instances: fewer bad columns.

On good instances, expected approximation ratio converges to 1 exponentially fast.

On bad instances, ratio at most 2.

Put together: expected approximation ratio converges to 1 exponentially fast as *r* grows.

One last kind of strength

Random motifs of expected strength *p*. Every entry in motif instances comes from independent coin flip with probability *p* of getting a zero bounded above .5

Can think of this in two steps:

1. Pick the number of zeros in the instance
2. Distribute them arbitrarily across all *nl* places for them.

Bad instances result from both steps.

Bad instances for this

Not enough zeros: If there are fewer than $(.5+p)/2$ zeros, it's a bad instance.

Bad distribution of zeros: If there are too many columns with more ones than zeros, it's a bad instance.

Then, same sort of probabilistic machinery as before.

Overall, our results

For a variety of definitions of **strong**, the simple PTAS described by Li, Ma and Wang gives performance much better (either in expectation or in guarantee) than is provable for general motifs.

Approximation guarantee converges to 1 exponentially fast as a function of the sample size *r* in all cases.

(One bad definition of “strong” is in the paper.)

Some last comments

- We computer scientists should be learning more probability than is in standard randomized algorithms textbooks.
  Both the Hoeffding bound and the Panconesi and Srinivasan paper are profoundly useful.
- Probably this can be extended to other models of motif finding. In general, motif finding is a much easier problem in practice than in theory.

Acknowledgments

- Kunsoo Park and the organizing staff for CPM 2005, in Jeju City, Korea, where I was happy enough to come up with some of the ideas of this work while asleep.
- Ming Li, for asking the right question, even when we were 12,000 km from home.
- NSERC for funding our work.