Designing Filters for Fast Protein and RNA Annotation

Yanni Sun
Dept. of Computer Science and Engineering
Advisor: Jeremy Buhler
Outline

- Background on sequence annotation
- Protein annotation acceleration
  - Filter design for sequence-model comparison
  - Filter design for model-model comparison
- Noncoding RNA annotation acceleration
- Research summary
Post-Genomic Era

- Sequencing a bacterial genome within 1 day and a human genome in several months [Wadman ‘08]
Urgent Task: Sequence Annotation

- Find biologically meaningful regions

part of the human chromosome 14

...GGGG...TCCA ATG

tRNA  
- a type of noncoding RNA

transcription

mRNA

translation

protein

protein domains
Sequence Annotation Method

- Comparative sequence analysis
  - Using sequence similarity

<table>
<thead>
<tr>
<th>Species 1</th>
<th>species 2</th>
<th>species 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>...YTCSYCGLKSFTQS...</td>
<td>...YTCPCYCDKRFQTQR...</td>
<td>...YLCLYCGLKTLSDR...</td>
</tr>
</tbody>
</table>

Part of protein family zf-C2H2

- Compare a newly obtained sequence to a database of annotated sequence families
Sequence Annotation is Expensive!

- Requires **CPU months** to find noncoding RNA signals in a bacterial genome
- Requires **CPU days** to classify all the proteins generated by a genome

Protein family DB
Pfam-A+Pfam-B: ~30,000

Metagenomic data sets

- microbial genomes from human gut, Sea water, etc.
- Typical size: $10^9$ bases [Chen & Pachter '05]
Preview of My Work

- Design filtration algorithms for
  - sequence-profile HMM (pHMM) comparison for protein annotation
    - ~20-35 times speedup on Pfam vs. Swiss-Prot protein database search
    - [Sun, Buhler] ECBB ’06 (Bioinformatics ’07), [Sun, Buhler] IEEE/ACM TCBB ’08
  - PHMM-pHMM comparison for protein annotation
    - ~21-31 times speedup on Pfam vs. Pfam search
  - Sequence-Stochastic Context Free Grammar comparison for noncoding RNA annotation
    - ~200 times speedup on Rfam vs. 65-Mbase genomic sequence database search
    - [Sun, Buhler] CSB’08
Outline

- Background on sequence annotation
- Protein annotation acceleration
  - Background
  - Filter design for sequence-pHMM comparison
  - Filter design for pHMM-pHMM comparison
- Noncoding RNA annotation acceleration
- Research summary
Protein (Domain) Family

- Sequences evolving from the same ancestor & having similar function
- Multiple sequence alignment
  - Identify conserved residues

seq1: A A A P P Q
seq2: - A - P P Q
seq3: G Y - P P A
seq4: A Y - P P P
seq5: G Y - P P P
seq6: G Y - P P Q
seq7: A Y - A A Q

XYPPX repeat

Profile HMM (pHMM): compute the generation probability for a sequence: [Rabiner ’89], [Durbin et al. ’98]
  - High generation probability → member sequence
Annotation - Case Study

- Human gut microbiome initiative (HGMI)
  - 100 bacterial genomes:
    - http://www.genome.wustl.edu/hgm
  - Protein annotation: compare $5 \times 10^5$ hypothetical proteins with Pfam (30,000)
    - Needs CPU-days
  - New protein families from HGMI
    - Compare to NCBI nr protein database
      - CPU-hours per family
Acceleration Strategy: Filtration

- Design filter for each protein family
- Does a sequence match the designed filter?
  - Yes → full comparison with the pHMM
  - No → ignore it
Filter Design Problem

- Sensitivity: \( \frac{\text{# of filter matched members}}{\text{# of members}} \)

- False positive rate (FP rate):
  \( \frac{\text{# of filter matched background sequences}}{\text{# of background sequences}} \)

- Significantly affects the speedup

- Goal: design a filter with maximum sensitivity while keeping its fp rate \( \leq \) a given threshold
Outline

- Background on sequence annotation
- Protein annotation acceleration
  - Background
  - Filter design for sequence-pHMM alignment
    - Profile-based filters
    - Pattern-based filters
  - Filter design for pHMM-pHMM alignment
- Noncoding RNA annotation acceleration
- Research interests and future work
Profile and Score Function [Gribskov,’87]

e.g. score ( AEEMRIGC,Profile) = ?

-2+(-3)+(-3)+(-4)+(-5)+(-3)+10+(-6)
Profile Design Based on PHMM

- **Profile match:** \( \text{score}(s,\text{Profile}) \geq T \)
  - Sensitivity: \( \Pr(\text{score}(s,\text{Profile}) \geq T \mid s \sim \text{pHMM } M) \)
  - False positive rate (FP rate):
    \[
    \Pr(\text{score}(s,\text{Profile}) \geq T \mid s \sim \text{background model } M^0)
    \]

- **Central design problem:** design \( P \) and \( T \) from pHMM, s.t. \( P \) has maximum sensitivity while keeping its fp rate \( \leq \) a given threshold \( r \)
Profile Design Framework

- Profile extraction from pHMM
- Compute score threshold $T$ to achieve given FP rate $r$
- Choose top profiles in order of sensitivity
- Compute profile’s sensitivity given $T$
Evaluate a Profile $P$

- Given a FP rate threshold $r$, find score threshold $T$ s.t.
  $$\Pr(\text{score}(s, \text{Profile}) \geq T \mid s \sim \text{background model } M^0) \leq r$$

- Evaluate $P$'s sensitivity under $T$:
  $$\Pr(\text{score}(s, \text{Profile}) \geq T \mid s \sim M)$$

- We have efficient algorithm to compute $T$ and sensitivity of $P$ given $r$
Outline

- Background on sequence annotation
- Protein annotation acceleration
  - Background
  - Filter design for sequence-pHMM alignment
    - Profile-based filters
    - Pattern-based filters
  - Filter design for pHMM-pHMM alignment
- Noncoding RNA annotation acceleration
- Research interests and future work
Design PROSITE-like Patterns for Highly Conserved Protein Families

- **Combinatorial problem:**
  - Pattern length not fixed a priori
  - Large design space: ~350 columns on average, ~$2^{20}$ sets of residues at each column

- **Map pattern design into a knapsack problem**
  - Design an approximation algorithm
  - Prove a bound: $sensitivity(P) \geq ((1 - \varepsilon)^m sensitivity(P^*))^{(1+\varepsilon)}$
    
    $m$: # of elements in $P^*$, $\varepsilon$: approximation ratio
Data and Experiments

- Design profiles and patterns for ~1000 pHMMs randomly chosen from Pfam-A
  - Sensitivity: fraction of true member sequences containing a match to the filter

- Speedup
  \[
  \frac{T_{original \ search}}{T_{filtered \ search}}
  \]
  time of searching for pHMM matches in Swiss-Prot DB
  
  time of searching for pHMM matches in Swiss-Prot DB after using a filter
Comparison of Patterns vs. Profiles

- Profiles have better sensitivity on average
- Patterns have better speedup for well-conserved family

Sensitivity comparison of patterns and profiles for 955 protein families

Speedup comparison of patterns and profiles for 630 protein families (sensitivity $\geq 0.9$)
Outline

- Background on sequence annotation
- Protein annotation acceleration
  - Background
  - Filter design for sequence-pHMM alignment
  - Filter design for pHMM-pHMM alignment
- Noncoding RNA annotation acceleration
- Research interests and future work
PHMM-PHMM Alignment

- Identify remotely related sequences

>query sequence (245 residues)
...LRYYPGSPELARRLTRAQD...

PSI-Blast
...VKKYPGSP...
...LRYYPGSP...
...SHYPGSP...
...GDLYPGSP...

build pHMM

Application: ~2000 pHMMs in Pfam are grouped into 283 clans [Finn '06] using pHMM-pHMM alignment

- 81: putative structure and function
PHMM-PHMM Alignment

- Identify remotely related sequences

>query sequence (245 residues)
.... VKKYPGSP LLARHLLR...

PSI-Blast

...VKKYPGSP...
...LRYYPGSP...
....-SHYPGSP...
...GDLYPGSP...

build pHMM

Application: ~2000 pHMMs in Pfam are grouped into 283 clans [Finn ‘06] using pHMM-pHMM alignment
- 81: putative structure and function
PHMM-PHMM Alignment: PRC
[Madera '05]

\[ \text{Similarity}(M_1, M_2) = \sum_{s \in \Sigma^*} (\Pr_{M_1}(s) \Pr_{M_2}(s)) \]

Need scoring function between states: dot product between their emission probability vectors

- PRC is \textcolor{red}{\textbf{7 times slower}} than sequence-pHMM alignment
- Filtration: represent each state by a symbol in a new alphabet
  - PHMM alignment \rightarrow sequence alignment
Degenerate Alphabet-based Filter

A filter $P$ contains 3 elements:

1. a degenerate alphabet $\Delta$

2. a function to map the residue emission probability vector in a state into a symbol in $\Delta$

3. a scoring function between two symbols in $\Delta$
Degenerate Alphabets for DNA

- IUPAC ($2^4-1=15$ elements); PHYLONET [Wang & Stormo 05]; Regulatory potential score [Kolbe et al. 03]

**DNA → protein:** # of residue combinations = $2^{20} - 1$
Preserve Well-Conserved States
Keep at most 3 Residues

- For a probability vector $e$
  1) Sort $e$ in decreasing order of emission probs.
  2) Choose least number of residues with sum probability $\geq$ a threshold (e.g. 0.5)
    - e.g. (P:0.89, A:0.02, ...) $\rightarrow$ P; (G:0.32, A:0.30, ...) $\rightarrow$ AG
  3) If more than 3 residues are needed, output *

- Alphabet size
  $$|\Delta| = C_0^{20} + C_1^{20} + C_2^{20} + C_3^{20} = 1531$$
Experimental Results

- Test set: 1388 pHMM pairs output by PRC with E-value 0.001
  - Sensitivity: percentage of test set matched by our filters
  - FP rate: # of pHMM pairs output by our filters but not PRC

Sensitivity = 0.938; Speedup = 21
PHMM-PHMM Alignment: Summary and Future Work

- Design the first degenerate alphabet for protein sequence comparison
  - Can be extended to other alignment tools besides PRC

Future work

- Ungapped → gapped alignment
- BLASTP-like alignment rather than full DP
Outline

- Motivation
- Protein function annotation acceleration
- Noncoding RNA annotation acceleration
- Research interests and future work
Noncoding RNA (ncRNA)

- Fold into secondary structure
- Base pair interaction: A-U, U-A, C-G, G-C

AUCCGAAAGGAU

start

AUCCGAAAGGAU

end

A A
G A
C-G
C-G
U-A
A-U

start

end

t-RNA
Modeling an ncRNA Family

- Describe both the sequence and secondary structure features shared by a group of ncRNA sequences
  - HMM? Cannot describe secondary structure
  - Context-free grammar

```
S Æ AS₁U | US₁A;
S₁Æ CS₂G | GS₂C | AS₂U;
S₂Æ GS₃C;
S₃Æ AS₄; S₄Æ GS₅.
```

- Stochastic context-free grammar (SCFG) [Eddy&Durbin '94]
  - Probability that a sequence $s$ is generated by a SCFG $M$: CYK algorithm $O(|s|^3|M|)$
NcRNA Annotation-Case Study

- Building full alignments in Rfam (500+)
  - Use BLAST to find initial set of family members [Eddy 02]

- Human gut microbiome initiative (HGMI)
  - 100 bacterial genomes: http://www.genome.wustl.edu/hgm
  - Compare 100 genomes to Rfam (500+): CPU years
Secondary Structure Profile (SSP)

Two components: seeds & profile

- Use multiple seeds to accommodate insertion and deletion
  - Compute length distribution inside a base pair
Data and Experiments

- NcRNA family database Rfam
  - Over 500 ncRNA families
  - Genomic database: 65 Mbase, containing human, mouse, bacterial genome, etc.
- Design SSP-based filter for each ncRNA family
- Evaluate SSP’s sensitivity and speedup
Filter Comparison on tRNA

- Real application: find tRNAs in E.Coli
  - Infernal: 8.23 hours
  - Using SSPs: sensitivity = 1.0; time = 3.3 minutes
Choose between SSP and Regular Profile

- Many ncRNA families have strong sequence conservation
  - Regular profiles have better speedups

  Design regular profiles for all ncRNA families;
  If (the theoretical sensitivity < a pre-determined threshold)
  Design SSP for this ncRNA family;

- 13 out of 233 ncRNA families choose SSP
Histogram of Sensitivity and Speedup of SSP

Sensitivity for 233 ncRNA families
84%: sensitivity $\geq 0.99$
98.7%: sensitivity $\geq 0.9$

Speedup for 88 random ncRNA families on 1M synthetic data set
average $T_{\text{cmsearch}} = 8701$ seconds
average speedup = 222x
Research Summary

- Similarity search in large scale database, pattern design, pattern match, sequence alignment
  - [Buhler, Keich, and Sun] RECOMB ’03
  - [Sun, Buhler] RECOMB ’04
  - [Sun, Buhler] Journal of Computational Biology ’05
  - [Sun, Buhler] BMC Bioinformatics ’06

- Protein sequence annotation, pattern and profile-based protein motif design
  - [Sun, Buhler] ECBB ’06 (Bioinformatics ’07)
  - [Sun, Buhler] IEEE/ACM Transactions on Computational Biology and Bioinformatics ’08

- RNA sequence annotation, RNA motif finding
  - [Sun, Buhler] csb’08
Acknowledgments

- Dr. Jeremy Buhler (advisor)
- Dr. Michael Brent, Dr. Gary Stormo, Dr. Sally Goldman, Dr. Tao Ju, Dr. Nan Lin
- Joe, Octav, Rachel, Lu, and Hongtao
- Funding: NSF DBI-0237902

Thank you!