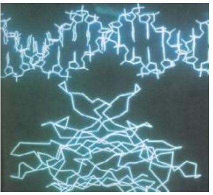


Discovery of non-induced patterns from sequences

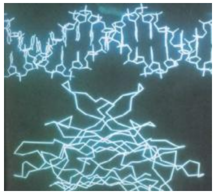
5th IAPR International Conference on Pattern Recognition in Bioinformatics September 23, 2010

Andrew K.C. Wong,
Dennis Zhuang,
Gary C.L. Li
Annie Lee

PAMI Group
University of Waterloo

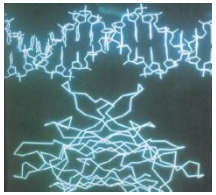
Presentation Outline

- Introduction
- Methodology
  - Statistically significant Patterns
  - Representative pattern
  - Statistically induced patterns
  - Use of Generalize Suffix Tree
  - Removal of statistically induced patterns
  - Algorithm and complexity
- Experiments and Results
- Conclusion



Introduction

- Discovering patterns from sequence data has significant impact in genomics, proteomics and business.
- A common problem encountered:
  A large number of fake patterns are usually induced by their statistically significant sub-patterns
- This paper presents an algorithm to identify and remove redundant patterns to yield a compact succinct set of *statistically significant patterns*.

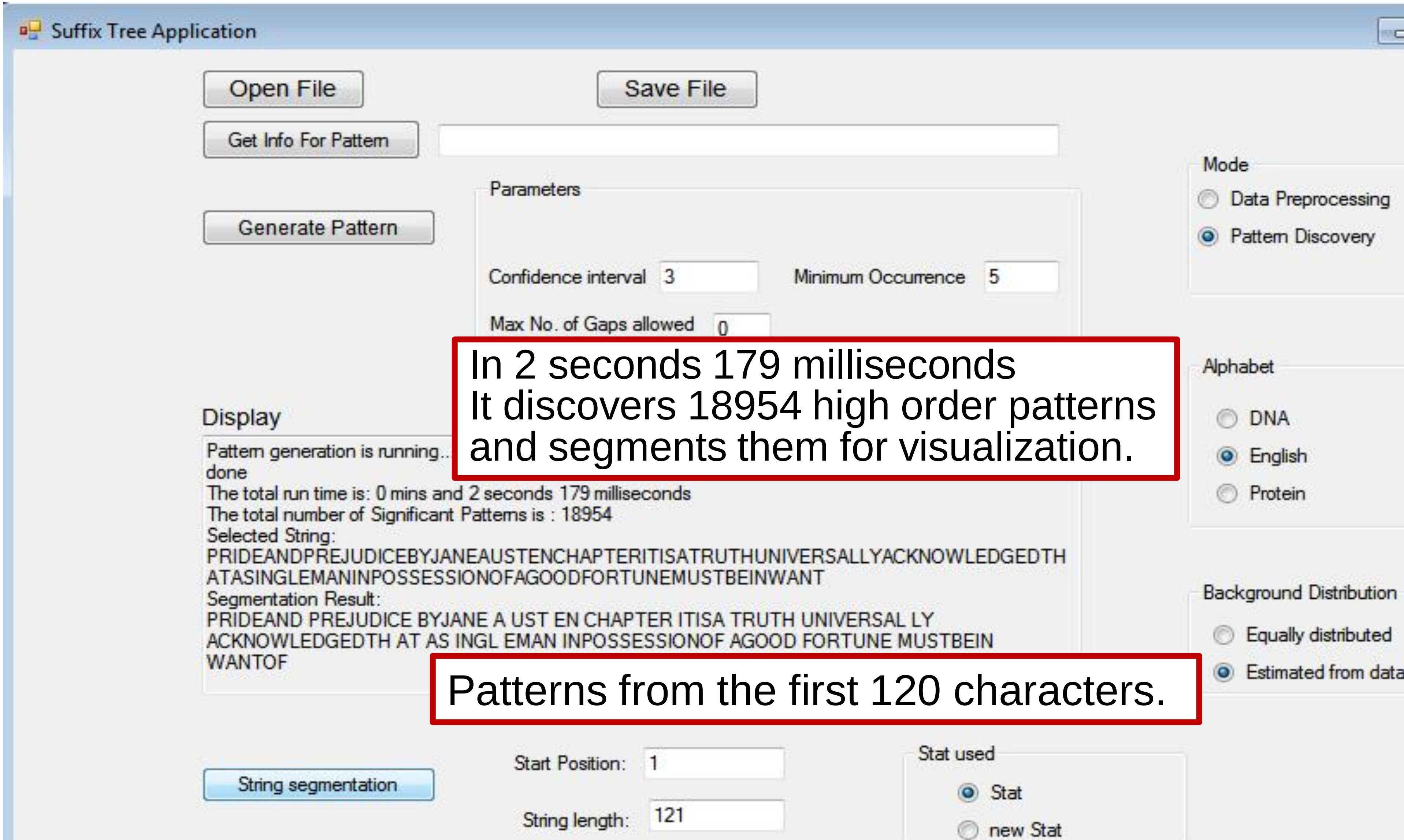


Sequence Patterns

- We define a *Sequence Pattern* as a statistically significant association of characters along a sequence.
- We use *Bernoulli scheme* as the default “random background model” to discover non-statistically induced patterns without relying on prior knowledge or training.
- We develop an algorithm to remove *statistically induced patterns* from their statistically significant sub-patterns.
- We relate the discovered patterns to *functional units* inherent in the biological sequences.

Can meaningful functional units be discovered from sequences ?

We use a text sequence taken from the entire book of "Pride and Prejudice" with punctuations/spaces removed .



Parameters

Confidence interval 3

Minimum Occurence 5

Max No. of Gaps allowed 0

Mode

○ Data Preprocessing
● Pattern Discovery

Alphabet

○ DNA
● English
○ Protein

Background Distribution

○ Equally distributed
● Estimated from data

Display

```
Pattern generation is running...
done
The total run time is: 0 mins and 2 seconds 179 milliseconds
The total number of Significant Patterns is : 18954
Selected String:
PRIDEANDPREJUDICEBYJANEAUSTENCHAPTERITISATRUTHUNIVERSALLYACKNOWLEDGEDTH
ATASINGLEMANINPOSSESSIONOFAGOODFORTUNEMUSTBEINWANT
Segmentation Result:
PRIDEAND PREJUDICE BYJANE A UST EN CHAPTER ITISA TRUTH UNIVERSAL LY
ACKNOWLEDGEDTH AT AS INGL EMAN INPOSSESSIONOF AGOOD FORTUNE MUSTBEIN
WANTOF
```

Start Position: 1

String length: 121

Stat used

● Stat
○ new Stat

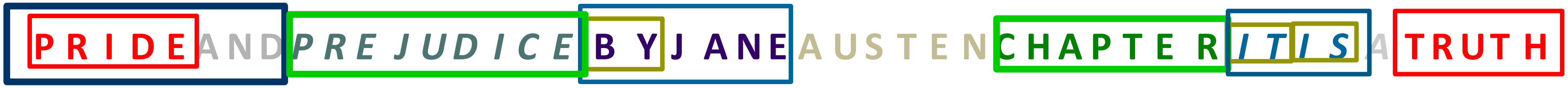
In 2 seconds 179 milliseconds
It discovers 18954 high order patterns and segments them for visualization.

Patterns from the first 120 characters.

Patterns discovered are highlighted in colors.

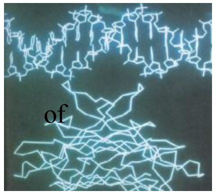
PRIDEANDPREJUDICE BYJANEAUSTENCHAPTERITISATRUTH
UNIVERSALLYACKNOWLEDGEDTHATASINGLEMANIN
PROCESSIONOFAGOODFORTUNEMUSTBEINWANTOF

The segmented patterns mostly correspond to English words and short phrases suggesting underlying language structure and meaning.



PRIDEANDPREJUDICEBYJANEAUSTENCHAPTERITISATRUTH

They are functional units in English Language.



Statistically significant pattern

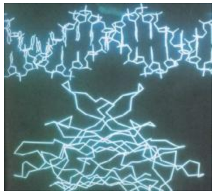
To measure how the frequency k_P of P deviates from its expected random model, we use the ***standard residual***

$$\boxed{z_P = \frac{k_P - E(X_P)}{\sqrt{E(X_P)}}}$$

A pattern is ***statistically significant*** *or over represented* if

$$\boxed{z_P \geq t}$$

where t is a *predefined minimum threshold.*



Methodology

Use of Generalized Suffix Tree

1) to *store multiple strings* data;

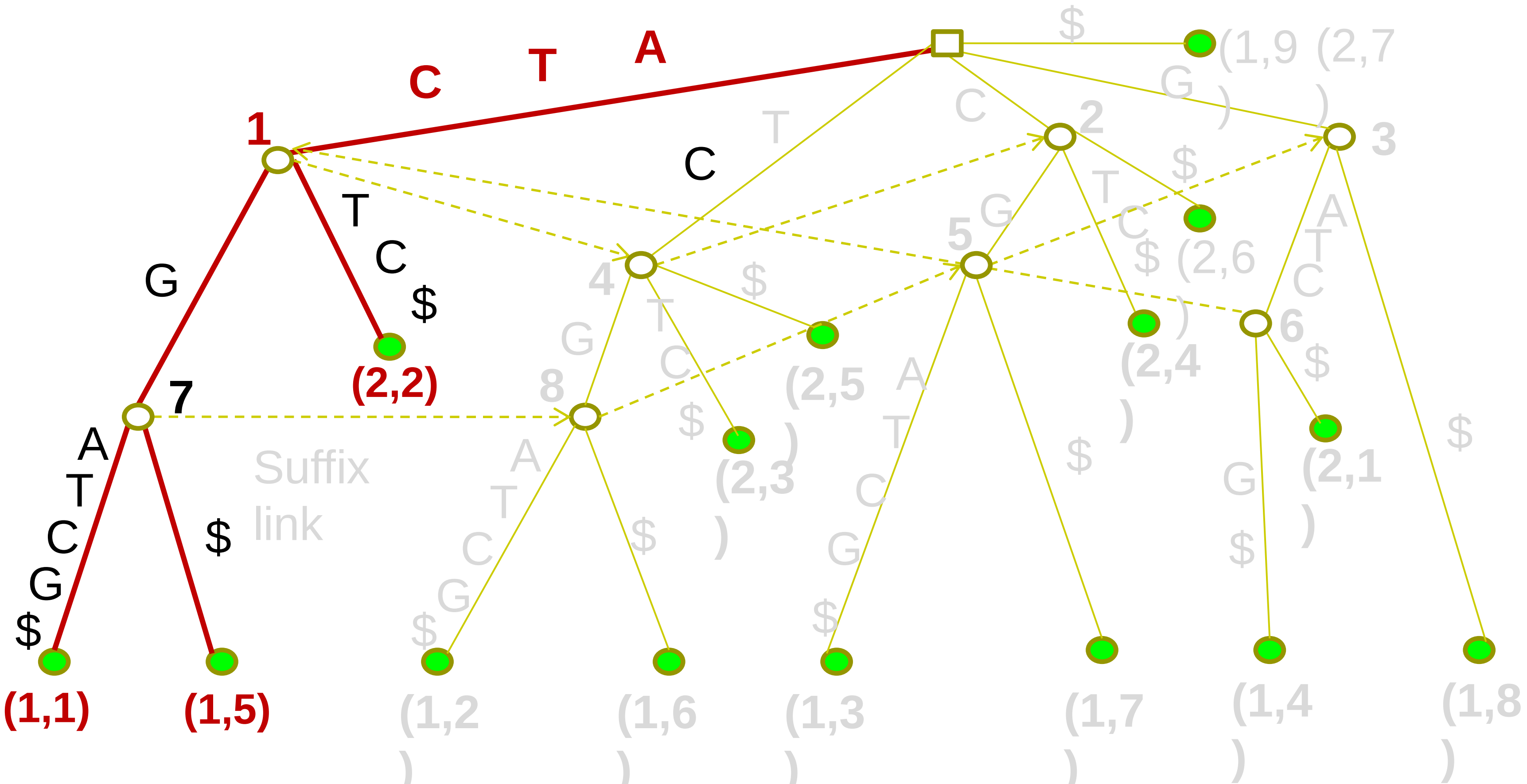
2) to *obtain statistics* for evaluating statistically significant patterns;

3) to use *suffix links* to identify *representative patterns.*

4) to use of *conditional statistical test* to *screen non-induced patterns* from representative patterns and remove *statistically induced redundant patterns.*

Patterns in GST

Position:	1	2	3	4	5	6	7	8	9
1	A	T	C	G	A	T	C	G	$
2	G	A	T	C	T	C	$		



Patterns are ***path labels*** with ***frequency*** $k(x) >= min_{occ}$

ATC occurs at positions **(1,1), (1,5)** and **(2,2)** with **k(x) = 3.**

It is then a ***significant pattern*** and can be found from the **GST.**

Statistically induced pattern

Let $X_P = \sum_i X_i$ be a random variable with binominal distribution and P as one of its outcomes. Let *P'* be a subpattern of *P*. The ***conditional statistical significance*** of *P given P'* is defined as:

$$z_{P|P'} = \frac{k_P - E(X_P|P')}{\sqrt{E(X_P|P')}}$$

where

$$E(X_P|P') = pr(P|P') \cdot k_{P'} = \frac{pr(P)}{pr(P')} \cdot k_{P'}$$

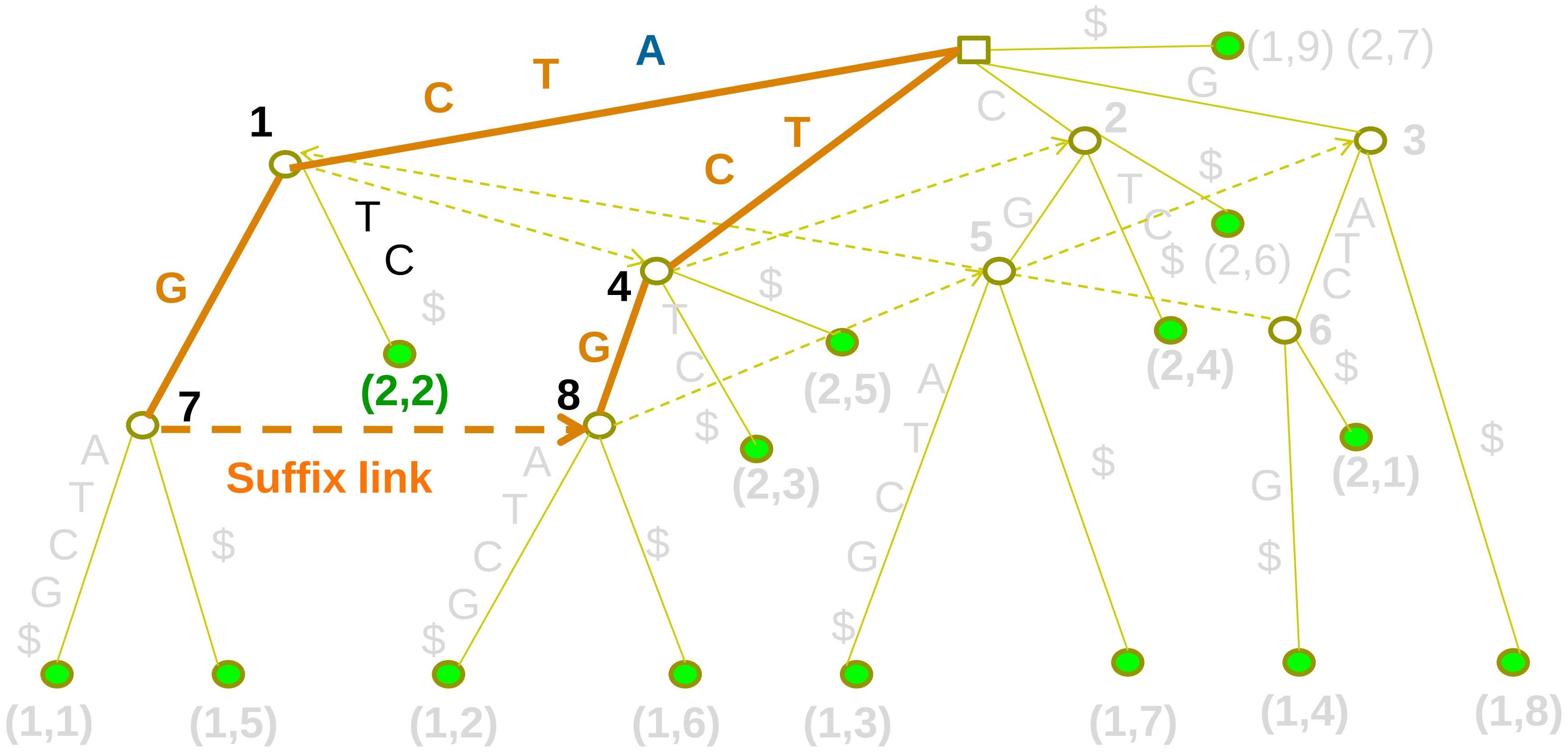
Given a set of significant representative patterns, a pattern in it is said to be ***statistically induced*** if there exists a *proper* subpattern *P'* of *P* such that

$$z_{P|P'} < t$$

Finding of Non-Statistically Induced Patterns

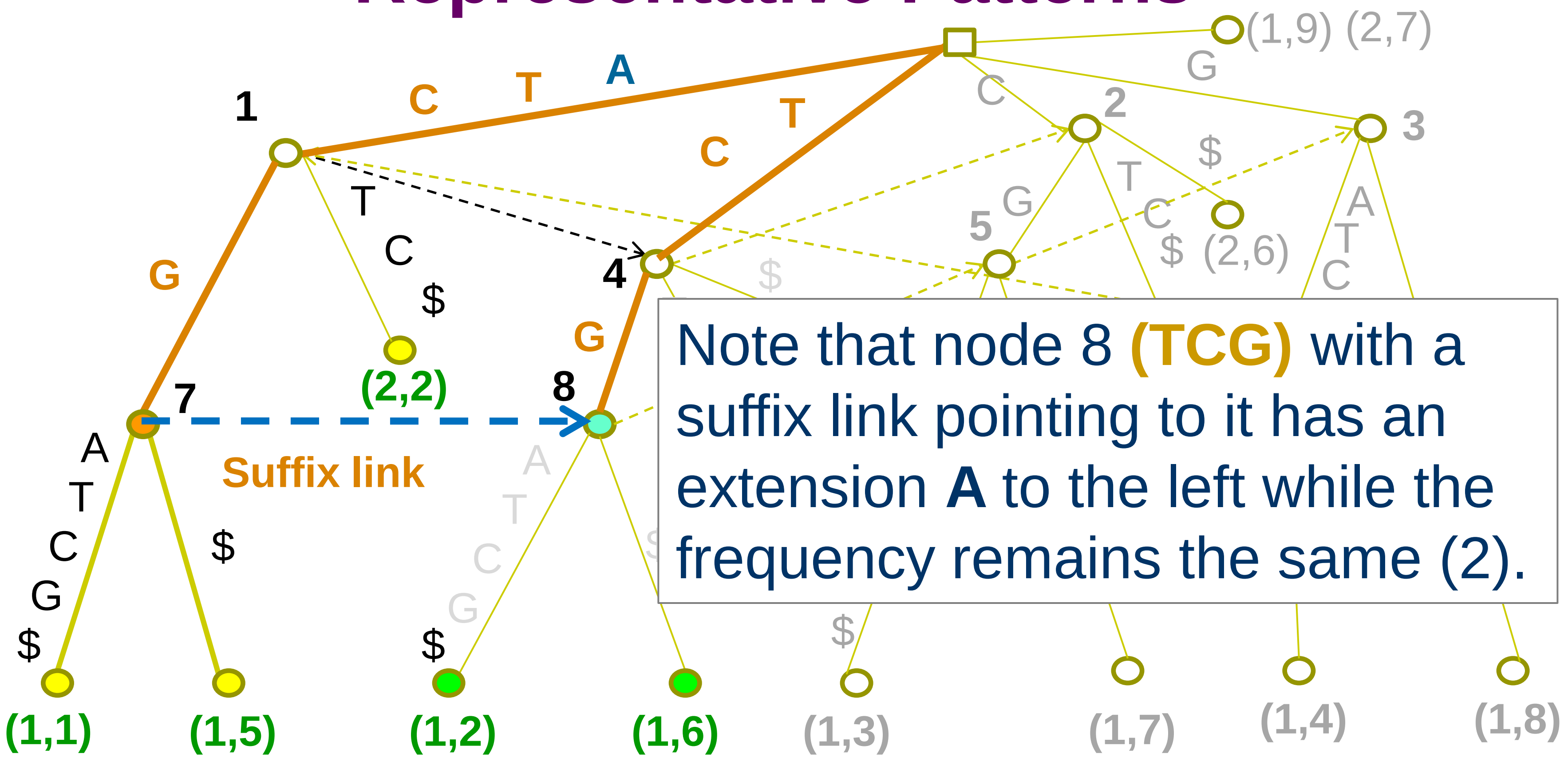
- Patterns having the same list of occurrence positions can be labeled as an equivalent group.
- *A representative patterns* in such group is one having highest statistical significance or order, i.e. *it cannot be extended either to the left or right without decreasing its frequency.*
- They can be found from the GST effectively by *suffix links .*
- Representative patterns passing the statistical significant test are called *significant representative patterns*

Suffix Link



Suffix link (node **7** to node **8**) indicates that the path label at **8** **(TCG)** is the second suffix of the path label **(ATCG)** at **7.**

Representative Patterns

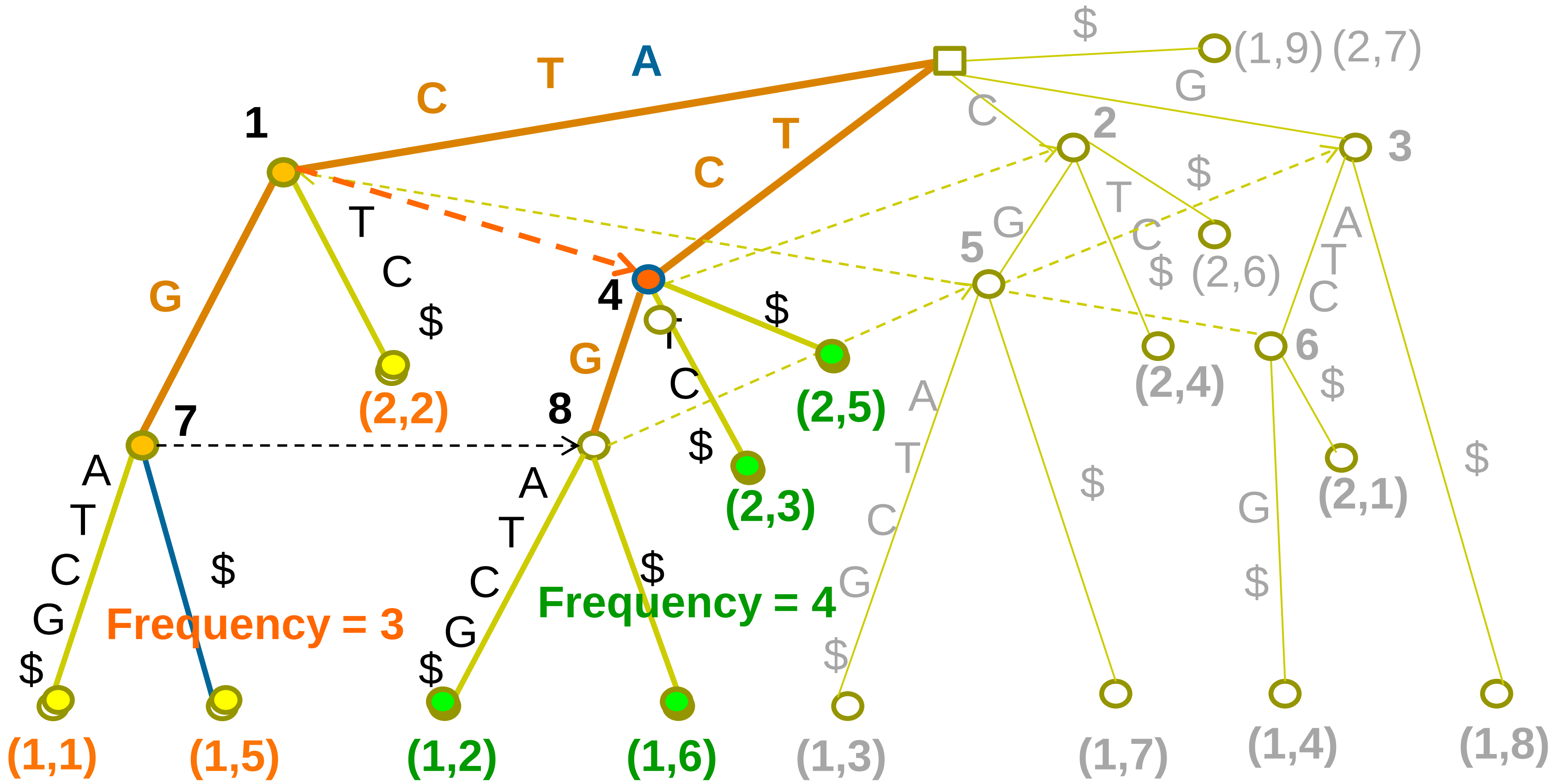


Note that node 8 **(TCG)** with a suffix link pointing to it has an extension **A** to the left while the frequency remains the same (2).

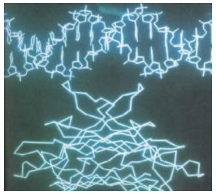
A representative pattern **(**say **ATCG** at **7)** is the path label of an internal node (**7**) such that:

1) no suffix link is pointing to it. (i.e no extension to the left; right extension could only have lower frequency).

Methodology:
representative patterns



2) with suffix link pointing to it, say **(T C)** at node **4,** yet whose position frequency **(i.e. 4)** is > that **(i.e. 3)** of its left hand extension (path label **ATC** at node **1**).



Removal of Fake Statistical patterns

- *Conditional statistical significance* is used to evaluate how strongly the statistical significance of a pattern is *attributed by* the occurrences of one of its proper *sub-patterns*.
- Patterns whose statistical significances are due to their strong proper sub-patterns by *mere chance* are *fake patterns*.

Now, from the representative patterns identified, we could apply the *conditional significance test* $z_{P|P'} < t$ to *screen out the fake patterns* so as to render a more succinct set of patterns.

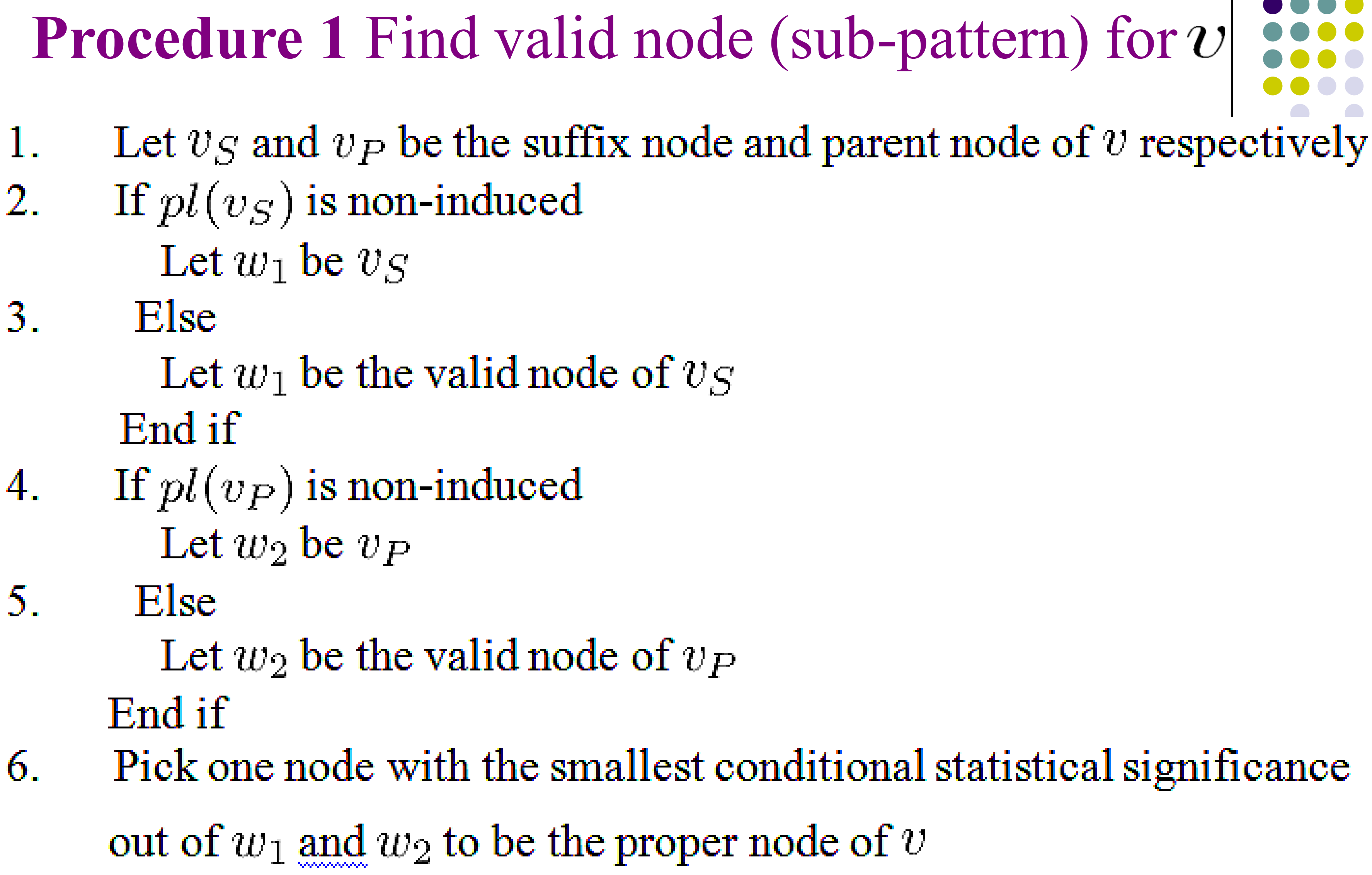
Algorithm 1 Discovery of non-induced patterns

```
[1] Construct a GST $T$ for the input sequences
[2] Annotate $k(v)$the number of positions under each of$v$ $T$
[3] Extract a set of nodes whose $k(v) \geq min_{occ}$
[4] Sort the above nodes in ascending order according to
    order of $pl(v)$ using counting sort.
[5] For each node $v$
    [a] Find the valid node$w$ (sub-pattern) of $v$ using **Procedure 1**
    [b] If $v$ is not a suffix node and $z_{pl(v)} \geq t$ ( i.e.$pl(v)$ is not
            induced by $pl(w)$ )
       Output $pl(v)$
                End if
       End for.
```

PRESNET ONLY THE BLOCKS

Algorithm

v



Procedure 1 Find valid node (sub-pattern) for v

```
1.  Let $v_S$ and $v_P$ be the suffix node and parent node of $v$ respectively
2.  If $pl(v_S)$ is non-induced
        Let $w_1$ be $v_S$
3.  Else
        Let $w_1$ be the valid node of $v_S$
    End if
4.  If $pl(v_P)$ is non-induced
        Let $w_2$ be $v_P$
5.  Else
        Let $w_2$ be the valid node of $v_P$
    End if
6.  Pick one node with the smallest conditional statistical significance
    out of $w_1$ and $w_2$ to be the proper node of $v$
```

Running time analysis for Algorithm 1

Step 1-3 (*search for frequent nodes*) achieved in linear time.

Step 4 using counting sort to *sort the nodes according to the path length*, done in linear time.

Steps 5a and 5b (*finding sub-patterns and checking non-induced condition*) take constant time. Hence, Step 5 can be done in linear time.

Therefore, non-induced patterns can be found in linear time.

Experiment on Synthetic Data

100 random sequences of 1000 bases each over the DNA alphabet are created.

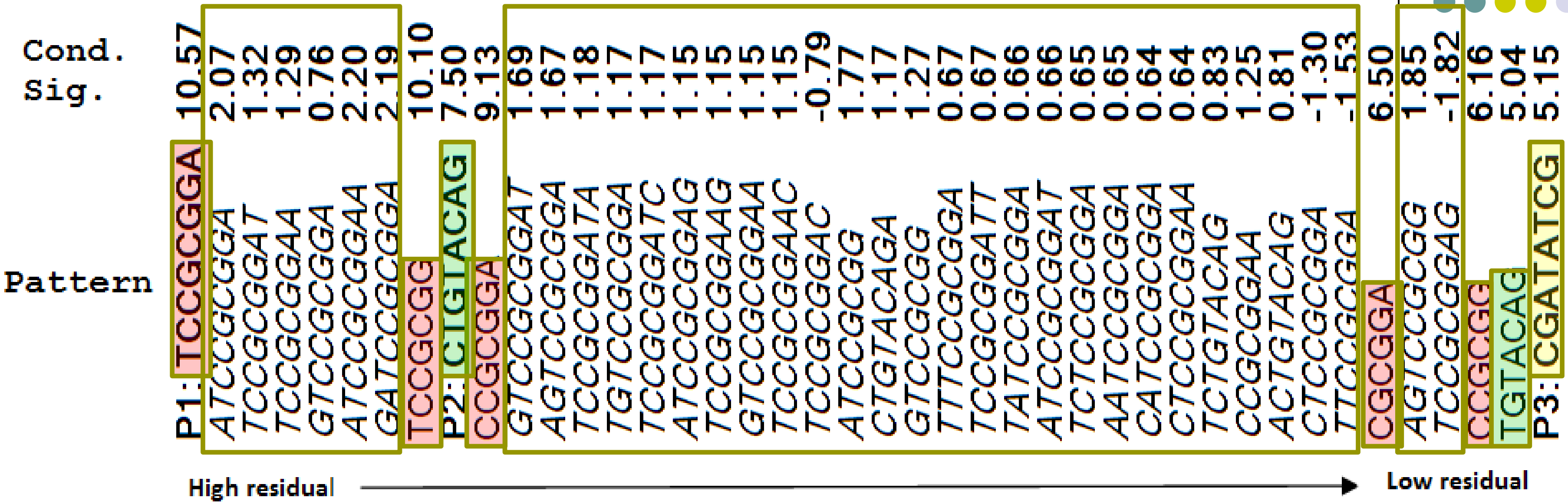
Three strong patterns

$$P_1 = \mathrm{TCCGCGGA}$$
$$P_2 = \mathrm{CTGTACAG}$$
$$P_3 = \mathrm{CGATATCG}$$

are implanted ∋ their standard residuals are 48, 24, and 12 respectively.

We apply the first method to *discover significant representative patterns* and the second method to *obtain non-induced patterns*.

Patterns ranked by standard residual



Pattern	Cond. Sig.
P1: TCCGCGGA	10.57
ATCCGCGGA	2.07
TCCGCGGAT	1.32
TCCGCGGAA	1.29
GTCCGCGGA	0.76
ATCCGCGGAA	2.20
GATCCGCGGA	2.19
TCCGCG	10.10
P2: CTGTACAG	7.50
CCGCGGA	9.13
GTCCGCGGAT	1.69
AGTCCGCGGA	1.67
TCCGCGGATA	1.18
TGTCCGCGGA	1.17
TCCGCGGATC	1.17
ATCCGCGGAG	1.15
TCCGCGGAAG	1.15
GTCCGCGGAA	1.15
TCCGCGGAAC	1.15
TCCGCGGAC	-0.79
ATCCGCGG	1.77
CTGTACAGA	1.17
GTCCGCGG	1.27
TTCCGCGGA	0.67
TCCGCGGATT	0.67
TATCCGCGGA	0.66
ATCCGCGGAT	0.66
TCTCCGCGGA	0.65
AATCCGCGGA	0.65
CATCCGCGGA	0.64
CTCCGCGGAA	0.64
TCTGTACAG	0.83
CCGCGGAA	1.25
ACTGTACAG	0.81
CTCCGCGGA	-1.30
TTCCGCGGA	-1.53
CGCGGA	6.50
AGTCCGCGG	1.85
TCCGCGGAG	-1.82
CCGCG	6.16
TGTACAG	5.04
P3: CGATATCG	5.15

High residual → Low residual

- Patterns in italic are *super patterns induced by P1, P2* and *P3* (with conditional significance < 3) and hence removed.
- Note that P3 is raised from the **42th** to the **7th.**
- The colored ones are all patterns/sub-patterns of binding sites.
- # of patterns reduced: from 527 to 315 (40% reduction rate).
- A more compact set of patterns is obtained.

Experiment on Transcription Factor Binding Sites

- Objective: to discover biological functional units such as transcription factor (TF) binding sites on Yeast (SCPD database) from the upstream promoter regions of genes.
- Conditions for choosing the 18 regulons:

(1) the number of genes at least 3 in the DB

(2) the consensus binding sites are available

(3) only consecutive patterns are chosen.

TF Ranking score

- To find TF amongst multiple sequences, patterns with higher support are more important than those with less. Hence, We use a combined score defined as:

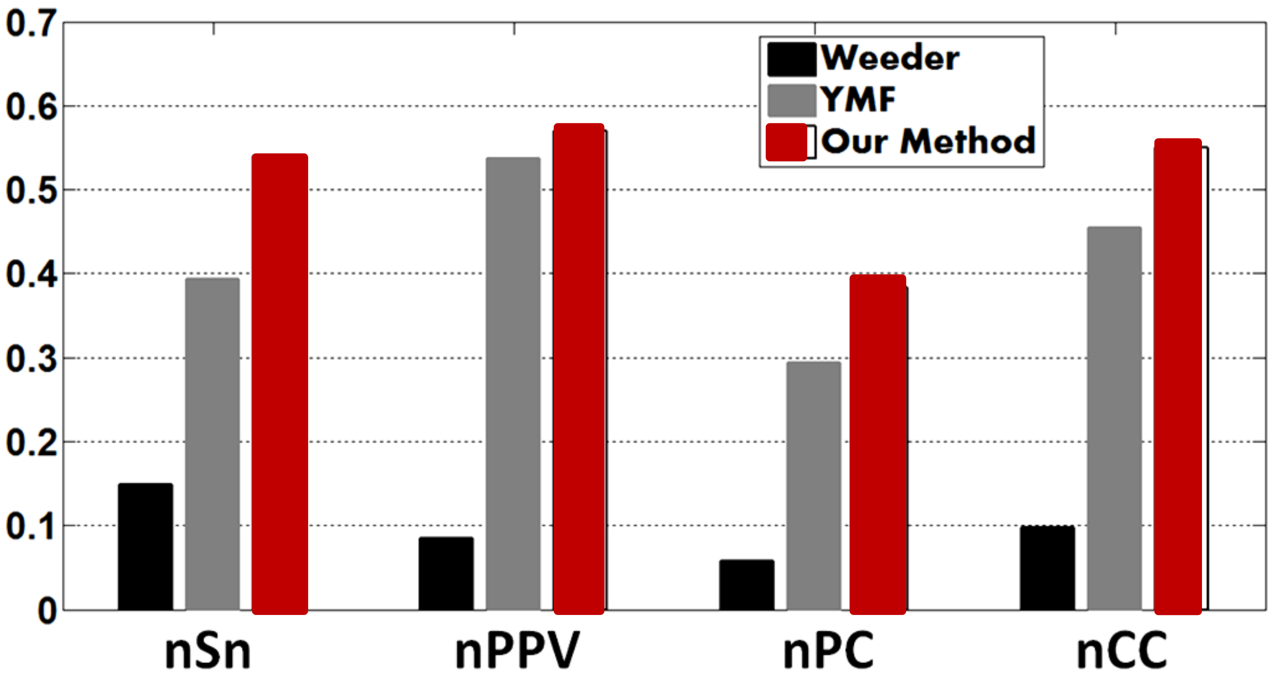
$$score = \frac{\text{Support}}{\text{No. of genes}} \cdot \text{Standard residual}$$

- For repeated sequences like AAAAAATTTTTT, (AAAA occurs at positions 1, 2 and 3 which overlap multiple times) are removed.

Results and Comparison

- We discovered the non-induced patterns for each dataset and compared the results with YMF and Weeder.
- For comparison, we use the measures
  - nSn (sensitivity)
  - nPPV (positive preditive value)
  - nPC (performance coefficient)
  - nCC (correlation coefficient)

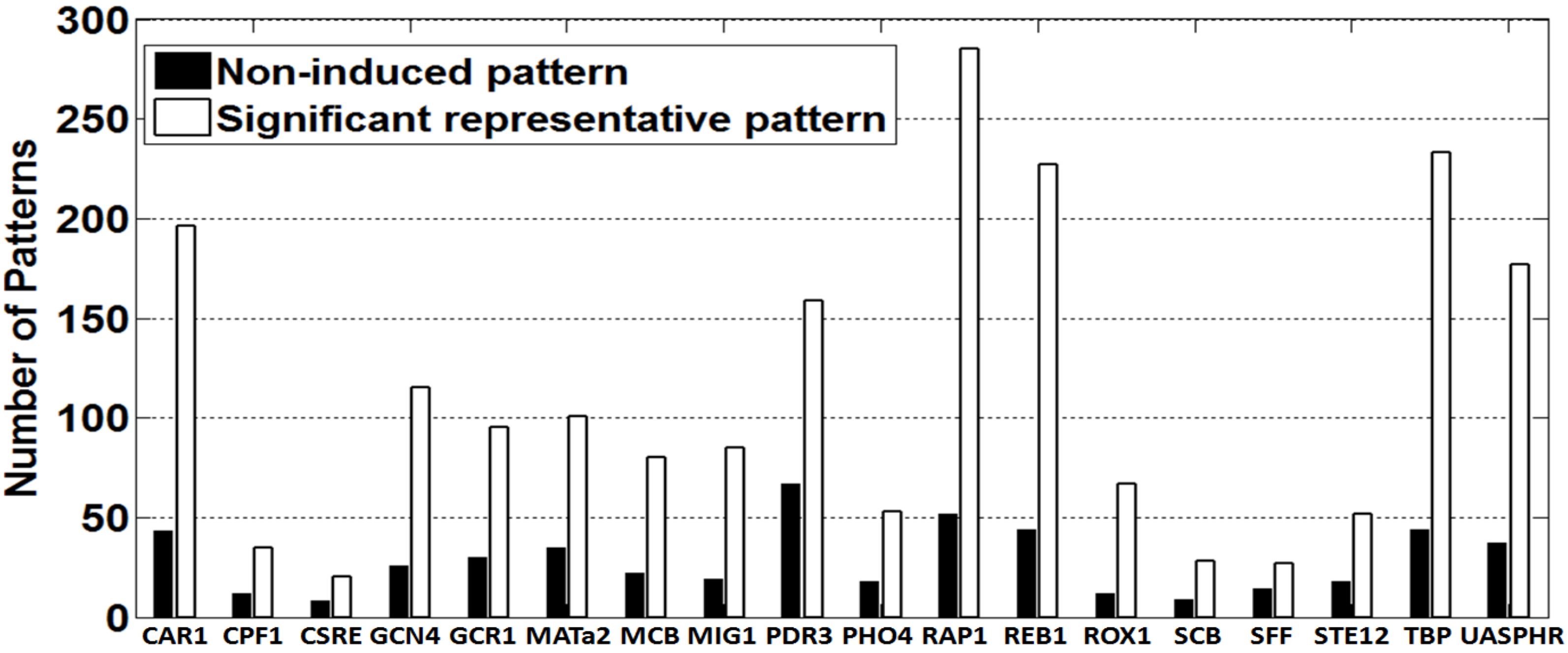
Results Comparison



TF		Motif/Pattern	nSn	nPPV	nPC	nCC
Average	Weeder		0.16	0.09	0.05	0.09
	YMF		0.48	0.51	0.37	0.47
	Our Method		**0.54**	**0.65**	**0.44**	**0.56**

Non-induced patterns vs significant representative patterns

- After removing induced patterns, a relative small set of non-induced patterns are retained (from 8 to 67), showing that our method is able to retain patterns associated with conserved functional units.



Note the significant reduction !

Discovering Patterns in Biological Sequences

- Among the 18 datasets, the top 13 patterns discovered by our method match the consensus binding sites in 14 datasets, and 4 ranked top.
- For the remaining four we missed:
  - The consensus binding sites of CPF1, CSRE and SFF consensus binding sites have fewer than 2 occurrences.
  - Though the consensus of MATalpha2 has 6 occurrences, but it has many substitutions that couldn't be handled by our present algorithm which is targeted on consecutive patterns.

Discovery of Transcription Factor Binding Sites

TF(s)	Discovered pattern	Rank	Consensus	Pattern #	Pattern # before pruning	Reduction rate
CAR1	AGCCGCC	6	**AGCCGCC**[GA]	57	213	0.74
CPF1 →	N/A (# ≤ 2)		TCACGTG	13	36	0.64
CSRE →	N/A (# ≤ 2)		[TC]CGGA[TC][GA][GA]A[AT]GG	8	20	0.6
GCN4	TGACT	13	**TGA**N**T**N	30	124	0.76
GCR1	CTTCCT	6	**C**[A**T**]**TCC**	34	100	0.66
MATalpha2	N/A *		C[GA]TGT[AT][AT][AT][AT]	37	105	0.63
MCB	ACGCGT	1	[**A**T]**CGCG**[A**T**]	26	86	0.7
MIG1 →	CCCCAG	2	**CCCC**[G**A**]NN[AT][AT][AT][AT][AT]	21	90	0.77
PDR3	TCCGCGGA	1	**TCCG**[**C**T]**GGA**	66	166	0.61
PHO4	ACGTG	1	C**ACGT**[T**G**]	17	53	0.68
RAP1	CACCCA	9	[G][A**C**]**ACCCA**	59	309	0.81
REB1	TTACCCG	7	[**T**C][**T**C]**ACCCG**	48	246	0.81
ROX1	ATTGTT	6	[TC][TC]N**ATTGTT**[TC]	14	71	0.81
SCB	CACGAAA	1	**C**N**CGAAA**	10	29	0.66
SFF →	N/A (# ≤ 2)		GT[AC]AACAA	15	29	0.49
STE12	TGAAACAA	1	**TGAAACA**	21	56	0.63
TBP	ATATAAA	14	T**ATA**[AT]**A**[**A**T]	54	252	0.79
UASPHR	TCTTCC	1	**CTTCCT**	42	188	0.78

Conclusion

- We have presented an efficient algorithm to discover non-induced patterns from a large sequence data.
- It uses a Generalized Suffix Tree to assist the identification of significant representative patterns and the removal of the fake statistically induced patterns
- By ensuring that each pattern discovered is non-induced, it produces a more compact pattern set.
- The Transcription Factor binding sites among its top ranking patterns confirm its ability to acquire a small set of patterns revealing interesting, unknown information inherent in the sequences.

Conclusion and Future Work

- While the algorithm drastically reduces the pattern size, it is still able to retain patterns associated with conserved functional units in the promoter regions without relying on any prior knowledge.
- Our future work will advance in the following directions:

  - Extending our method to discover patterns with gaps;

  - Discovering distance patterns in DNA, RNA and protein sequences and relating them to 3-D conformation and sequence pattern synthesis.

明月幾時有

Thank You

and

Wishing You All

A Happy Mid-Autumn Festival