High Confidence Rule Mining for Microarray Analysis

Kang Deng
University of Alberta

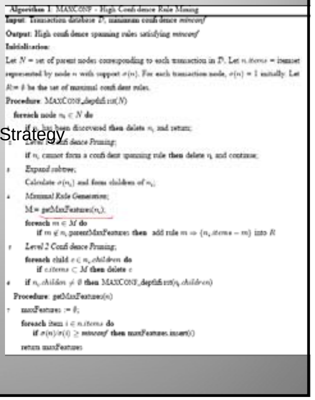
Outline

- Introduction
- Row Enumeration
- Confidence-based Prune Strategy
- MAXCONF Algorithm
- Evaluation
- References

“High Confidence Rule Mining for Microarray Analysis”, by Tara McIntosh, Sanjay Chawla, 2006

- 27 pages
- 14 definitions
- 2 lemmas
- 4 tables
- Figures, formulas, etc

Confidence-based Strategy



Algorithm 1: MAXCONF - High Confidence Rule Mining

Input: Transaction database D, minimum confidence *minconf*

Output: High confidence spanning rules satisfying *minconf*

Initialization:

Let N = set of parent nodes corresponding to each transaction in D. Let $n.items$ = itemset represented by node n with support $\sigma(n)$. For each transaction node, $\sigma(n) = 1$ initially. Let $R = \emptyset$ be the set of maximal confident rules.

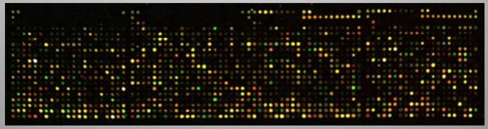
```
Procedure: MAXCONF_depthfirst($N$)
  foreach node $n_i \in N$ do
      if $n_i$ has been discovered then delete $n_i$ and return;
      Level 1 Confidence Pruning;
      if $n_i$ cannot form a confident spanning rule then delete $n_i$ and continue;
      Expand subtree;
      Calculate $\sigma(n_i)$ and form children of $n_i$;
      Minimal Rule Generation;
      M = getMaxFeatures($n_i$);
      foreach $m \in M$ do
          if $m \notin n_i$.parentMaxFeatures then add rule $m \Rightarrow \{n_i.items - m\}$ into $R$
      Level 2 Confidence Pruning;
      foreach child $c \in n_i.children$ do
          if $c.items \subseteq M$ then delete $c$
      if $n_i.children \neq \emptyset$ then MAXCONF_depthfirst($n_i.children$)
  Procedure: getMaxFeatures($n$)
    maxFeatures := $\emptyset$;
    foreach item $i \in n.items$ do
        if $\sigma(n)/\sigma(i) \geq minconf$ then maxFeatures.insert($i$)
    return maxFeatures
```

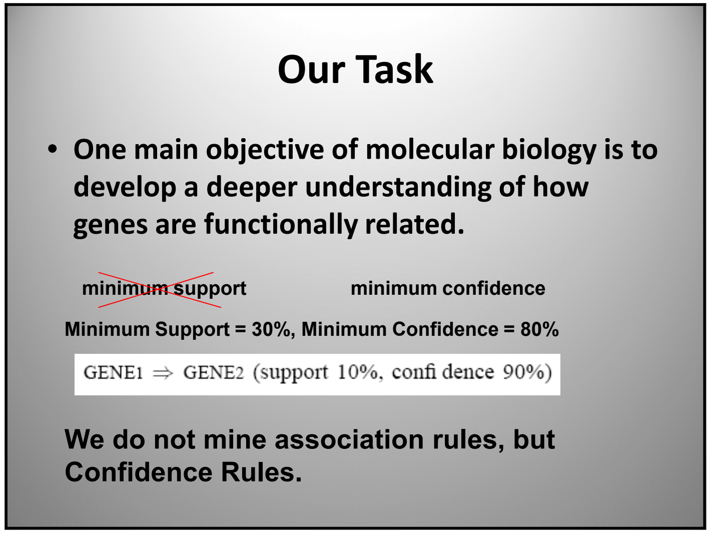
What is Microarrays?

- A DNA microarray is a collection of microscopic DNA spots, commonly representing single genes, arrayed on a solid surface by covalent attachment to a chemical matrix.

~~genes~~ items

~~samples~~ transactions





Our Task

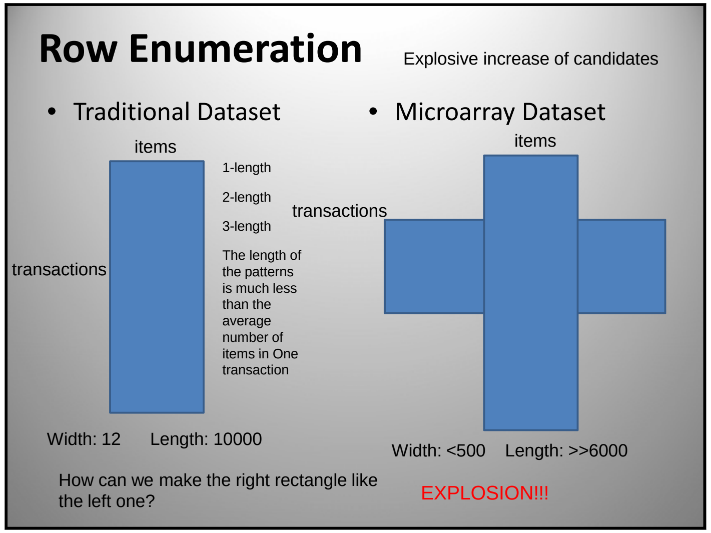
- **One main objective of molecular biology is to develop a deeper understanding of how genes are functionally related.**

~~minimum support~~ **minimum confidence**

Minimum Support = 30%, Minimum Confidence = 80%

GENE1 ⇒ GENE2 (support 10%, confi dence 90%)

We do not mine association rules, but Confidence Rules.



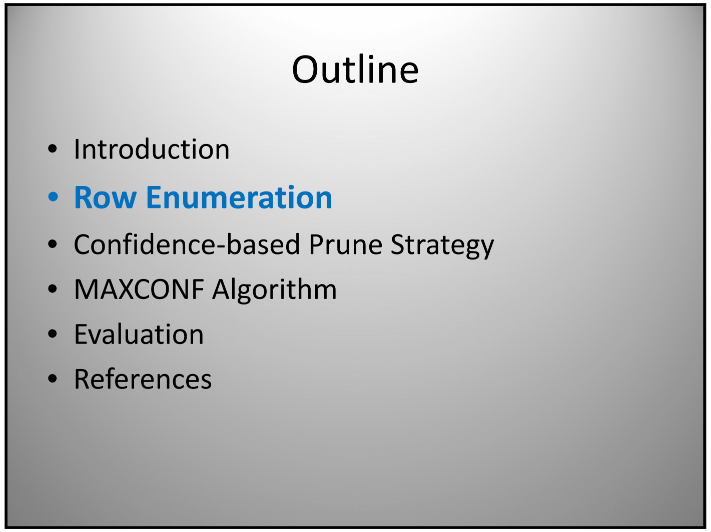
Row Enumeration

Explosive increase of candidates

- Traditional Dataset
- Microarray Dataset

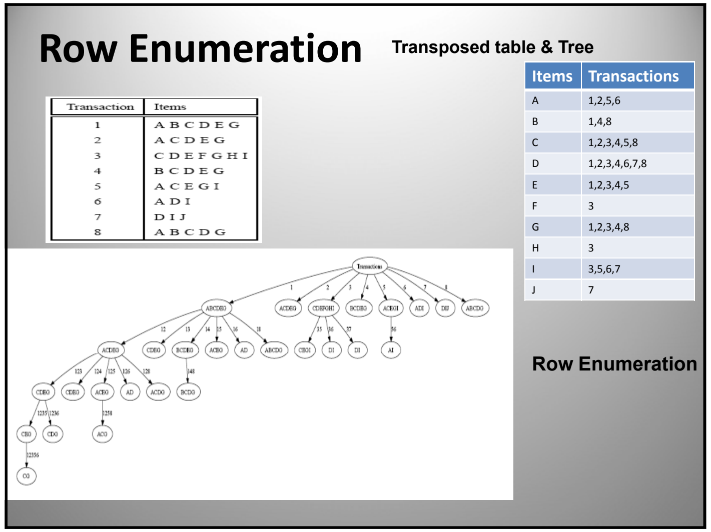
How can we make the right rectangle like the left one?

EXPLOSION!!!



Outline

- Introduction
- **Row Enumeration**
- Confidence-based Prune Strategy
- MAXCONF Algorithm
- Evaluation
- References



Row Enumeration

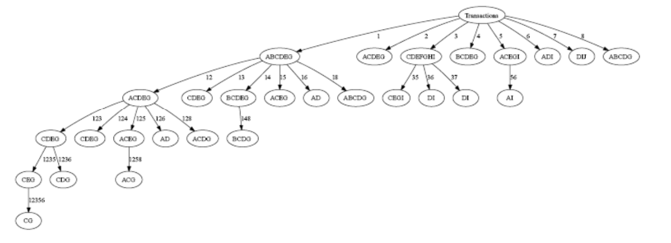
Transposed table & Tree

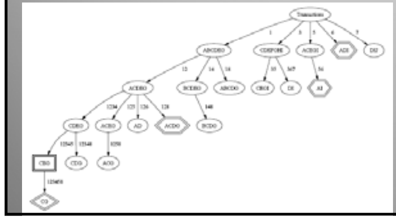
Transaction	Items
1	A B C D E G
2	A C D E G
3	C D E F G H I
4	B C D E G
5	A C E G I
6	A D I
7	D I J
8	A B C D G

Items	Transactions
A	1,2,5,6
B	1,4,8
C	1,2,3,4,5,8
D	1,2,3,4,6,7,8
E	1,2,3,4,5
F	3
G	1,2,3,4,8
H	3
I	3,5,6,7
J	7

Row Enumeration

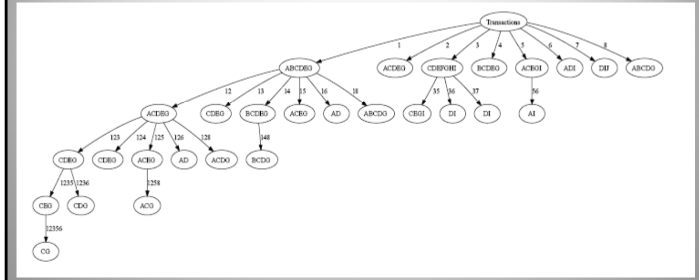
Row Enumeration Tree





If the current parent node n, is completely contained within a sibling node, a child node is not constructed.
For Example, node 2.

Confidence-based Strategy



RER II, "Mining frequent closed patterns in microarray data." by G. Cong, K.-L. Tan, A. Tung, and F. Pan, 2004

Support-based pruning strategy

Minimum Support = 30%, Minimum Confidence = 80%

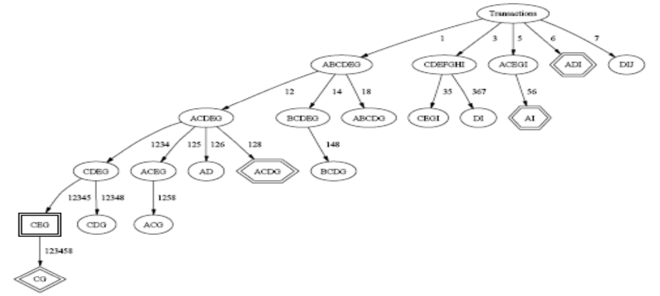
GENE1 ⇒ GENE2 (support 10%, confi dence 90%)

In Biology, we care confidence rules, but not support

Outline

- Introduction
- Row Enumeration
- **Confidence-based Prune Strategy**
- MAXCONF Algorithm
- Evaluation
- References

Confidence-based Strategy Prune #1

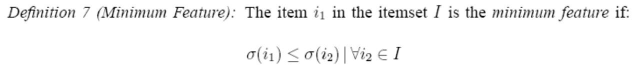


Definition 5 (Maximum Support [4]): Given a node n with k sibling nodes, the *maximum support* of the itemset at n, represented as $\sigma_{max}(n)$, or any of n's potential child nodes is:

$$\sigma_{max}(n) = n.\text{initial\_support} + k \quad (5)$$

$$\sigma_{max}(5) = 1 + 2 = 3$$

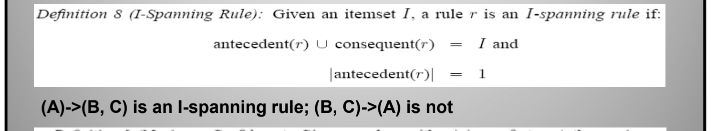
Confidence-based Strategy Prune #1



Definition 7 (Minimum Feature): The item i_1 in the itemset I is the *minimum feature* if:

$$\sigma(i_1) \leq \sigma(i_2) \mid \forall i_2 \in I$$

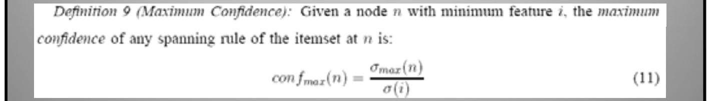
In the itemset {A,B,C}, Support(A)<=Support(B), Support(A)<=Support(C)
So, A is the minimum feature in {A,B,C}



Definition 8 (I-Spanning Rule): Given an itemset I, a rule r is an *I-spanning rule* if:

$$\begin{aligned} \text{antecedent}(r) \cup \text{consequent}(r) &= I \text{ and} \\ |\text{antecedent}(r)| &= 1 \end{aligned}$$

(A)->(B, C) is an I-spanning rule; (B, C)->(A) is not



Definition 9 (Maximum Confidence): Given a node n with minimum feature i, the *maximum confidence* of any spanning rule of the itemset at n is:

$$conf_{max}(n) = \frac{\sigma_{max}(n)}{\sigma(i)} \quad (11)$$

Confidence-based Strategy Prune #1

$(I) \rightarrow (ACEG)$ **This rule has the highest confidence**

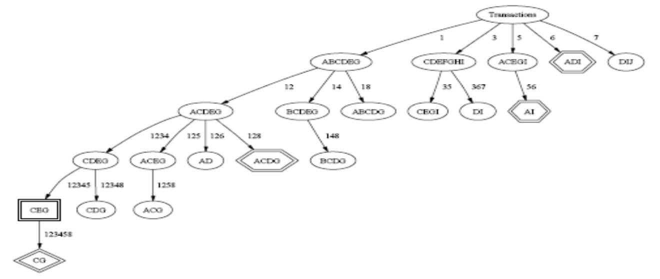
$(AI) \rightarrow (CEG)$ **What about this one?** √

Itemset becomes larger, the support of it will not change or even become smaller

$$\sigma(AI) \leq \sigma(I)$$

$$\uparrow \; confidence = \frac{\sigma(itemset)}{\sigma(antecedent)} \; \downarrow$$

Confidence-based Strategy Prune #1



Maximum Support of 5 $\sigma_{max}(5) = 1 + 2 = 3$

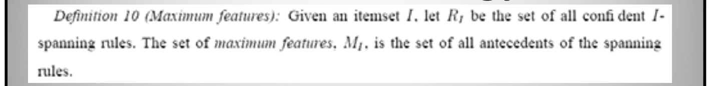
Minimum Feature in this itemset is I $\sigma(I) = 4$

Maximum Confidence of 5: $conf_{max}(5) = \frac{\sigma_{max}(5)}{\sigma(I)} = 3/4$

If minimum confidence is 4/5, the child of node #5 will be pruned

$\sigma(antecedent) \downarrow \qquad \sigma(I) \uparrow$

Confidence-based Strategy Prune #2



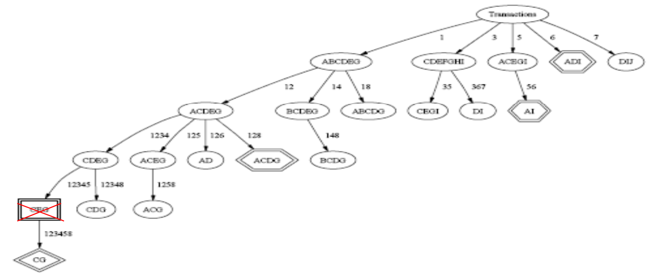
Definition 10 (Maximum features): Given an itemset I, let R_I be the set of all confident I-spanning rules. The set of *maximum features*, M_I, is the set of all antecedents of the spanning rules.

Itemset: {CDEG} $C \rightarrow DEG, E \rightarrow CDG, G \rightarrow CDE$

The maximum feature of CDEG is CEG

Prune Strategy #2:
If maximum feature set M of an itemset at node n is not empty, we can prune all child nodes of n whose itemsets are subsets of M.

Confidence-based Strategy Prune #2

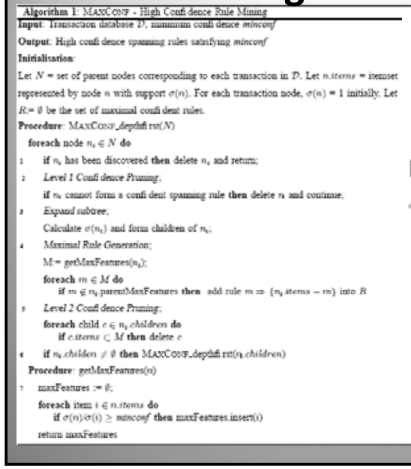


Itemset: (1234){CDEG} $C \to DEG, E \to CDG, G \to CDE$

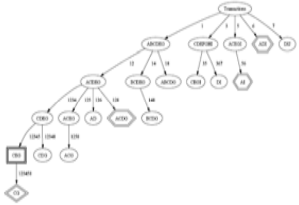
The maximum feature of CDEG is CEG

Node (12345)generates: $C \to EG, E \to CG, G \to CE$ **Sub-rules**

MAXCONF Algorithm



```
Algorithm 1: MAXCONF - High Confidence Rule Mining
Input: Transaction database $D$, minimum confidence $minconf$
Output: High confidence spanning rules satisfying $minconf$
Initialisation:
Let $N$ = set of parent nodes corresponding to each transaction in $D$. Let $n.items$ = itemset
represented by node $n$ with support $\sigma(n)$. For each transaction node, $\sigma(n) = 1$ initially. Let
$R := \emptyset$ be the set of maximal confident rules.
Procedure: MAXCONF_depthfirst($N$)
  foreach node $n_t \in N$ do
      if $n_t$ has been discovered then delete $n_t$ and return;
      Level 1 Confidence Pruning;
      if $n_t$ cannot form a confident spanning rule then delete $n_t$ and continue;
      Expand subtree;
      Calculate $\sigma(n_t)$ and form children of $n_t$;
      Maximal Rule Generation;
      M := getMaxFeatures($n_t$);
      foreach $m \in M$ do
          if $m \notin n_t$.parentMaxFeatures then  add rule $m \Rightarrow \{n_t.items - m\}$ into $R$
      Level 2 Confidence Pruning;
      foreach child $c \in n_t.children$ do
          if $c.items \subseteq M$ then delete $c$
      if $n_t.childen \neq \emptyset$ then MAXCONF_depthfirst($n_t.children$)
  Procedure: getMaxFeatures(n)
    maxFeatures := $\emptyset$;
    foreach item $i \in n.items$ do
        if $\sigma(n)/\sigma(i) \geq minconf$ then maxFeatures.insert(i)
    return maxFeatures
```



Pruning #2

Itemset: (1234){CDEG}

$conf_{max}(5) = \frac{\sigma_{max}(5)}{\sigma(I)} = 3/4$

$C \to DEG, E \to CDG, G \to CDE$

Maximum Feature {CEG}

Itemset of child node {CEG}

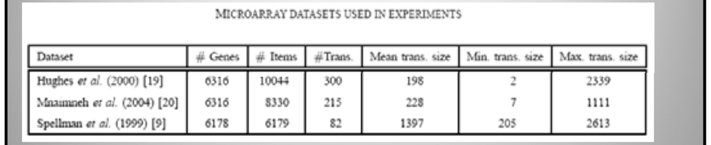
Outline

- Introduction
- Row Enumeration
- Confidence-based Prune Strategy
- **MAXCONF Algorithm**
- Evaluation
- Conclusion

Outline

- Introduction
- Row Enumeration
- Confidence-based Prune Strategy
- MAXCONF Algorithm
- **Evaluation**
- References

Evaluation



MICROARRAY DATASETS USED IN EXPERIMENTS

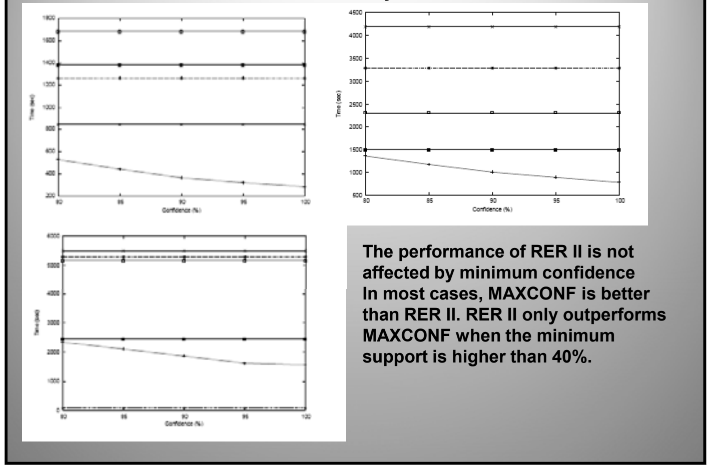
Dataset	# Genes	# Items	#Trans.	Mean trans. size	Min. trans. size	Max. trans. size
Hughes *et al.* (2000) [19]	6316	10044	300	198	2	2339
Mnaimneh *et al.* (2004) [20]	6316	8330	215	228	7	1111
Spellman *et al.* (1999) [9]	6178	6179	82	1397	205	2613

MAXCONF vs RER II
Two Aspect:

1. **Rule Generation**
2. **Scalability**

Evaluation

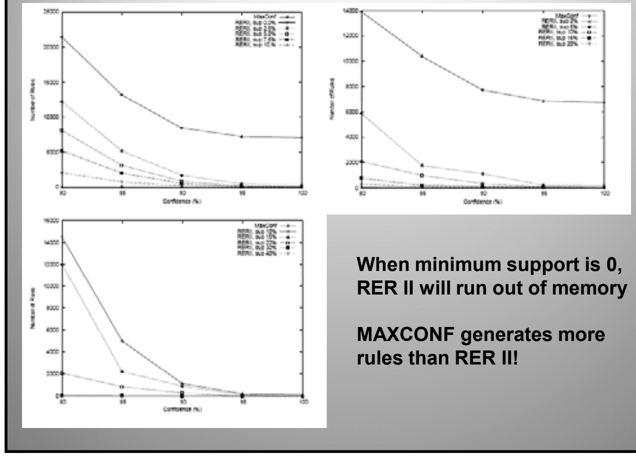
Scalability



The performance of RER II is not affected by minimum confidence
In most cases, MAXCONF is better than RER II. RER II only outperforms MAXCONF when the minimum support is higher than 40%.

Evaluation

Rule Generation



When minimum support is 0, RER II will run out of memory

MAXCONF generates more rules than RER II!

References

- **MAXCONF, "High Confidence Rule Mining for Microarray Analysis", by Tara McIntosh, Sanjay Chawla, 2006**
- **RER II, "Mining frequent closed patterns in microarray data." by G. Cong, K.-L. Tan, A. Tung, and F. Pan, 2004**

Any Question?

Thanks for your Attentions