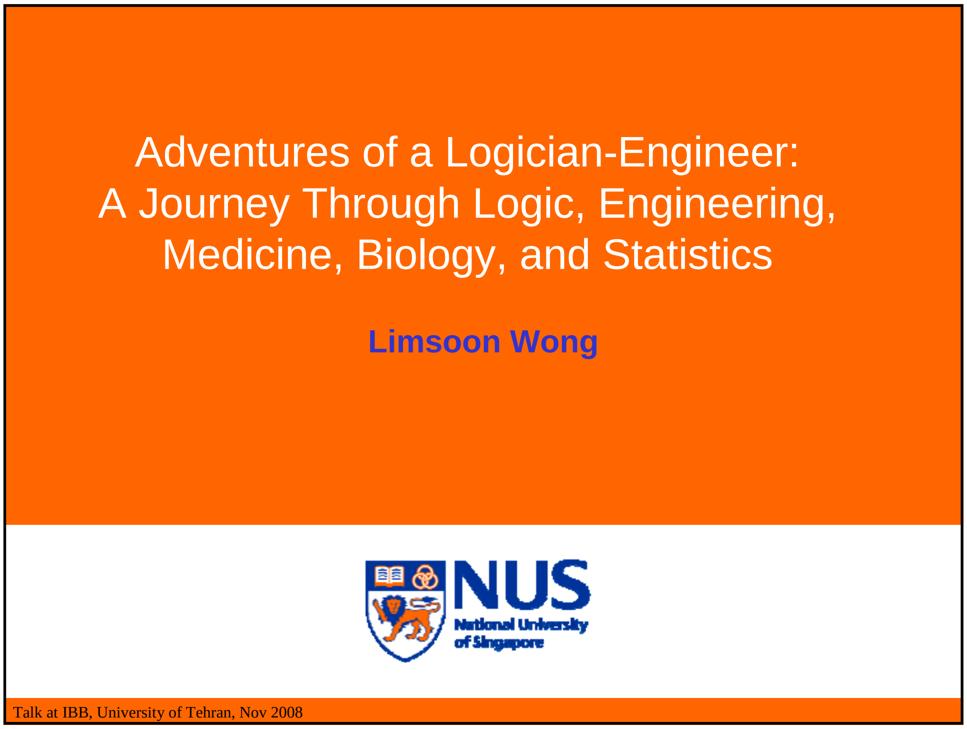
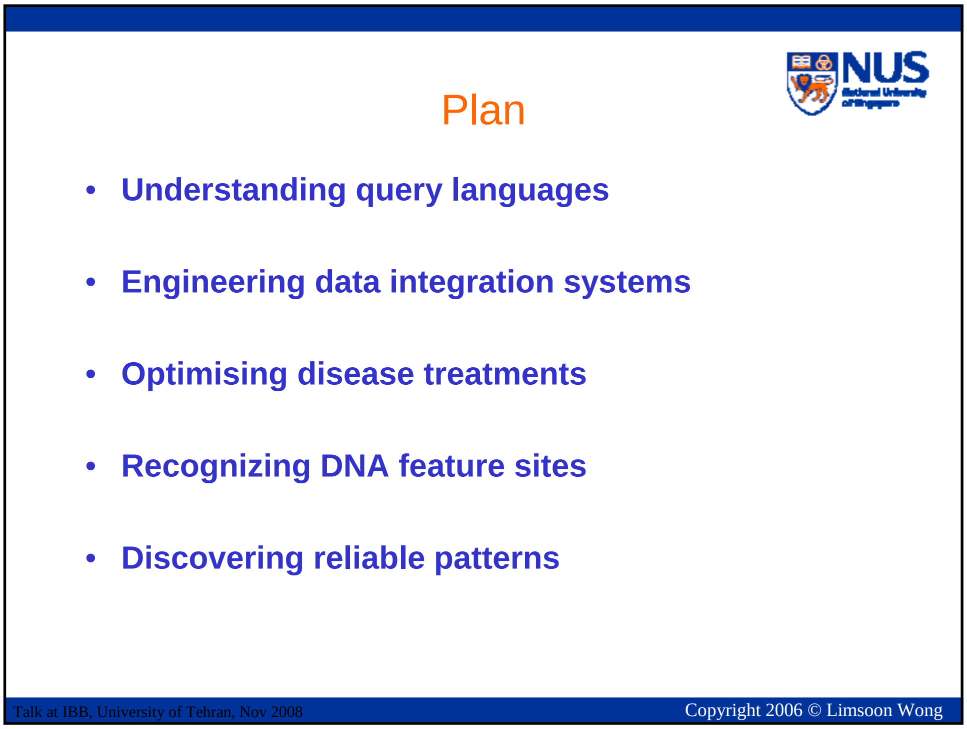


Adventures of a Logician-Engineer: A Journey Through Logic, Engineering, Medicine, Biology, and Statistics

Limsoon Wong

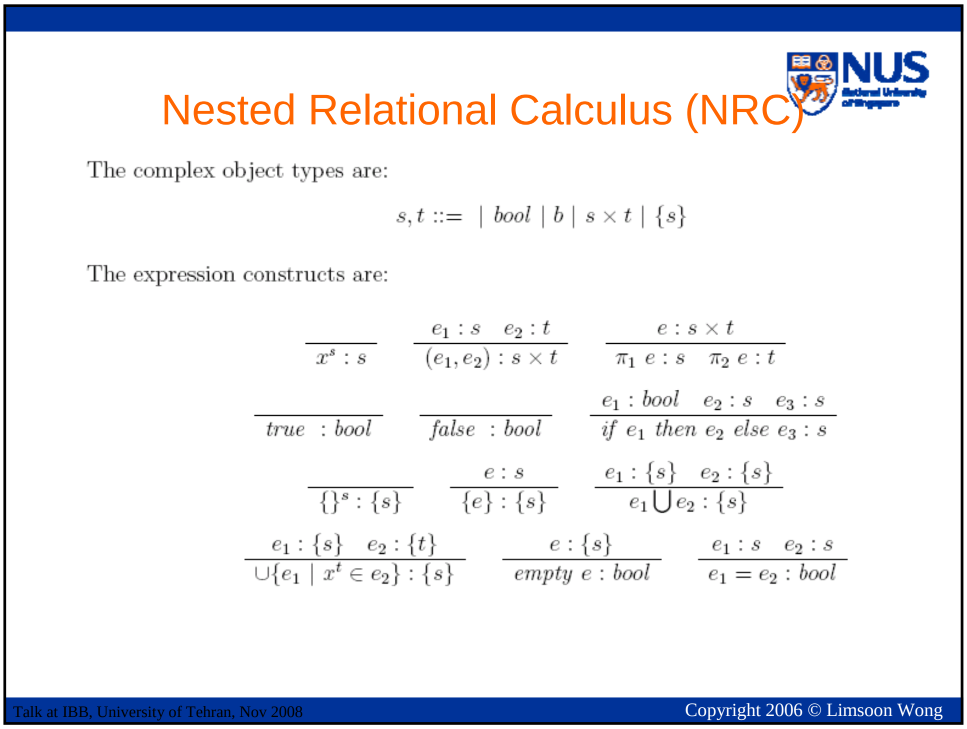


Plan

- **Understanding query languages**
- **Engineering data integration systems**
- **Optimising disease treatments**
- **Recognizing DNA feature sites**
- **Discovering reliable patterns**

Copyright 2006 © Limsoon Wong

Understanding Query Languages



Nested Relational Calculus (NRC)

The complex object types are:

$$s, t ::= \ \mid bool \mid b \mid s \times t \mid \{s\}$$

The expression constructs are:

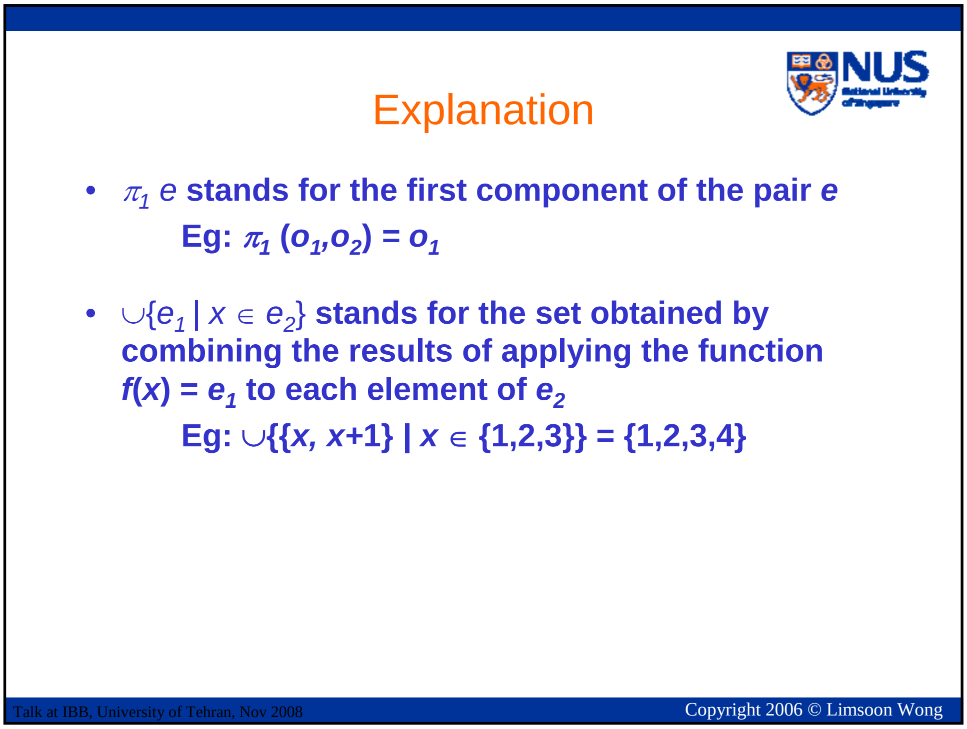
$$\frac{}{x^s : s} \qquad \frac{e_1 : s \quad e_2 : t}{(e_1, e_2) : s \times t} \qquad \frac{e : s \times t}{\pi_1\ e : s \quad \pi_2\ e : t}$$

$$\frac{}{true\ \ : bool} \qquad \frac{}{false\ \ : bool} \qquad \frac{e_1 : bool \quad e_2 : s \quad e_3 : s}{if\ e_1\ then\ e_2\ else\ e_3 : s}$$

$$\frac{}{\{\}^s : \{s\}} \qquad \frac{e : s}{\{e\} : \{s\}} \qquad \frac{e_1 : \{s\} \quad e_2 : \{s\}}{e_1 \bigcup e_2 : \{s\}}$$

$$\frac{e_1 : \{s\} \quad e_2 : \{t\}}{\cup\{e_1 \mid x^t \in e_2\} : \{s\}} \qquad \frac{e : \{s\}}{empty\ e : bool} \qquad \frac{e_1 : s \quad e_2 : s}{e_1 = e_2 : bool}$$

Copyright 2006 © Limsoon Wong



Explanation

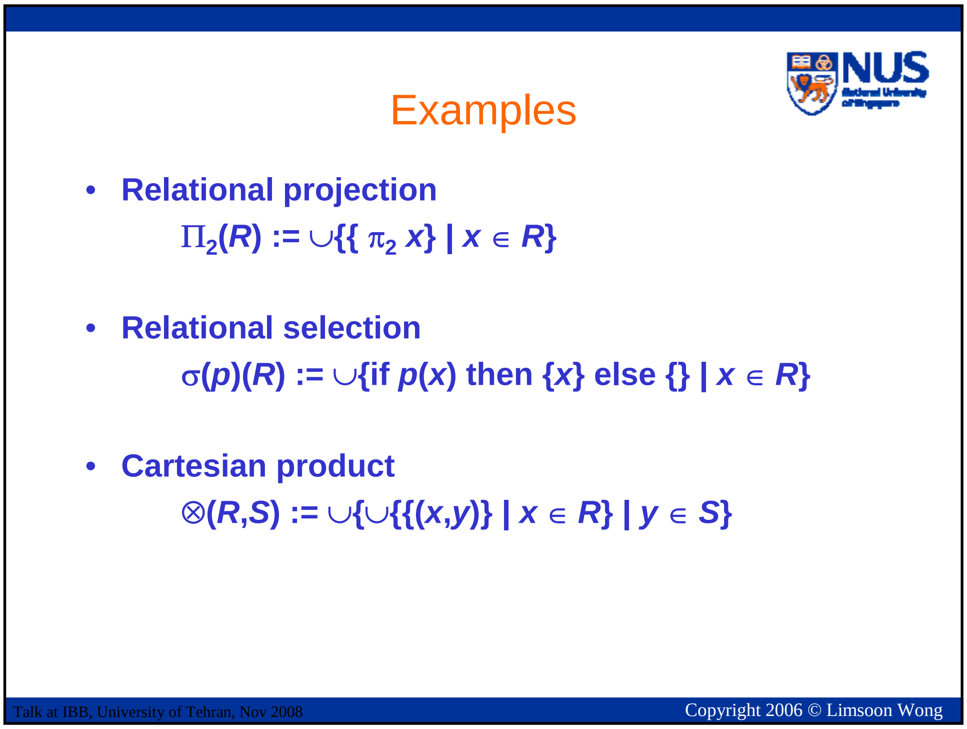
- $\pi_1\ e$ **stands for the first component of the pair *e***

  Eg: $\pi_1\ (o_1,o_2) = o_1$

- $\cup\{e_1 \mid x \in e_2\}$ **stands for the set obtained by combining the results of applying the function** $f(x) = e_1$ **to each element of** e_2

  Eg: $\cup\{\{x, x+1\} \mid x \in \{1,2,3\}\} = \{1,2,3,4\}$

Copyright 2006 © Limsoon Wong



Examples

- **Relational projection**

  $\Pi_2(R) := \cup\{\{\ \pi_2\ x\} \mid x \in R\}$

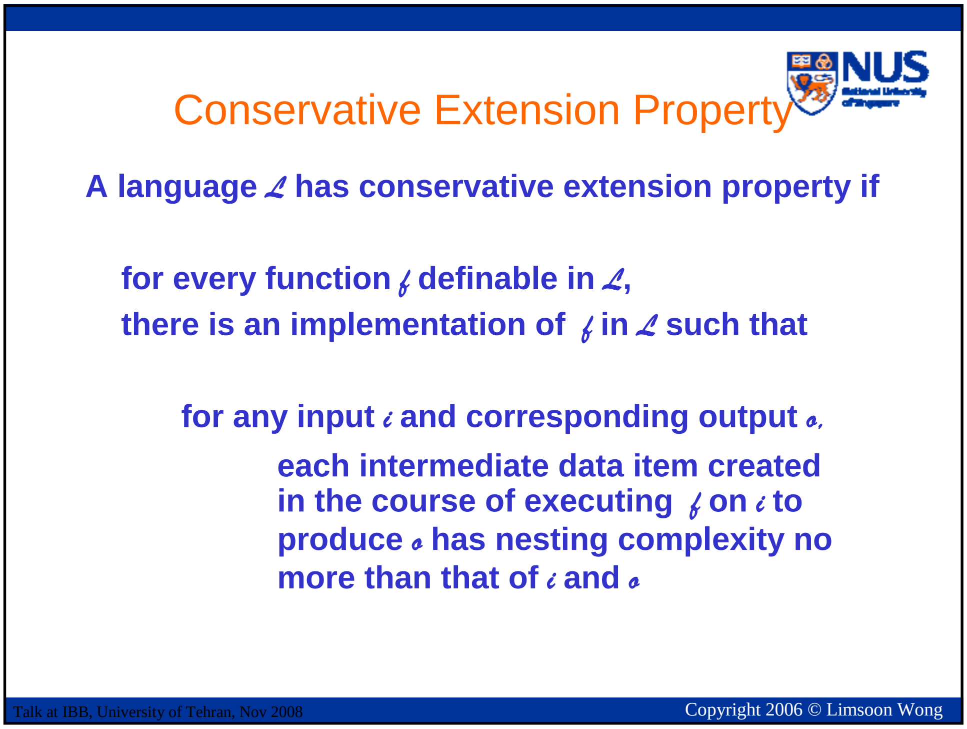
- **Relational selection**

  $\sigma(p)(R) := \cup\{\text{if } p(x) \text{ then } \{x\} \text{ else } \{\} \mid x \in R\}$

- **Cartesian product**

  $\otimes(R,S) := \cup\{\cup\{\{(x,y)\} \mid x \in R\} \mid y \in S\}$

Copyright 2006 © Limsoon Wong



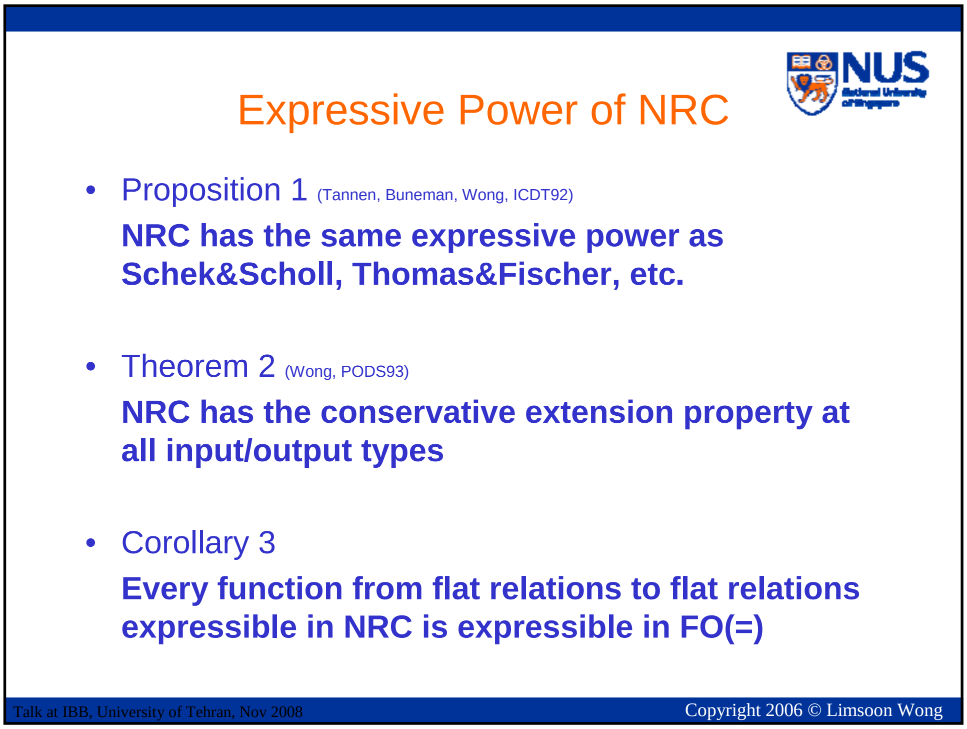
Conservative Extension Property

A language $\mathcal{L}$ has conservative extension property if

for every function f definable in $\mathcal{L}$,
there is an implementation of f in $\mathcal{L}$ such that

for any input i and corresponding output o,

each intermediate data item created in the course of executing f on i to produce o has nesting complexity no more than that of i and o



Expressive Power of NRC

- Proposition 1 (Tannen, Buneman, Wong, ICDT92)

  NRC has the same expressive power as Schek&Scholl, Thomas&Fischer, etc.

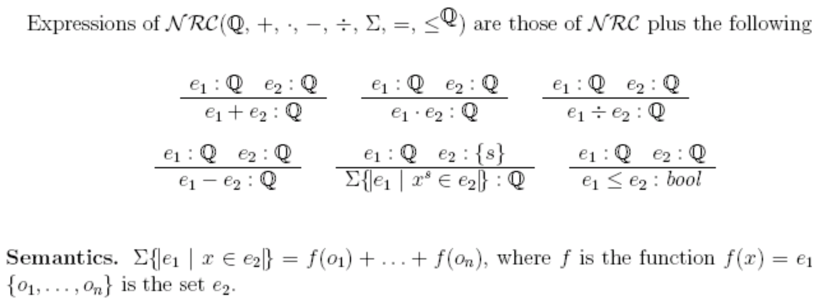
- Theorem 2 (Wong, PODS93)

  NRC has the conservative extension property at all input/output types

- Corollary 3

  Every function from flat relations to flat relations expressible in NRC is expressible in FO(=)

Theoretical Reconstruction of SQL



Expressions of $\mathcal{NRC}(\mathbb{Q}, +, \cdot, -, \div, \Sigma, =, \leq^{\mathbb{Q}})$ are those of $\mathcal{NRC}$ plus the following

$$\frac{e_1 : \mathbb{Q} \quad e_2 : \mathbb{Q}}{e_1 + e_2 : \mathbb{Q}} \qquad \frac{e_1 : \mathbb{Q} \quad e_2 : \mathbb{Q}}{e_1 \cdot e_2 : \mathbb{Q}} \qquad \frac{e_1 : \mathbb{Q} \quad e_2 : \mathbb{Q}}{e_1 \div e_2 : \mathbb{Q}}$$

$$\frac{e_1 : \mathbb{Q} \quad e_2 : \mathbb{Q}}{e_1 - e_2 : \mathbb{Q}} \qquad \frac{e_1 : \mathbb{Q} \quad e_2 : \{s\}}{\Sigma\{\!| e_1 \mid x^s \in e_2 |\!\} : \mathbb{Q}} \qquad \frac{e_1 : \mathbb{Q} \quad e_2 : \mathbb{Q}}{e_1 \leq e_2 : bool}$$

Semantics. $\Sigma\{\!| e_1 \mid x \in e_2 |\!\} = f(o_1) + \ldots + f(o_n)$, where f is the function $f(x) = e_1$ $\{o_1, \ldots, o_n\}$ is the set e_2.

Example Aggregate Functions

- **Count the number of records**

  count(*R*) := Σ{| 1 | *x* ∈ *R* |}

- **Total the first column**

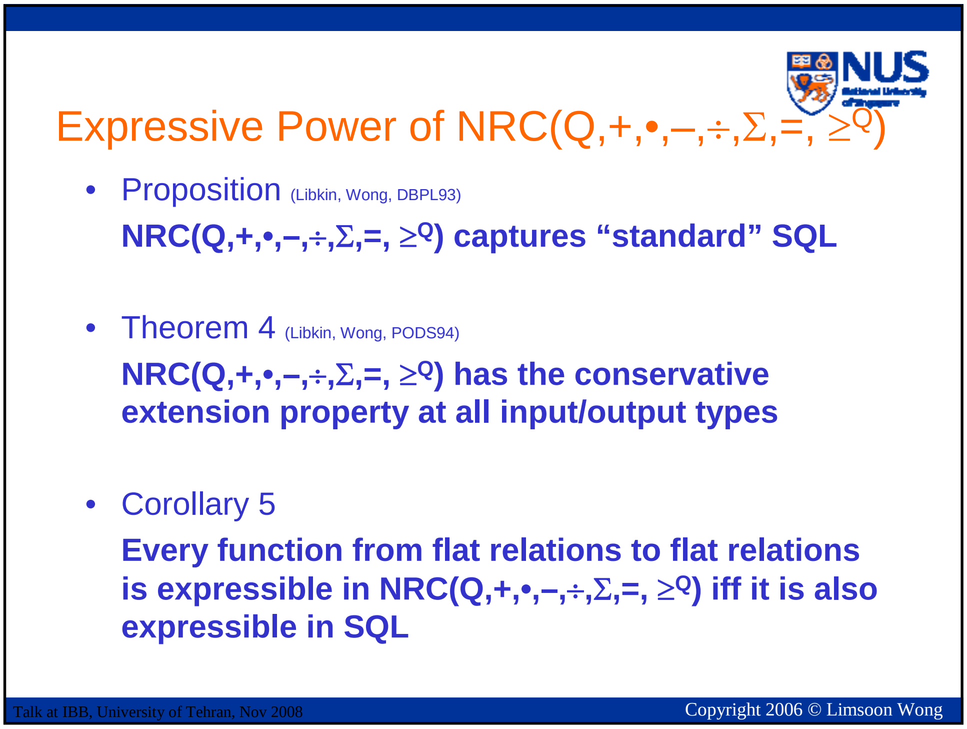
  $\text{total}_1(R) := \Sigma\{\!|\ \pi_1\ x \mid x \in R\ |\!\}$

- **Average of the first column**

  $\text{ave}_1(R) := \text{total}_1(R) \div \text{count}(R)$

- **A totally generic query expressible in SQL but inexpressible in FO(=)**

  eqcard(*R*,*S*) := count(*R*) = count(*S*)



Expressive Power of NRC(Q,+,•,–,÷,Σ,=, $\geq^Q$)

- Proposition (Libkin, Wong, DBPL93)

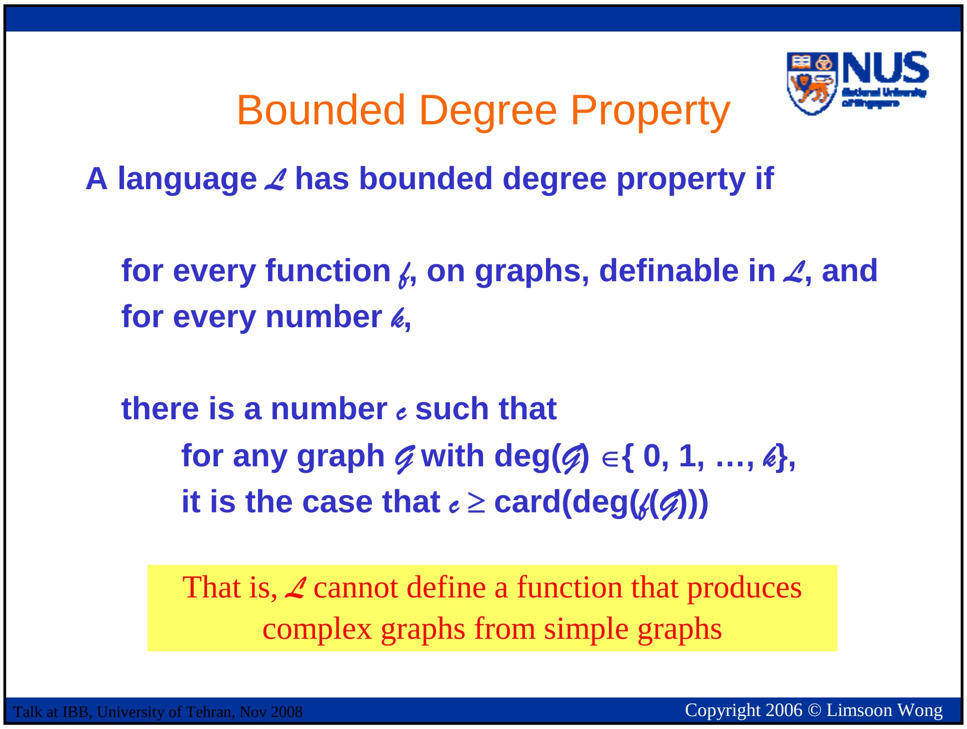
  NRC(Q,+,•,–,÷,Σ,=, $\geq^Q$) captures "standard" SQL

- Theorem 4 (Libkin, Wong, PODS94)

  NRC(Q,+,•,–,÷,Σ,=, $\geq^Q$) has the conservative extension property at all input/output types

- Corollary 5

  Every function from flat relations to flat relations is expressible in NRC(Q,+,•,–,÷,Σ,=, $\geq^Q$) iff it is also expressible in SQL



Bounded Degree Property

A language $\mathcal{L}$ has bounded degree property if

for every function f, on graphs, definable in $\mathcal{L}$, and for every number k,

there is a number c such that

for any graph $\mathcal{G}$ with deg($\mathcal{G}$) $\in$ { 0, 1, …, k},

it is the case that $c \geq$ card(deg(f($\mathcal{G}$)))

That is, $\mathcal{L}$ cannot define a function that produces complex graphs from simple graphs

Expressive Power of NRC(Q,+,•,–,÷,Σ,=, $\geq^Q$)

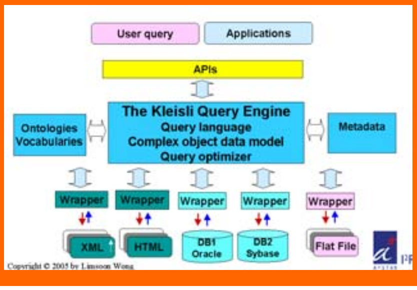
- Theorem 6 (Dong, Libkin, Wong, ICDT97)

  NRC(Q,+,•,–,÷,Σ,=, $\geq^Q$) has the bounded degree property

- Corollary 7
  - Transitive closure of unordered graphs cannot be expressed in SQL
  - Parity test on cardinality of unordered graphs cannot be expressed in SQL
  - Transitive closure of linear chains cannot be expressed in SQL
  - ...

Copyright 2006 © Limsoon Wong

Engineering Data Integration Systems



Copyright © 2005 by Limsoon Wong

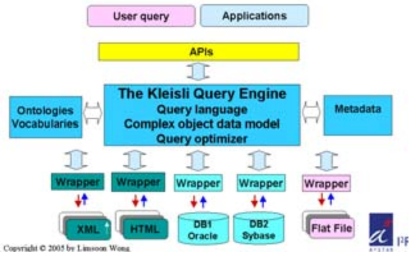
Integration: What are the problems?

A US DOE “impossible query”, circa 1993:

For each gene on a given cytogenetic band, find its non-human homologs.

source	type	location	remarks
GDB	Sybase	Baltimore	Flat tables SQL joins Location info
Entrez	ASN.1	Bethesda	Nested tables Keywords Homolog info

Integration Solution: Kleisli

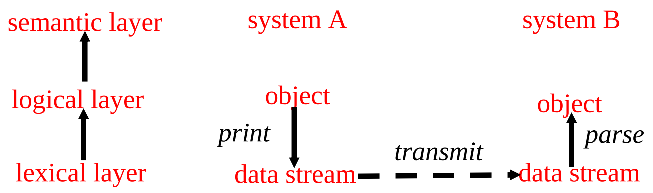

User query
Applications
APIs
The Kleisli Query Engine
Query language
Complex object data model
Query optimizer
Ontologies
Vocabularies
Metadata
Wrapper
Wrapper
Wrapper
Wrapper
Wrapper
XML
HTML
DB1
Oracle
DB2
Sybase
Flat File

Copyright © 2005 by Limsoon Wong

Buneman, Davidson, Hart, Overton, Wong, VLDB95
Wong, ICFP00

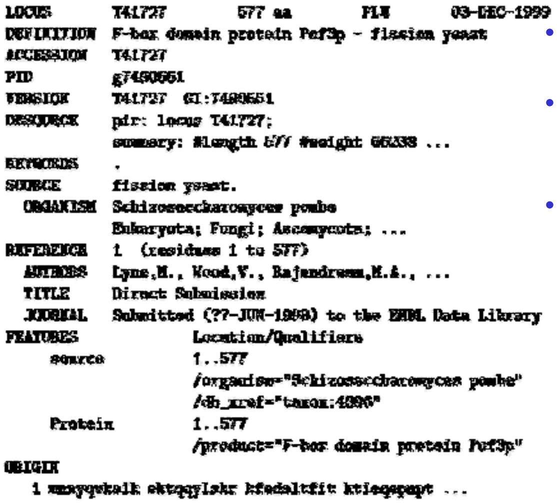
- Nested relational data model
- Self-describing data exchange format
- Thin wrappers & lots of them
- High-level query languages
- Powerful query optimizer
- Open “database” connectivity & API
- Nested relational data store

Self-Describing
Data Exchange Format



- Logical & lexical layers are important aspects
- “Print” & “parse” to move between layers
- “Transmit” to move between systems
- Clear separation ⇒ generic parsers & “printers”

GenPept: E.g. of Poor Format



```
LOCUS       T41727          577 aa          PLN       03-DEC-1999
DEFINITION  F-box domain protein Pof3p - fission yeast
ACCESSION   T41727
PID         g7490551
VERSION     T41727  GI:7490551
DBSOURCE    pir: locus T41727;
            summary: #length 577 #weight 65338 ...
KEYWORDS    .
SOURCE      fission yeast.
  ORGANISM  Schizosaccharomyces pombe
            Eukaryota; Fungi; Ascomycota; ...
REFERENCE   1  (residues 1 to 577)
  AUTHORS   Lyne,M., Wood,V., Rajandream,M.A., ...
  TITLE     Direct Submission
  JOURNAL   Submitted (??-JUN-1998) to the EMBL Data Library
FEATURES             Location/Qualifiers
     source          1..577
                     /organism="Schizosaccharomyces pombe"
                     /db_xref="taxon:4896"
     Protein         1..577
                     /product="F-box domain protein Pof3p"
ORIGIN
   1 msnyqvkalk ektqqylskr kfedaltfit ktieqspqpt ...
```

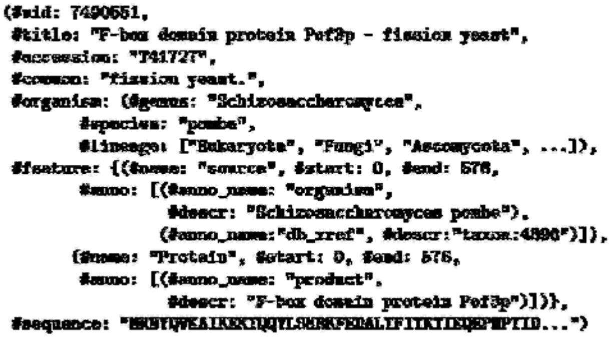
- Deeply nested structure
- No separation of logical vs. lexical layers
- Specialized parser is a must

Kleisli’s Data Exchange Format

logical layer	lexical layer	remarks
Booleans	True, false	
Numbers	123, 123.123	Positive numbers
	~123, ~123,123	Negative numbers
strings	“a string”	String is inside double quotes
records	(#l$_1$: v$_1$, …, #l$_n$: v$_n$)	Record is inside round brackets
sets	{v$_1$, …, v$_n$}	Set is inside curly brackets

- **Lexical layer matches logical layer**
- **Mirrors nested relational data model**
- **Avoids impedance mismatch**
- **Easier to write wrappers**

GenPept: In a Better Format



```
(#uid: 7490651,
 #title: "F-box domain protein Pof3p - fission yeast",
 #accession: "T41727",
 #common: "fission yeast.",
 #organism: (#genus: "Schizosaccharomyces",
         #species: "pombe",
         #lineage: ["Eukaryota", "Fungi", "Ascomycota", ...]),
 #feature: {(#name: "source", #start: 0, #end: 576,
         #anno: [(#anno_name: "organism",
                   #descr: "Schizosaccharomyces pombe").
                  (#anno_name:"db_xref", #descr:"taxon:4896")]},
        {#name: "Protein", #start: 0, #end: 576,
         #anno: [(#anno_name: "product",
                   #descr: "F-box domain protein Pof3p")])},
 #sequence: "MKSYTQWEAIKEKIQQYLSERKFEDALIFITIKTIEQEPMPTID...")
```

- **Boundaries of different nested structures are explicit**
- **Logical vs. lexical layers no longer mixed up**
- **Specialized parser no longer needed**

Data Integration Results

- **Using Kleisli:**
  - Clear
  - Succinct
  - Efficient
- **Handles**
  - heterogeneity
  - complexity

```
sybase-add (#name:”GDB", ...);
create view  L from locus_cyto_location using GDB;
create view E from object_genbank_eref using GDB;
select
    #accn: g.#genbank_ref,   #nonhuman-homologs: H
from
    L as c,  E as g,
    {select u
     from g.#genbank_ref.na-get-homolog-summary as u
     where not(u.#title string-islike "%Human%") &
             not(u.#title string-islike "%H.sapien%")} as H
where
    c.#chrom_num = "22” &
    g.#object_id = c.#locus_id &
    not (H = { });
```

Optimising Disease Treatments

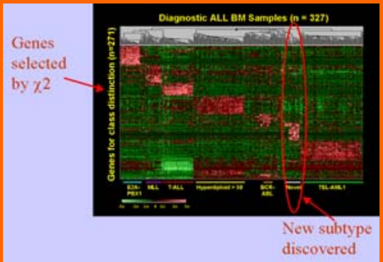
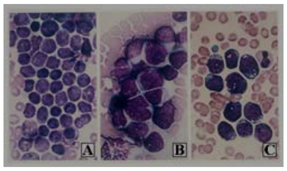


Image credit: Yeoh et al, 2002

Childhood ALL

- **Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid>50,**
- **Diff subtypes respond differently to same Tx**
- **Over-intensive Tx**
  - Development of secondary cancers
  - Reduction of IQ
- **Under-intensiveTx**
  - Relapse

- **The subtypes look similar**



- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- **Unavailable in most ASEAN countries**

Single-Test Platform of Microarray & Knowledge Discovery

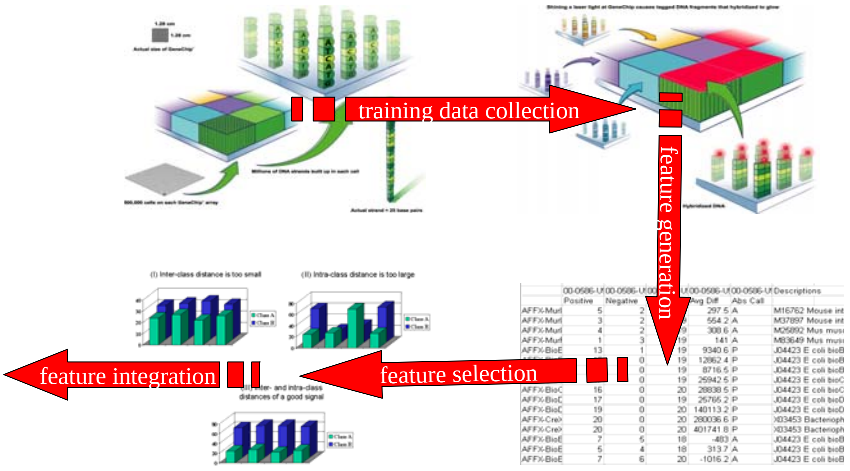
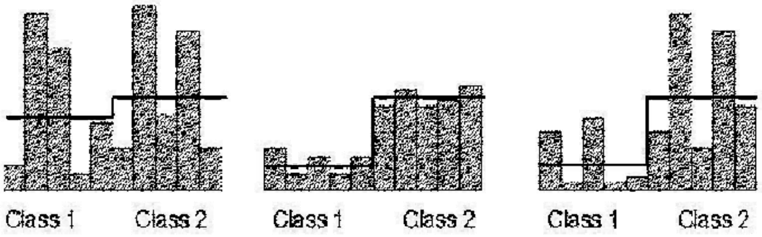


Image credit: Affymetrix

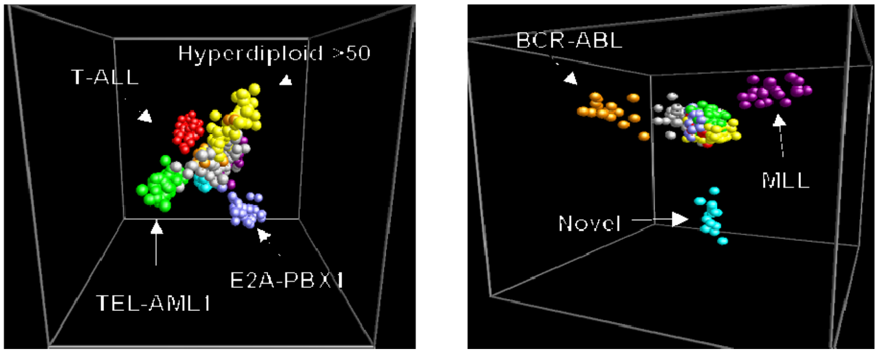
Signal Selection Basic Idea

- Choose a signal w/ low intra-class distance
- Choose a signal w/ high inter-class distance

⇒ An invariant of a disease subtype!

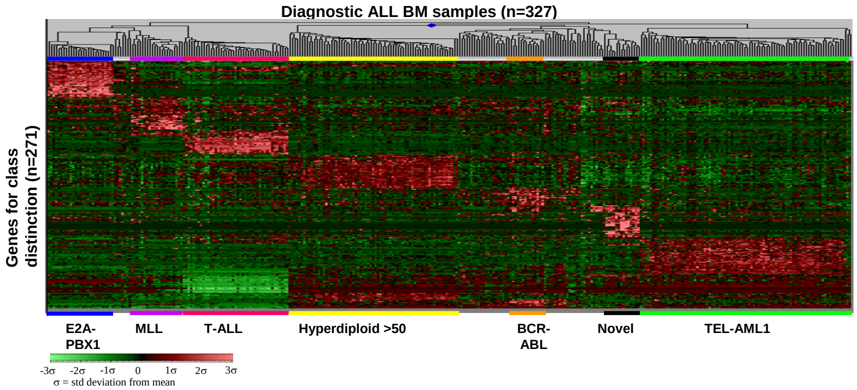


Multidimensional Scaling Plot for ALL Subtype Diagnosis

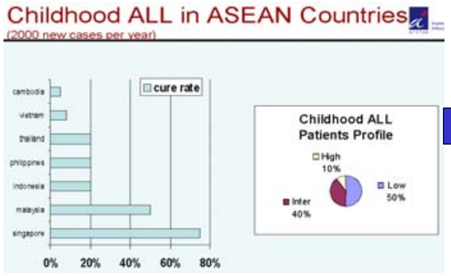


Obtained by performing PCA on the 20 genes chosen for each level

Gene Expression Profiles for ALL Subtype Diagnosis



Impact



Conventional Tx:
- intermediate intensity to all

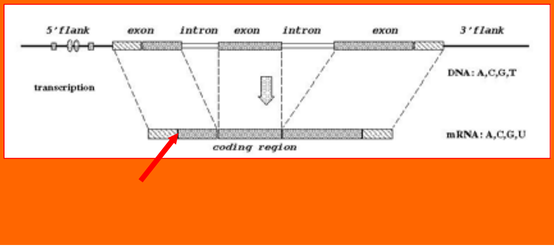
⇒ 10% suffers relapse
⇒ 50% suffers side effects
⇒ costs US$150m/yr

Our optimized Tx:
- high intensity to 10%
- intermediate intensity to 40%
- low intensity to 50%
- costs US$100m/yr

- **High cure rate of 80%**
- **Less relapse**
- **Less side effects**
- **Save US$51.6m/yr**

Yeoh et al, *CANCER CELL*, 2002

Recognizing DNA Feature Sites



Talk at IBB, University of Tehran, Nov 2008

A Sample cDNA

```
 299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA     160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA     240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
............................................................      80
................................iEEEEEEEEEEEEEEEEEEEEEEEEEEE     160
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE     240
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

- **What makes the second ATG the TIS?**

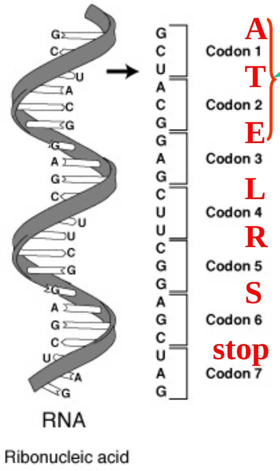
Talk at IBB, University of Tehran, Nov 2008

Copyright 2006 © Limsoon Wong

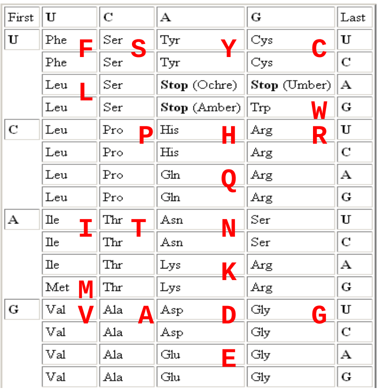
Approach

- **Training data gathering**
- **Signal generation**
  - k-grams, distance, domain know-how, ...
- **Signal selection**
  - Entropy, $\chi 2$, CFS, t-test, domain know-how...
- **Signal integration**
  - SVM, ANN, PCL, CART, C4.5, kNN, ...

mRNA→protein

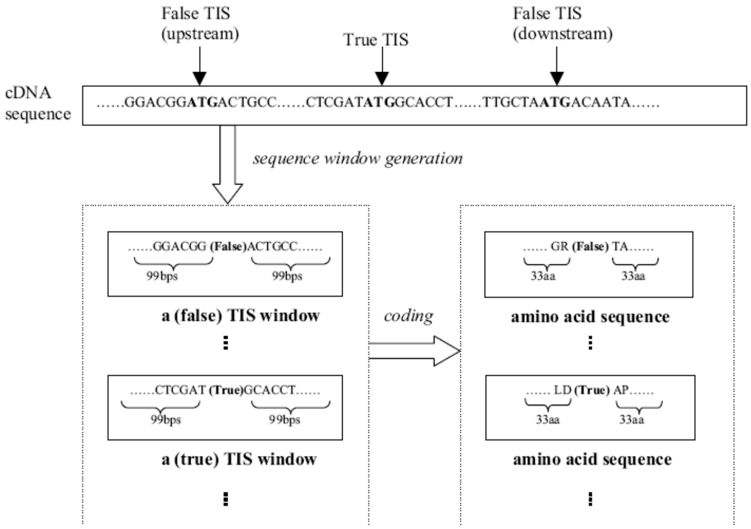


How about using k-grams from the translation?



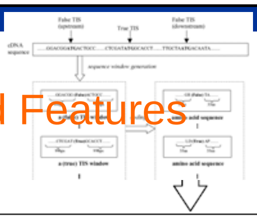
First	U	C	A	G	Last
U	Phe F	Ser S	Tyr Y	Cys C	U
	Phe	Ser	Tyr	Cys	C
	Leu L	Ser	**Stop** (Ochre)	**Stop** (Umber)	A
	Leu	Ser	**Stop** (Amber)	Trp W	G
C	Leu	Pro P	His H	Arg R	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln Q	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile I	Thr T	Asn N	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys K	Arg	A
	Met M	Thr	Lys	Arg	G
G	Val V	Ala A	Asp D	Gly G	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu E	Gly	A
	Val	Ala	Glu	Gly	G

Amino-Acid Features



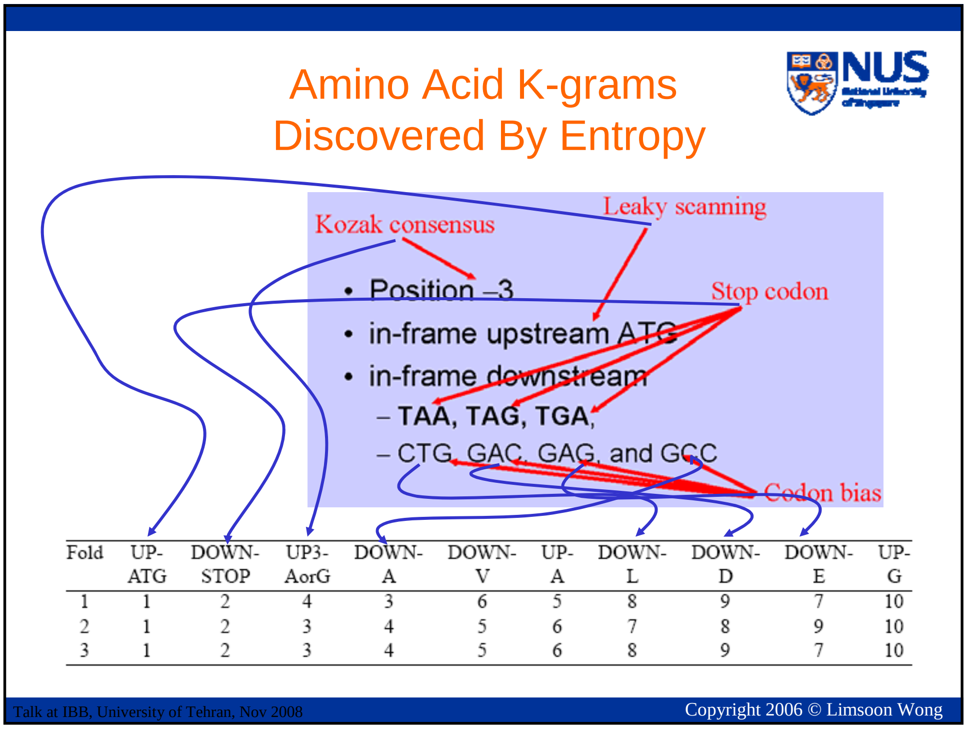
Copyright 2006 © Limsoon Wong

Amino-Acid Features



New feature space (total of 927 features + class label)			
42 1-gram amino acid patterns	882 2-gram amino acid patterns	3 bio-know-ledge patterns	class label
UP-A, UP-R, ….,UP-N, DOWN-A, DOWN-R, …., DOWN-N (numeric type)	UP-AA, UP-AR, …., UP-NN, DOWN-AA, DOWN-AR , …, DOWN-NN (numeric type)	DOWN4-G UP3-AorG, UP-ATG (boolean type, Y or N)	True, False
Frequency as values			
1, 3, 5, 0, 4, …	6, 2, 7, 0, 5, …	N, N, N,	False
⋮	⋮	⋮	⋮
6, 5, 7, 9, 0, …	2, 0, 3, 10, 0, …	Y, Y, Y,	True
⋮	⋮	⋮	⋮

Copyright 2006 © Limsoon Wong



Amino Acid K-grams Discovered By Entropy

Kozak consensus

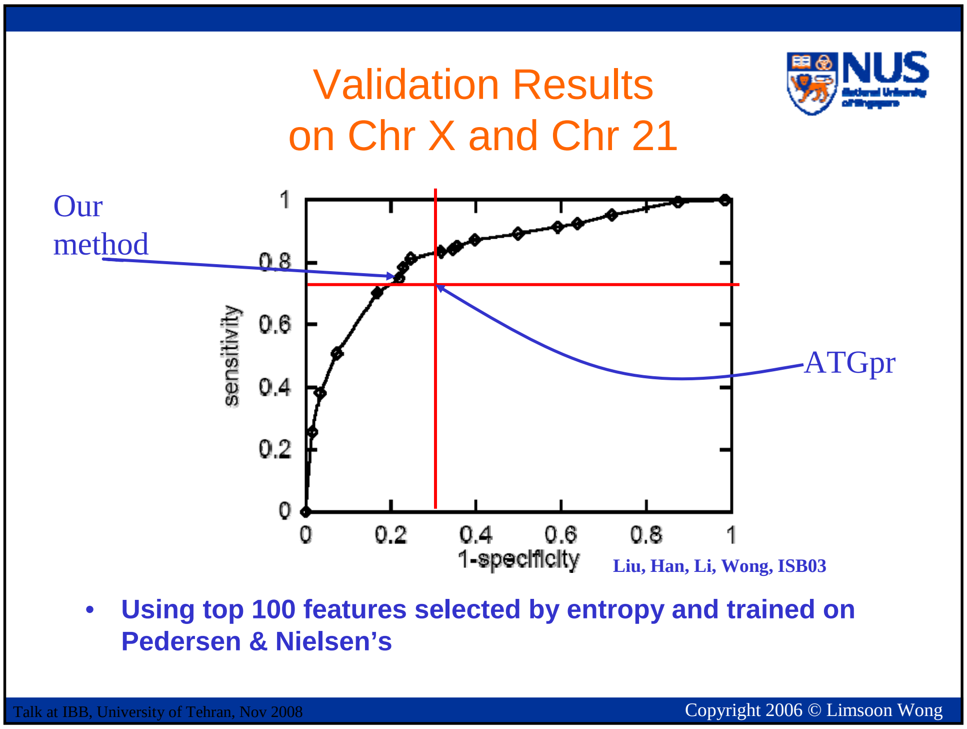
Leaky scanning

Stop codon

Codon bias

- Position –3
- in-frame upstream ATG
- in-frame downstream
  - TAA, TAG, TGA,
  - CTG, GAC, GAG, and GCC

Fold	UP-ATG	DOWN-STOP	UP3-AorG	DOWN-A	DOWN-V	UP-A	DOWN-L	DOWN-D	DOWN-E	UP-G
1	1	2	4	3	6	5	8	9	7	10
2	1	2	3	4	5	6	7	8	9	10
3	1	2	3	4	5	6	8	9	7	10

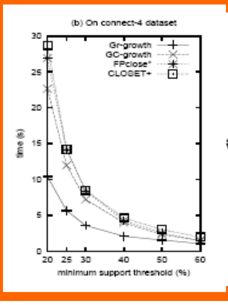


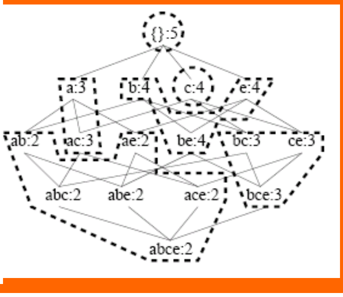
Validation Results on Chr X and Chr 21

Liu, Han, Li, Wong, ISB03

- Using top 100 features selected by entropy and trained on Pedersen & Nielsen’s

Discovering Reliable Patterns





Discovering Invariants

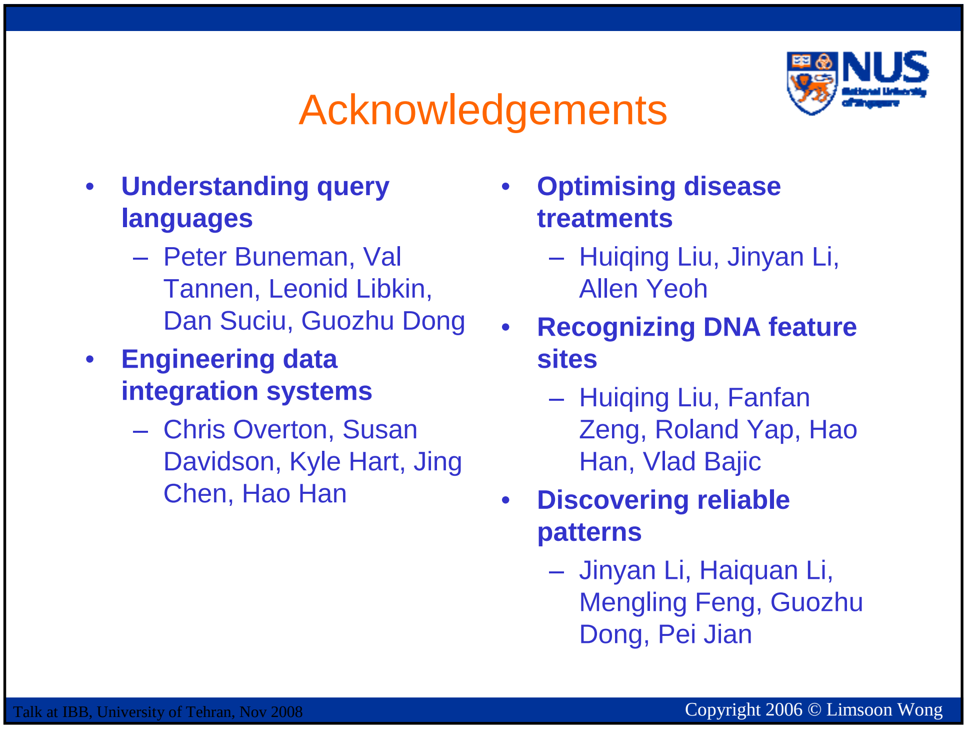
- Conservative extension property
- Bounded degree property
- Logical layer of self-describing exchange formats

Insights of an expert

- Diagnosis patterns of ALL subtypes
- Signals for protein translation initiation

Identified using existing machine learning methods

- Next Goal: Improve capability of machines to discover useful invariants



Acknowledgements

- **Understanding query languages**
  - Peter Buneman, Val Tannen, Leonid Libkin, Dan Suciu, Guozhu Dong
- **Engineering data integration systems**
  - Chris Overton, Susan Davidson, Kyle Hart, Jing Chen, Hao Han
- **Optimising disease treatments**
  - Huiqing Liu, Jinyan Li, Allen Yeoh
- **Recognizing DNA feature sites**
  - Huiqing Liu, Fanfan Zeng, Roland Yap, Hao Han, Vlad Bajic
- **Discovering reliable patterns**
  - Jinyan Li, Haiquan Li, Mengling Feng, Guozhu Dong, Pei Jian

Talk at IBB, University of Tehran, Nov 2008

Copyright 2006 © Limsoon Wong