DIMACS Summer School
Tutorial on New Frontiers in Data Mining

Pattern Discovery in Computational Biology

Isidore Rigoutsos
Bioinformatics Group
IBM TJ Watson Research Center
rigoutso@us.ibm.com

The Context

one or more *streams* like:

```
.....
SILDIENXYAJVDWLXDLYFTOSHWUBSKUTFBNGLEKPQCUDKFNYXJDHVQJAUNJEGAMLAYQWUUVGAFNF
XOIXYQEDYDTKGFEPUQGNLXPSYKJNPNGYJIDXDIXKXFAQISVTFAKAZNENIDSENABCSVCUSRYURAS
TOEOSAKCIGKYQVUANESCGQBUSUKSKFDRAUSGCUUQSQEFYTSYMORFWAPDBNHXFNHSFZFUGTMIUZP
NRXLRARXIJGBOLDMZQFOLGCUEVOOZIHEITNHWXHQITVDTQOGZKRNJMDHSVONIRHEBERXBSTJTUJ
FYHCZWFSIJGBONICFAJSUSTJDUNCXHTGQGWVQXVARREOPZNZZEPMIZVCMZVUFLAYUYHTYXTIXBX
HQXEVFNHFZDGEWXFAYQZGGQZXVCAXINEBLRKKLHPRHPQYJFVIFTBNBUFRVDPAWGPSVQFYCPIPTA
GWWZURSOXFHUAPJVEINSWCUQCITQPIXUAHNXLESWPILEJEVMDIKVMXVWWRMEKFXWOATNXNDJRSH
VGAENFTLITCJAOGRPEDODFVEBBWVPRJYHRPCAQBESMUPTEVOEEWJEWIRADLUGVKEYYEIEZFPRQJ
BHSWOYABEDSHCODLQHYTPUTBXQCXQVRCQHRQAZKSUBEVPRSDCSPZWKCUNVYQIXCMMZOUASKWPWD
WHLSLSEFMCBQIOXLIYGYWHYIMFAENSETWOMAFCTNCDYFYAHCCZCUDRISRADTBEVTZGNAIFSRIAT
UKDHCXTWDJVMVHEANVRPLEBFAYGSJTJTEALHTNGDBCDAHCCDCUHRRPRAETKEJCRAKDKCEGARPWF
NLMBXPJYYKOQRFLXCTMDINTEQWDLTJHJSCXSZARMWSWVZZOULNERMYZSUXQJZATUWIHDTPYYORO
TGBIIKFAFPYXQQAMIZTILXLCUOGSYGRSLLDIMHGREUUVJUZPTEQHRNPUGABJUJYAITJFIZXBADJ
WQFALZASKQVAXIRUTSCGWPZGQZWHWTWQULYCVBGXAELZIIYXTUVQSEAMPXTEMDABBAGXAEHCBKV
GDXQAVQTDSXFCDBCKUMIBOLFAHPSETCBBAGGCNIDMFXAPCGZCUARQHRAUTAEEJTCKTNRAEBGAXB
DLQNKDNAOZUTHHEHTWADKQBDWKFQCULRXPBJENTAWJSQBEPUONFFZUMNGNXVQRAUXHHRIEHGPOZ
POYVEEBAVHCUEEYEVYIQOAKKJLIYPEJGDCAARGDMLBDHFLZWYQTNTPPIZYNAYKCDGANNNDEHZLQ
IWNCDMKXIAOANERJKVHRHQHPZCZKFMMXCXTWWSCVASDPUGYCBLPUXUVDPYZCQZPQEKMKTIVOJFE
CSLHCGKUBCGZIMSYNYCDWUBEHPDLAEROUGXEUDSBNNJGZDOYJKRDQMNHHSXANQYPWPPCARDPXDK
WMCEVURGSLVJCTEMEOMBWTGIHLLMDTLLFZVAZGUFJFNIZCLGRMKIGRFERLDYAQSEEZCZYNPTCMD
JNRASNWDEOGKNIUMRIIEPTOLZMHHNWCOLFITFNLQASCHUAQVOXLWJOJKKWIDKBTGWHZDQGAOWXR
HFGFJRNWBKPZWTCYPSGJCLTIYEIUBDBNJQIKCFGSMUBBGHIDGLVDUVYKLGXQOOXSOUDKIEVUVTK
YKBJNBKJHBZCFXUMLLI
.........
...
```

(c) I. Rigoutsos / August 17, 2001

Some Solutions

(c) I. Rigoutsos / August 17, 2001

Solutions Depend On...

- the representation scheme
- the nature of the input dataset
- *minimum* allowed *local density*
- *minimum* required *support*
- *form* of sought expressions, e.g.

  $\Sigma (\Sigma \cup '.')^* \Sigma$

  $(\Sigma \cup [\Sigma\Sigma^*\Sigma]) (\Sigma \cup [\Sigma\Sigma^*\Sigma] \cup '.')^* (\Sigma \cup [\Sigma\Sigma^*\Sigma])$

- permitted *equivalence classes*:
  - A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
  - {A,G}, C, {D,E}, {F,Y}, H, {I,L,M,V}, {K,R}, {N,Q}, P, {S,T}, W
  - scoring-matrix + distance threshold
  - ...

(c) I. Rigoutsos / August 17, 2001

What Do Solutions Look Like?

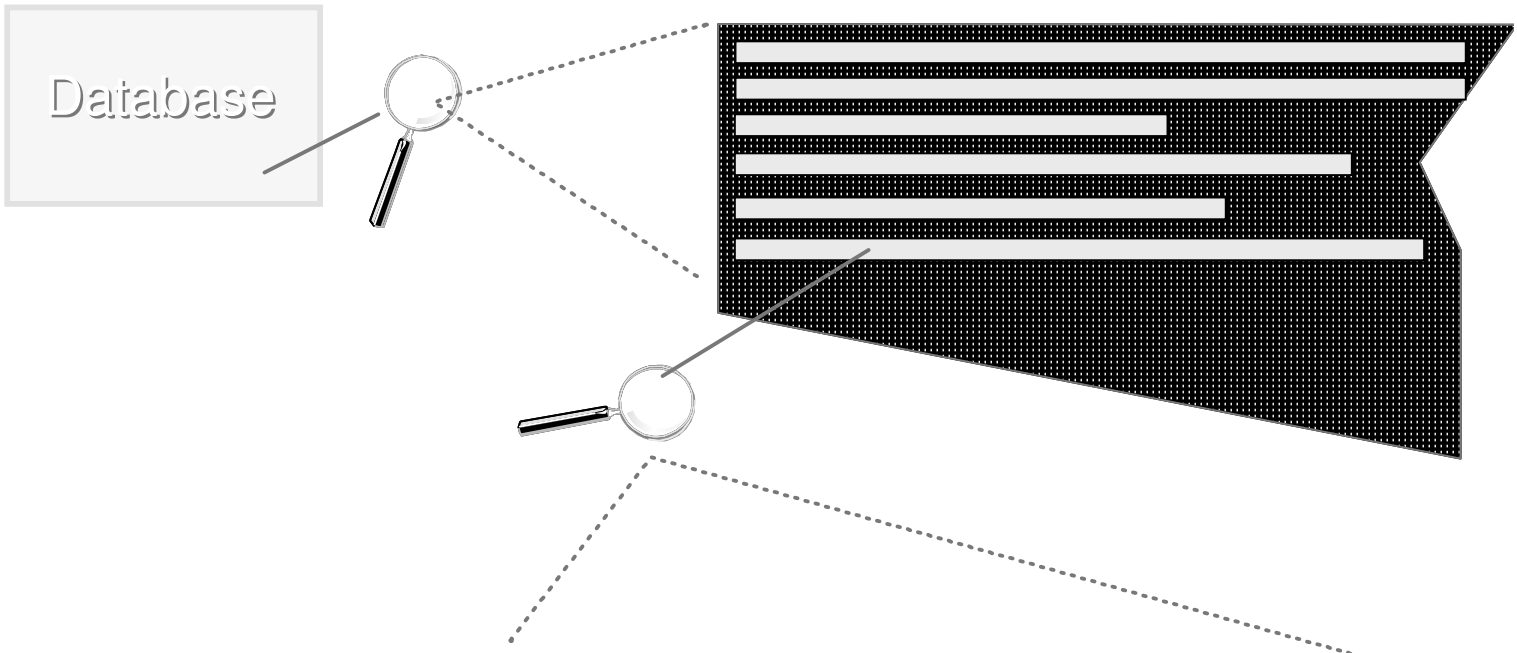
- Typical classes of patterns include:
  - simple strings:
    ABCDEFG
  - rigid gaps: AB...C.....D.E....F
  - flexible gaps:
    A-x(1,3)-B-x-C-x(4)-D-x(3,5)-E
  - unrestricted gaps:
    A * BCD * E * FG

    and their versions with expressions involving brackets:
    [APQs][BbKm]...C * D.[EW]....F

(c) I. Rigoutsos / August 17, 2001

The General Problem

Given:



$E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10} E_{11} E_{12} E_{13} E_{14} E_{15} E_{16} E_{17} E_{18} \ldots$

- may be given: real-valued function $F(E_i, E_j)$

(c) I. Rigoutsos / August 17, 2001

Database Examples

```
025  5  1  1  1  2  1  3  1  1  2
945  5  4  4  5  7  10 3  2  1  2
425  3  1  1  1  2  2  3  1  1  2
277  6  8  8  1  3  4  3  7  1  2
023  4  1  1  3  2  1  3  1  1  2
122  8  10 10 8  7  10 9  7  1  2

099  1  1  1  1  2  10 3  1  1  2
561  2  1  2  1  2  1  3  1  1  2
078  2  1  1  1  2  1  1  1  1  2
078  4  2  1  1  2  1  2  1  1  2
283  1  1  1  1  1  1  3  1  1  2
172  2  1  1  1  2  1  2  1  1  2
801  5  3  3  3  2  3  4  1  1  2
99   1  1  1  1  2  2  1  1  1  2
```

```
>Aeropyrum pernix
AAATAATAATAAAAATTAAGTGACT
ATTTAGGGACAATAGTGTTTAAGG
AGTTATAGTTGCGTGGATTTAGAT
ATTAACATTGTTACAACGAATAGA
ATATACACCTATAGCTATGAGAGAT
GTATTACATATGCTTCCGGCCTTTT
AACGTTCCGTATGCACCAATACTA
TATAATAGTCAAATATTTTAATTTAA
CGGCGAGCCTCATAGAATATTCCA
ATACCATATGCTGCACCAACCACA
>HBP_CANLI 62686306
MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNM
>HBP1_CASGL 1
ALTEKQEALLKQSWEVLKQNIPAHSLRLFALIIEAAPESKYVFSFLK
>HBP2_CASGL 1
MSTLEGRGFTEEQEALVVKSWSAMKPNAGELGLKFFLKIFEIAPS
>HBPL_PARAD 1
SSSEVNKVFTEEQEALVVKAWAVMKKNSAELGLQFFLKIFEIAPS
>HBPL_TRETO 1
MSSSEVDKVFTEEQEALVVKSWAVMKKNSAELGLKFFLKIFEIAP
>LGB1_LUPLU 1
GVLTDVQVALVKSSFEEFNANIPKNTHRFFTLVLEIAPGAKDLFSF
>LGB1_MEDSA 1
MSFTDKQEALVNSSWEAFKQNLPRYSVFFYTVVLEKAPAAKGLF
>LGB1_MEDTR 1
MSFTDKQEALVNSSYEAFKQNLSGYSVFFYTVILEKAPAAKGLFS
>LGB1_PEA 1
GFTDKQEALVNSSSEFKQNLPGYSILFYTIVLEKAPAAKGLFSFLK
>LGB1_SOYBN 1
GAFTEKQEALVSSSFEAFKANIPQYSVVFYNSILEKAPAAKDLFSF
>LGB1_VICFA 1
GFTEKQEALVNSSSQLFKQNPSNYSVLFYTIILQKAPTAKAMFSFL
>LGB2_LUPLU 1
GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSF
>LGB2_MEDTR 1
MGFTEKQEALVNSSWELFKQNPGNSVLFYTIILEKAPAAKGMFSF
>LGB2_SESRO 1
GFTEKQEALVNASYEAFKQNLPGNSVLFYSFILEKAPAAKGMFSF
>LGB2_SOYBN 1
GAFTEKQEALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF
>LGB3_MEDSA 1
MGFTDKQEALVNSSWESFKQNPGNSVLFYTIILEKAPAAKGMFSF
>LGB3_SESRO 1
GFTEKQEALVNASYEAFKQNLPGNSVLFYSFILEKAPAAKGMFSF
>LGB3_SOYBN 1
```

SECTION 2. THE HOUSE OF REPRESENTATIVES SHALL BE
SEVERAL STATES, AND THE ELECTORS IN EACH STATE SH
OF
THE STATE LEGISLATURE.

NO PERSON SHALL BE A REPRESENTATIVE WHO SHALL N
CITIZEN OF THE UNITED STATES, AND WHO SHALL NOT, W

REPRESENTATIVES AND DIRECT TAXES SHALL BE APPOR
ACCORDING TO THEIR RESPECTIVE NUMBERS, WHICH SH
THOSE BOUND TO SERVICE FOR A TERM OF YEARS, AND
ENUMERATION SHALL BE MADE WITHIN THREE YEARS AF
SUBSEQUENT TERM OF TEN YEARS, IN SUCH MANNER AS
FOR EVERY THIRTY THOUSAND, BUT EACH STATE SHALL
STATE OF NEW HAMPSHIRE SHALL BE ENTITLED TO CHUS
CONNECTICUT FIVE, NEW YORK SIX, NEW JERSEY FOUR,
CAROLINA FIVE, SOUTH CAROLINA FIVE, AND GEORGIA TH

WHEN VACANCIES HAPPEN IN THE REPRESENTATION FRO
SUCH VACANCIES.

THE HOUSE OF REPRESENTATIVES SHALL CHOOSE THEIR

SECTION 3. THE SENATE OF THE UNITED STATES SHALL B
THEREOF, FOR SIX YEARS; AND EACH SENATOR SHALL HA

IMMEDIATELY AFTER THEY SHALL BE ASSEMBLED IN CONS
THREE CLASSES. THE SEATS OF THE SENATORS OF THE
CLASS AT THE EXPIRATION OF THE FOURTH YEAR, AND TH
CHOSEN EVERY SECOND YEAR; AND IF VACANCIES HAPPE
STATE, THE EXECUTIVE THEREOF MAY MAKE TEMPORARY
SUCH VACANCIES.

NO PERSON SHALL BE A SENATOR WHO SHALL NOT HAVE
UNITED STATES AND WHO SHALL NOT, WHEN ELECTED, B

THE VICE PRESIDENT OF THE UNITED STATES SHALL BE P
DIVIDED.

THE SENATE SHALL CHOOSE THEIR OTHER OFFICERS, AN
HE SHALL EXERCISE THE OFFICE OF PRESIDENT OF THE U

THE SENATE SHALL HAVE THE SOLE POWER TO TRY ALL
AFFIRMATION. WHEN THE PRESIDENT OF THE UNITED STA
WITHOUT THE CONCURRENCE OF TWO THIRDS OF THE M

JUDGMENT IN CASES OF IMPEACHMENT SHALL NOT EXTE
ANY OFFICE OF HONOR, TRUST OR PROFIT UNDER THE U
INDICTMENT, TRIAL, JUDGMENT AND PUNISHMENT, ACCO

SECTION 4. THE TIMES, PLACES AND MANNER OF HOLDIN
STATE BY THE LEGISLATURE THEREOF; BUT THE CONGRE
PLACES OF CHOOSING SENATORS.

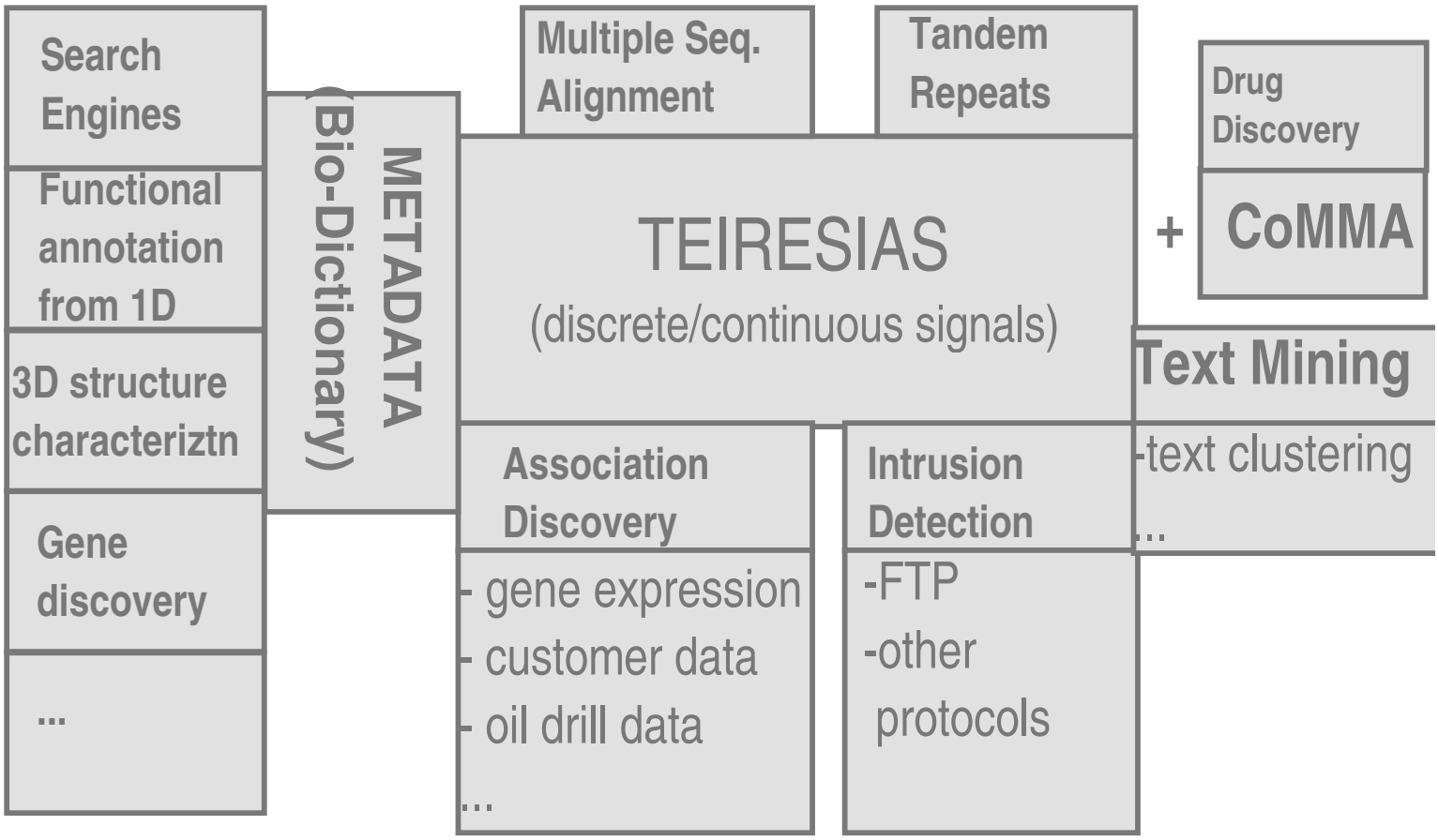
THE CONGRESS SHALL ASSEMBLE AT LEAST ONCE IN EVE
THEY SHALL BY LAW APPOINT A DIFFERENT DAY.

SECTION 5. EACH HOUSE SHALL BE THE JUDGE OF THE E
EACH SHALL CONSTITUTE A QUORUM TO DO BUSINESS; B

The context that provides such an input is typically not known and thus the only thing one can rely upon is the number of times that something (anything!) appears.

So basically, one has to come up with all the 'questions' that the dataset supports and also the 'answers' for them.

Research Work in the IBM Bioinformatics Group



Irredundant Motif Discovery

Algorithm //ization Shared Mem. / Message Passing

Error Propagation Analysis

New Approaches To Regression With Applications

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery (cont.)

Problem Definition

Given an input involving one or more streams of variable length, and an algorithm supporting one of the pattern classes mentioned above, find all the patterns with K or more instances in the input

(K is user-defined)

Additional constraints may be imposed on the patterns that should be reported.

Pattern Discovery - Basic Methods

The basic approaches:

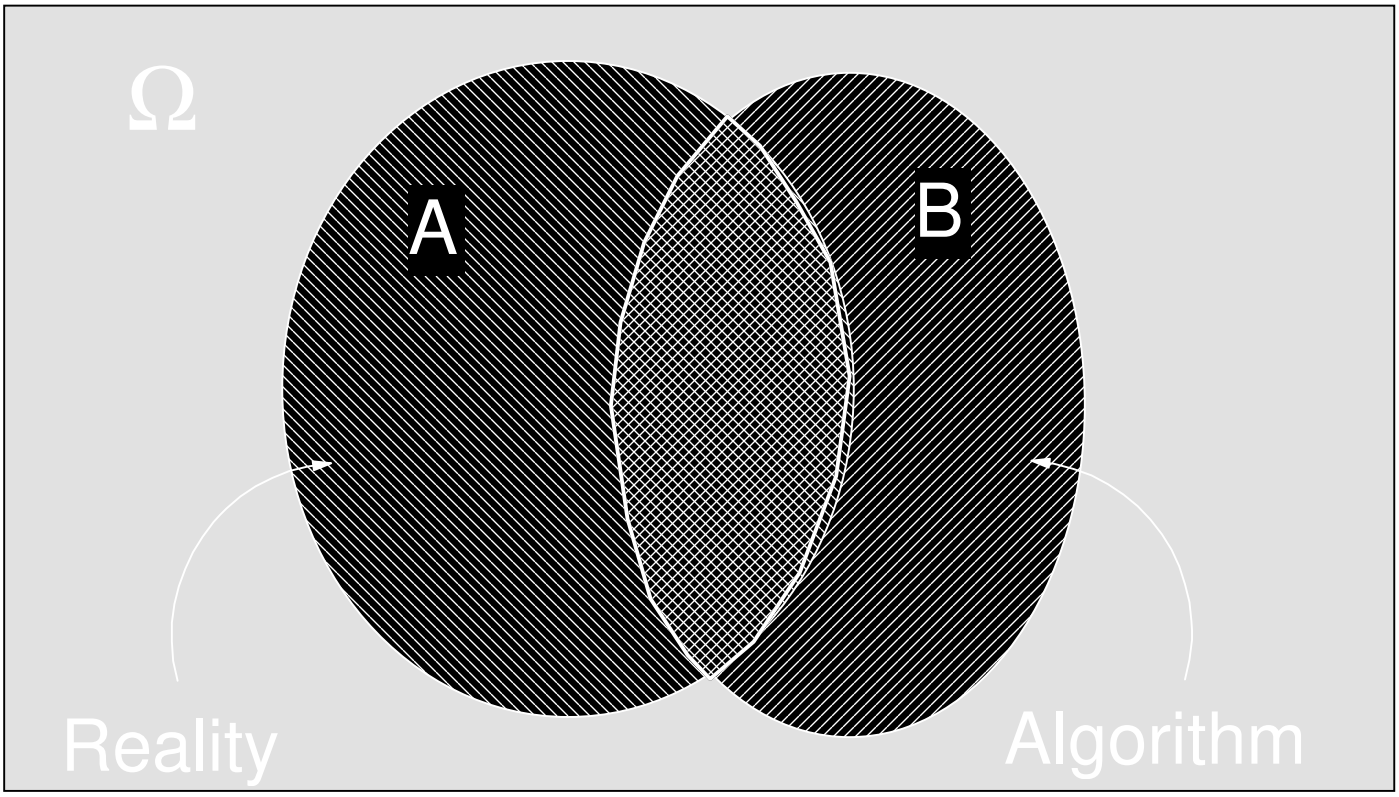
Bottom-up / Enumeration
- begin with a trivial pattern
- extend in all possible ways
- check if extension has sufficient support
  - if yes, repeat
  - otherwise, backtrack

Top-down / Alignment-based
- build alignment
- determine motifs

Hybrid approaches
- enumerate *all* patterns of relatively small size
- use patterns as alignment anchors & extend left and right

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - Considerations



- true positives: A
- true negatives: $\Omega - A$
- false positives: $B \setminus A \cap B$
- false negatives: $A \setminus A \cap B$

sensitivity: $|A \cap B| / |A|$

specificity: $|A \cap B| / |B|$

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery (cont.)

McCain Attacks Bush, Christian Right

By Alan Elsner, Political Correspondent

WASHINGTON (Reuters) - In a big political gamble, Republican presidential candidate John McCain (news - web sites) on Monday condemned two top Christian conservatives as bigots and slanderers and his rival George W. Bush (news - web sites) as their willing acolyte. Speaking on the eve of Republican primaries in Virginia and the state of Washington and caucuses in North Dakota, McCain, a senator from Arizona, nailed his colors to the mast in one of the most provocative speeches in recent presidential campaigns, delivered in the heart of Christian Coalition country. ``Political intolerance by any political party is neither a Judeo-Christian nor an American value. The political tactics of division and slander are not our values," McCain said. In a news conference in Bellevue, Washington, Bush accused McCain of ``shamelessly playing on religious fear." Christian Coalition vice president Roberta Combs said her organization would rise above McCain's attack and draw citizens to the polls in record numbers. Addressing supporters in Virginia Beach, a few miles from Christian Coalition headquarters, McCain appeared to burn his bridges to that organization permanently, calling their leaders intolerant, divisive and un-American. He condemned as an ``agent of intolerance" Pat Robertson, the founder of the Christian Coalition and of the Christian Broadcasting Network, who has been a potent force in the evangelical conservative community for many years. McCain also targeted Jerry Falwell, founder of the Moral Majority in the 1980s and of Liberty University, a fundamentalist institution in Virginia, and Bob Jones, who runs an even more fundamentalist college in South Carolina which bans inter-racial dating among students. McCain dismissed his rival Bush, the governor of Texas, in scathing terms as a creature of these religious leaders who had done little to distance himself from them. ``My friends, I am a Reagan Republican who will defeat Al Gore (news - web sites). Unfortunately, Gov. Bush is a Pat Robertson Republican who will lose to Al Gore," McCain said, referring to the vice president who is likely to become the Democratic Party's presidential nominee. McCain said the Republican Party had lost its way and compared Falwell and Robertson to controversial black leaders Louis Farrakhan and Al Sharpton, whom he said were intolerant figures in the Democratic Party camp. Reagan Didn'T Point Fingers ``It sounds like Sen. McCain has taken to name calling, needless name calling," Bush responded. ``Ronald Reagan didn't point fingers. He never played to people's religious fears like Sen. McCain has shamelessly done." ``I'm a problem solver. It sounds like Sen. McCain is a finger pointer," he said. McCain said Robertson and Falwell had declared him unacceptable as a Republican presidential candidate. ``They distort my pro-life positions and smear the reputations of my supporters. Why? Because I don't pander to them, because I don't ascribe to their failed philosophy that money is our message," McCain said. Polls last week had shown McCain only 10-12 percentage points behind Bush in Virginia and even with him in Washington. Both men are frantically scrambling for advantage ahead of "Super Tuesday" on March 7, when 12 states including New York and California hold Republican primaries or caucuses. Political scientist Robert Holsworth of Virginia Commonwealth University, said McCain's words would reverberate for a long time and would rally all those who disliked the influence of the Christian Coalition to his banner. ``He was seeking to draw a contrast between his courage in facing intolerance and Bush's reticence in failing to condemn Bob Jones University for so long," he said. McCain's upstart campaign has rocked Bush who started the campaign as the overwhelming favorite for the Republican nomination having collected more money than any candidate in U.S. history -- over $70 million. Now, most of that money is spent and Bush is in a fight for survival. But McCain's victories, in the New Hampshire and Michigan primaries, have come largely due to the support of independents and Democrats who were allowed to vote in the Republican contests in those states. The primaries in Virginia and Washington are also open to all-comers but many of the contests on March 7 will be Republican only, giving Bush an edge.

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery (cont.)

4	4	IN VIRGINIA AND
3	3	LIKE SEN MCCAIN
3	3	NEWS WEB SITES
3	2	OF THE CHRISTIAN
2	2	FOUNDER OF THE
2	2	REPUBLICAN WHO WILL
2	2	THE . OF THE CHRISTIAN COALITION
2	2	REPUBLICAN PRESIDENTIAL CANDIDATE
2	2	LIKE SEN MCCAIN HAS
2	2	PRIMARIES IN VIRGINIA AND
2	2	IT SOUNDS LIKE SEN MCCAIN
2	2	OF . . ON MARCH 7
2	2	THE . . . OF THE
2	2	REAGAN DIDNT POINT FINGERS
2	2	CAMPAIGN FEBRUARY 27
2	2	THAT MONEY IS
2	2	MCCAIN SAID . . . HAD
2	2	OUR . MCCAIN SAID
2	2	THE . OF . AND
2	2	WHO . THE . . THE
2	2	IN . . FEBRUARY 28
2	2	MCCAIN . . CHRISTIAN RIGHT

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery (cont.)

Declaration Des Droits De L' Homme

17	17	TOUTE PERSONNE A
17	16	A DROIT à
12	12	$$ TOUTE PERSONNE
11	11	DROIT à LA
11	11	TOUTE PERSONNE A DROIT
10	10	A DROIT à LA
10	10	$$ TOUTE PERSONNE A
9	9	LE DROIT DE
8	8	TOUTE PERSONNE A DROIT à
7	7	A LE DROIT
7	7	$$ . . A DROIT
7	7	NE PEUT êTRE
6	6	$$ . . A DROIT à
6	6	DROIT ET LIBERTéS
6	6	TOUTE PERSONNE A LE DROIT DE
6	6	DE . . ET DE
5	5	DES NATION UNIES
5	5	à LA . DE
5	5	$$ TOUTE PERSONNE A LE DROIT DE
5	5	A DROIT à . . DE
5	5	DROIT DE L'HOMME
5	5	$$ TOUTE PERSONNE A DROIT
5	5	DE . . . ET DE
5	5	LA PRéSENTE DéCLARATION
5	5	DES DROIT DE
6	4	DE . DE . . DE
5	4	LA . DE LA
5	4	DROIT . LA . DE
...		

3	3	DE LA . HUMAINE
3	3	LA . . DE LA
3	3	A DROIT à LA LIBERTé
3	3	DE LA . ET DE
3	3	A DROIT à LA PROTECTION
3	3	à LA . DES
3	3	DE . . ET . LA
3	2	LES . ET . LES
3	2	DE . . . DE . DE
3	2	TOUS LES . ET
3	2	DROIT . LA LIBERTé DE
3	2	DROIT . LA . . . DE
3	2	DE . . . OU DE
2	2	A DROIT à LA PROTECTION . . . ET
2	2	DE . . DE LA . ET DE LA
2	2	A DROIT à LA PROTECTION DE LA . DE
2	2	LES . ET TOUS LES
2	2	$$ . . A DROIT à UNE
2	2	$$ . . NE PEUT êTRE . . LE
2	2	TOUTE PERSONNE A DROIT AU
2	2	ONT DROIT à UNE
2	2	ONT DROIT SANS
2	2	DROIT ET LIBERTéS . . LA
2	2	TOUTE PERSONNE A DROIT à UN
2	2	LES DROIT ET
2	2	DROIT à LA . EN
2	2	$$ TOUTE PERSONNE A DROIT à . . DE
2	2	LES êTRES HUMAINS . LIBRES
...		

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - Continuous Streams

THECONSTITUTIONOFTHEUNITEDSTATES
SEARCHTHECONSTITUTIONBROWSEBILLOFRIGHTSOTHERAMENDMENTSA
BOUTTHECONSTITUTIONWETHEPEOPLEOFTHEUNITEDSTATESINORDERTO
FORMAMOREPERFECTUNIONESTABLISHJUSTICEINSUREDOMESTICTRANQ
UILITYPROVIDEFORTHECOMMONDEFENSEPROMOTETHEGENERALWELFAR
EANDSECURETHEBLESSINGSOFLIBERTYTOOURSELVESANDOURPOSTERIT
YDOORDAINANDESTABLISHTHISCONSTITUTIONFORTHEUNITEDSTATESOFA
MERICAARTICLEISECTION1ALLLEGISLATIVEPOWERSHEREINGRANTEDSHAL
LBEVESTEDINACONGRESSOFTHEUNITEDSTATESWHICHSHALLCONSISTOFA
SENATEANDHOUSEOFREPRESENTATIVESSECTION2THEHOUSEOFREPRESE
NTATIVESSHALLBECOMPOSEDOFMEMBERSCHOSENEVERYSECONDYEARBY
THEPEOPLEOFTHESEVERALSTATESANDTHEELECTORSINEACHSTATESHALL
HAVETHEQUALIFICATIONSREQUISITEFORELECTORSOFTHEMOSTNUMEROU
SBRANCHOFTHESTATELEGISLATURENOPERSONSHALLBEAREPRESENTATIV
EWHOSHALLNOTHAVEATTAINEDTOTHEAGEOFTWENTYFIVEYEARSANDBEEN
SEVENYEARSACITIZENOFTHEUNITEDSTATESANDWHOSHALLNOTWHENELEC
TEDBEANINHABITANTOFTHATSTATEINWHICHHESHALLBECHOSENREPRESE
NTATIVESANDDIRECTTAXESSHALLBEAPPORTIONEDAMONGTHESEVERALST
ATESWHICHMAYBEINCLUDEDWITHINTHISUNIONACCORDINGTOTHEIRRESPE
CTIVENUMBERSWHICHSHALLBEDETERMINEDBYADDINGTOTHEWHOLENUME
EROFFREEPERSONSINCLUDINGTHOSEBOUNDTOSERVICEFORATERMOFYEA
RSANDEXCLUDINGINDIANSNOTTAXEDTHREEFIFTHSOFALLOTHERPERSONS
THEACTUALENUMERATIONSHALLBEMADEWITHINTHREEYEARSAFTERTHEFI
STMEETINGOFTHECONGRESSOFTHEUNITEDSTATESANDWITHINEVERYSUBS
EQUENTTERMOFTENYEARSINSUCHMANNERASTHEYSHALLBYLAWDIRECTTH
ENUMBEROFREPRESENTATIVESSHALLNOTEXCEEDONEFOREVERYTHIRTY
THOUSANDBUTEACHSTATESHALLHAVEATLEASTONEREPRESENTATIVEAND
UNTILSUCHENUMERATIONSHALLBEMADETHESTATEOFNEWHAMPSHIRE
SHALLBEENTITLEDTOCHUSETHREEMASSACHUSETTSEIGHTRHODEISLAN
DANDPROVIDENCEPLANTATIONSONE....

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - Continuous Streams (cont.)

191 148 SHALL
133 109 STATE
129 93 OFTHE
90 80 STATES
80 76 SHALLB
78 74 SHALLBE
55 55 UNITEDSTATES
55 55 THEUN
54 54 THEUNI
52 52 THEUNITEDSTATES
52 43 EPRES
43 42 SSHALL
...
42 37 THECO
37 37 OFTHEUNI
39 36 CTION
36 35 ESAND
35 35 OFTHEUNITEDSTATES
...
29 28 PRESENT
36 27 PRESID
36 27 RESIDE
35 27 RESIDENT
30 27 SOFTH
28 27 HOUSE
35 26 PRESIDE
34 26 PRESIDENT
...

29 26 ENTOF
28 26 IONOF
29 25 OFTHES
28 25 SOFTHE
27 25 NTOFT
29 24 SENAT
28 24 TOFTHE
26 24 TIONS
26 24 NTOFTH
25 24 EPRESEN
25 23 TIONOF
...
6 6 THESEVERALSTATES
6 6 OFTWO
6 6 SHALLHAVEP
6 6 ANDBE
6 6 ANYOF
6 6 OFTHELEGISLATURE
6 6 ANDMA
6 6 SHALLNOTBE
6 6 ANDWI
6 6 STATESW
6 6 EAUTHORI
6 6 ECTIONS
6 6 YEARSAND
6 6 THESUPREME
6 6 IMPEACHMENT
...

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - "Teiresias"

...RRYQKSTELLIRKLPFQRLVREIAQDFKTDLRFQSAAIGA...

11100

11010

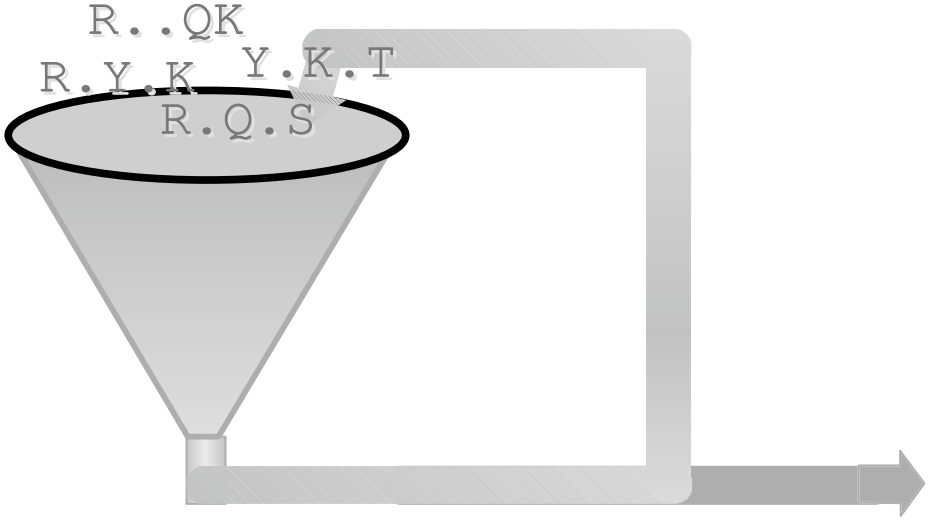
11001

Enumeration
(scanning phase)

k = 3 / w = 5

11100
11010
11001
10110
10101
10011

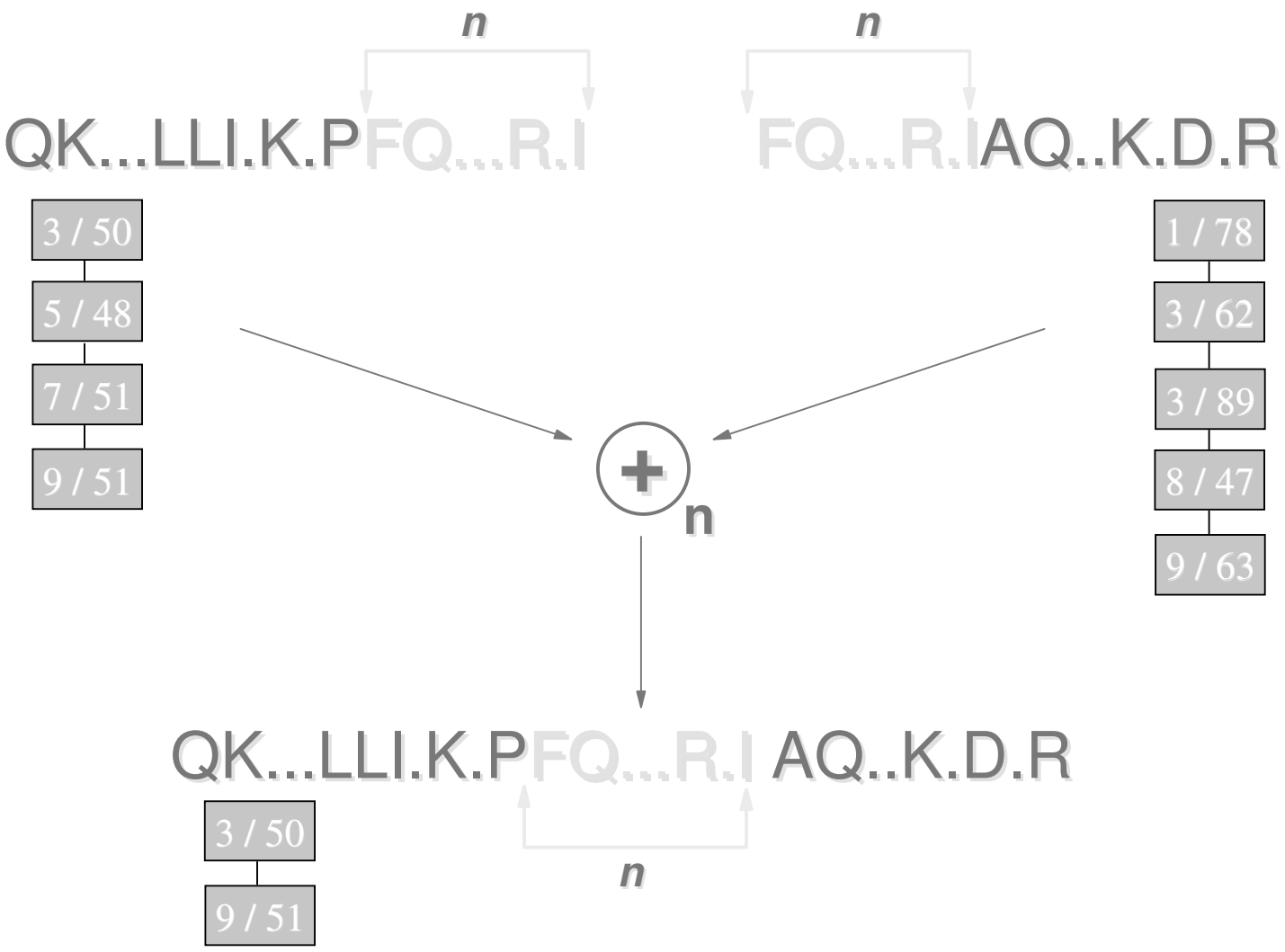
RRY YQK Y.KS
RY.K YQ.S
RR.Q Y..ST
RR..K RY..S
RYQ YQ..T
R.QK
R.YQ R..KS
R..QK
R.Y.K Y.K.T
R.Q.S



(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - "Teiresias" (cont.)

Convolution



Pattern Discovery - "Teiresias" (cont.)

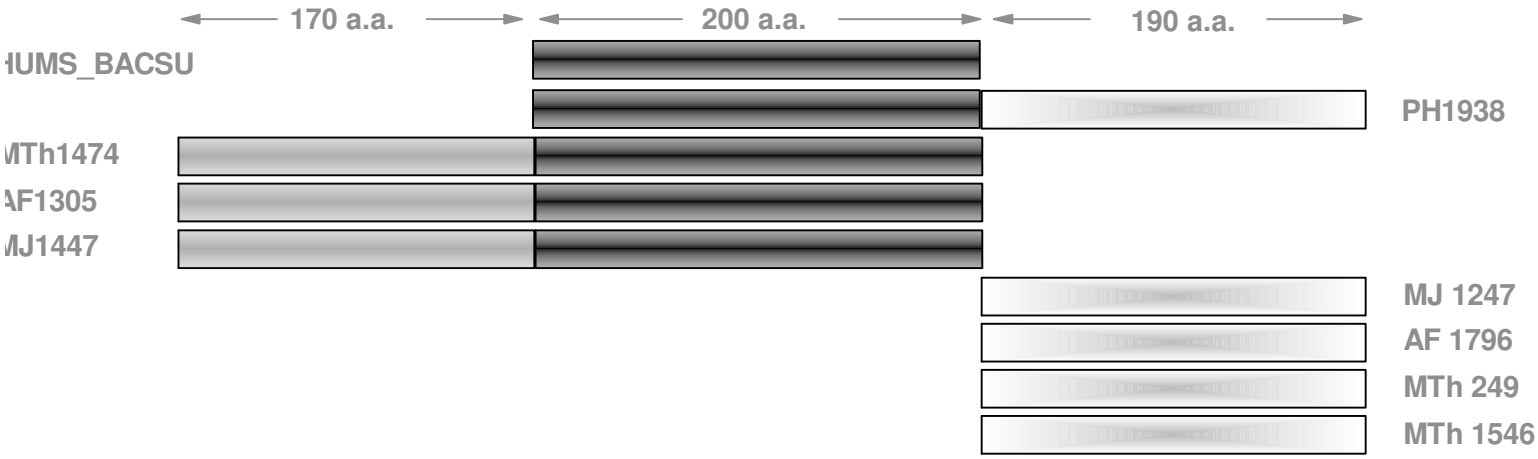
- Algorithm will report
  - *all* patterns appearing at least *k* times in the input
  - patterns that in most general form consist *entirely* of bracketed expressions, e.g. [NS][LIMYT][FYDN].[DNT][IMVY].[STGDN][DN]..[SGAP]
  - patterns *maximal* in *composition*
  - patterns *maximal* in *length*
- Time complexity: *quasi-linear* in the size of the generated output

About The "MUSCA" Algorithm

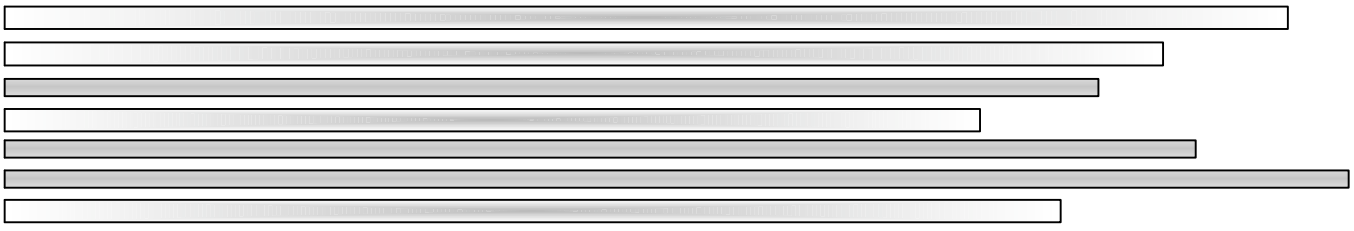
(c) I. Rigoutsos / August 17, 2001

About "MUSCA" (cont.)

- Can easily align inputs of the type:

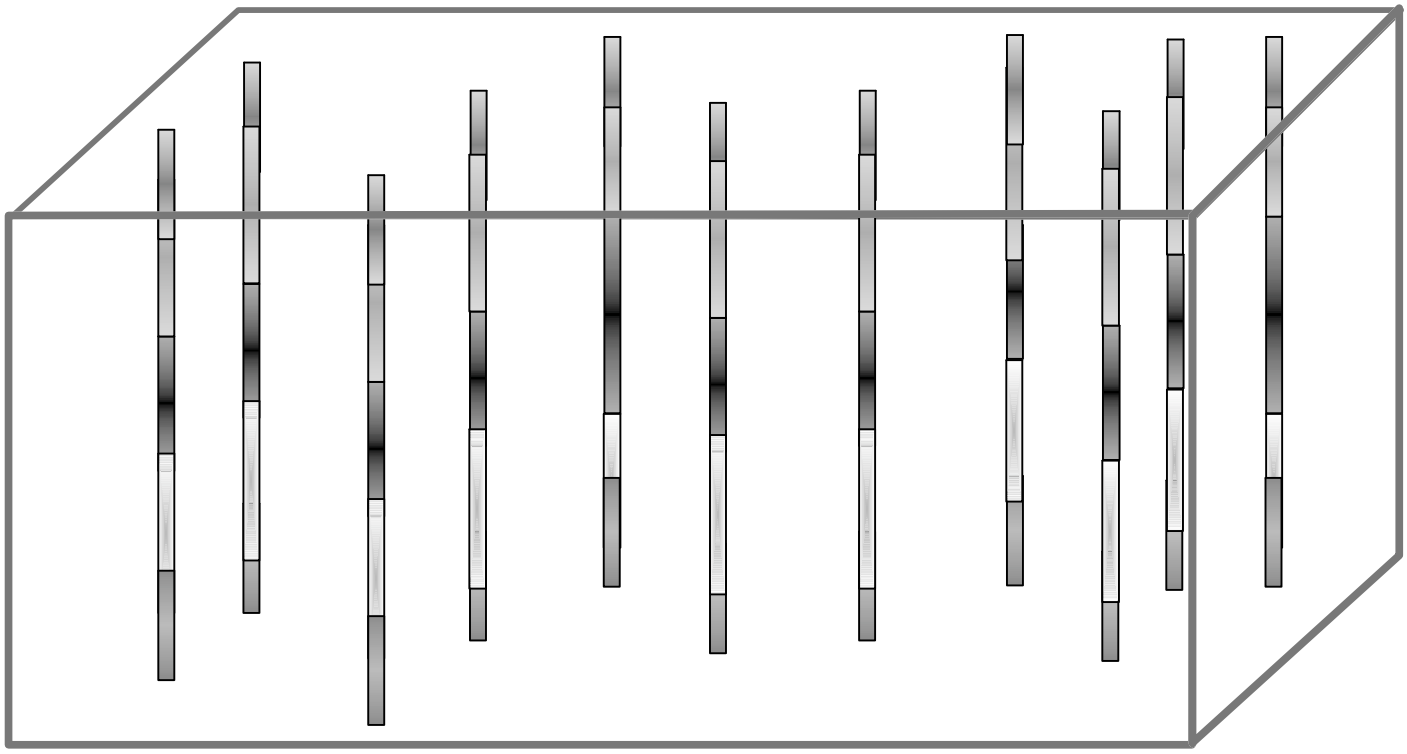


- Can easily align "mixed" inputs:



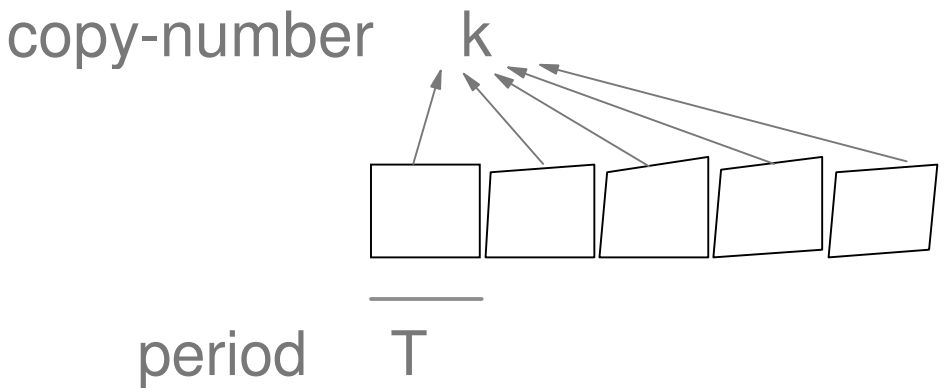
(c) I. Rigoutsos / August 17, 2001

Multiple Sequence Analysis & Oil Exploration



(c) I. Rigoutsos / August 17, 2001

Tandem-Repeat Discovery



Similarity Parameters:
μ (adjacent similarity)
p (overall similarity)

Central Idea:

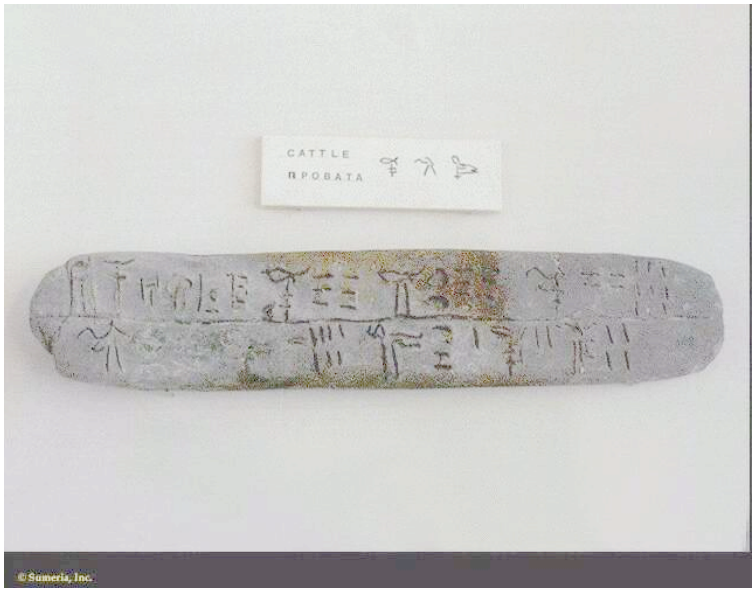
Tandem Repeats are pattern "nucleation sites"

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - Small vs. Large Databases

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - Small vs. Large Databases (cont.)



Linear B

Linear A

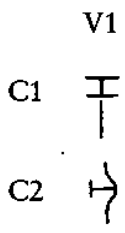
Pattern Discovery - Small vs. Large Databases (cont.)

We can see the inflection more clearly if we highlight the word endings:

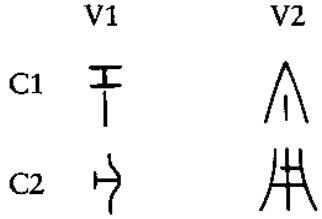
An English parallel might be:

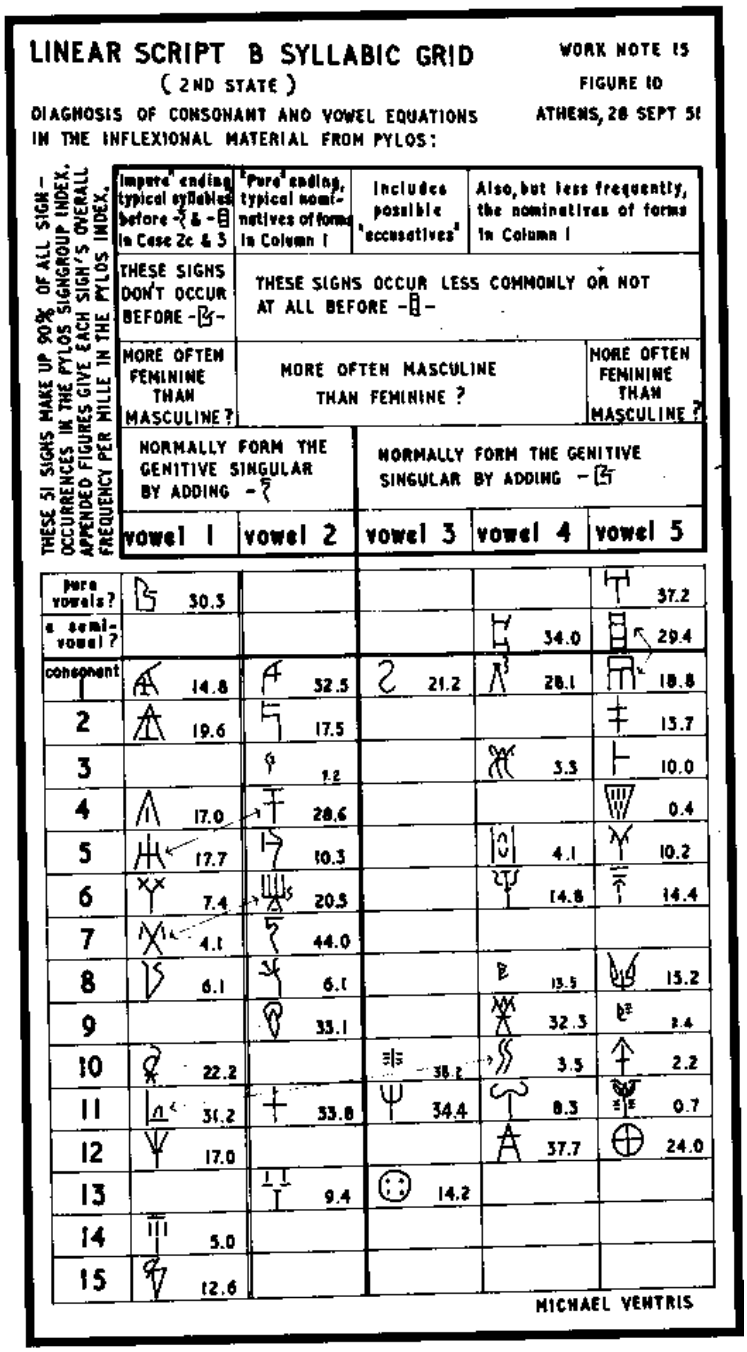
Ca-na-da	Ar-ge(n)-ti-na
Ca-na-di-a(n)	Ar-ge(n)-ti-ni-a(n)
Ca-na-di-a-(ns)	Ar-ge(n)-ti-ni-a(ns)

If such parallels were right (assuming Linear B was syllabic, like the Cypriot script), and would have different consonants (C) but the same vowel (V), like *da* and *na* above, i.e.:



So would and like *di* and *ni* above:

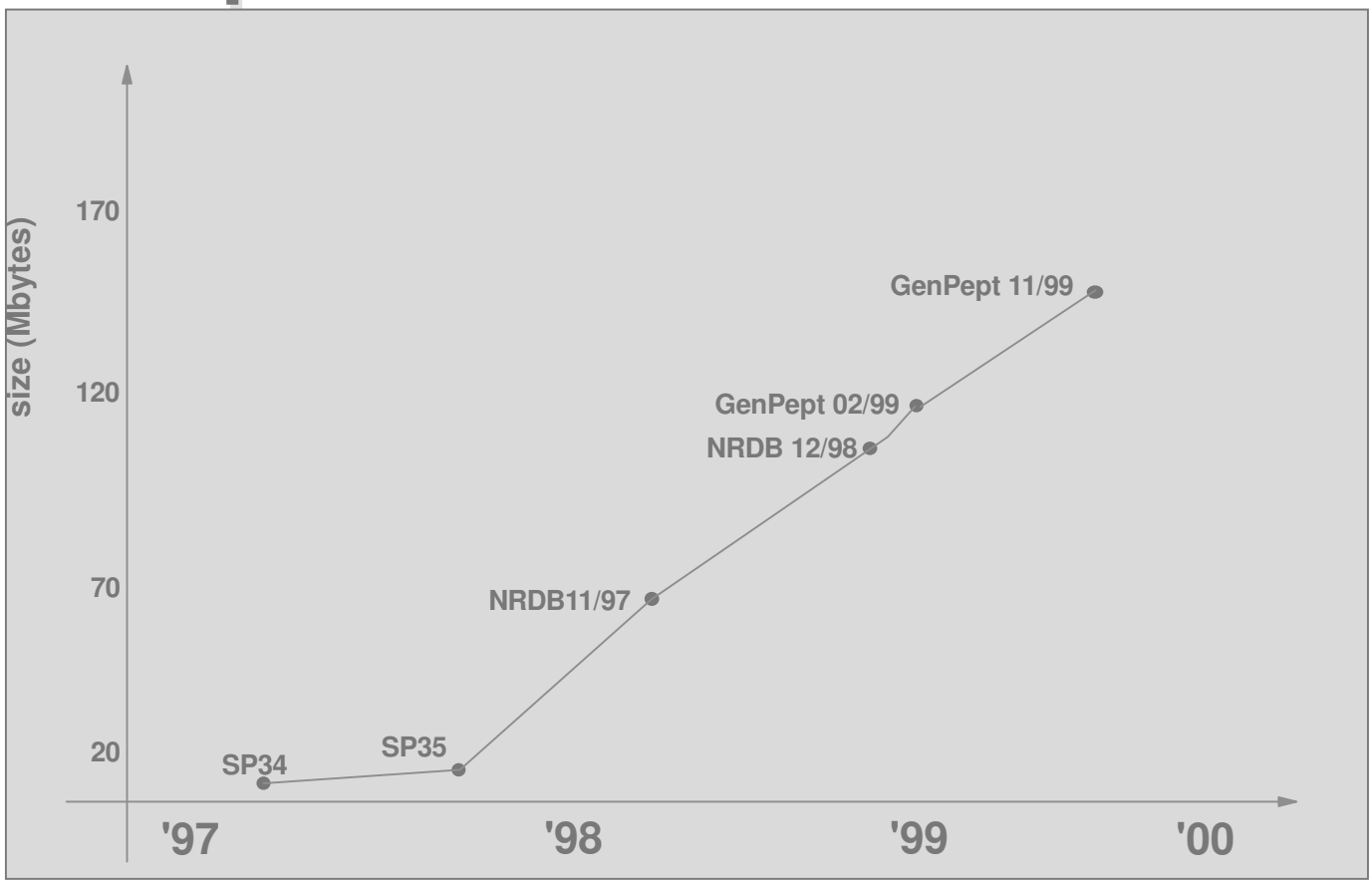


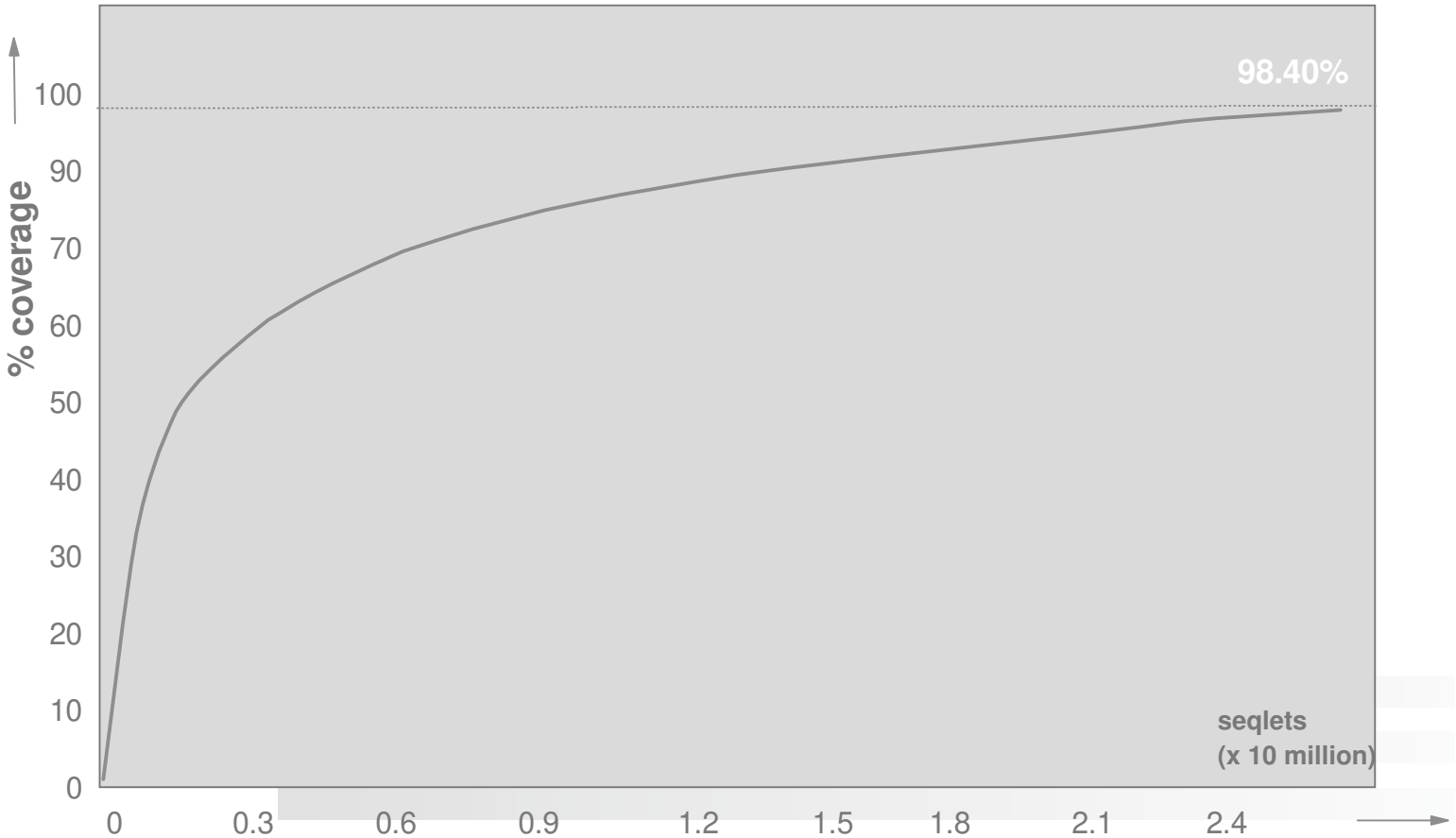


Figures from: The Story of Writing by Andrew Robinson pp. 114-115

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - The Case of GenPept





(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - A Webster's For Proteins

#	*SEQLET*	*SEQLET's* "Functional Meaning"
1	G..G.GK[STG]TL	ATP/GTP binding P-loop
2	H.....HRD.K..N	Ser/Thr-protein kinases
3	SGG[QEMRY]..R[VLIA].[IGLMV]R.L	ABC transporters
4	V.I.G.G..G...A	NAD/FAD-binding, Flavoproteins
6	G.GLGL.I	Sensory transd. His-prot. kinases
..		
10	GA.DY[LIV].KP	2-component sensory transductio
11	HR.GR..R....G	DEAD-box helicases
12	GDG[IVAMTD]ND[AILV][PEAS][AMV][LMIF]..A	Cation-transporting ATPases
13	D.FK.[IYVFL]N[DE].[YLFWR]GH..GD.[CLVF]L	Bacterial-type regulators*
14	DKT[GV]TLT	Cation-transporting ATPases
..		
16	KMSKS[LKDIR][GNDFQ]N	Amino-acyl-tRNA synthetases I
17	PTREL..Q	DEAD-box helicases
18	Q..GRAGR	DEAD-box helicases
19	F.[ASDN].[MIVTLA][SAT]HE[LIF]RTP	Sensory transd. His-prot. kinases
..		
27	DL[IVL][LIMVF]LD[ILVW].[ML]P..[DNST]G	2-component sensory transd.
..		
30	LD.GCG.G	Various methyltransferases
..		
34	T.[IVL][FLYMI]VTHD[QLIVP].[ELV]A	ABC transporters
..		
.		

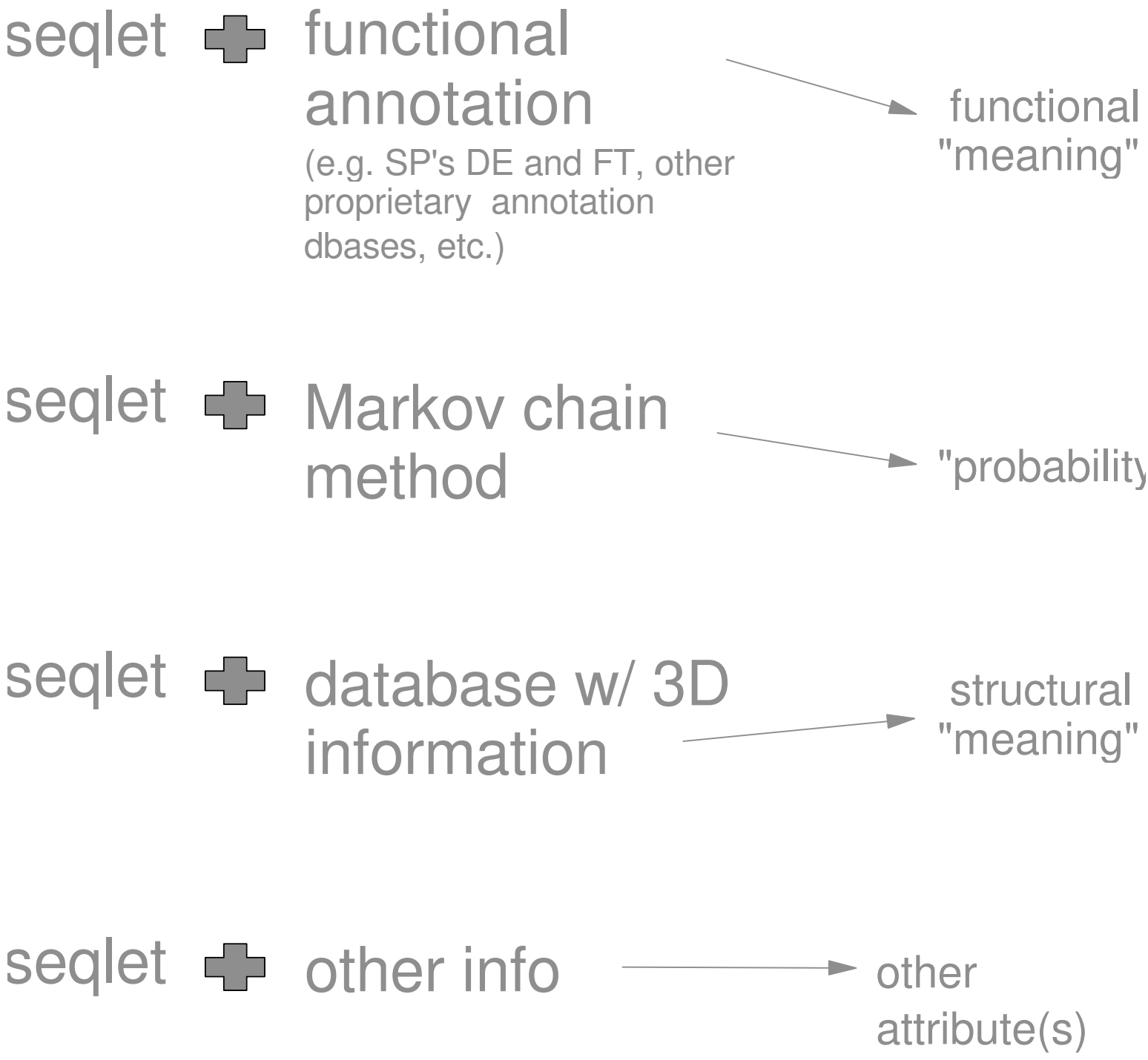
(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - An Example Swiss-Prot Entry

```
ID   1B07_HUMAN     STANDARD;      PRT;   362 AA.
AC   P30462;
DT   01-APR-1993 (Rel. 25, Created)
DT   01-APR-1993 (Rel. 25, Last sequence update)
DT   01-APR-1993 (Rel. 25, Last annotation update)
DE   HLA CLASS I HISTOCOMPATIBILITY ANTIGEN, B-14 B*1401 ALPHA CHAIN
DE   PRECURSOR.
GN   HLA-B OR HLAB.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 89235215.
RA   PARHAM P., LAWLOR D.A., LOMEN C.E., ENNIS P.D.;
RT   "Diversity and diversification of HLA-A,B,C alleles.";
RL   J. Immunol. 142:3937-3950(1989).
CC   -!- FUNCTION: INVOLVED IN THE PRESENTATION OF FOREIGN ANTIGENS TO
CC       THE IMMUNE SYSTEM.
CC   -!- SUBUNIT: DIMER OF ALPHA CHAIN AND A BETA CHAIN (BETA-2-
CC       MICROGLOBULIN).
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use  by  non-profit  institutions as long  as its content  is  in  no  way
CC   modified and this statement is not removed.  Usage  by  and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M24040; AAA59661.1; -.
DR   HSSP; P30460; 1AGB.
DR   MIM; 142830; -.
DR   PFAM; PF00129; MHC_I; 1.
DR   PROSITE; PS00290; IG_MHC; 1.
KW   MHC I; Transmembrane; Glycoprotein; Signal.
FT   SIGNAL        1     24
FT   CHAIN        25    362       HLA CLASS I HISTOCOMPATIBILITY ANTIGEN,
FT                                 B-14 B*1401 ALPHA CHAIN.
FT   DOMAIN       25    114       EXTRACELLULAR ALPHA-1.
FT   DOMAIN      115    206       EXTRACELLULAR ALPHA-2.
FT   DOMAIN      207    298       EXTRACELLULAR ALPHA-3.
FT   DOMAIN      299    309       CONNECTING PEPTIDE.
FT   TRANSMEM    310    333
FT   DOMAIN      334    362       CYTOPLASMIC TAIL.
FT   CARBOHYD    110    110       BY SIMILARITY.
FT   DISULFID    125    188       BY SIMILARITY.
FT   DISULFID    227    283       BY SIMILARITY.
SQ   1B07_HUMAN 450513
MLVMAPRTVLLLLSAALALTETWAGSHSMRYFYTSVSRPGRGEPRFISVGYVDDTQFVRFDSDAASPREEPRAPWIEQEG
PEYWDRNTQICKTNTQTDRESLRNLRGYYNQSEAGSHTLQWMYGCDVGPDGRLLRGYNQFAYDGKDYIALNEDLSSWTAA
DTAAQITQRKWEAAREAEQLRAYLEGTCVEWLRRHLENGKETLQRADPPKTHVTHHPISDHEATLRCWALGFYPAEITLT
WQRDGEDQTQDTELVETRPAGDRTFQKWAAVVVPSGEEQRYTCHVQHEGLPKPLTLRWEPSSQSTVPIVGIVAGLAVLAV
VVIGAVVAAVMCRRKSSGGKGGSYSQAASSDSAQGSDVSLTA
```

(c) I. Rigoutsos / August 17, 2001

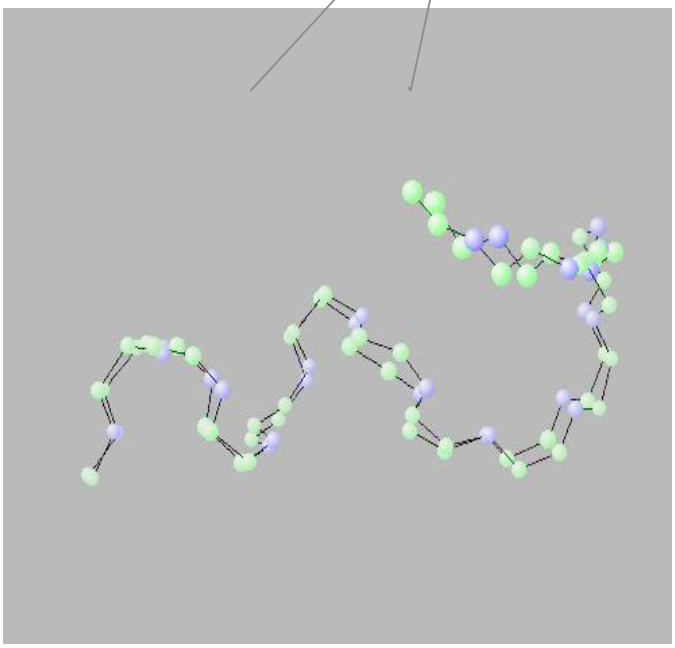
Pattern Discovery - A Webster's For Proteins (cont.)

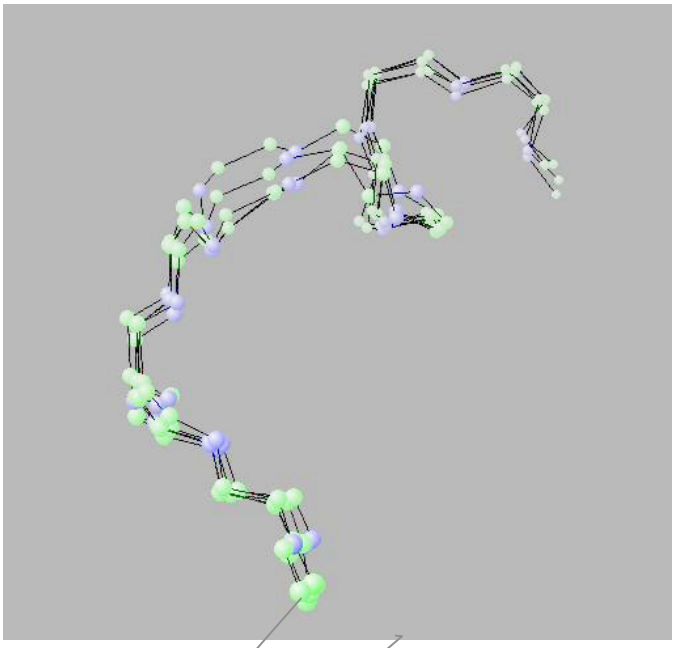


(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - A Webster's For Proteins (cont.)

Seqlet = G..G.[GAI]KST....L
Log Probability= -29.443913
# of occurences= 27 # of sequences = 27
(SeqId,Offset)s= [2917 27], [2918 20], [2919 12] ...
PDB Rel 38 Hits= 1nip_B_ 1gky___
PDB RMS Error = 0.649908 Angstr.
3D Struct File = G..G.[GAI]KST....L.mol2
Seqlet Annotat.= 229 NP_BIND ATP / 11 NP_BIND GTP





Seqlet = C..[CVY]G.C..VCP
Log Probability= -33.055840
# of occurences= 56 # of sequences = 51
(SeqId,Offset)s= [1598 415], [5102 215], [5938 416] ....
PDB Rel 38 Hits= 1bc6___15 1fca___ 1fdn___ 1clf___1 1fdx___
PDB RMS Error = 0.349209 Angstr.
3D Struct File = C..[CVY]G.C..VCP.mol2
Seqlet Annotat.= 159 Iron-Sulfur (4FE-4S) binding

(c) I. Rigoutsos / August 17, 2001

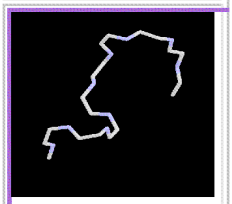
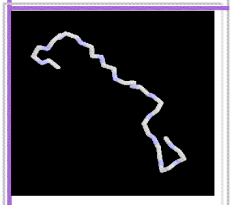
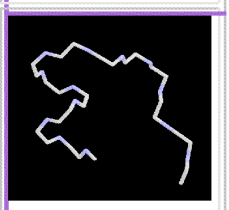
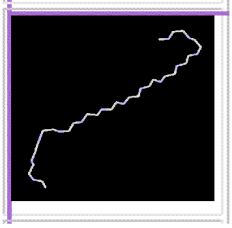
Pattern Discovery - Toward Saturation?

DATABASE	Release	SIZE	# SEQLETS	
SP36	05/97	~27M aa	~3.5 M	94.0%
GenPept	12/98	~119M aa	~26 M	
GenPept	11/99	~137M aa	~28 M	~98.2 %
SP/TrEMBL	07/00	~151M aa	~29 M	
SP/TrEMBL	05/01	~175M aa	~43 M	~98.2 %

(c) I. Rigoutsos / August 17, 2001

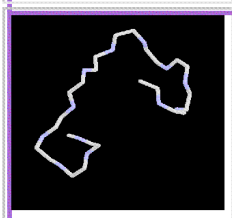
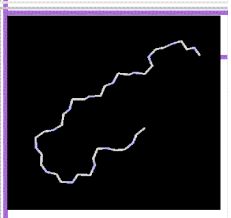
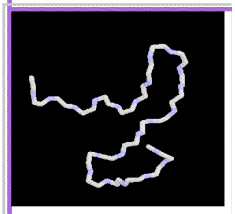
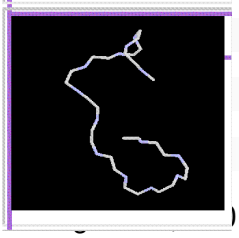
Pattern Discovery - Examples

file:///D|/Www/Group/ISL StandAlone Demo 122499/f

SEQLET	Its Functional Meaning	Its 3D Structure
G..G.GK[STG]TL	ATP/GTP binding P-loop	
H.....HRD.K..N	Ser/Thr-protein kinases	
SGG[QEMRY]..R[VLIA].[IGLMV]R.L	ABC transporters	
V.I.G.G..G...A	NAD/FAD-binding, Flavoproteins	
G.GLGL.I	Sensory transd. His-prot. kinases	
GA.DY[LIV].KP	2-component sensory transduction	
D.G.Y.V...N..G	IG-like C2-type domain	

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - Examples (cont.)

GDG[IVAMTD]ND[AILV][PEAS][AMV][LMIF]..A	Cation-transporting ATPases	
D.DGNG.....EL	Ca binding site	
D.FK.[IYVFL]N[DE].[YLFWR]GH..GD.[CLVF]L	Bacterial-type regulators*	
L.GG.G.D...G..G	??????	
Y.G....T..G..CQ.W....PH	Kringle domain type IV	
DKT[GV]TLT	Cation-transporting ATPases	
KMSKS[LKDIR][GNDFQ]N	Amino-acyl-tRNA synthetases I	
CG.D...Y...C.L	Kazal-type domain	

We have essentially compiled a comprehensive collection of patterns (=seqlets) for this 'unknown' language.

Some of these seqlets (a relatively small percentage) have been encountered before by other researchers and annotations exist.

But similar to words in natural languages, some words serve as connecting or modifying agents and appear in sentences (resp. proteins) that convey different meaning (resp. protein function).

In such cases, the seqlets can only be annotated with the union of the functions that correspond to the proteins that contain them.

(c) I. Rigoutsos / August 17, 2001

Similar to the grammatical variations in natural languages, the Bio-Dictionary contains variations of the same 'coherent entity'

On the next page, we are showing an example of variations for the ATP/GTP-binding P-loop

Ideally we would like to be able to find the "one regular expression" to use instead of all these variations.

Such tight, variation-encompassing expressions can be thought of as "irredundant motifs" and are useful not only in the context we are examining but in all data-mining applications.

(c) I. Rigoutsos / August 17, 2001

Variations On A Theme

G..G.GKST	G.GK.TL...L	V.L.G..G.GK
G..G.GK[STGDAR]TL	L.G....GKST	L.G....GKTT
G..G.GKTT	G...SGKST	I.G....GKS.L
G..G.GKS.L	G....GK[STGK]TLL	G....GKT.LL
G..G.GKT.L	G.GKTT....L	L.G..G.G.T.L
L.G..G.GK[STGDK]T	G..G.GKT..A	G....GKTT.L
L.G..G.GKT	G..GSGKT	GP...GKT.L
L.G..G.GK..L	V...G..G.GKT	G.SG.G.ST
G..GSGKS	G....GKTTL	G..G.G.S.LL
GP.G.GKT	I.G..G.GKT	L.G....GKS.L
G..GSGK[STGDKY]T	G....GKST.L	G....GKSTL
G.GKT..A...A	I...G..G.GKT	I.G.SG.GK
L.G..G.GKS	G.GKT.....LA	G.GKTTL
G..G.GKS..L	G.GKS.L...L	L.[GASMDL]P.G.GKT
L..G..G.GKT	G....GKS.LL	GP...GKT..A
G.GKST....L	L.G....GKT.L	V.I.G..G.GK
G.SG.GKS	GSGKST	TGSGKT
I.G..G.GKS	G..GSG.ST	I.G....GKTT
G.GKSTL	I.G....GKST	G...SGKS..L

Variations On A Theme (cont.)

- **Irredundant motif: maximal and not redundant**
  - A maximal motif is redundant if its position list is the union of the position lists of some *m* **maximal motifs**

- **Permitted Operations:**

  "Prefix:" Pre(A..F.G..K) = A..F.G
  "Union:" A..F.G..K A.M....LK = A.MF.G.LK
  "Intersection:" A..F.G..K A.M....LK = A.....K

- **Result: "For any input string of size n there can be no more than 3n irredundant motifs"**

Pattern Discovery - Bio-Dictionary Uses

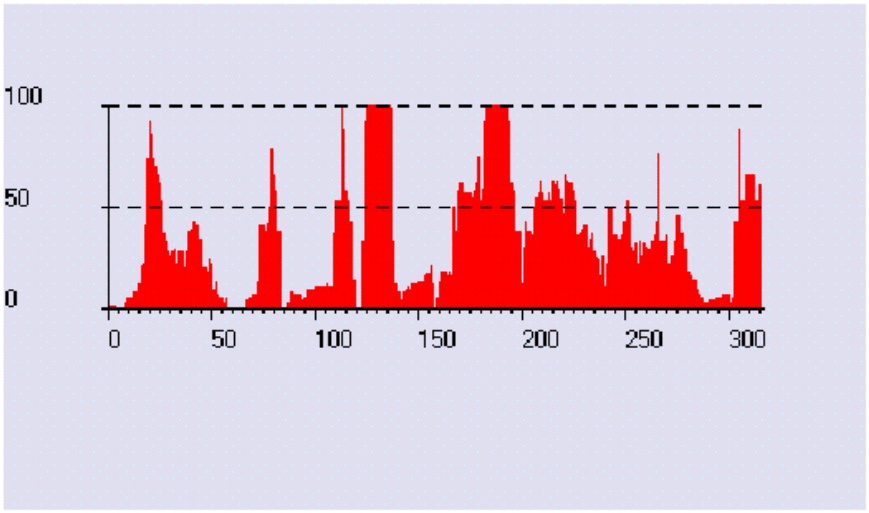
- homology searching
- functional annotation
- structural annotation
- phylogenetic domain analysis
- local structure characterization
- mini B-D's

  etc.

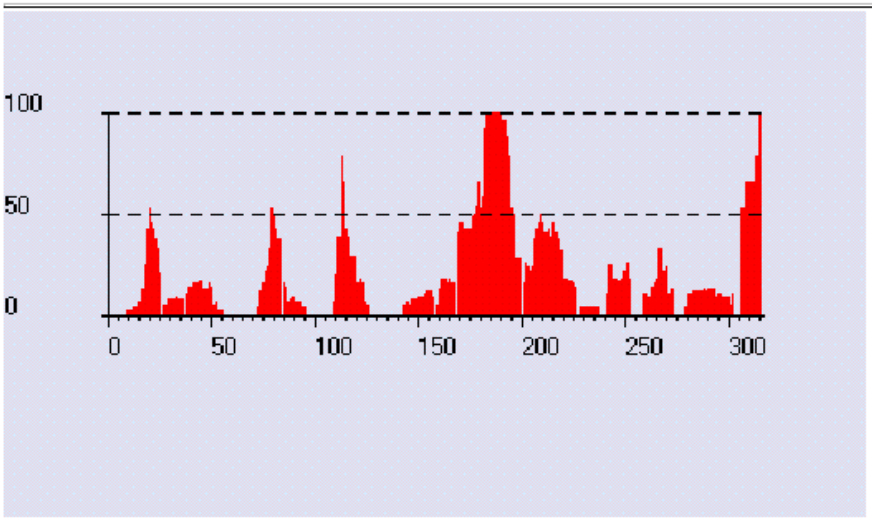
- visit: http://cbcsrv.watson.ibm.com/Tpa.html

  use "SAMPLE" as input...

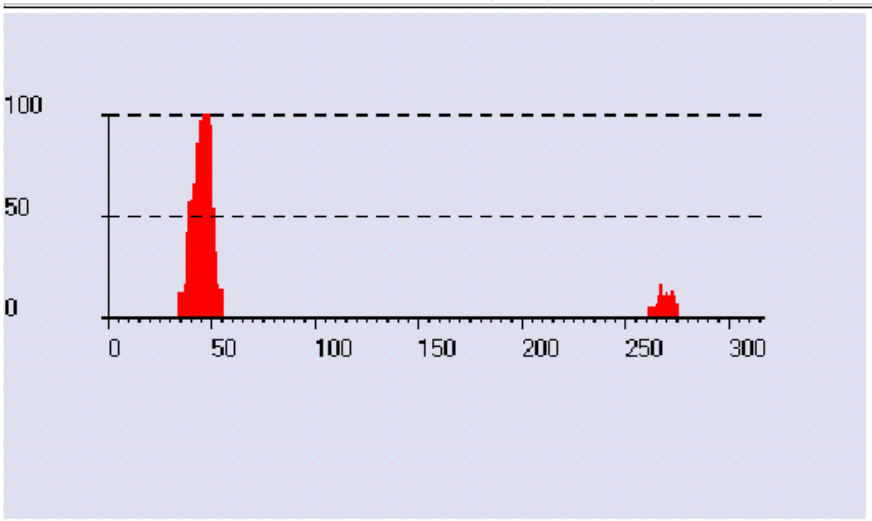
Pattern Discovery - Bio-Dictionary Uses (cont.)



[DE-1] aldose reductase (ec 1.1.1.21) (ar) (aldehyde reductase).



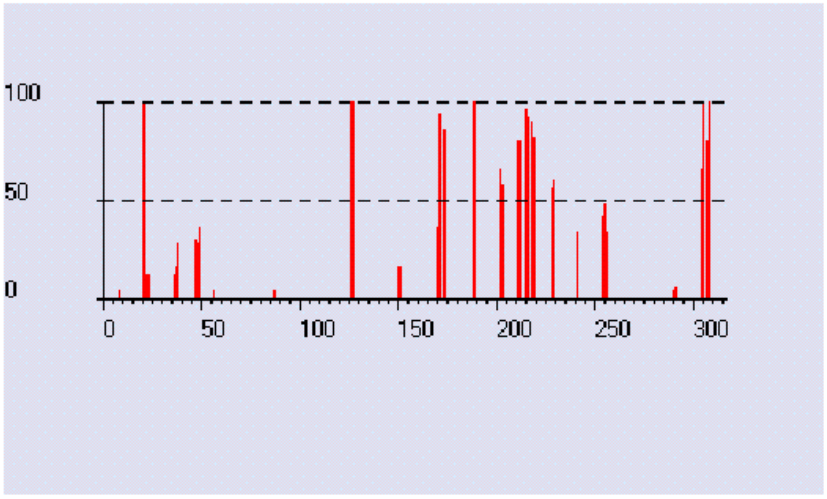
[DE-2] aldose reductase (ec 1.1.1.21) (ar) (aldehyde reductase) (20-alpha-



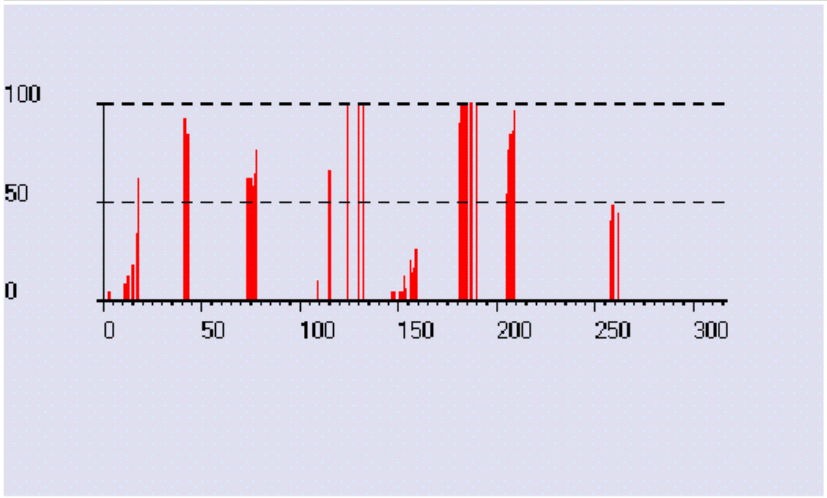
[DE-3] 2,5-diketo-d-gluconic acid reductase (ec 1.1.1.-) (2,5-dkg reductase).

(c) I. Rigoutsos / August 17, 2001

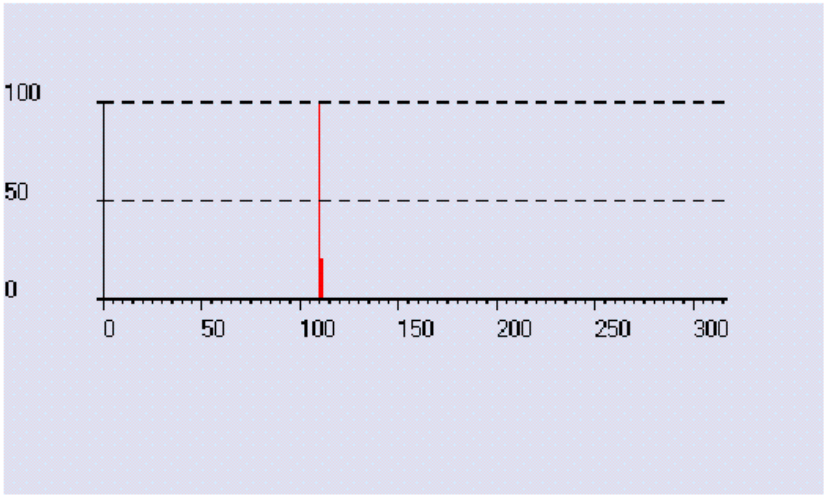
Pattern Discovery - Bio-Dictionary Uses (cont.)



[FT-1] turn
(Minimum, average, standard deviation) of this feature in related sequ



[FT-2] strand
(Minimum, average, standard deviation) of this feature in related sequ

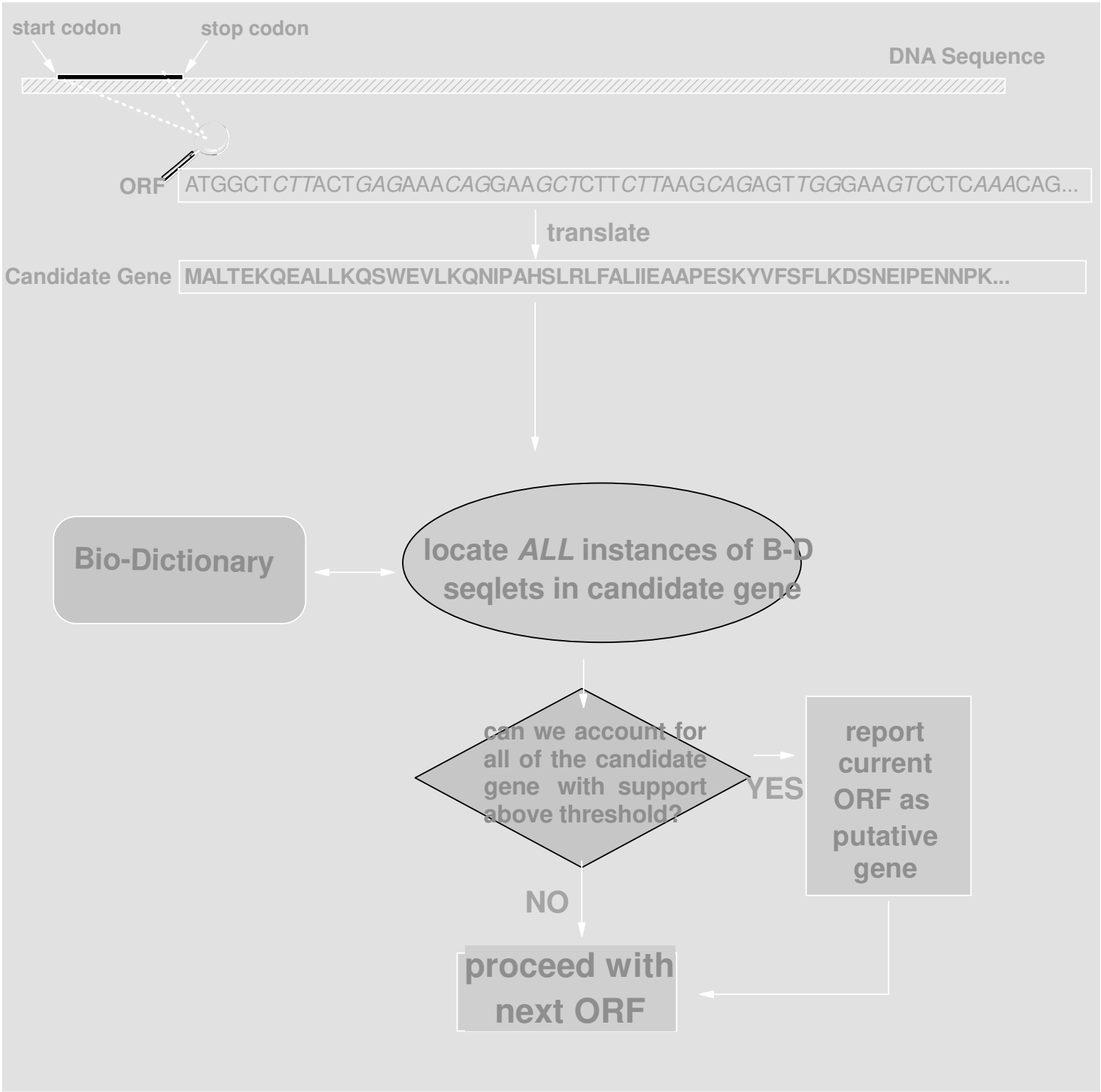


[FT-3] act_site hydrogen-bond donor (by similarity).
(Minimum, average, standard deviation) of this feature in related sequ

(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - Bio-Dictionary Uses (cont.)

Gene Identification



(c) I. Rigoutsos / August 17, 2001

Pattern Discovery - Statistics

"Something that is statistically significant/insignificant is not necessarilly biologically significant/insignificant"

Patricia Babitt

[ST]..[ST]C[ST][ST][NQ]..[AG]

glyceraldehyde 3-phosphate dehydrogenases

GDISYSLYLIHW → 29 instances / LogProb= -62.88

AREGRKFGVGL → 6 instances / LogProb= -54.78

C. elegans

Meth. thermoautotrophicum
Archaeoglobus fulgidus
Methanococcus jannaschii
Pyrococcus horikoshii

Pattern Discovery - Annotation Pitfalls

HUMS_BACSU

D-arabino 3-hexulose 6-phosphate formaldehyde lyase

(c) I. Rigoutsos / August 17, 2001

Association Discovery

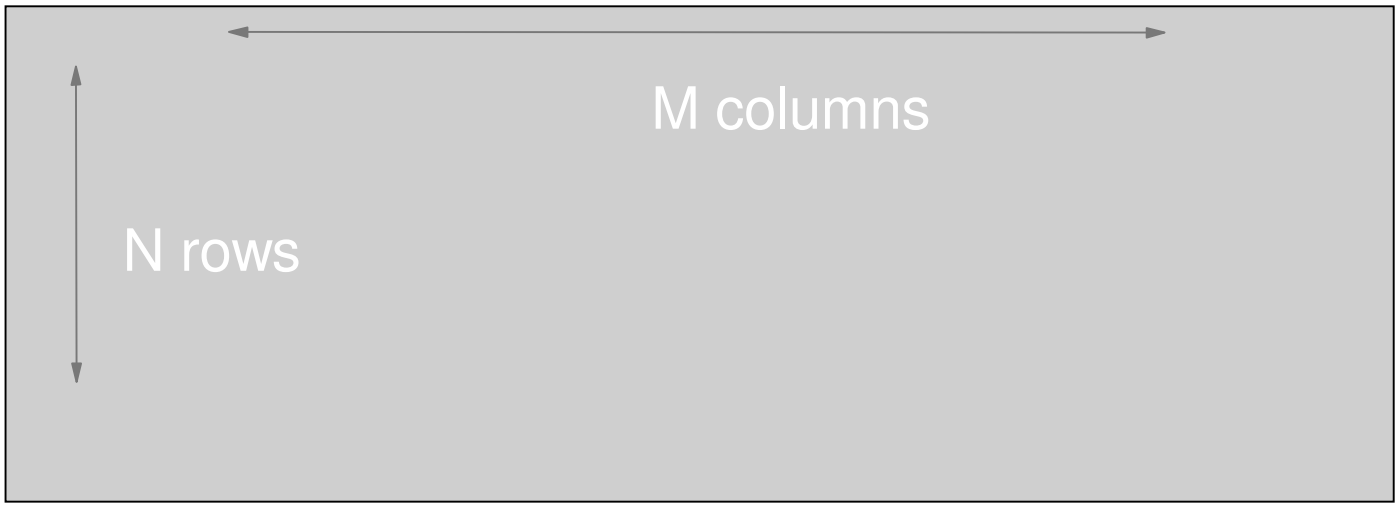
- Actual supermarket data:
  - 267 baskets
  - 2 week-period
  - 24 customers
  - a choice of 100 possible products

- Task:
  - What types of products do customers buy together?

45 [Toy car]
41 [Lemonade]
...
6 [Mineral water][Antifreeze][Soap A][Toilet paper]
6 [Mineral water][Apple juice][Antifreeze][Soap A]
....
4 [Orange juice][Lime juice][Washing-up liquid][Toilet paper][Colour slide film]
4 [Apple juice][Toilet paper][Toy car][Puzzle (1000 p.)]
4 [Tonic water][Apple juice][Washing-up liquid][Toilet paper]
....
3 [C-Beer][Mineral water][Antifreeze][Car light bulb H4][Soap A][Toilet paper] [Battery][Toy car]
3 [C-Beer][Disp. nappies Q][Puzzle (1000 p.)]
...
2 [Mineral water][Car tyres 175-14][Oil - brand B][Oil filter XYZ] [Windscreen wipers][Spark plugs ABC][Antifreeze][Car light bulb H4] [Soap A][Detergent][Toilet paper][Battery][Battery charger][Toy car]
2 [Mineral water][Lemonade][Orange juice][Windscreen wipers][Soap A] [Brandy][Colour slide film][Battery charger]
...

(c) I. Rigoutsos / August 17, 2001

Association Discovery - General Form



Instances:

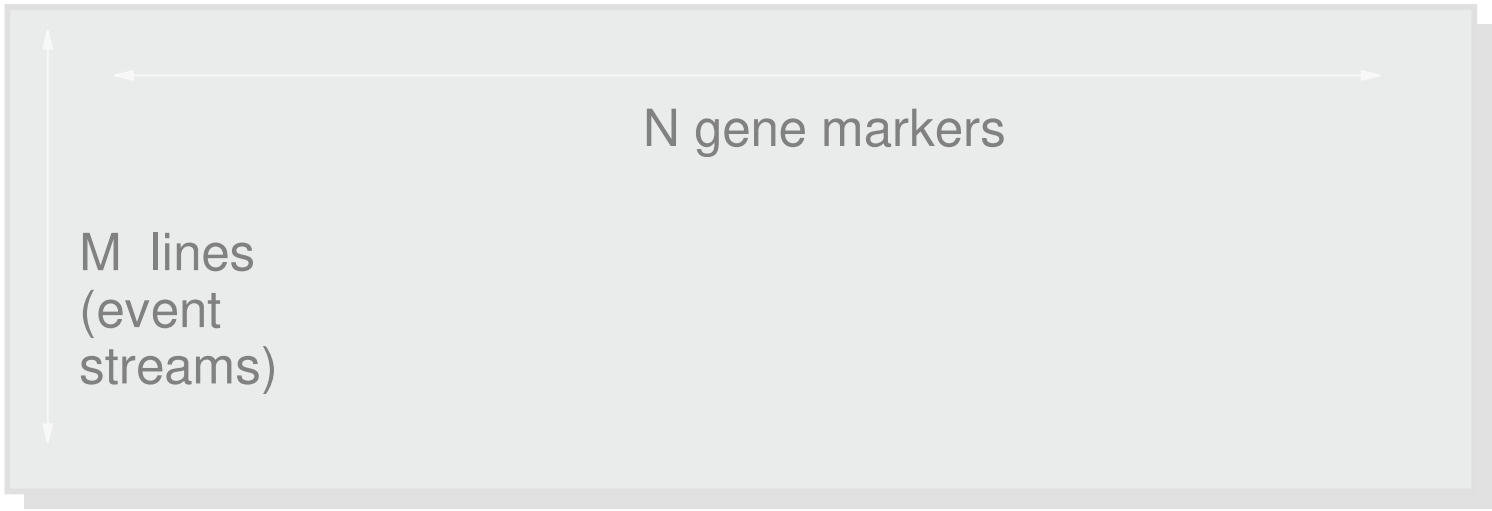
s_1 =	rent	R&B	romance	BSc	$30K	4 dr. sedan
s_2 =	rent	rock	fiction	MSc	$50K	2 dr. hatchback
s_3 =	own	jazz	science-fiction	PhD	$70K	sports util. veh.
s_4 =	own	jazz	romance	PhD	$70K	4 dr. hatchback
s_5 =	rent	R&B	fiction	MSc	$30K	2 dr. sedan

or, equivalently:

s_1 =	1	3	6	9	12	15
s_2 =	1	4	7	10	13	16
s_3 =	2	5	8	11	14	17
s_4 =	2	5	6	11	14	18
s_5 =	1	3	7	10	14	19

(c) I. Rigoutsos / August 17, 2001

Association Discovery - Gene Expression



- Given:
  - "quality" values for *P* macroscopic (phenotypic) properties available for each of the inputs
- Task:
  - cluster the values of the markers
  - evaluate clusters using "quality" values

(c) I. Rigoutsos / August 17, 2001

Association Discovery - Gene Expression (cont.)

- Results:
  - over 90,000 association patterns comprising 7 or more streams
  - largest cluster involved 30 streams
  - patterns involved up to 10 markers

Line Identifier	*RMS error*	*Corresponding Markers W/ Their Values*
18 20 22 31 35	*2.50*	35 (= 1) 37 (= 1) 39 (= 1) 41 (= 1) 42 (= 1)
1 14 17 47 48	*2.79*	41 (= 2) 43 (= 1) 45 (= 1)
1 11 17 37 47	*3.51*	21 (= 1) 25 (= 1) 27 (= 2)
1 11 17 39 47	*3.56*	21 (= 1) 24 (= 1) 27 (= 2)
3 14 15 18 36 37	*3.77*	49 (= 1) 53 (= 1) 55 (= 1) 56 (= 1)
3 14 15 22 36 37 …	*3.77*	46 (= 1) 49 (= 1) 53 (= 1) 56 (= 1)
5 13 14 15 47	*4.17*	38 (= 2) 43 (= 1)

Association Discovery - Gene Expression (cont.)

Line Identifier	*RMS error*	*Corresponding Markers W/ Values*
3 6 11 23 32	*20.56*	2 (=1) 5 (= 1) 8 (=1) 10 (=1) 13 (= 1)

20.56 = 4.78 + **0.36** + 4.58 + **0.64** + **0.75** + 8.99 + **0.33** + **0.10** + **0.02**

3	98.15	100.15	98.90	**99.86**	**100.16**	99.16	**99.15**	**100.04**	**99.87**
6	95.82	100.19	96.46	**99.89**	**100.05**	96.31	**99.20**	**99.62**	**99.83**
11	99.44	99.94	99.94	**99.22**	**99.78**	99.60	**99.84**	**99.93**	**100.00**
23	97.37	101.02	100.02	**100.79**	**101.31**	94.73	**98.99**	**99.55**	**100.08**
32	99.53	**100.**09	99.89	**100.14**	**99.91**	98.92	**99.84**	**99.96**	**99.96**

Moisture **Test Weight** **Stand** **Stalk Lodging** **Root Lodging** **Dropped Ears**

(c) I. Rigoutsos / August 17, 2001

Association Discovery - Gene Expression

Types Of DNA arrays

- the "DNA chip"

a large number of oligonucleotides each of which is synthesized in situ on solid support using known sequence information

- the "microarray"

a collection of DNA targets usually PCR products from cDNA or genomic clones arrayed on solid support

- the "gel-chip"

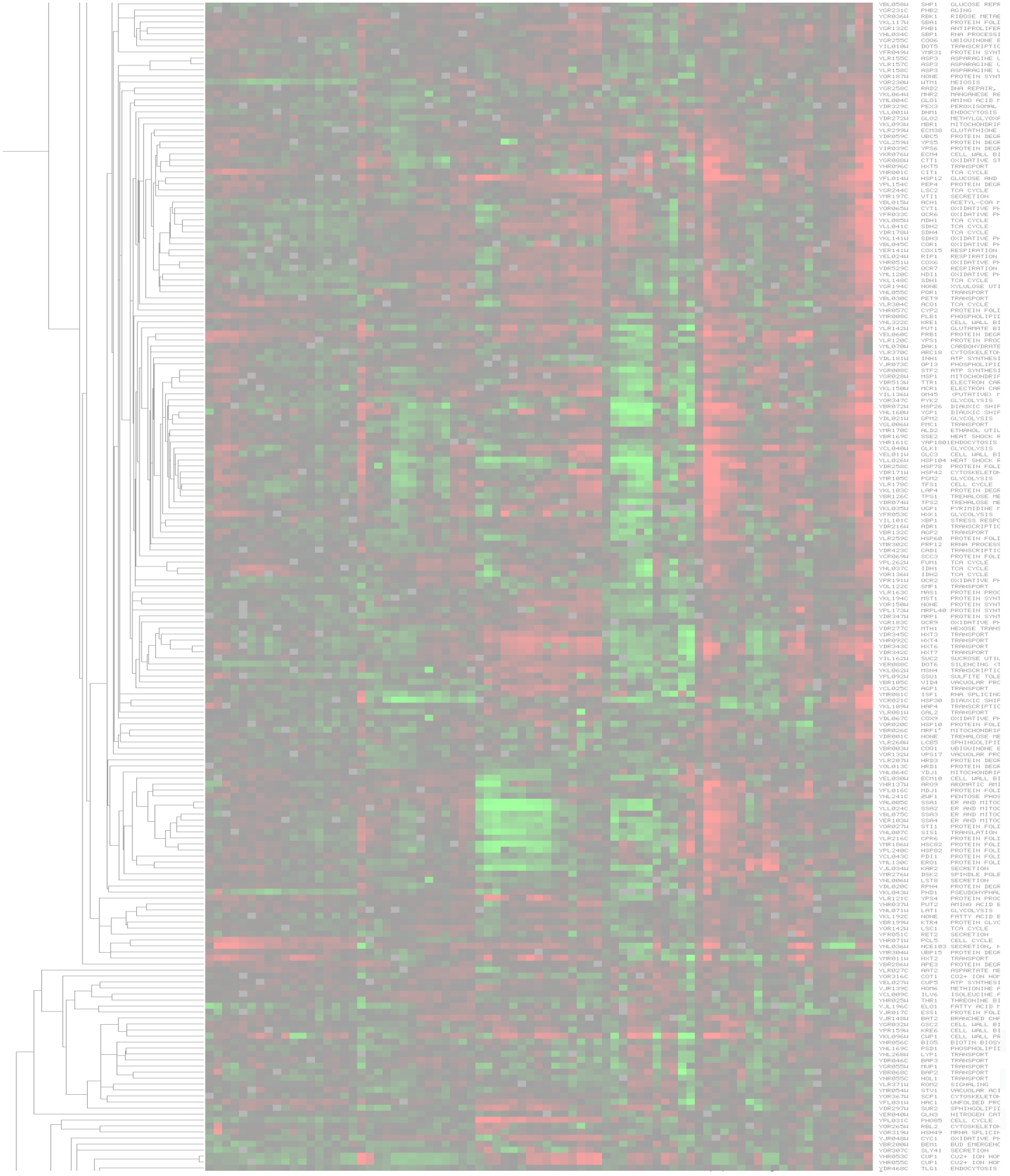
The Method

mRNA purified from experimental samples under study is labeled with a fluorescent dye (e.g. 'red') / a reference sample is labeled with a different dye (e.g. 'green').

Both samples are then allowed to hybridize with the targets of the DNA array. After hybridization any unbound mRNA probe is washed away and the array scanned with a laser. The emitted light is captured by an appropriate detector such as a confocal microscope and the light's intensity is recorded for each of the two dyes: more bound probe means stronger fluorescense.

(c) I. Rigoutsos / August 17, 2001

Association Discovery - Gene Expression (cont.)

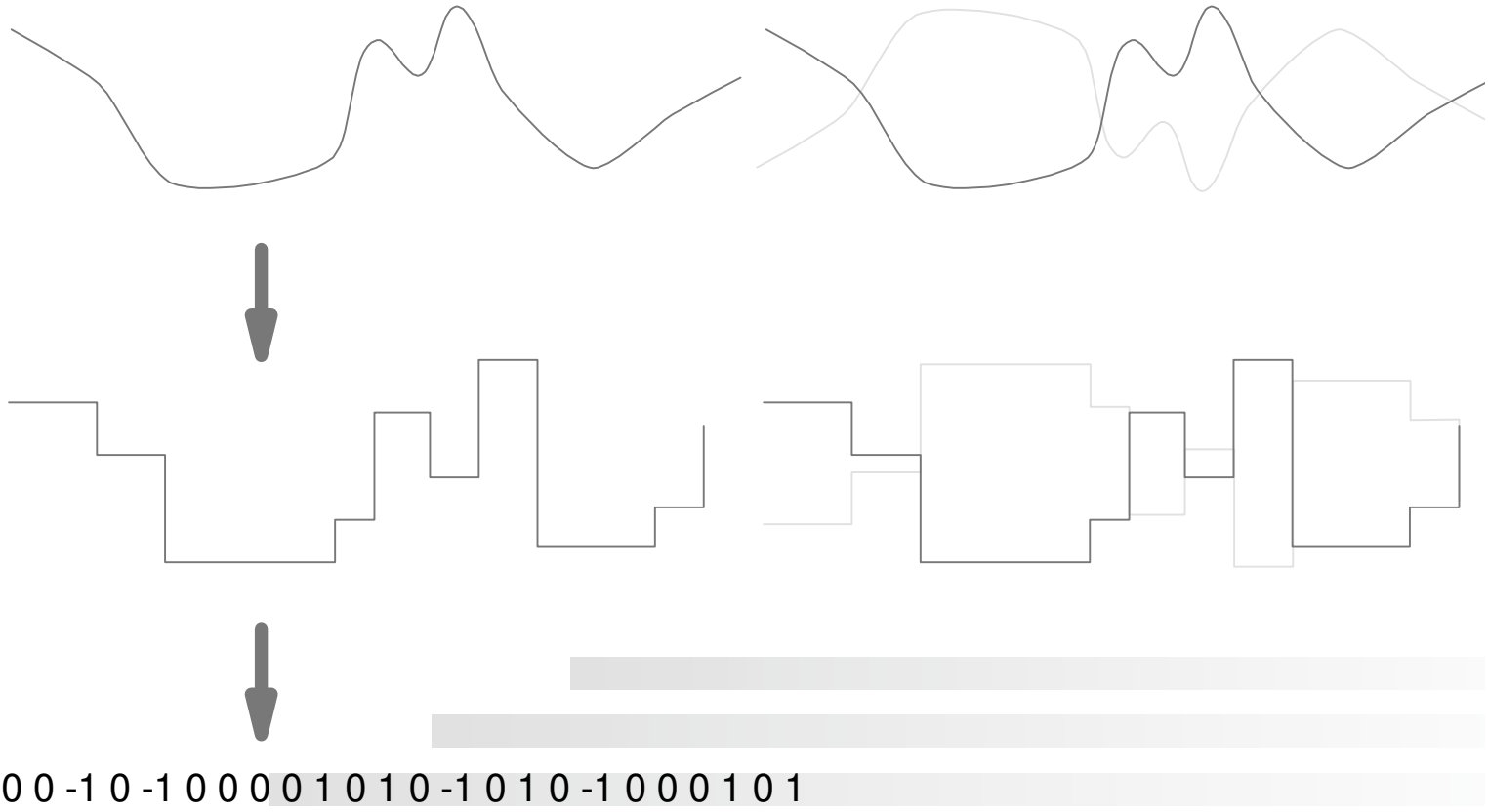


Association Discovery - Gene Expression (cont.)

Representation Scheme

- quantize expression ratios, e.g. 3 steps per unit
- optionally apply hysteresis
- replace quantized values by the sign of their derivatives

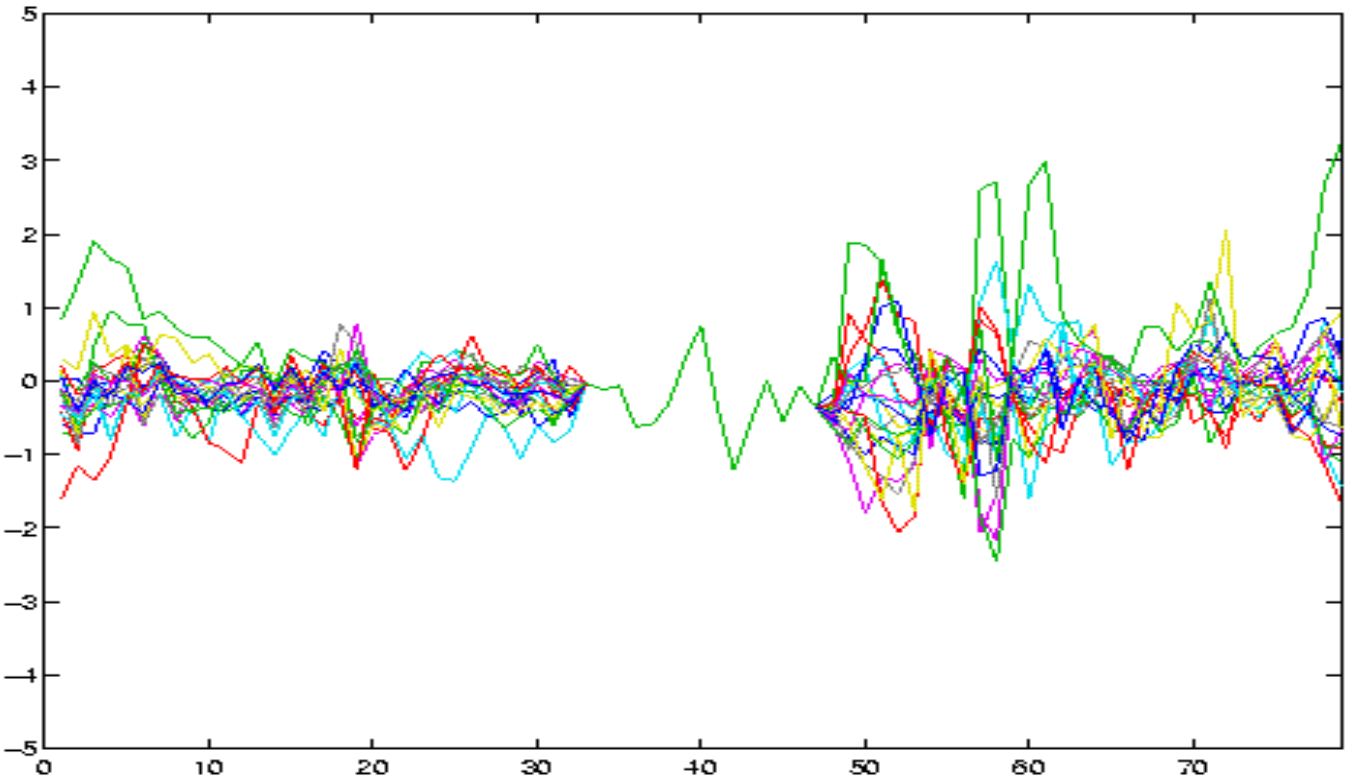
- invert all of the M input rows and append to original input for a total of 2M rows



(c) I. Rigoutsos / August 17, 2001

Association Discovery - Gene Expression (cont.)

Yeast Data: values @ CDC15

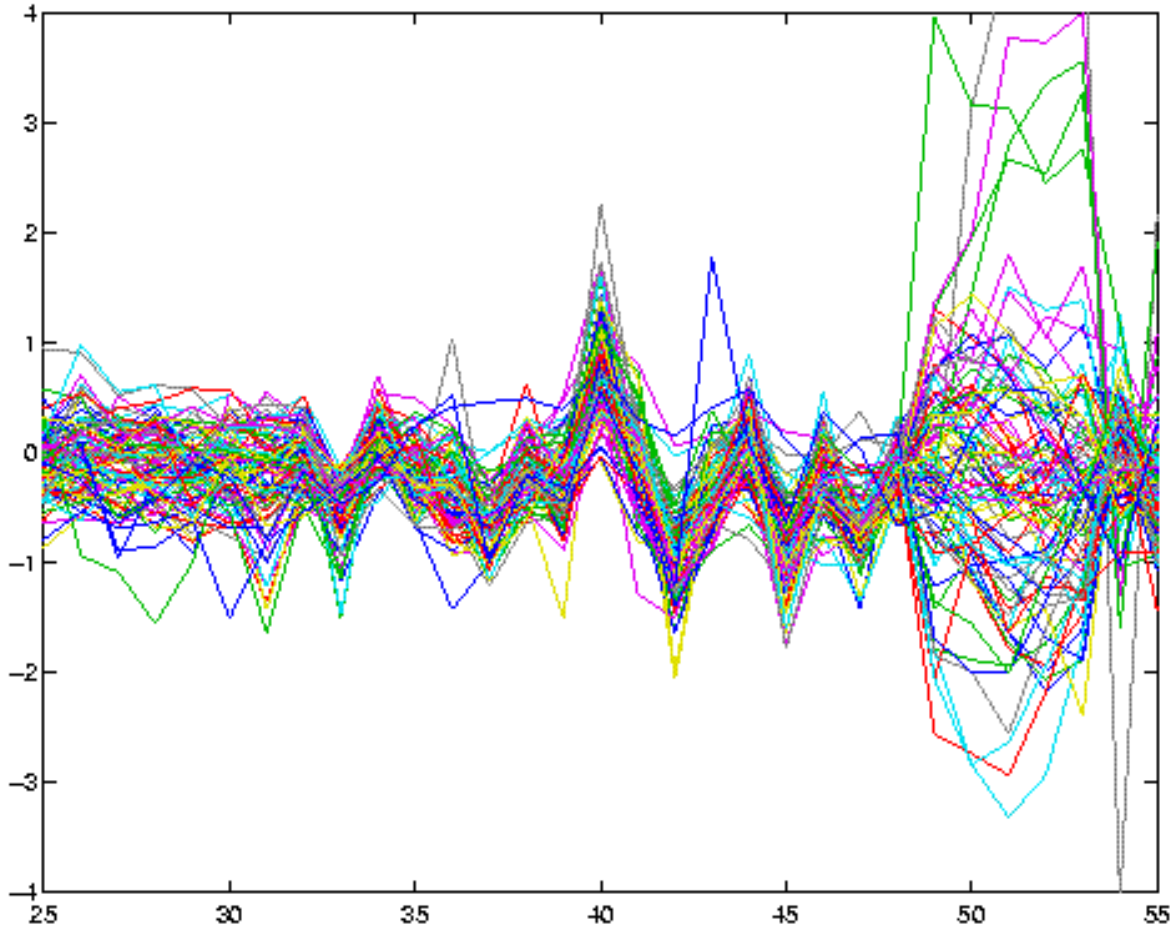


NL330C	RPD3	CHROMATIN STRUCTURE
DL012C	HTA3	CHROMATIN STRUCTURE
NL312W	RFA2	DNA REPAIR
NL290W	RFC3	DNA REPLICATION
DL018C	TLG2	ENDOCYTOSIS
NR017W	MAS6	MITOCHONDRIAL PROTEIN TA
NL286W	CUS2	MRNA SPLICING, PUTATIVE
NL316C	PHA2	PHENYLALANINE BIOSYNTHES
NL306W	NONE	PROTEIN SYNTHESIS
NR015W	SMM1	PROTEIN SYNTHESIS, MITOC
NL282W	POP3	RRNA AND TRNA PROCESSING
NR049C	MSO1	SECRETION

YNR019W	ARE2	STEROL METABOLISM
YNR043W	MVD1	STEROL METABOLISM
YNR001C	CIT1	TCA CYCLE
YNL268W	LYP1	TRANSPORT
YOL020W	TAT2	TRANSPORT
YNL292W	PUS4	TRNA PROCESSING
YNR041C	COQ2	UBIQUINONE BIOSYNTHES
YNL332W	THI12	PYRIMIDINE BIOSYNTHES
YNR023W	SNF12	TRANSCRIPTION
YNR045W	PET494	PROTEIN SYNTHESIS
YNL284C	MRPL10	PROTEIN SYNTHESIS

Association Discovery - Gene Expression (cont.)

Yeast Data: derivatives @ CDC15



ARG Biosynthesis
ATP Synthesis
Biotins Biosynthesis
Bud Emergence
Bud Site Selection
Catabolite Repression
Cell Cycle
Cell Wall Biogenesis
Chromatin Structure
Cytokenesis
Cytoskeleton

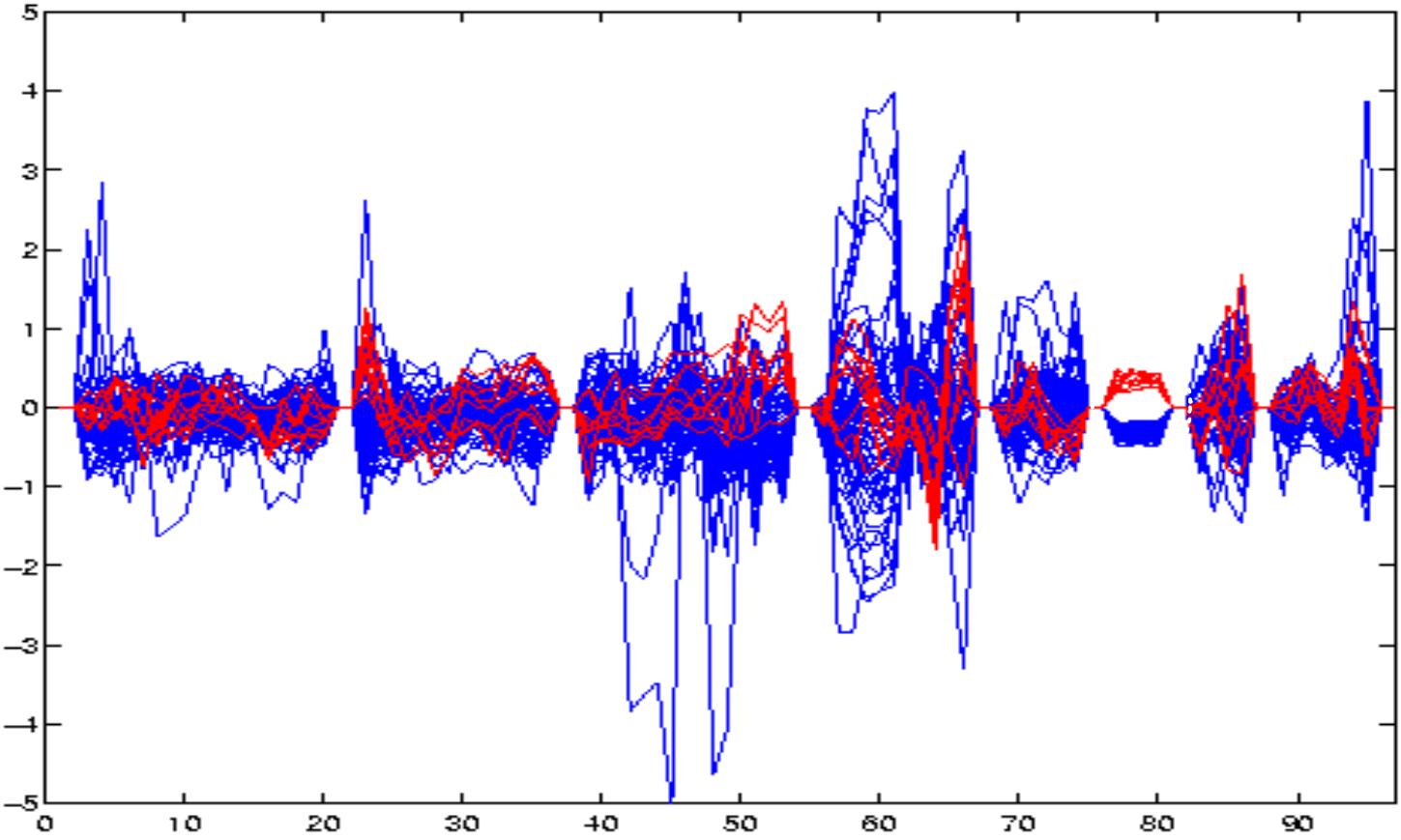
DNA Repair
DNA Recombination
DNA Replication
Fermentation
Glucose Repression
Glycolysis
Heme Biosynthesis
Mating
Meiosis
Mitosis
MRNA Splicing

Oxidative Stress
Phospholipid Biosynth.
Phospholipid Metabol.
Protein Degradation
Protein Folding
Protein Processing
Protein Synthesis
Protein Targeting
Respiration
Secretion
Silencing

TCA Cycle
Transcription
Transport
Trehalose Utilization
Vacuolar Protein Targeting
etc.

Association Discovery - Gene Expression (cont.)

Yeast Data: derivatives @ Cold Shock



Aging
Antiprolif. Agent
Protein Degradation
Transcription (Global Regulator)

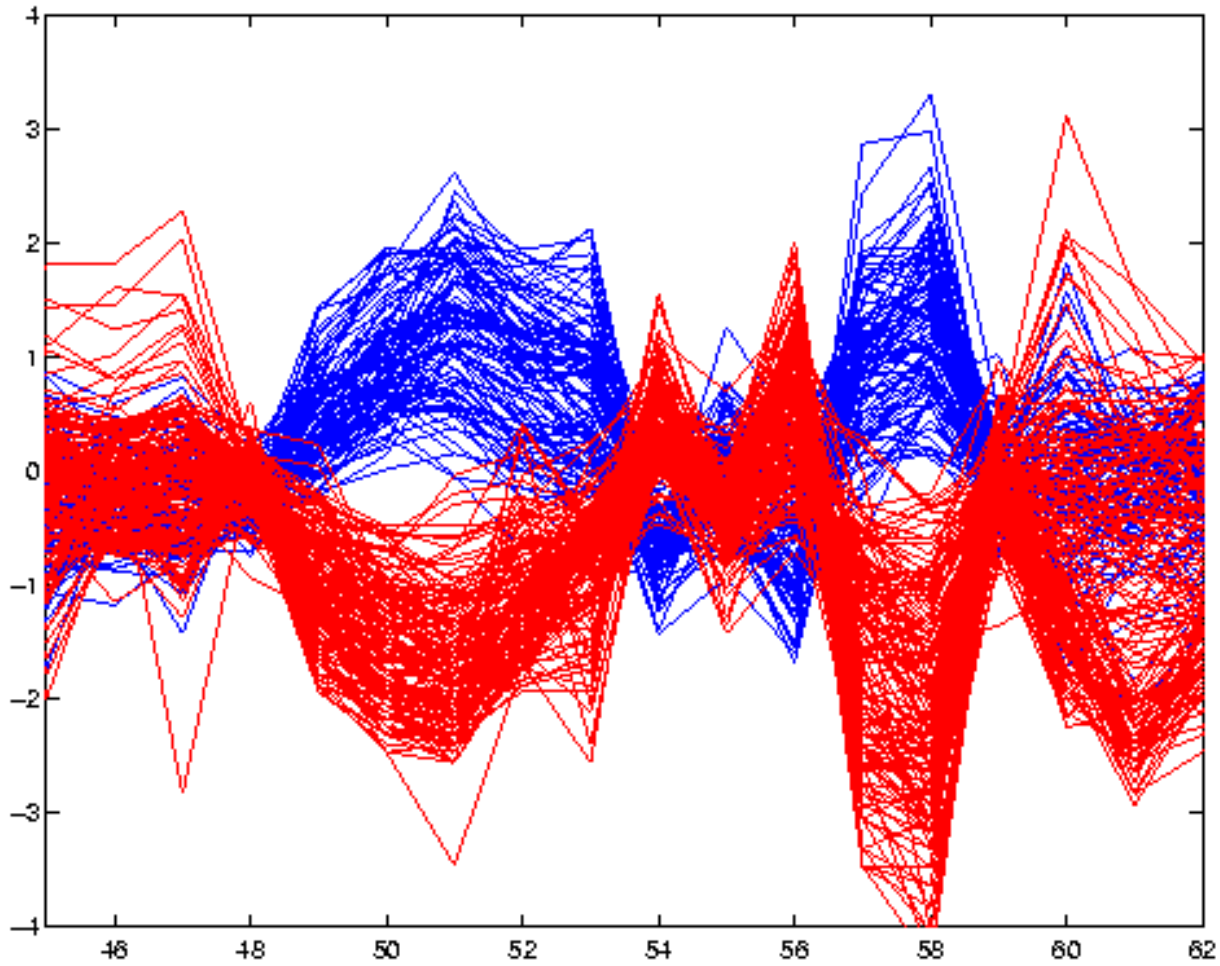
Cell Cycle
Cell-Wall Biogeneis
Chromatin Structure
Cytoskeleton
DNA Repair

DNA Replication
MRNA Splicing
Nitrogen Transport
Nuclear Prot. Targ.
Protein Synthesis

Respiration
Secretion
Transcription
Transport
TRP Biosynthesis

Association Discovery - Gene Expression (cont.)

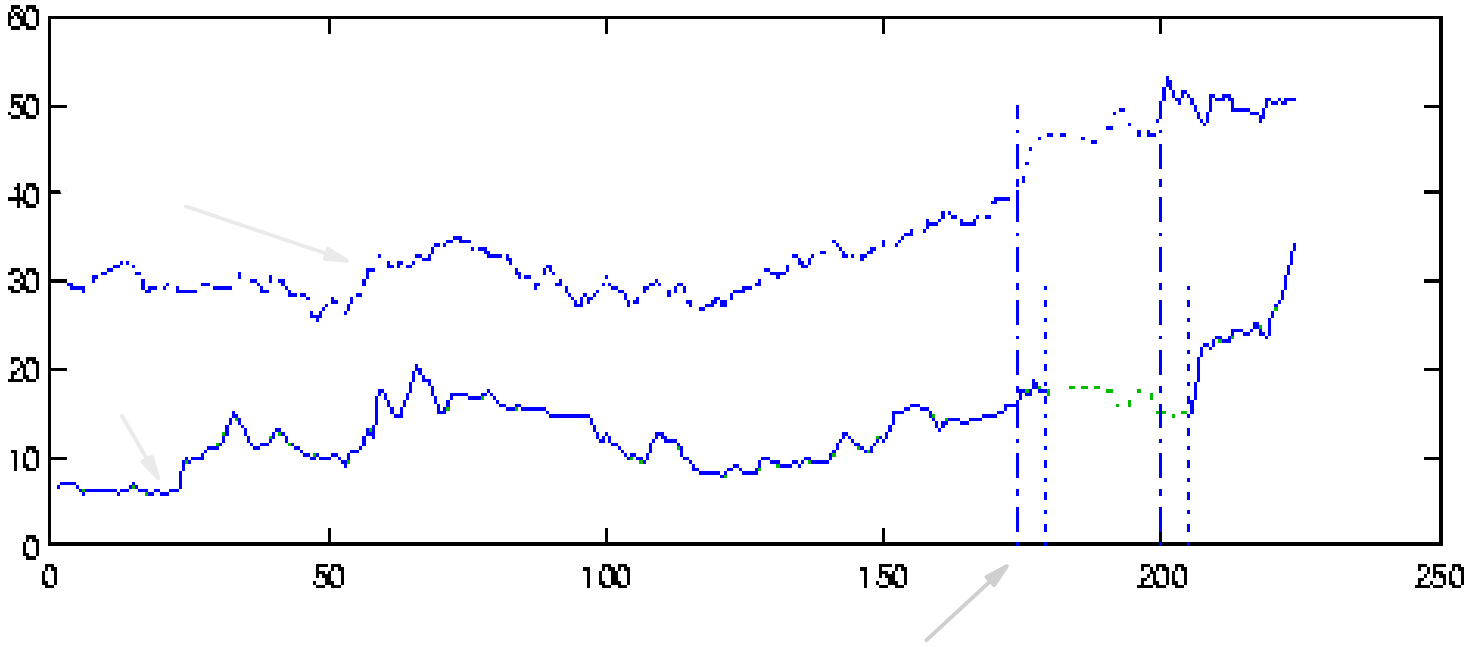
Yeast Data: derivatives @ Sporulation

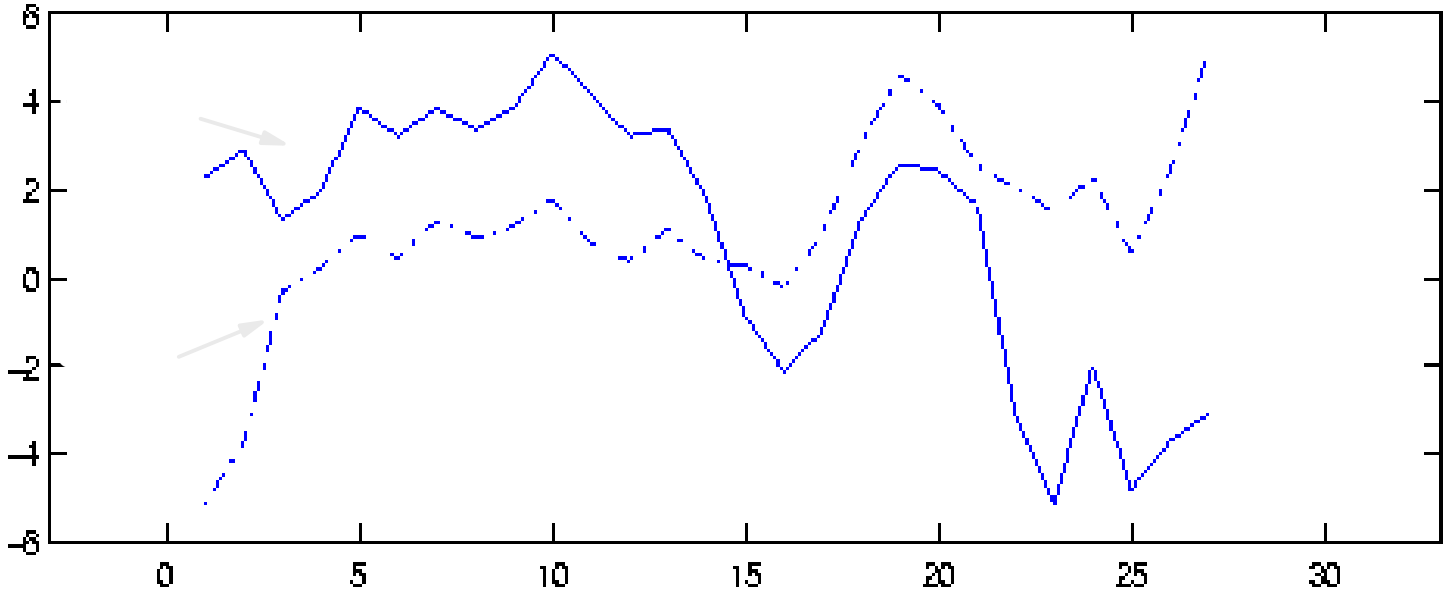


Arginine Metabolism	MRNA 3'end Processing	ATP Synthesis	MRNA Splicing
ATP Synthesis	MRNA Splicing	Cell Cycle	Nuclear Protein Targeting
Bud Site Selection	Nuclear Fusion	Cell Wall Biogenesis	Oxidative Stress Response
Carbohydrate Metabolism	Organelle Communic.	Chromatin Structure	Peroxisome Biogenesis
Cell Cycle	Osmotic Stress Respon.	Chromosome Maintenance	Peroxisome Proliferation
Cell Wall Biogenesis	Peroxisome Biogenesis	Diphtamide Biosynthesis	Phospholipid Biosynth.
Chromatin Structure	Phelylacrylic Acid Resist.	DNA Repair / Replication	Polyamine Transport
Cytoskeleton	Protein Degradation	Fatty Acid Metabolism	Protein Degradation
DNA Repair/Replication	Protein Folding	Glycose Repression	Protein Folding
Drug Resistance	Protein Glycosylation	Glutamate Biosynthesis	Protein Glycosylation
Endocytosis of STE3P	Protein Targeting	Glutathione Biosynthesis	Protein Processing
Glycolysis	Respiration	Glycolysis	Protein Synthesis
Heme Biosynthesis	Secretion	Histidine Biosynthesis	Purine/Pyrimidine Biosynth.
Mating	Silencing	Methionine Biosynthesis	Transcription / Transport
Meiosis	Sterol Metabolism	Mitochondrial Protein	Vacuolar Prot. Target.
Methionine Biosynthesis	Trancsription / Transport	MRNA Processing	

The Issue Of Representation

7 stocks: SUNW, SGI, MSFT, IBM, XCIT, LCOS, MOT / 1 year





(c) I. Rigoutsos / August 17, 2001

The Issue Of Representation (cont.)

- the correct choice is CRITICAL to the success or failure of any approach
- should be invariant under an allowed transformation
- should allow for redundancy
- should allow for missing erroneous data
- should be general enough...

(c) I. Rigoutsos / August 17, 2001

Web Access: Tools & Downloadable Codes

General information about the group:

http://www.research.ibm.com/bioinformatics/

Downloadable Code (executable):

(IBM/AIX, Linux/i586, SUN/Solaris,Win95/98)

http://www.research.ibm.com/bioinformatics/download.phtml.html

Web-accessible Tools:

http://cbcsrv.watson.ibm.com/Tspd.html

Complete Genome Annotations:

http://cbcsrv.watson.ibm.com/Annotations/

Complete Genome Comparisons:

http://cbcsrv.watson.ibm.com/TsearchAllGenomes.html

Downloadable Bio-Dictionaries:

http://www.research.ibm.com/bioinformatics/metadata.phtml.html

(c) I. Rigoutsos / August 17, 2001