10.555
Bioinformatics

Principles, Methods, Applications

Multiple Sequence Alignment

- Definition of MSA:

  given a set of sequences s_1, s_2, …, s_k over some alphabet insert spaces into (or at either end of) each of the sequences to make them all have the same length

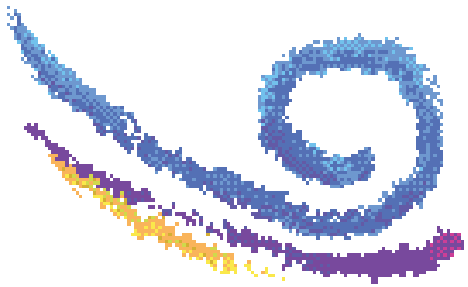
```
MQPILLL          MQPILLL
MLRLL      ⇒     MLR-LL-
MKILLL           MKI-LLL
MPPVLIL          MPPVLIL
```

- *"One or two homologous sequences whisper … a full multiple sequence alignments shouts out loud"*

Arthur Lesk

Multiple Sequence Alignment (cont.)

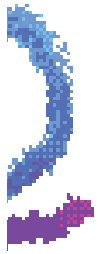
- Why bother?
  - representation of families & superfamilies
  - motifs / profiles / combinations
  - secondary structure inference
  - deduction of evolutionary history



Multiple Sequence Alignment (cont.)

{A,G}, C, {D,E}, {F,Y}, H, {I,L,M,V}, {K,R}, {N,Q}, P, {S,T}, W

```
            folded Alignments   (max extent = 171 columns)
         ***************                      **********************     ***************
-----mgAFSEkQEsLVksSWeafkqnvphhsavfytl--IlEkAPaAqnMFSFLsngvdp--NNPKLKaHAekVF
-------AlTEkQEaLLkqSWevlkqnipahs------LRLFaLiieaapeskyLKDSneipENNPKLKaHAavIF
mstlegrGFTEeQEaLVvkSWsamkpnagelglkff--LKIFEIAPsAqkLFSFLKDSnvplErNPKLKsHAmsVF
sssevnkvFTEeQEaLVvkawavmkknsaelglqff--LKIFEIAPsAknLFSYLKDSpvplEQNPKLKpHAttVF


*******                          **********  ****************************************
DSAvqLRakgevvladptlgsvhvqkg------------VlDpHFlVvKeALLkTfKEAVgDkWndelgnawevay
ESAteLRqkghavwdnnt---------LKRLGsiHlKnkItDpHFEVmKgALLgTIKEAIkEnWSdEMgQAWteAY
ESAvqLRkagkvtvress---------LKKLGasHfKhgVaDeHFEVtKfALLeTIKEAVpEtWSpEMkNAWgeAY
ESAvqLRkagkatvkesd---------LKRIGaiHfKtgVvneHFEVtRfALLeTIKEAVpEmWSpEMkNAWgvAY


aaikkamgsa--
atikaemke---
aaiklemkpss-
aaikfemkpsst
```



Multiple Sequence Alignment (cont.)

- Definition
  the Sum-Of-Pairs (SP) *function* is defined as:

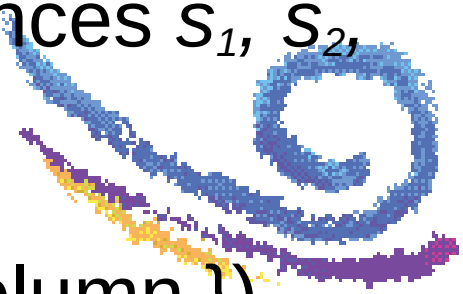
  if a column contains $\ell_1, \ell_2, \ell_3, \ldots \ell_m$ then

  $$SP(\{\ell_1, \ell_2, \ell_3, \ldots \ell_m\}) = \Sigma_i \Sigma_{j>i} \text{ score}(\ell_i, \ell_j)$$

  where: score(- , -) = 0

- the SP-score for multiply aligned sequences $s_1, s_2, \ldots, s_k$ is defined as:

  $$\text{SP-score}(s_1, s_2, \ldots, s_k) = \Sigma_{\text{all collumns } i} \text{ SP}(\{\text{column}_i\})$$

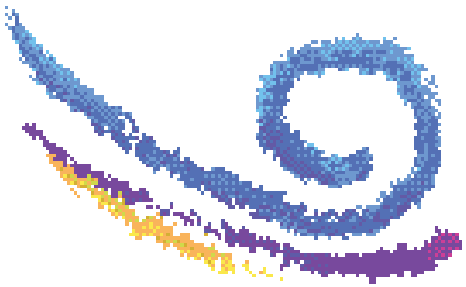


Multiple Sequence Alignment (cont.)

- Dynamic Programing for 3 sequences:

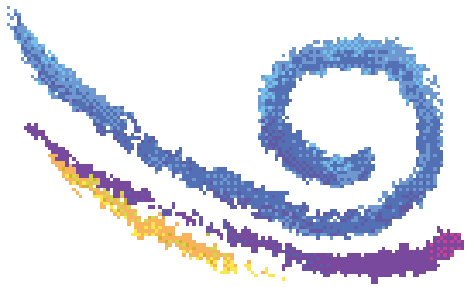
$$D(i,j,k) = \min \begin{cases} D(i-1,j-1,k-1)+cost(i->j)+cost(j->k)+cost(i->k) \\ D(i-1,j-1,k)+cost(i->j)+2*cost(space) \\ D(i-1,j,k-1)+cost(i->k)+2*cost(space) \\ D(i,j-1,k-1)+cost(j->k)+2*cost(space) \\ D(i-1,j,k)+2*cost(space) \\ D(i,j-1,k)+2*cost(space) \\ D(i,j,k-1)+2*cost(space) \end{cases}$$

$$D(i,j,0) = pairwise(i,j) + (i+j)*cost(space)$$
$$D(i,0,k) = pairwise(i,k) + (i+k)*cost(space)$$
$$D(0,j,k) = pairwise(j,k) + (j+k)*cost(space)$$



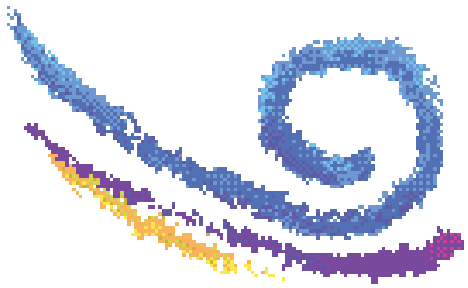
Multiple Sequence Alignment (cont.)

- **MSA through pair-wise alignment:**
  - start w/ *best* pair
  - incrementaly add the rest of the S_i according to some *schedule*
    E.g. next add the s_i that gives the best score when compared with some s_j already in the collection / MSP
- star-alignment-based MSA
- Also: MSA through clustering-imposed order
- One major assumption...



Multiple Sequence Alignment (cont.)

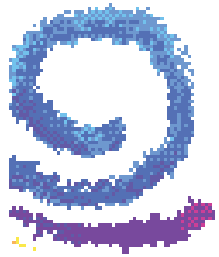
- Gibbs Sampling-based MSA
- Motif-based MSA
- other…



Pretty Print An Alignment

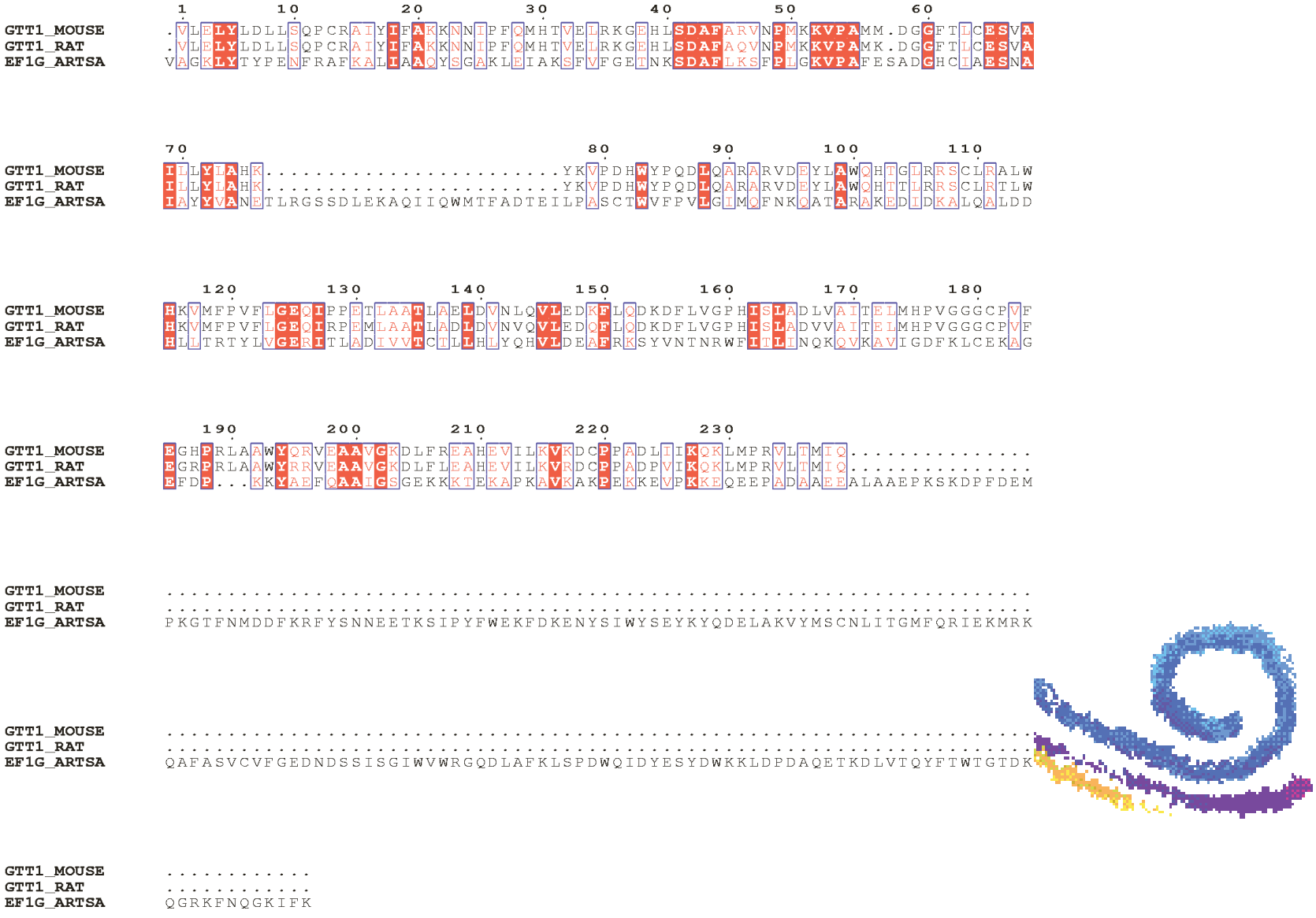
- http://www.ch.embnet.org/software/BOX_form.html

```
GTT1_MOUSE     1 -VLELYLDLLSQPCRAIYIFAKKNNIPFQMHTVELRKGEHLSDAFARVNPMKKVPAMM-D
GTT1_RAT       1 -VLELYLDLLSQPCRAIYIFAKKNNIPFQMHTVELRKGEHLSDAFAQVNPMKKVPAMK-D
EF1G_ARTSA     1 VAGKLYTYPENFRAFKALIAAQYSGAKLEIAKSFVFGETNKSDAELKSFPLGKVPAFESA
consensus      1      **                * *          .   .      ****    *. ****

GTT1_MOUSE    59 GGFTLCESVAILLYLAHK----------------------YKVPDHWYPQDLQARARV
GTT1_RAT      59 GGFTLCESVAILLYLAHK----------------------YKVPDHWYPQDLQARARV
EF1G_ARTSA    61 DGHCIAESNAIAYYVANETLRGSSDLEKAQIIQWMTFADTEILPASCTWVFPVLGIMQFN
consensus     61  *   . **  **   *.*                                 *     *

GTT1_MOUSE    95 DEYLAWQHTGLRRSCLRALWHKVMFPVFLGEQIPPETLAATLAELDVNLQVLEDKFLQDK
GTT1_RAT      95 DEYLAWQHTTLRRSCLRTLWHKVMFPVFLGEQIRPEMLAATLADLDVNVQVLEDQFLQDK
EF1G_ARTSA   121 KQATARAKEDIDKALQALDDHLLTRTYLVGERITLADIVVTCTLLHLYQHVLDEAFRKSY
consensus    121      *  .  . .        * .     .** *    .  *   *  .   **.. *

GTT1_MOUSE   155 DFLVGPHISLADLVAITELMHPVGGGCPVFEGHPRLAAWYQRVEAAVGKDLFREAHEVIL
GTT1_RAT     155 DFLVGPHISLADVVAITELMHPVGGGCPVFEGRPRLAAWYRRVEAAVGKDLFLEAHEVIL
EF1G_ARTSA   181 VNTNRWFITLINQKQVKAVIGDFKLCEKAGEFDP---KKYAEFQAAIGSGEKKKTEKAPK
consensus    181         *.*     .  ..           *  *     *    **.*

GTT1_MOUSE   215 KVKDCPPADLIIKQKLMPRVLTMIQ-----------------------------------
GTT1_RAT     215 KVRDCPPADPVIKQKLMPRVLTMIQ-----------------------------------
EF1G_ARTSA   238 AVKAKPEKKEVPKKEQEEPADAAEEALAAEPKSKDPFDEMPKGTFNMDDFKRFYSNNEET
consensus    241  *.  *    . *

GTT1_MOUSE       ------------------------------------------------------------
GTT1_RAT         ------------------------------------------------------------
EF1G_ARTSA   298 KSIPYFWEKFDKENYSIWYSEYKYQDELAKVYMSCNLITGMFQRIEKMRKQAFASVCVFG
consensus    301

GTT1_MOUSE       ------------------------------------------------------------
GTT1_RAT         ------------------------------------------------------------
EF1G_ARTSA   358 EDNDSSISGIWVWRGQDLAFKLSPDWQIDYESYDWKKLDPDAQETKDLVTQYFTWTGTDK
consensus    361

GTT1_MOUSE       ------------
GTT1_RAT         ------------
EF1G_ARTSA   418 QGRKFNQGKIFK
consensus    421
```



Pretty Print An Alignment (cont.)

- http://www-pgm1.ipbs.fr:8080/cgi-bin/nph-ESPript_exe.cgi



Putting It All Together

Step 1: start with your DNA sequence of interest

- Search PubMed, a public version of full Medline for topics of interest (US).
- Search Genbank for sequences of interest (US).
- Search protein and nucleic acid databases
- Search a variety of sequence and structure databases using the SRS server at EMBL/EBI

Step 2. Identify ORFs and translate into protein

- GeneMark (US) or GENSCAN at MIT (US) or GRAIL at ORNL (US), etc.
- Gene feature searches at Baylor College of Medicine (US).
- "DNA sequence translation into protein" tool at ExPaSy (Switzerland).

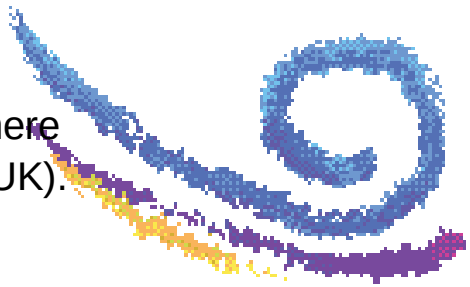
Step 3. Find similar sequences in the databases

- Nucleotide sequence: Search the database of your choice using Blast/Fasta/S-W
- Protein sequence: Search the database of your choice using Blast/Fasta/S-W

Step 4. Do a global alignment of your sequence vs (allegedly) similar sequences

Step 5. Look for gene families

- Multiple sequence Alignment query at Baylor College of Medicine (US), or elsewhere
- Analyze multiple sequence alignments at the AMAS server at Oxford University (UK).
- ...



Modified from http://www.3rdmill.com/discovery_main_panel.htm

Putting It All Together (cont.)

Color/Visualize the aligned regions with BOXSHADE or other
Multiple sequence Alignmnt with phylogenetic tree capabilities using CLUSTAL@EBI (UK)

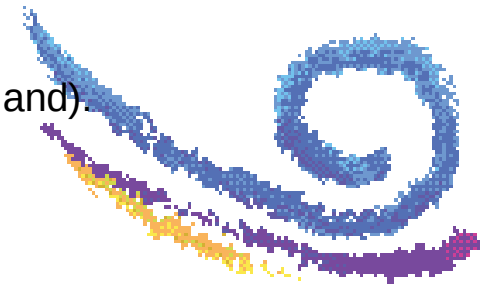
Step 6. Look for the presence of specific patterns in your protein
Search your sequence with InterPro/Blocks / Use PHI-BLAST

Step 7. Find similar sequences in other species
Search several proteomes using the FASTA search interface at the Univ. of Virginia (US).
Search several species-specific databases with Blast at NCBI (US)

Step 8. Determine the putative structure of your protein
Predict the secondary structure of your protein w/ the GOR/JPred/PredictProtein Server
Look for tm/coil-coil/other regions using the servers at ISREC (Switzerland).

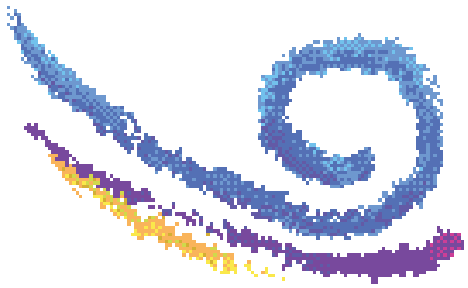
Step 9. Search PubMed for information about function of related proteins

Step 10. Input your sequence into an "alert" server
Sequence Alerting server at the EMBL (Germany).
MIPS alert server at MIPS (Germany). / Swiss-Shop server at ExPaSy (Switzerland).

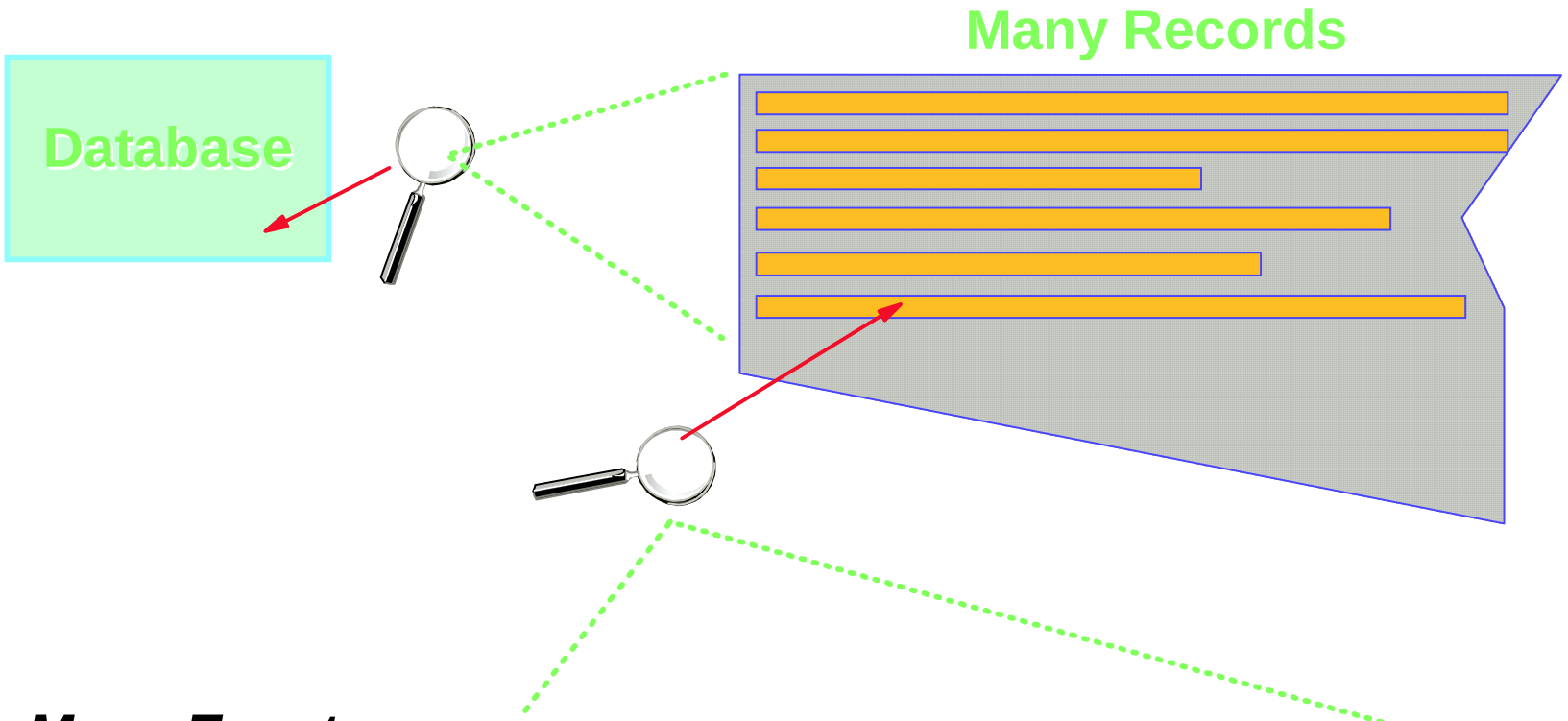


Modified from http://www.3rdmill.com/discovery_main_panel.htm

Pattern Discovery



When Given An Input Like…



Many Events

$$E_1E_2E_3E_4E_5E_6E_7E_8E_9E_{10}E_{11}E_{12}E_{13}E_{14}E_{15}E_{16}E_{17}E_{18}\ldots$$

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

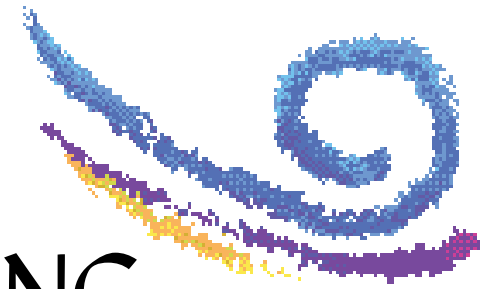
Our Task Is To...

- FIND, and EXPLOIT

interesting combinations of events that are contained in the database in the presence or absence of domain specific knowledge

- But how do we define interesting?
- We make the Assumption:

FREQUENT = INTERESTING



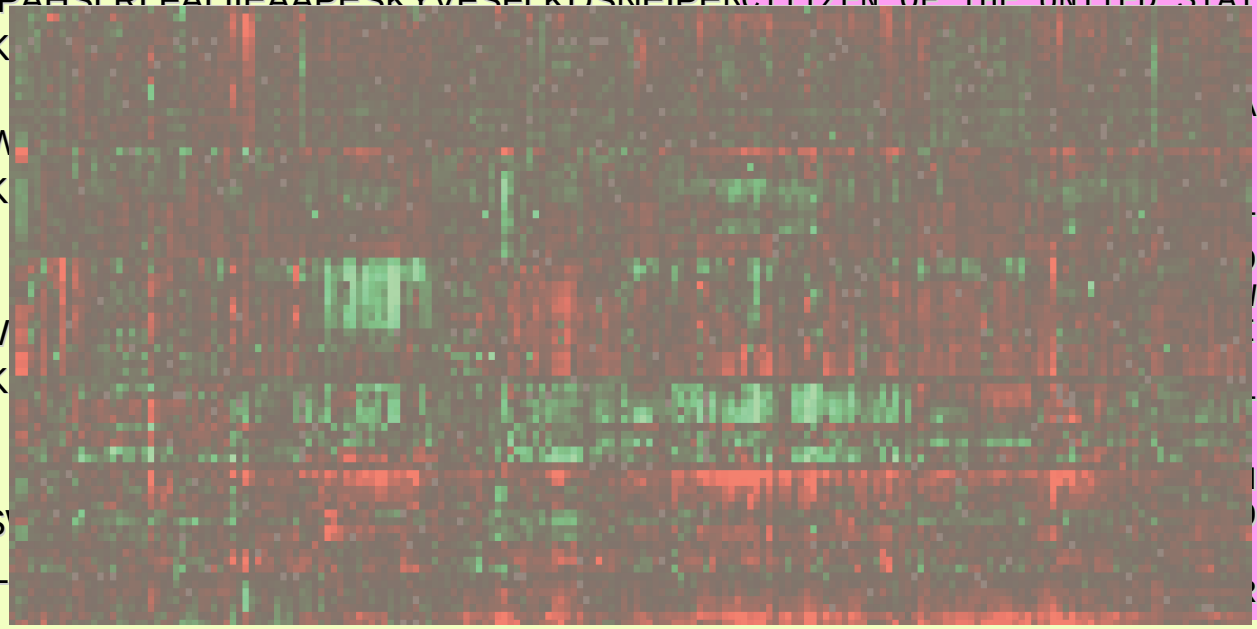
Database Examples

```
1000025 5 1 1 1 2 1 3 1 1 2
1002945 5 4 4 5 7 1
1015425 3 1 1 1 2 2
1016277 6 8 8 1 3 4
1017023 4 1 1 3 2 1
1017122 8 10 10 8 7
1018099 1 1 1 1 2 1
1018561 2 1 2 1 2 1
```

```
>Aeropyrum pernix
AAATAATAATAAAAATTAAGTGACTCATGCATTATCCT
ATCCCAGACTACCATCAATTTAGGGACAATAGTGTTTA
GCTCGCGGGTTCAAACTCGCGTAGGGCCCGAGTTCTAG
AAATTGAGTATGATCTCTCAGTTTTATATCAATACTTA
TTGTTACAACGAATAGAGTGGTCACTCCCGCCAACAGC
TGCTAAAATCATATATACACCTATAGCTATGAGAGATA
```

```
>HBP_CANLI 62686306
MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSNGV
QLRAKGEVVLADPTLGSVHVQKGVLDPHFLVVKEALLKTFKEAVGDKWNDEL
>HBP1_CASGL 1
ALTEKQEALLKQSWEVLKQNIPAHSLRLFALIIEAAPESKYVFSFLKDSNEIPEN
GHAVWDNNTLKRLGSIHLKNK
>HBP2_CASGL 1
MSTLEGRGFTEEQEALVVKSW
CESAVQLRKAGKVTVRESSLK
LEMKPSS
>HBPL_PARAD 1
SSSEVNKVFTEEQEALVVKAW
CESAVQLRKAGKATVKESDLK
FEMKPSST
>HBPL_TRETO 1
MSSSEVDKVFTEEQEALVVKS
TCESAVQLRKAGKVTVRESNL
SEMKPSST
```

```
SECTION 2. THE HOUSE OF
REPRESENTATIVES SHALL BE
COMPOSED OF MEMBERS CHOSEN
EVERY SECOND YEAR BY THE PEOPLE
OF THE SEVERAL STATES, AND THE
ELECTORS IN EACH STATE SHALL
HAVE THE QUALIFICATIONS
REQUISITE FOR ELECTORS OF THE
MOST NUMEROUS BRANCH OF
THE STATE LEGISLATURE.

NO PERSON SHALL BE A
REPRESENTATIVE WHO SHALL NOT
HAVE ATTAINED TO THE AGE OF
TWENTY FIVE YEARS, AND BEEN
SEVEN YEARS A
CITIZEN OF THE UNITED STATES,
```



Let's Begin With An Example

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

Things We Can Find In It

MORE Things We Can Find In It

Questions

What is the difference between these two phrases?

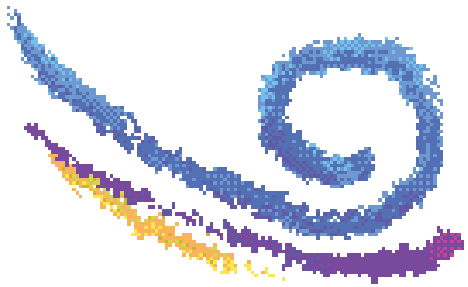
Are there more hidden phrases?

Are they longer or shorter than what was shown?

Can we enumerate them?

KEY OBSERVATION:

if a message is hidden *K* times then
its pieces will appear ***at least K*** times



Solutions Depend On...

- the nature of the input dataset
- *minimum* allowed *local density*
- *minimum* required *support*
- *form* of sought regular 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
  - ...

What Do Solutions Look Like?

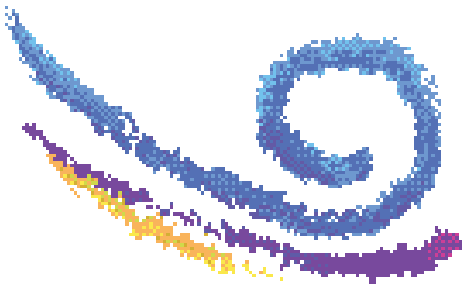
- Typical classes of patterns include:
  - simple strings: ABCDEFG
  - rigid gaps: AB...C.....D.E....F
  - flexible gaps: A-x(1,3)-B-x-C-x(4)-D-x(3,5)-E
  - unrestricted gaps: A * BCD * E * FG

and their versions with expressions involving brackets:

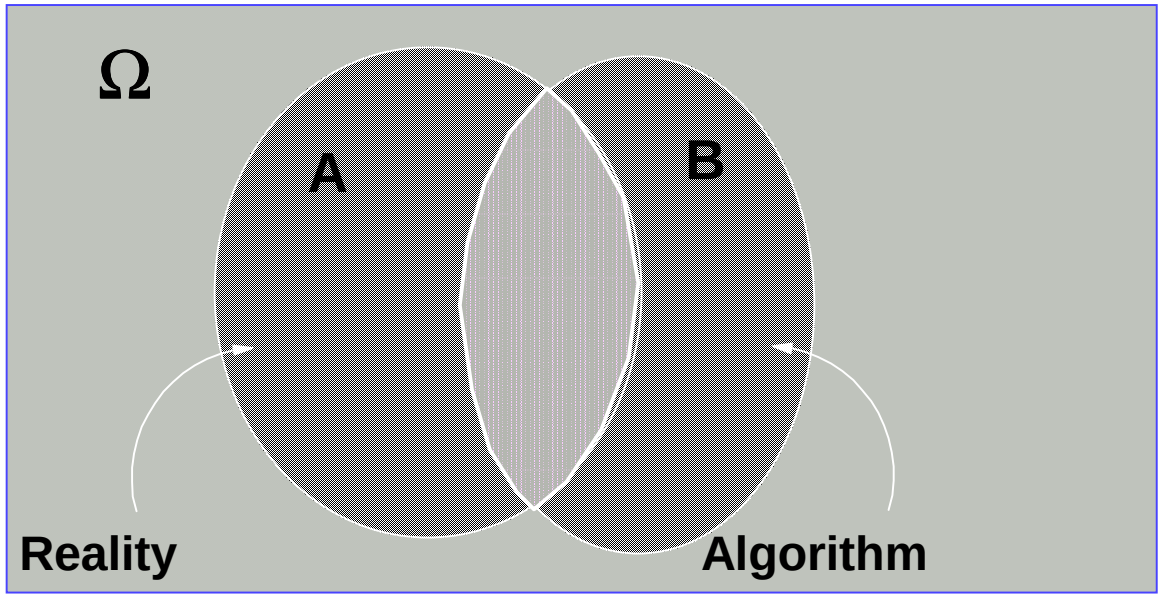
[APQs][BbKm]...C * D.[EW]....F

General 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



Considerations



- positives: A
- true positives: A∩B
- negatives: Ω – A

- false positives: B \ A ∩ B
- false negatives: A \ A ∩ B

sensitivity: |A ∩ B| / |A|

specificity: |A ∩ B| / |B|

Understanding How Teiresias Works

UALHSQBREMHBPOVFBVOAPPROACHHBWFYJNXLHNXGCUIBRGVDCBITESOJWZGR
BXFESKSQLELSVNVEPOACHXHAAUWTKYODFFZHCGSGNFSFJKHVPYBYKRNYBPEW
AQXWODCZISRBWOLSTWWFPGKGYRTMIUNKZBAHQMWWVOHIDOTBGPKAMDHHPVYQ
KWCZBRNMGKSAPPLEDPRSDGCRLXWPKCVSOIWALTCYOSQPOACHNAKMKLQBISFF
XQXTDWWNOVLOUPOACHWGSRNZEJSKHRPOLILRTAINVDXCAOABWZAXQFAMSKGT
IWOALALKYBJXGQXWYUUZUPOACHEIKAPQQBITFMUINZPGMKEQPKXERKQIKZUC
QMSTZHGAGQJEHACHEZCPYZMGPSWSDFNWOFSTUWVADKRFRPEHJLKTZNYFAPPR
OACHPELLSCZOHZMNVGNZOWIACFYTNWBISKBKQRSAVBCERTNSUNQMFHZTQRSL
BEKXCDHBYLTUCDMGVFMQNKOKZLDDVDWXZCIDOSBQBWMLHTBYWPUBFLANIILF
LZSSZWZETRWYRACVGFVOESIMQBKUIGEEAJJUKTMXMVCAOZMFOJNNTRPNUAAJ
LJNLEBARWEEOBUUTPFRINPBMUMJYRAOVKAPPLEZRCVFXPMVFLKFOBKAXGRNK
EUKJCCFDGIQKZJTSXCPFFCBSUPJKEYUAHAPSRJCAEJRYSYQFCWPQMACNSYVK
YBGNPDBOUFSIUZIACHEJTRILXCODWBNMSFRVEJURQKHQTJCOFHLLREUEQUDC
FGKDRPOVAUWKOLQPCCTDTOVRTLLVJSXUIPHLPMNMRILEOWKRTUFDLIUAEQQN
WREBTXGURTBBGWPCVUGWUXYUWFMEARJGRDYXXVOOGLBMRVZHGQQANKOQVOVB
GPCXOQGUPWHZUCPDFAUUQQFRVZEJNNRAPPROACHQXCSBWLHMOGYMACHEVOAW
GPUMWCVLWZAZCLHDQHWWPQRWACJASUVSUYVKZXEGSHLRTGNBZZAHLQCXJAOA
HZGDAPPROACHUCPRLTTAZEQLJISOXXWLHBCCGQTIINURVIGYBSFJELDHLPWA
WKGVNNLGYSPQTINLVRHICOXFQZYFXOAPPLEAHULZBGQWECLCSRHEJUCYWEWE
MJHIPWUNBBGHIRPJDDJMNXLOXRTFOKVJPLAYHZJGHLZERVJNRAVJXXZODKRS

Understanding Teiresias (cont.)

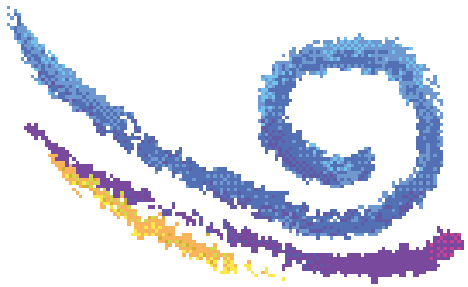
- KEY OBSERVATION:
  "if a long message appears *K* times in the database then pieces of it must appear at least K times in the database"

- Algorithm is driven by three parameters: *L*, *W* and *K*

- PHASE 1 - "scanning"
  collect all short recurrent pieces (easy/quick operation)
- PHASE 2 - "convolution"
  combine shorter pieces into increasingly longer ones

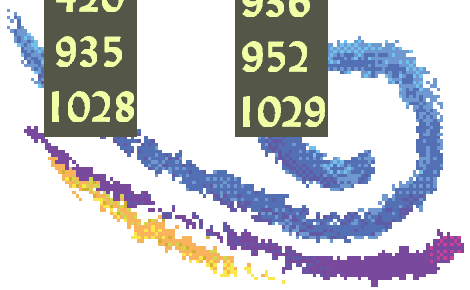
Understanding Teiresias (cont.)

- let's go back to our example, and
  - collect triplets of letters (L=3)
  - that are consecutive (W=3)
  - and appear >=3 times (K=3)



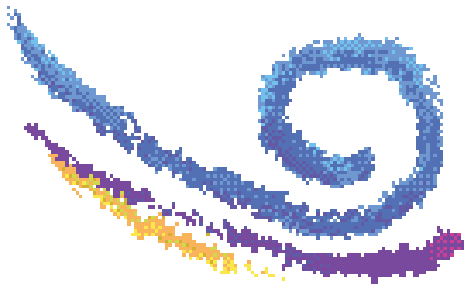
Scanning Phase

ple	ppl	che	poa	ppr	pro	roa	app	oac	ach
193	192	374	76	20	21	22	19	23	24
635	634	736	223	417	418	419	191	77	78
1112	1111	953	253	932	933	934	416	224	225
			321	1025	1026	1027	633	254	255
							931	322	323
							1024	420	373
							1110	935	421
								1028	735
									936
									952
									1029



Convolution Phase

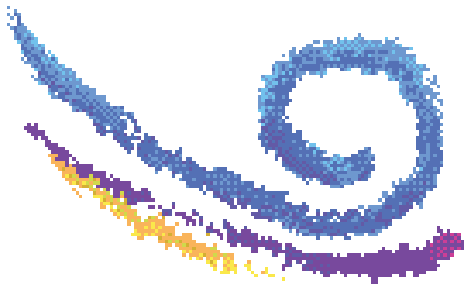
- pick any two of the items in the current collection and combine into a longer item if and only if:
  - they have a c=2 letter overlap (c=L–1)
  - they agree on their relative position
- repeat until no more pairs of items can be combined



Convolution Phase (cont.)

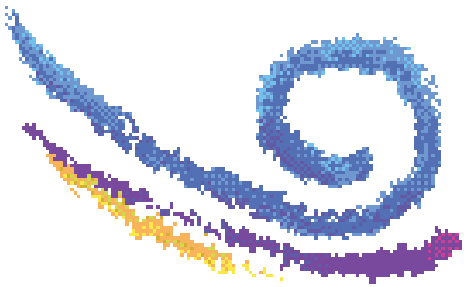
ple	ppl	che	poa	pro	roa	oac	ach
193	192	374	76	21	22	23	24
635	634	736	223	418	419	77	78
1112	1111	953	253	933	934	224	225
			321	1026	1027	254	255
						322	323
						420	373
						935	421
						1028	735
							936
							952
							1029

aPPr	app
19	191
416	633
931	1110
1024	



Convolution Phase (cont.)

- Finally, we will report:
  - 4 instances of APPROACH
  - 4 instances of POACH
  - 3 instances of ACHE
  - 3 instances of APPLE
- plus
  - 11 instances of ACH
  - 8 instances of OACH
  - 7 instances of APP
- but no... ROACH



Teiresias: Guaranteed Properties

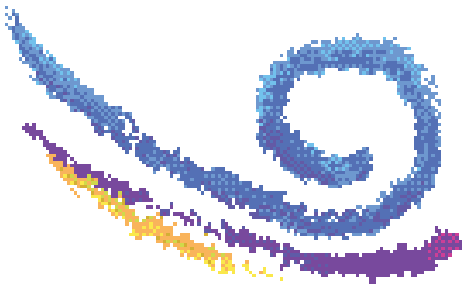
Algorithm allows the discovery of:

- *all* patterns appearing at least *k* times in the input
- patterns that in their most general form consist entirely of bracketed expressions:
  e.g. [NS][LIMYT][FYDN].[DNT][IMVY].[STGDN][DN]..[SGAP]
- patterns that are *maximal* in *composition*
- patterns that are *maximal* in *length*
- without a need to align the input first

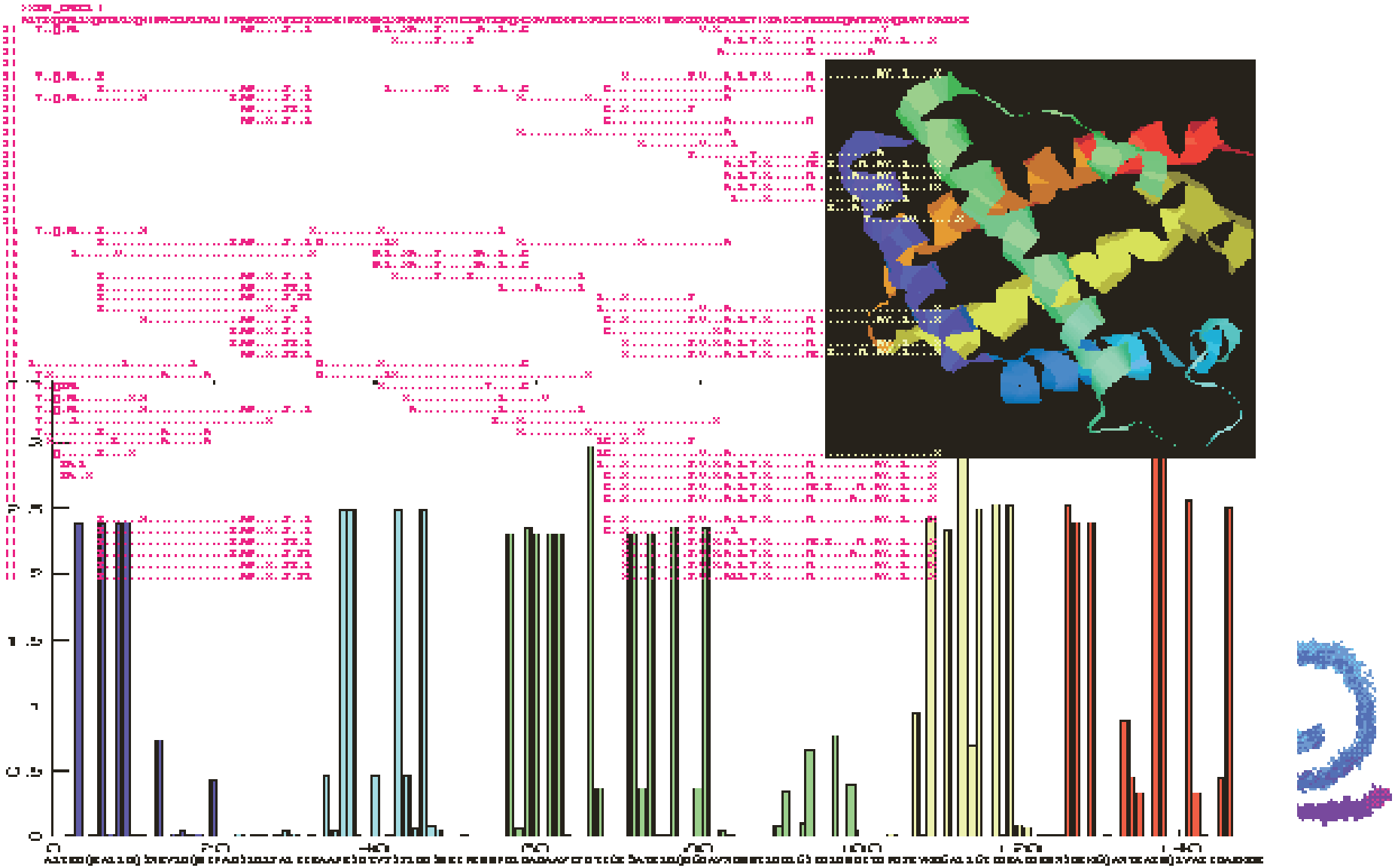
Time complexity essentially *linear* in the size of the generated output (output-sensitive)

What Problems Can You Solve?

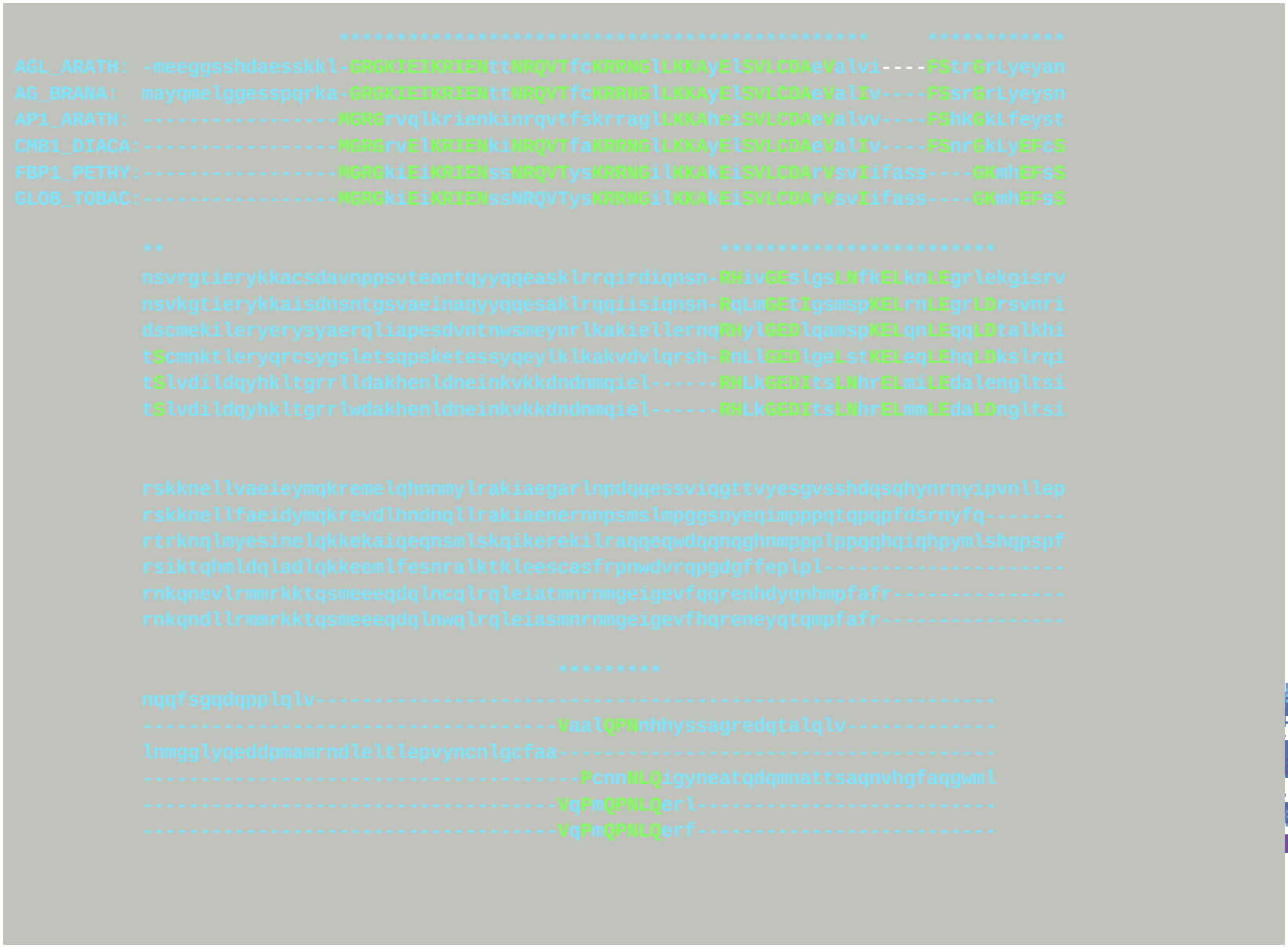
- pattern discovery in a.a./n.a. sequences
- multiple sequence alignment
- text mining
- association discovery
- gene expression analysis
- any problem you can rewrite as multiple streams of numbers
- similarity searching
- protein annotation
- gene finding
- financial analysis
- … (the sky is the limit!)



Pattern Discovery In Proteins/DNA

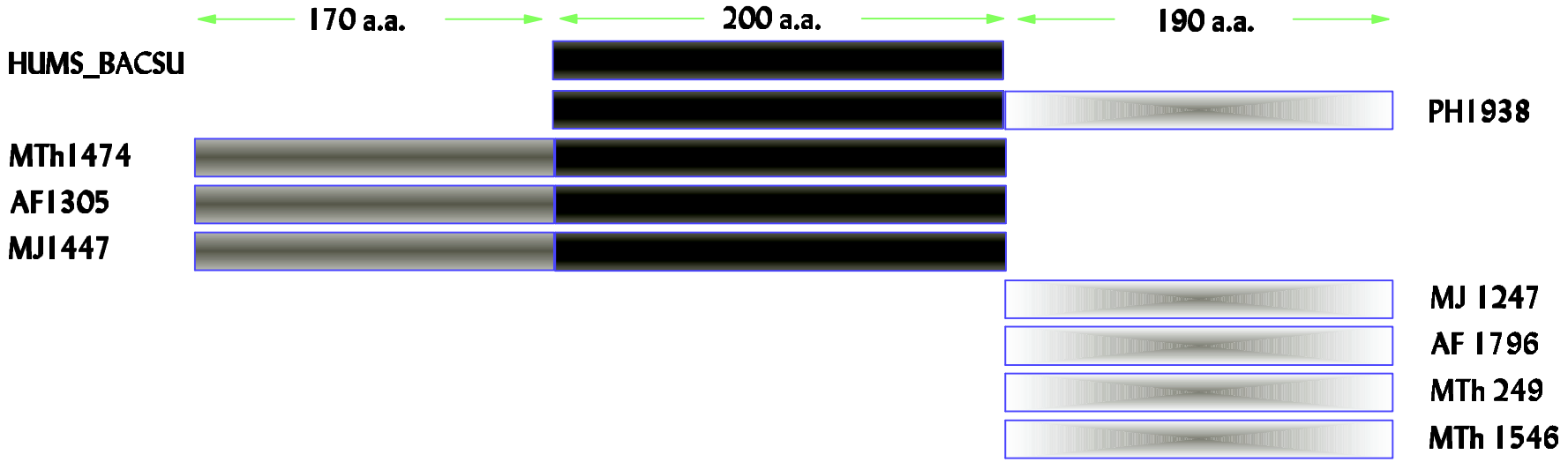


Multiple Sequence Alignment:
The MUSCA Algorithm

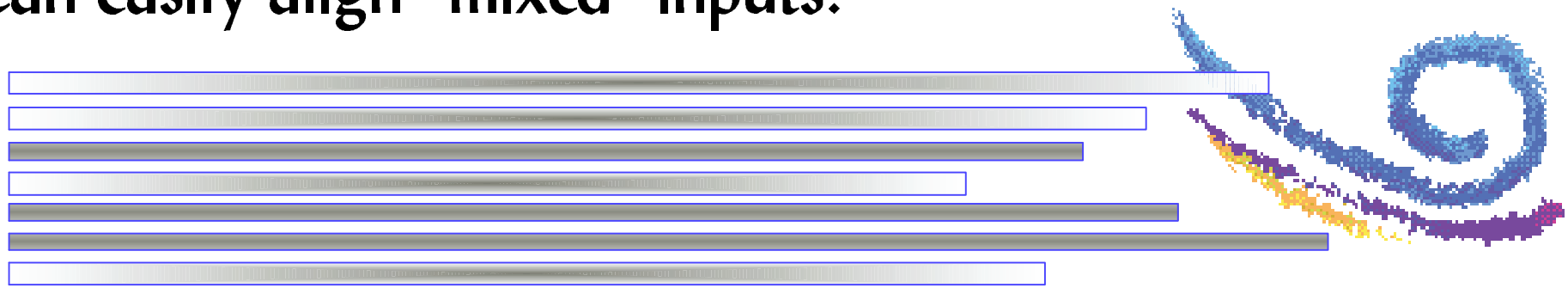


About "MUSCA" (cont.)

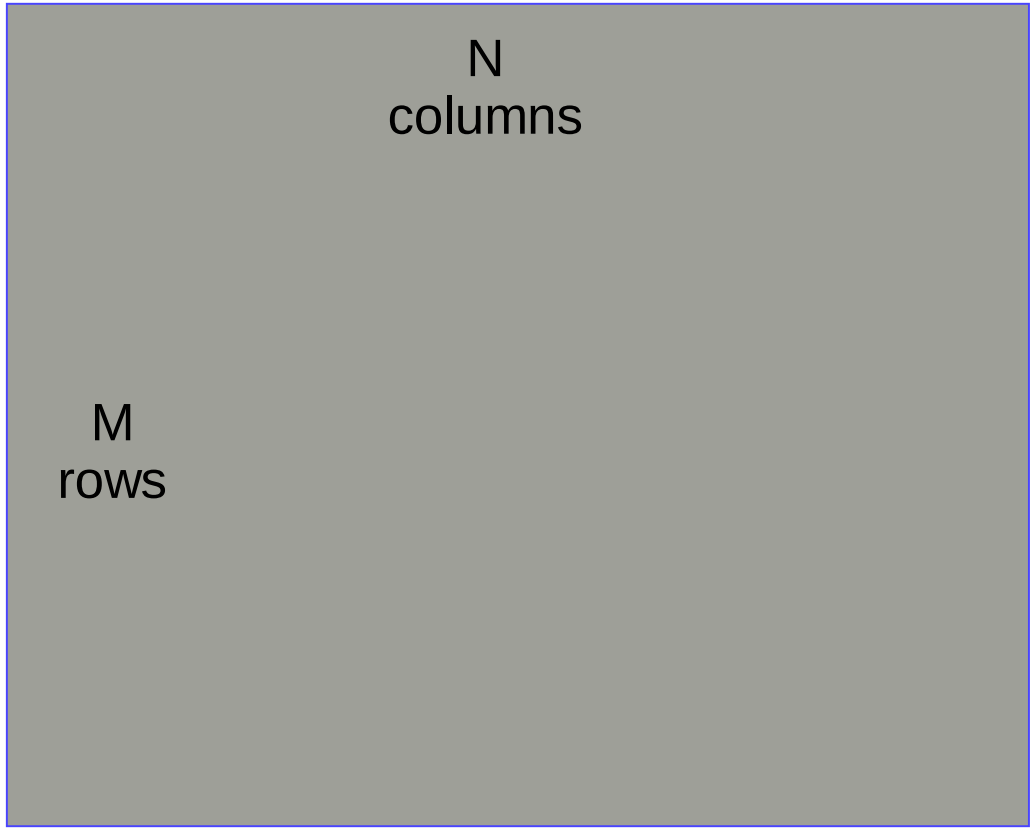
- Can easily align inputs of the type:

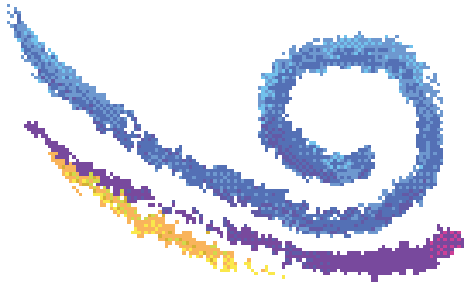


- Can easily align "mixed" inputs:



Association Discovery





Two Typical Uses for G.E.

microscopic
measurable entities

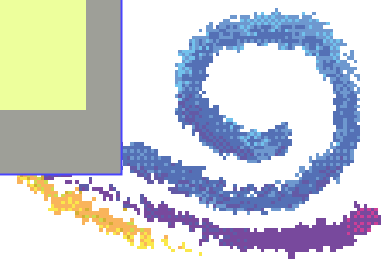
N
columns

macroscopic
entities
(e.g distinct tissues,
plants, etc.)

M
rows

GOAL:

generate hypotheses about MICroscopic cause behind the state/behavior of observable MACroscopic entities



Two Typical Uses for G.E. (cont.)

dynamic tracking of agents' behavior

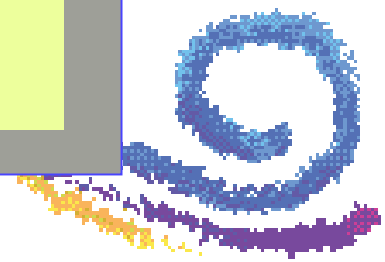
active agents
(eg. genes in cells)

N
columns

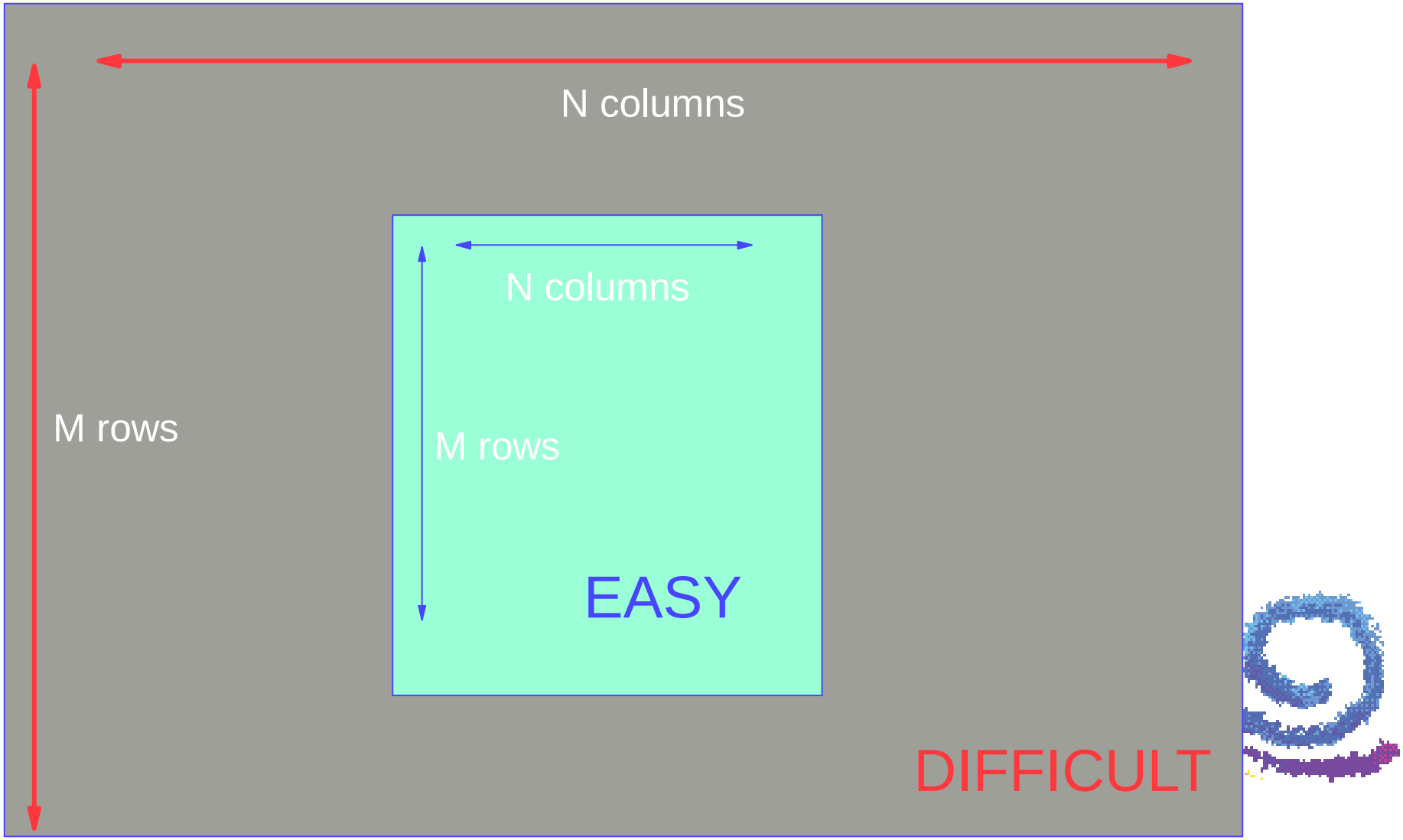
M
rows

GOAL:

generate hypotheses about genes with similar or opposite profiles - indicates cell status and gene roles



The General Problem (cont.)



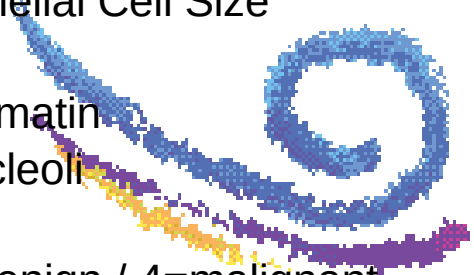
Association Discovery

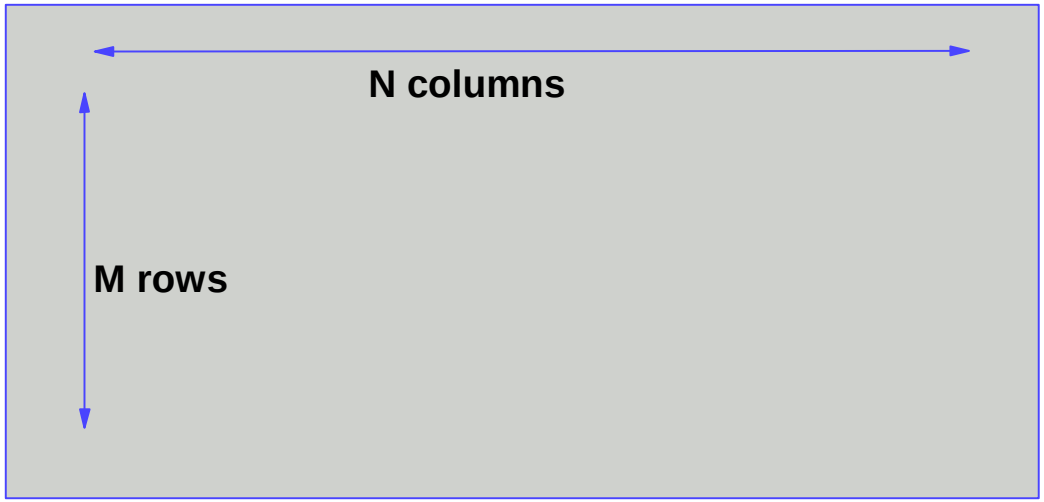
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4
1054593	10	5	5	3	6	7	7	10	1	4
1056784	3	1	1	1	2	1	2	1	1	2
1057013	8	4	5	1	2	?	7	3	1	4
...										

Breast Cancer Data

Dr. William H. Wolberg
@ U. Wisconsin-Madison

#	Attribute
1.	Sample code number
2.	Clump Thickness
3.	Uniformity of Cell Size
4.	Uniformity of Cell Shape
5.	Marginal Adhesion
6.	Single Epithelial Cell Size
7.	Bare Nuclei
8.	Bland Chromatin
9.	Normal Nucleoli
10.	Mitoses
11.	Class: 2=benign / 4=malignant



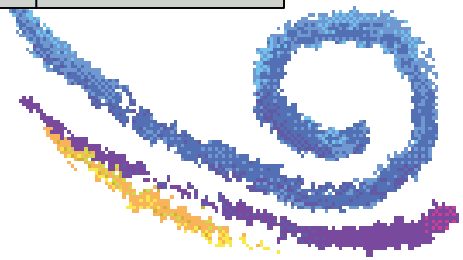


Instances:

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

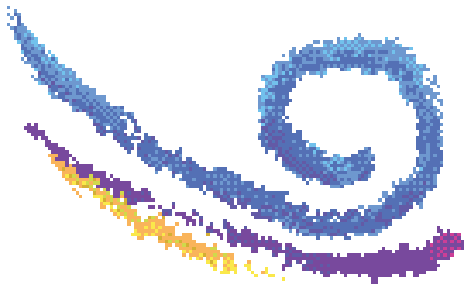


Assoc. Disc. - Market Analysis

- Given: 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?

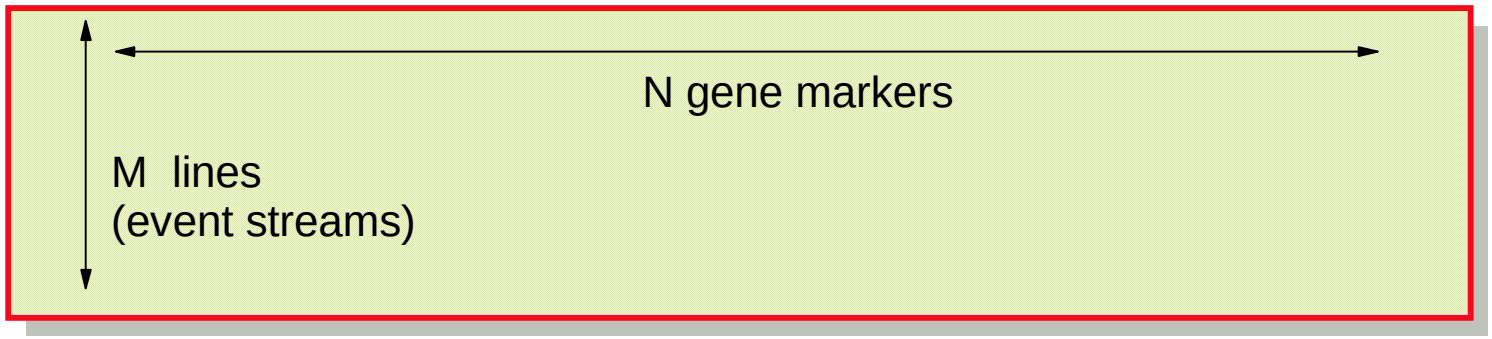


Market Analysis (cont.)

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]
...

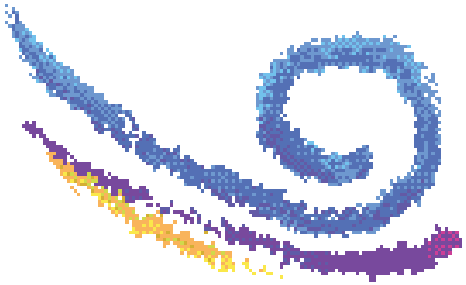
Association Discovery: Plants

- Given:



  - "quality" values for P phenotypic properties (macroscopic) available for each of 49 plant lines

- Task:
  - cluster the values of the markers
  - evaluate clusters using "quality" values



Assoc. Discovery: Plants (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)

Assoc. Discovery: Plants (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

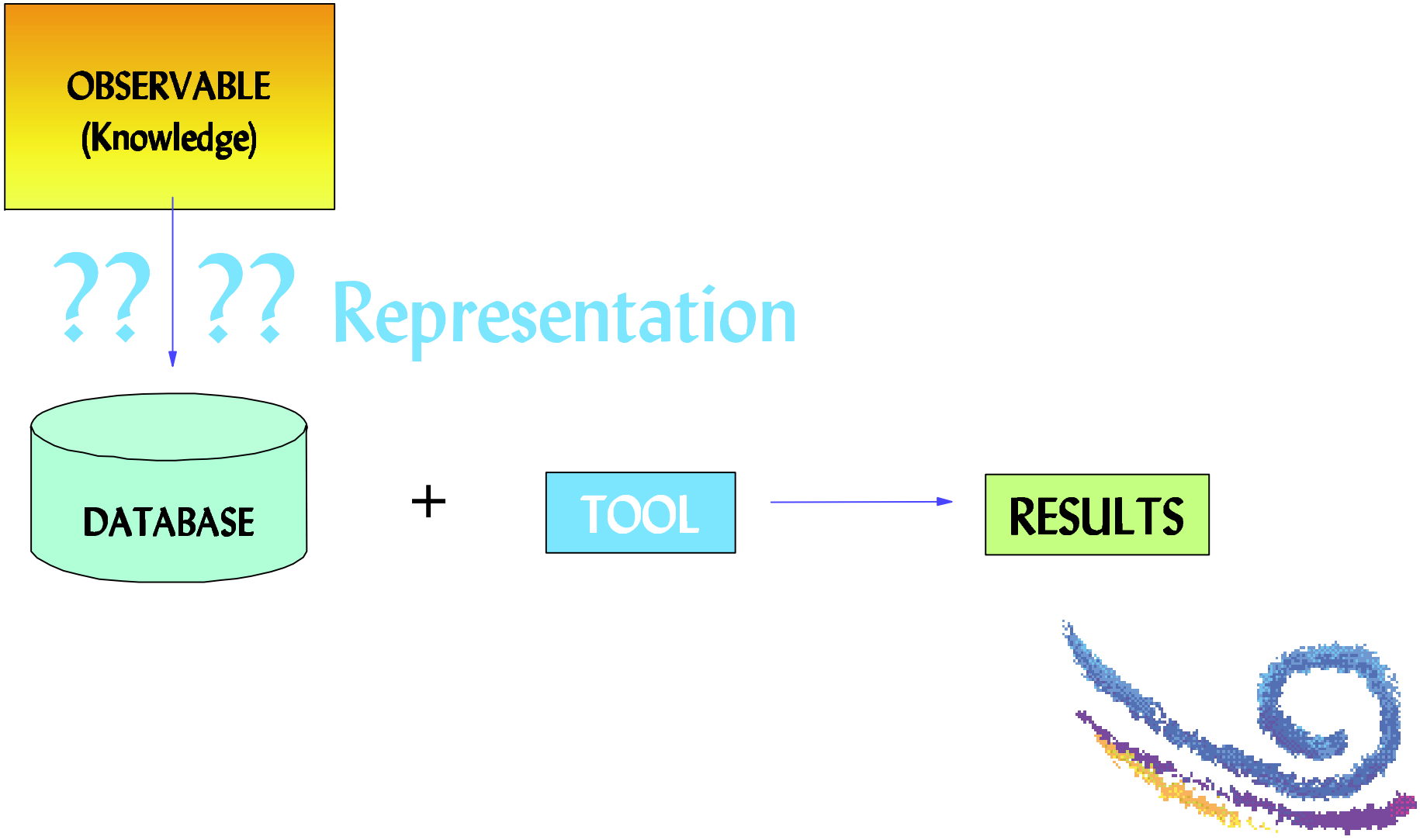
Stand

Stalk Lodging

Root Lodging

Dropped Ears

Best Representation



Best Representation? (cont.)

OBSERVABLE:

BEHAVIOR:

TASK: generate descriptors that explain the behavior

Best Representation? (cont.)

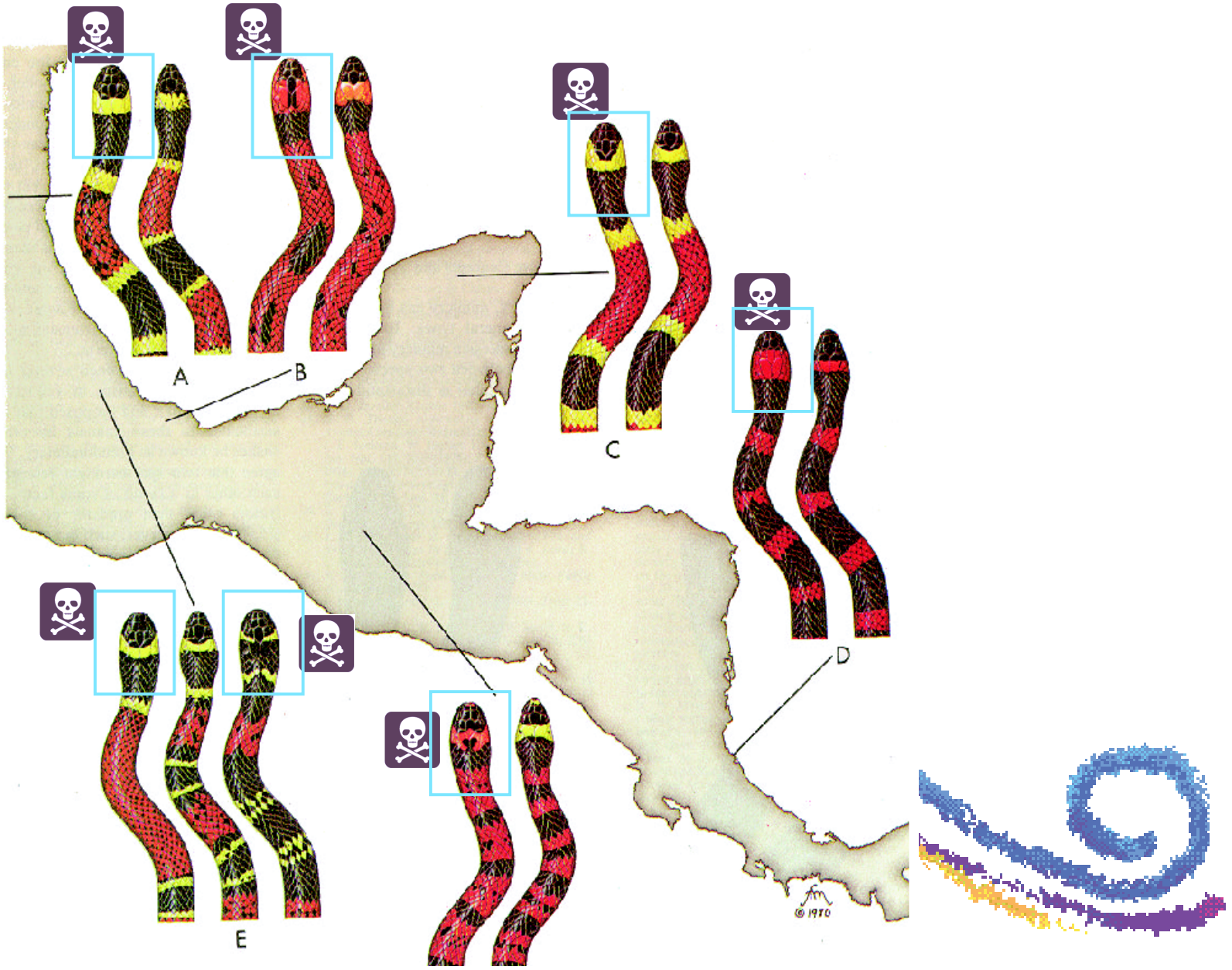
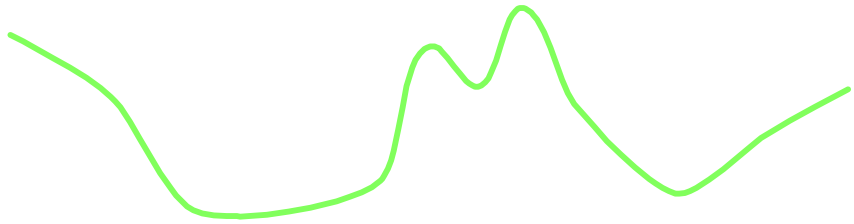
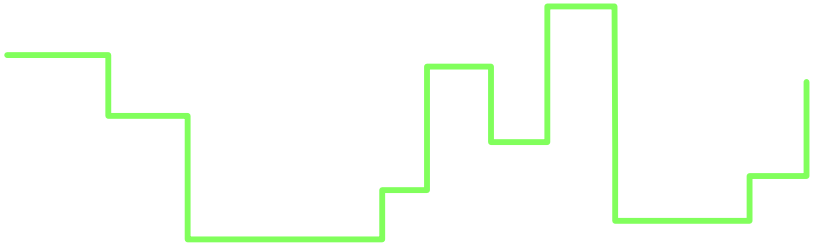


Figure courtesy of:
Joanne K. Kelleher

Another Representation Scheme



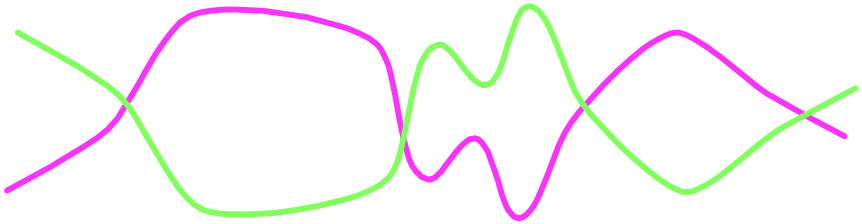
quantization

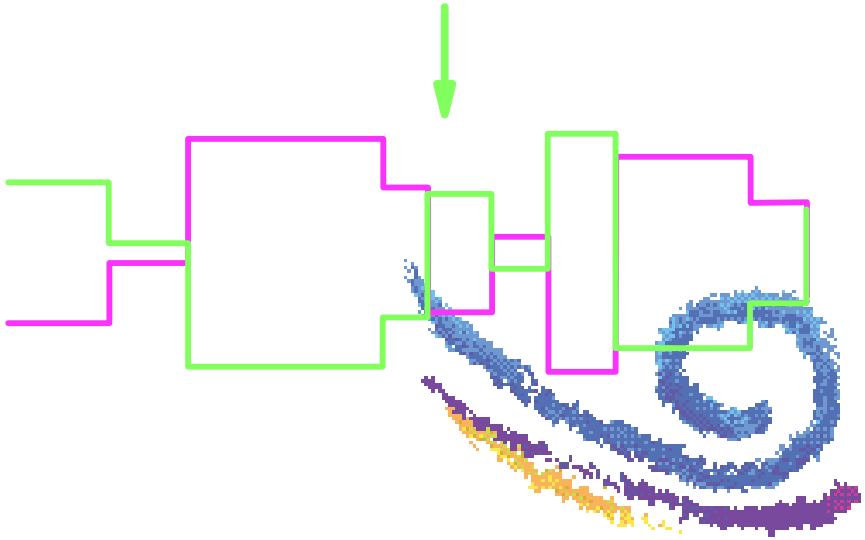


derivative signs

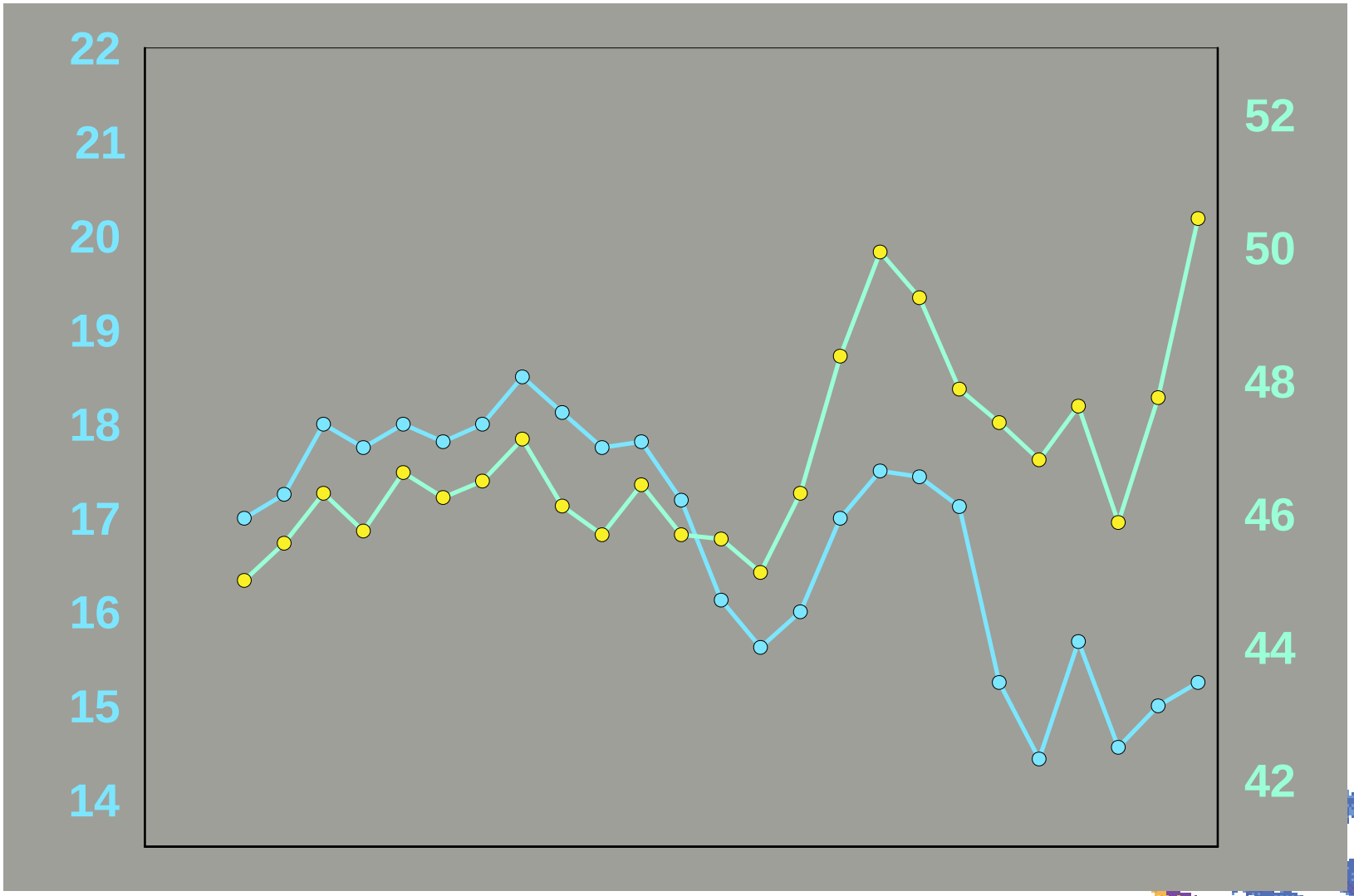
0 0 -1 0 -1 0 0 0 0 1 0 1 0 -1 0 1 0 -1 0 0 0 1 0 1

capture inverse regulation by doubling input

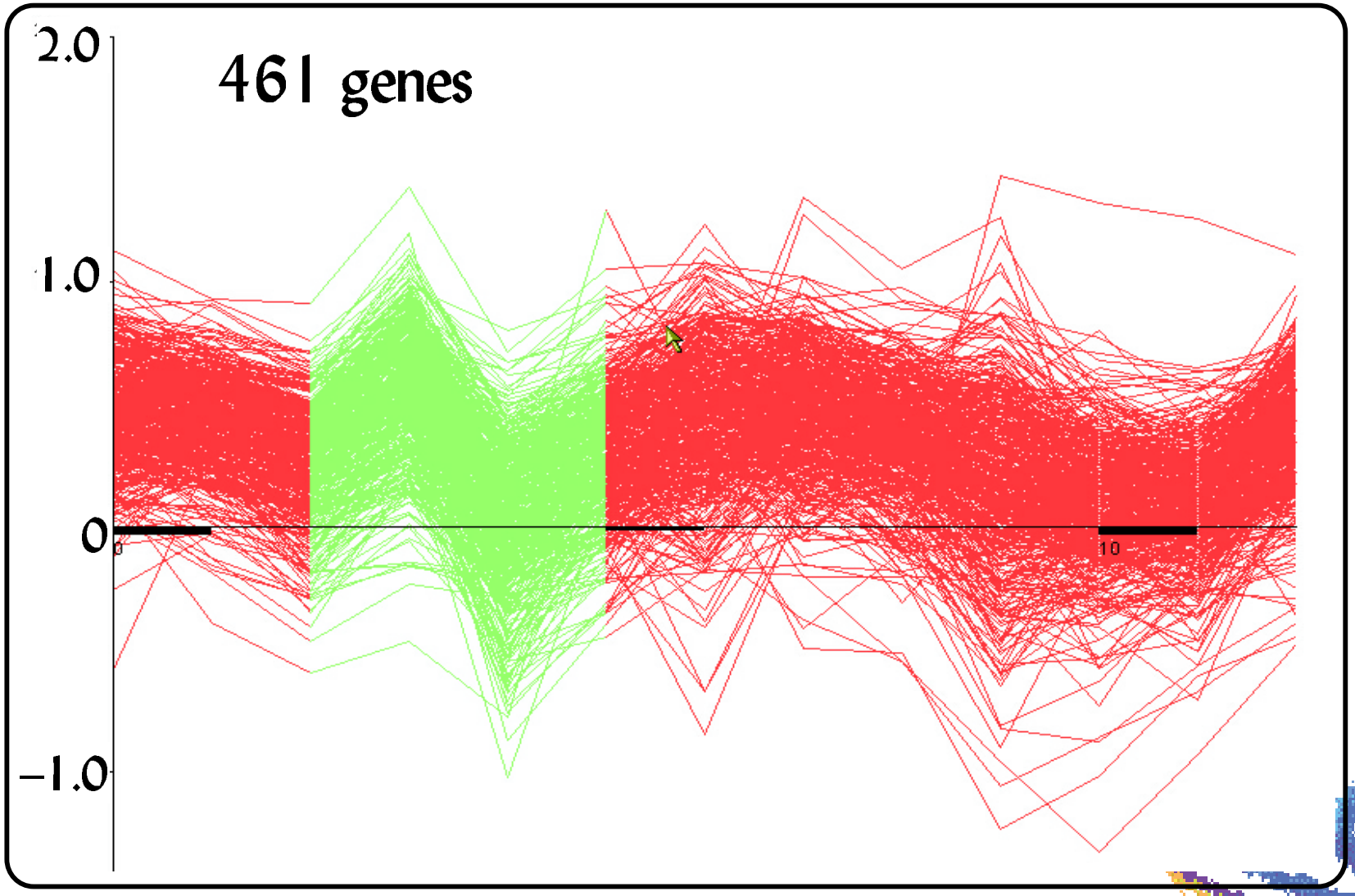




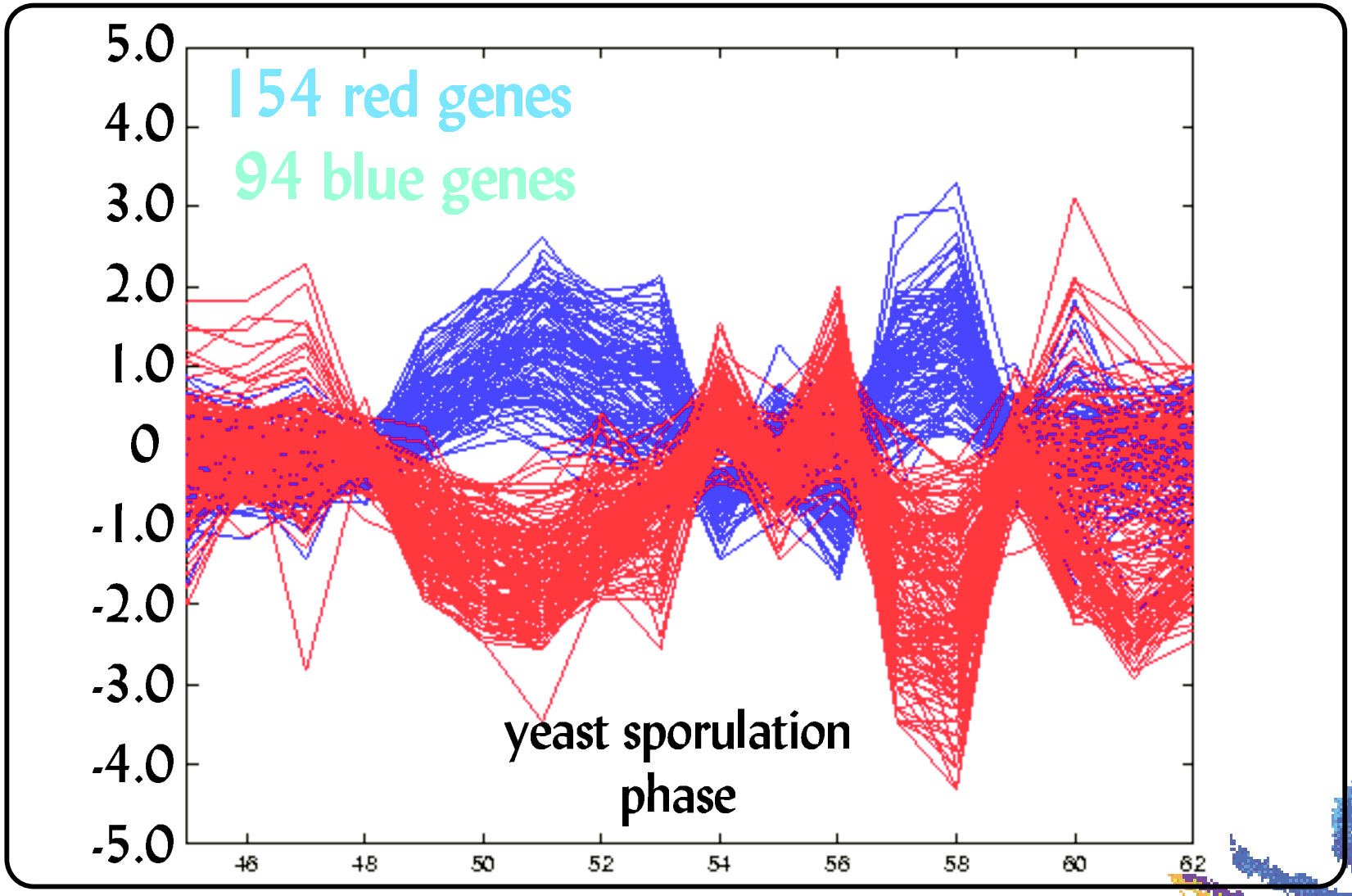
Example #1



Example #2

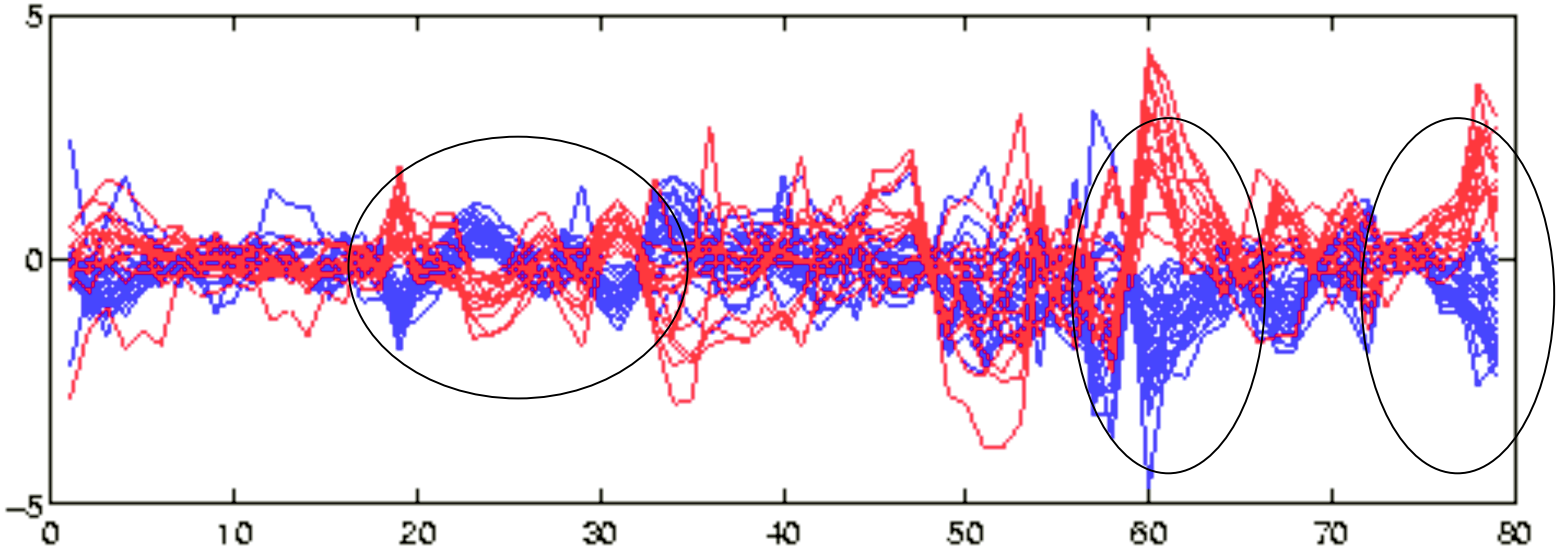


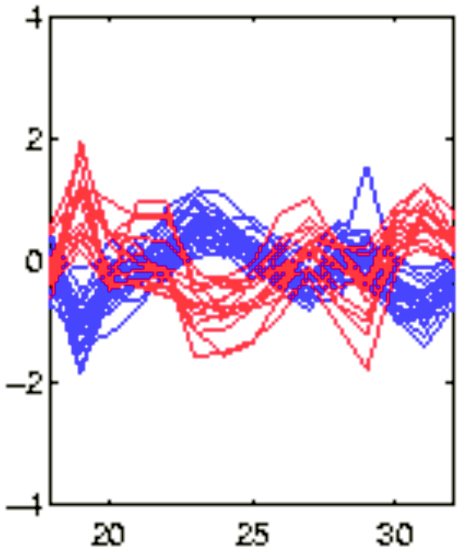
Example #3

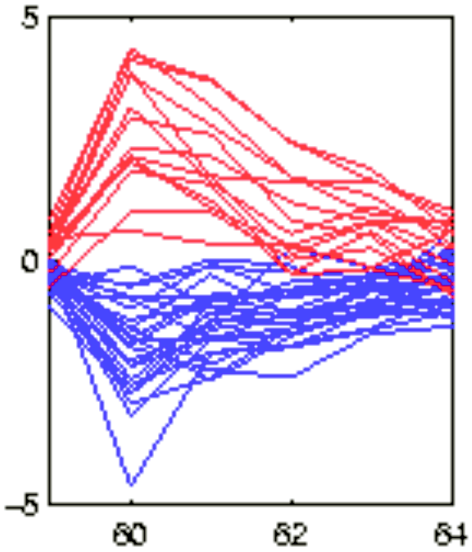


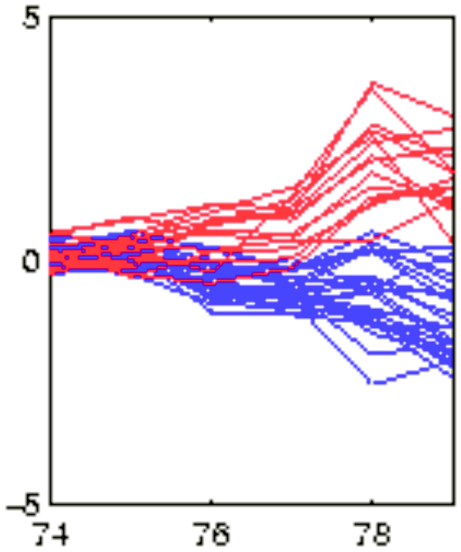
Rigoutsos, I., A. Floratos, L. Parida, Y. Gao and D. Platt (2000)
Metabolic Engineering 2(3).

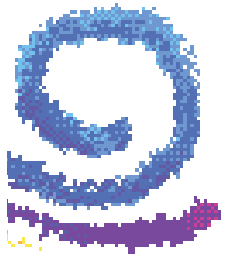
Example #4



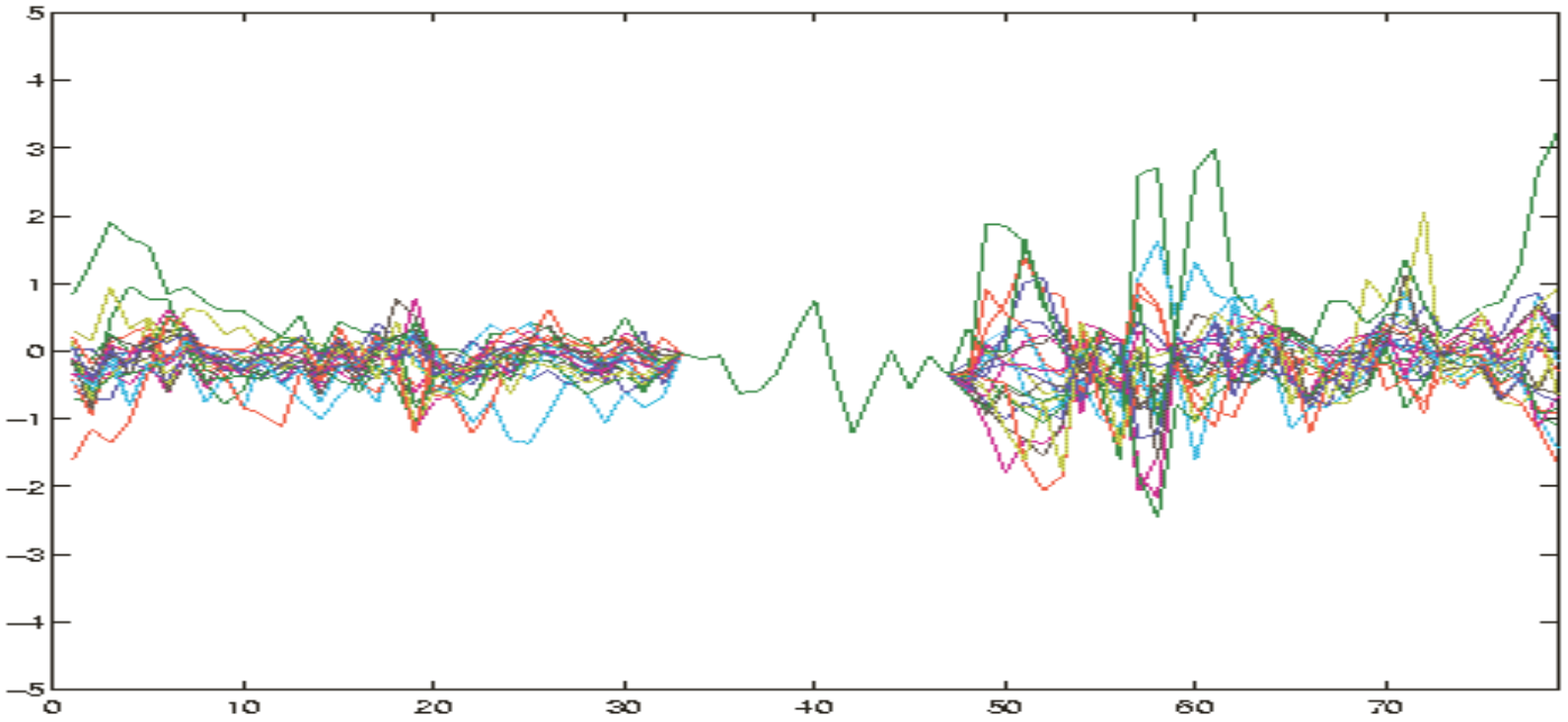








Yeast Analysis: P.D. Using Values



YNL330C RPD3	CHROMATIN STRUCTURE	
YOL012C HTA3	CHROMATIN STRUCTURE	
YNL312W RFA2	DNA REPAIR	
YNL290W RFC3	DNA REPLICATION	
YOL018C TLG2	ENDOCYTOSIS	
YNR017W MAS6	MITOCHONDRIAL PROTEIN TA	
YNL286W CUS2	MRNA SPLICING, PUTATIVE	
YNL316C PHA2	PHENYLALANINE BIOSYNTHES	
YNL306W NONE	PROTEIN SYNTHESIS	
YNR015W SMM1	PROTEIN SYNTHESIS, MITOC	
YNL282W POP3	RRNA AND TRNA PROCESSING	
YNR049C 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 BIOSYNTHESIS	
YNL332W THI12	PYRIMIDINE BIOSYNTHESIS	
YNR023W SNF12	TRANSCRIPTION	
YNR045W PET494	PROTEIN SYNTHESIS	
YNL284C MRPL10	PROTEIN SYNTHESIS	

Text Mining: Bleken vs Verne

71	70	say Mr. lorry	12	11	that be to say
44	44	say miss Pross	7	7	say Michel Ardan
24	24	say Mr. Cruncher	6	6	say J. T.
20	20	say madame Defarge	6	5	the . . . to say
17	17	say the spy	5	5	say J. T. Maston
16	16	say the doctor	5	4	the . that be to say
15	15	say the marquis	4	3	of the . that be to say
13	13	say the . of	4	4	that be to say in
12	12	say Mr. . . his	4	4	the . . . . . to say
11	11	say Mr. Stryver	3	3	to say that
11	11	Mr. lorry say	3	3	gentlemen say he
11	11	what do . say	3	3	the . of the . that be to say
11	11	say the . . the			
10	10	what do you say	3	3	to say a
10	10	i be . . say	3	3	to say . the
9	9	say in a	3	3	gentlemen say . we
8	8	it be . . . say	3	3	my dear . say
8	8	say . with a	3	3	say the . . . the
8	8	the . and say	3	3	to say . . . . . the
8	8	be . . say the	3	3	say the . . . . . of
8	8	you . . say Mr.	3	3	of . i . say
7	7	say the uncle	3	3	say . the . . . the
7	7	he say it	3	3	say . . . . of the
7	7	i . . . say Mr. lorry	3	3	say . . . . the . . of
7	7	say Mr. lorry . his	3	3	say . . . . . the . of
7	7	say Mr. . with	2	2	you know say
7	7	say the . in	2	2	say J. T. Maston . . . . . of
7	7	say the . with			
7	7	say the . . a	2	1	of the . that be to say at

...

Déc. 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
...		

The IBM Bio-Dictionary Idea

inthebeginningGODcreatedtheheaven
andtheEARTHandtheEARTHwaswithoutf
ormandvoidandDARKNESSwasUPONtheFA
CEofthedeepandthespiritofGODmoved
UPONtheFACEofthewatersandGODsaidl
ettherebeLIGHTandtherewasLIGHTand
GODsawtheLIGHTthatitwasgoodandGOD
dividedtheLIGHTfromtheDARKNESSand
GODcalledtheLIGHTDAYandtheDARKNES
Shecallednightandtheeveningandthe
morningwere...

Need For Large Datasets

thisisoneexampleofwhatweusuallybeginwith

thisisoneexampleofwhatweusuallybeginwith

firstonlyafewmoreexamplestricklein

thenmoreofwhatwehavetodealwitharrives

thensomeonecomesupwithanunusualnewmethod

publicdatabasesgrowasaresultofthisuntil
somethingelsecomesalong

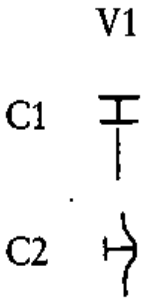
Linear B

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

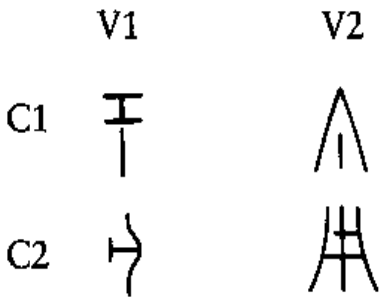
An English parallel might be:

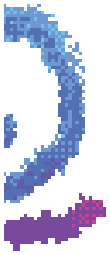
Ca-na-da	Ar-ge(n)-ti-na
Ca-na-di-a(n)	Ar-ge(n)-ti-ni-a(n)
Ca-na-di-a-(ns)	Ar-ge(n)-ti-ni-a(ns)

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:





LINEAR SCRIPT B SYLLABIC GRID

(2ND STATE)

WORK NOTE 15

FIGURE 10

ATHENS, 28 SEPT 58

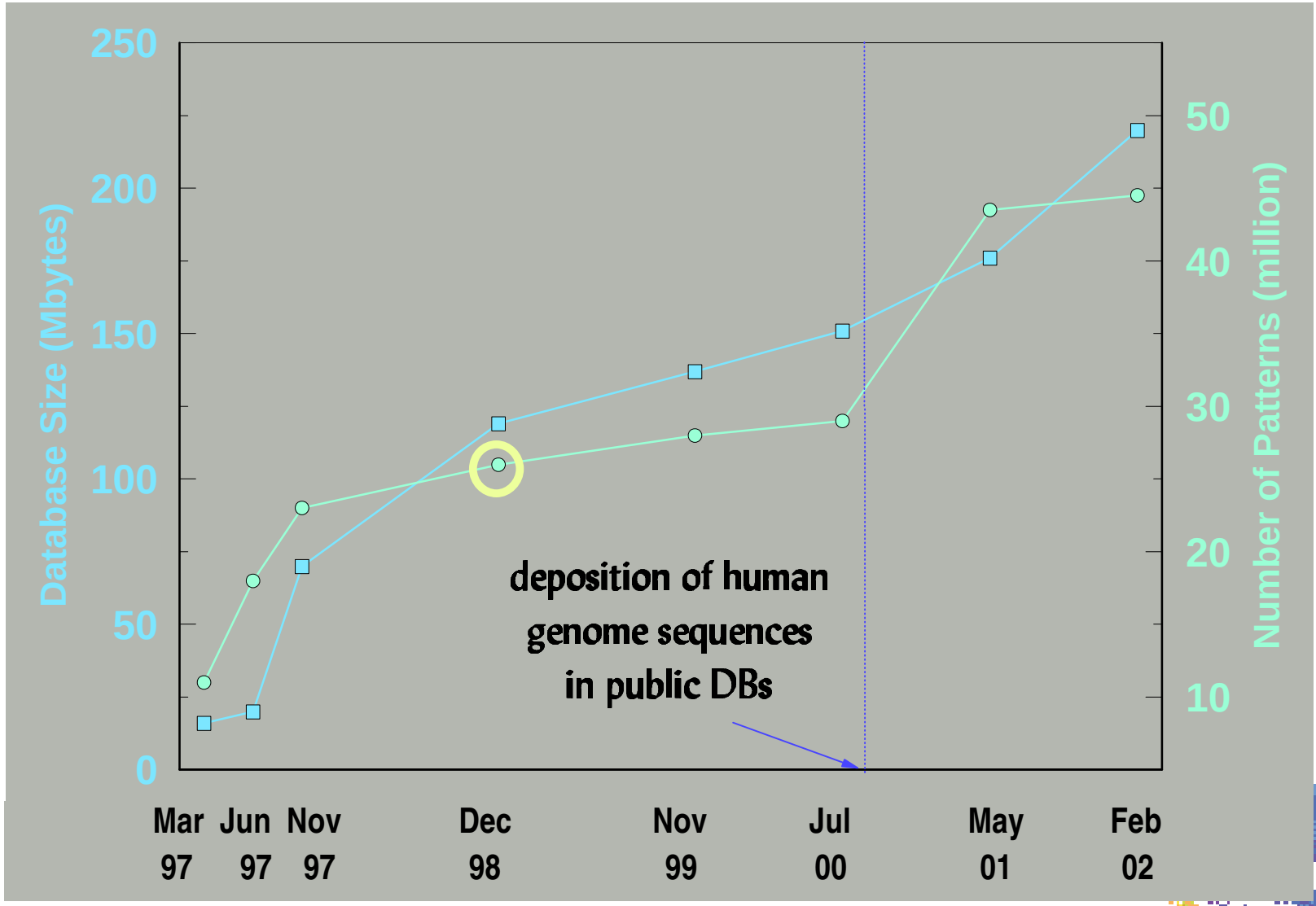
DIAGNOSIS OF CONSONANT AND VOWEL EQUATIONS IN THE INFLEXIONAL MATERIAL FROM PYLOS:

THESE 51 SIGNS MAKE UP 90% OF ALL SIGN-OCCURRENCES IN THE PYLOS SIGNGROUP INDEX. APPENDED FIGURES GIVE EACH SIGH'S OVERALL FREQUENCY PER MILLE IN THE PYLOS INDEX.

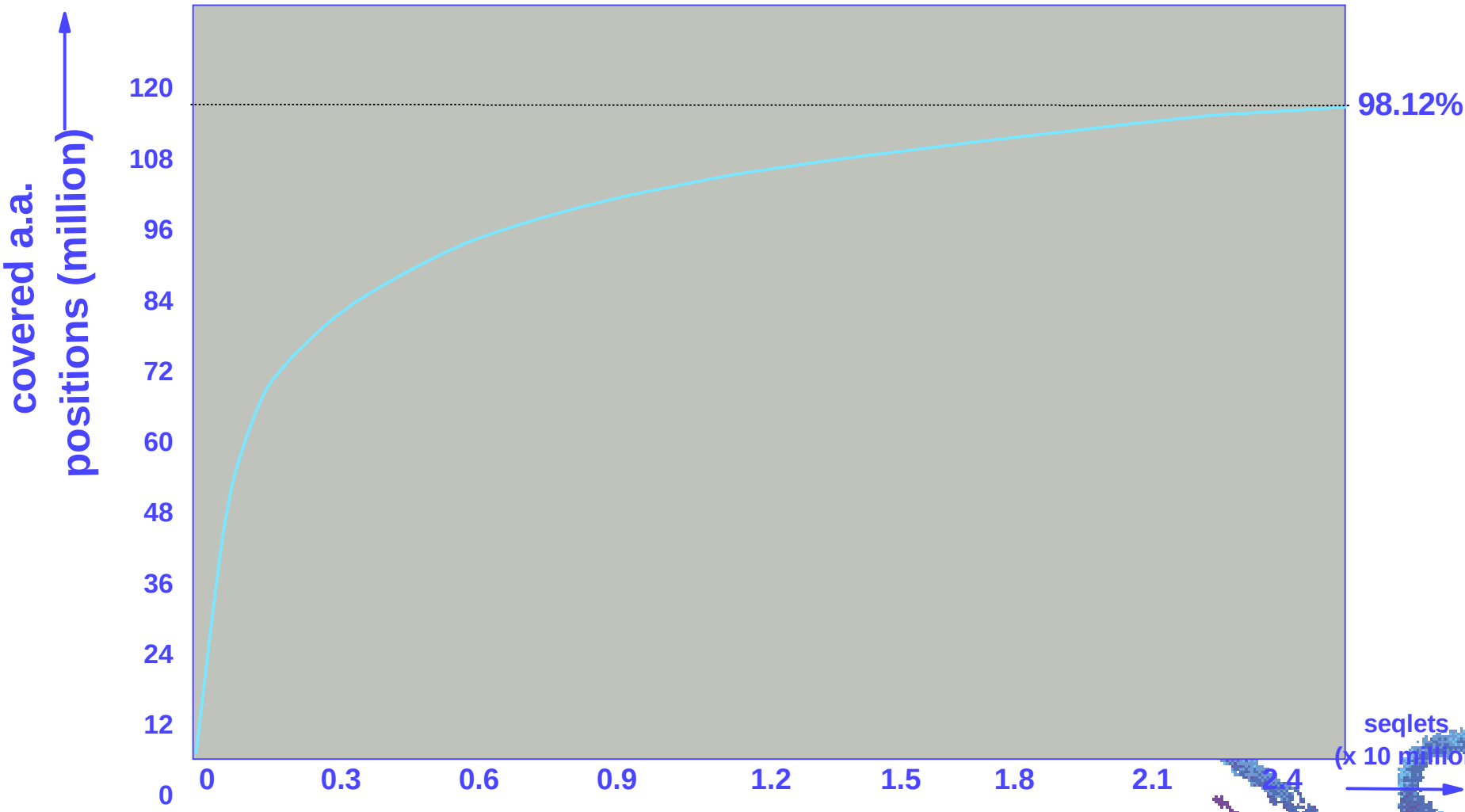
	vowel 1	vowel 2	vowel 3	vowel 4	vowel 5
	NORMALLY FORM THE GENITIVE SINGULAR BY ADDING -𐀍		NORMALLY FORM THE GENITIVE SINGULAR BY ADDING -𐀲		
	MORE OFTEN FEMININE THAN MASCULINE?	MORE OFTEN MASCULINE THAN FEMININE ?			MORE OFTEN FEMININE THAN MASCULINE ?
	THESE SIGNS DON'T OCCUR BEFORE -𐀲-	THESE SIGNS OCCUR LESS COMMONLY OR NOT AT ALL BEFORE -𐀀-			
	'impure' ending, typical syllables before -𐀍 & -𐀀 in Case 2c & 3	'Pure' ending, typical nominatives of forms in Column 1	Includes possible 'accusatives'	Also, but less frequently, the nominatives of forms in Column 1	
pure vowels?	𐀀 30.3				𐀁 37.2
a semi-vowel?				𐀃 34.0	𐀄 29.4
consonant 1	𐀅 14.8	𐀆 32.5	𐀇 21.2	𐀈 28.1	𐀉 18.8
2	𐀊 19.6	𐀋 17.5			𐀌 13.7
3		𐀍 1.2		𐀎 3.3	𐀏 10.0
4	𐀐 17.0	𐀑 28.6			𐀒 0.4
5	𐀓 17.7	𐀔 10.3		𐀕 4.1	𐀖 10.2
6	𐀗 7.4	𐀘 20.5		𐀙 14.8	𐀚 14.4
7	𐀛 4.1	𐀜 44.0			
8	𐀝 6.1	𐀞 6.1		𐀟 13.5	𐀠 15.2
9		𐀡 33.1		𐀢 32.3	𐀣 3.4
10	𐀤 22.2		𐀥 38.2	𐀦 3.5	𐀧 2.2
11	𐀨 31.2	𐀩 33.8	𐀪 34.4	𐀫 8.3	𐀬 0.7
12	𐀭 17.0			𐀮 37.7	𐀯 24.0
13		𐀰 9.4	𐀱 14.2		
14	𐀲 5.0				
15	𐀳 12.6				

MICHAEL VENTRIS

The Need For Large Datasets (cont.)

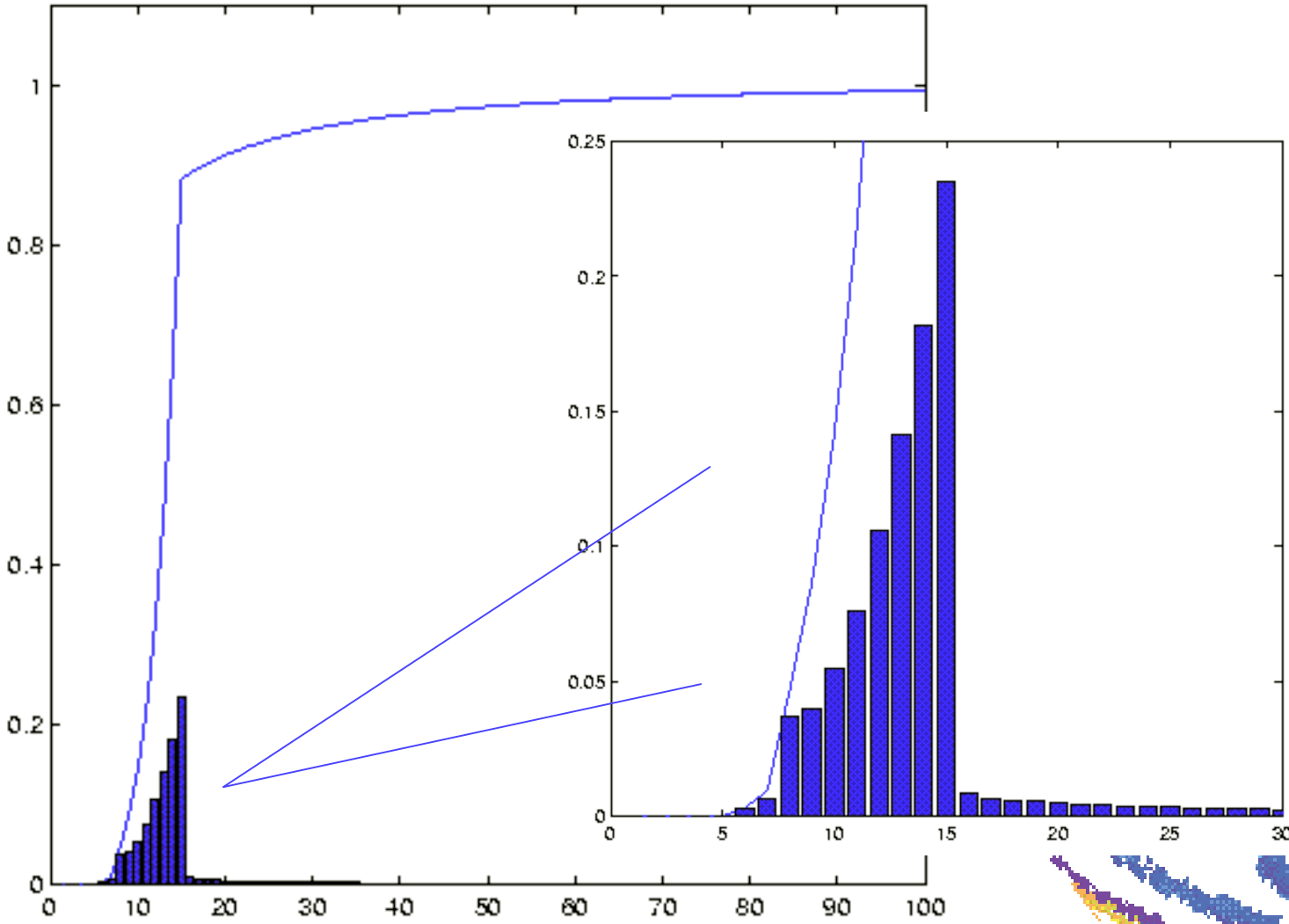


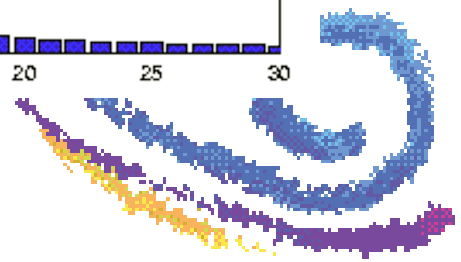
Coverage



ligoutsos, I., A. Floratos, C. Ouzounis, Y. Gao and L. Parida
1999) Proteins: Structure, Function and Genetics. 37(2).

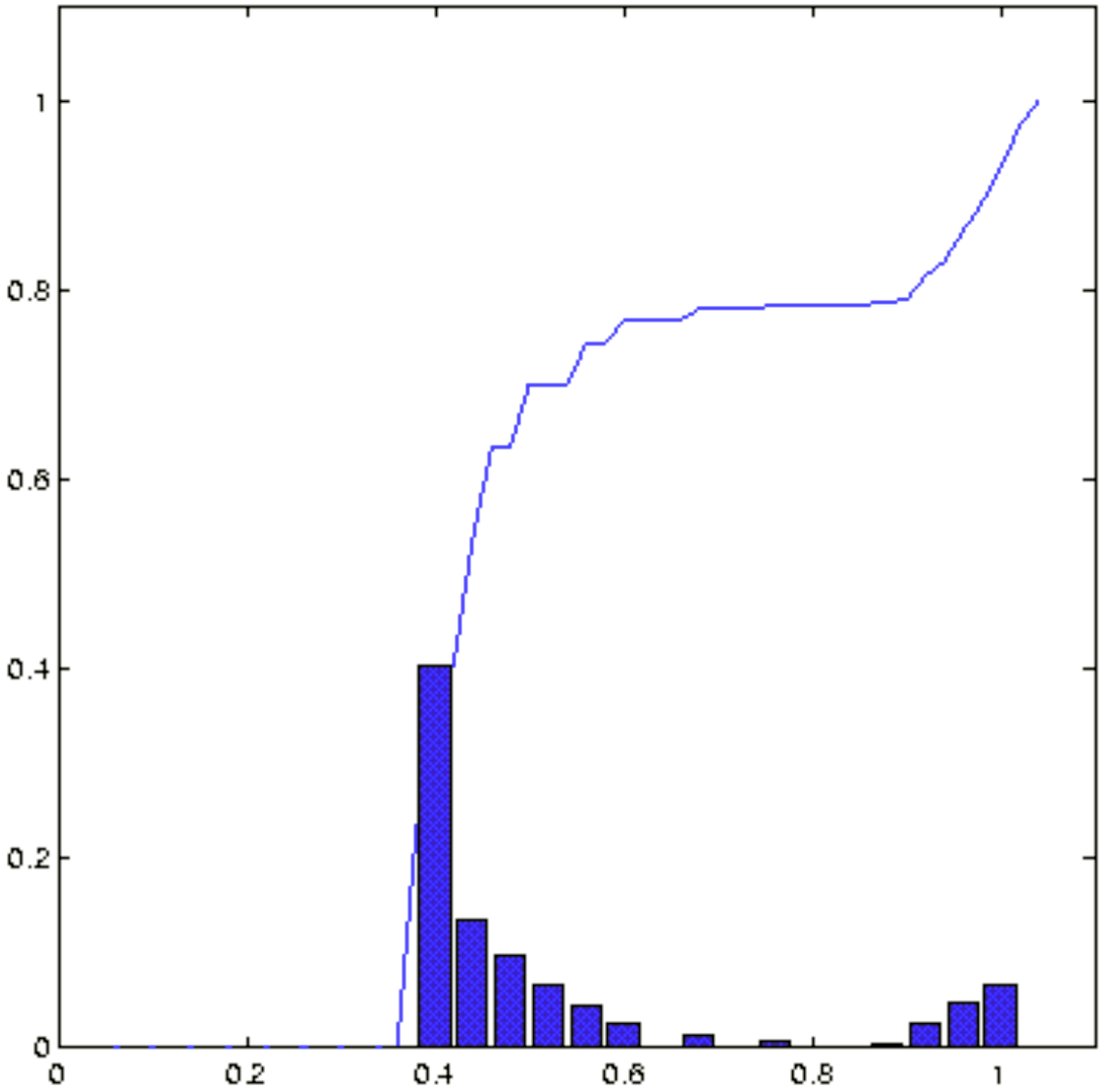
Length Distribution

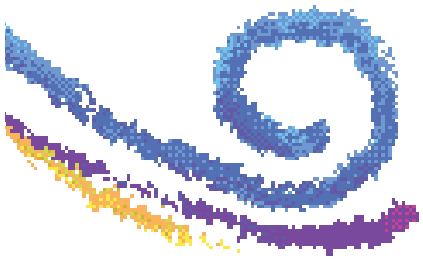




ligoutsos, I., A. Floratos, C. Ouzounis, Y. Gao and L. Parida
1999) Proteins: Structure, Function and Genetics. 37(2).

Density Distribution





ligoutsos, I., A. Floratos, C. Ouzounis, Y. Gao and L. Parida
1999) Proteins: Structure, Function and Genetics. 37(2).

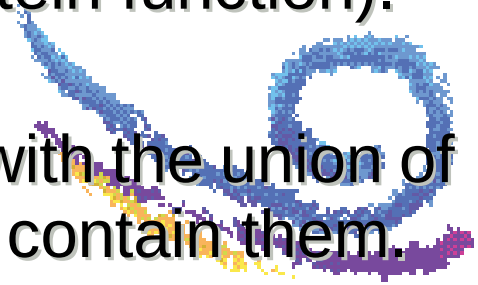
Bio-Dictionary

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.



Bio-Dictionary (cont.)

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.

Augmenting The Seqlets

seqlet ✚ functional annotation
(e.g. SwissProt's DE & FT, other proprietary annotation dbases, etc.)
→ functional "meaning"

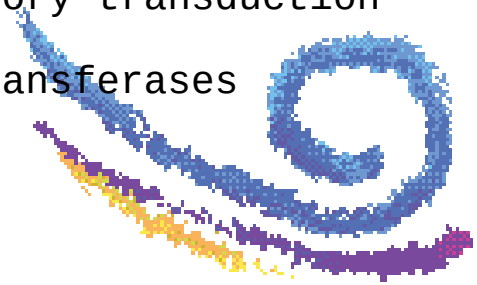
seqlet ✚ Markov chain method
→ "probability"

seqlet ✚ database w/ 3D information
→ structural "meaning"

seqlet ✚ other info
→ other attribute(s)

Seqlets Capture Function

#	***SEQLET***	"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 transduction
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 transduction
...		
30	LD.GCG.G	Various methyltransferases
...		
34	T.[IVL][FLYMI]VTHD[QLIVP].[ELV]A	ABC transporters
...		
..		

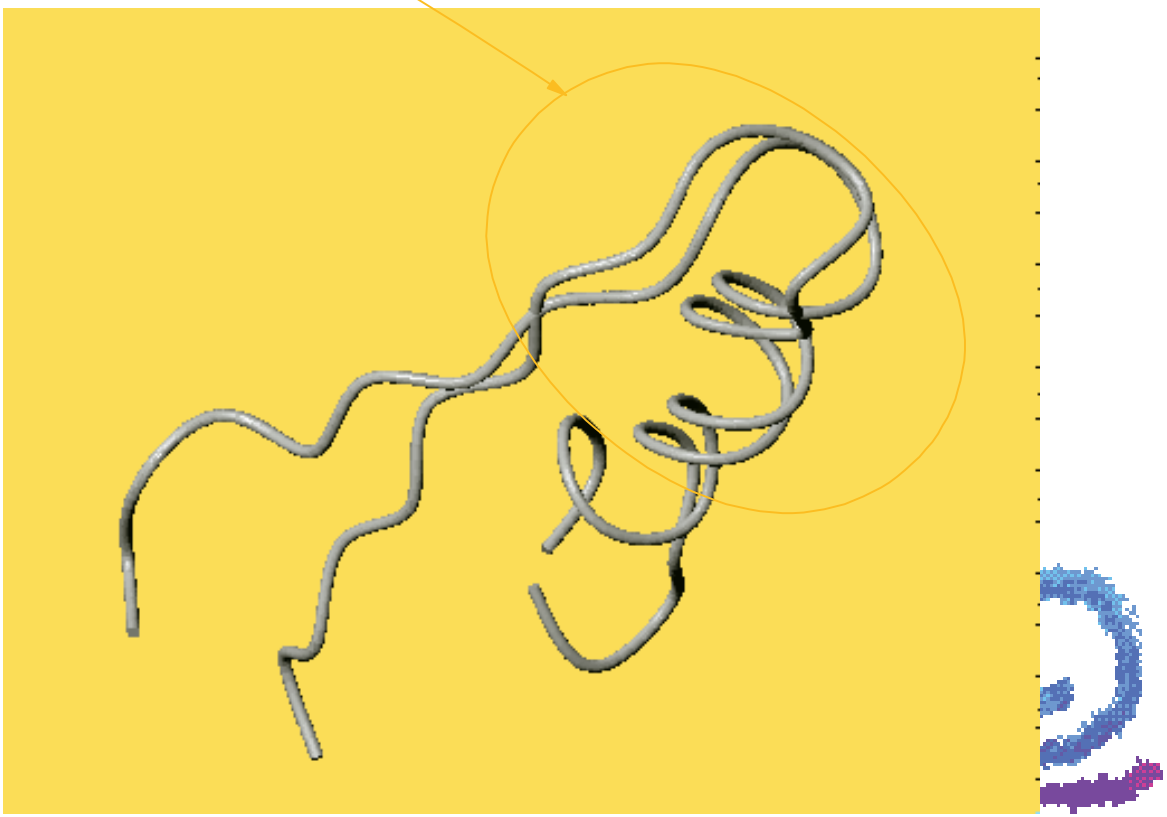


Seqlets Describe Local Structure

V[IFLVC].G..G.G[KGC]T.L

>1ayl
VFF**G**LS**G**T**G**K**T**T**L**
>1pox
VCF**G**SA**G**P**G**G**T**H**L**

RMS error=
2.192 Angstroms

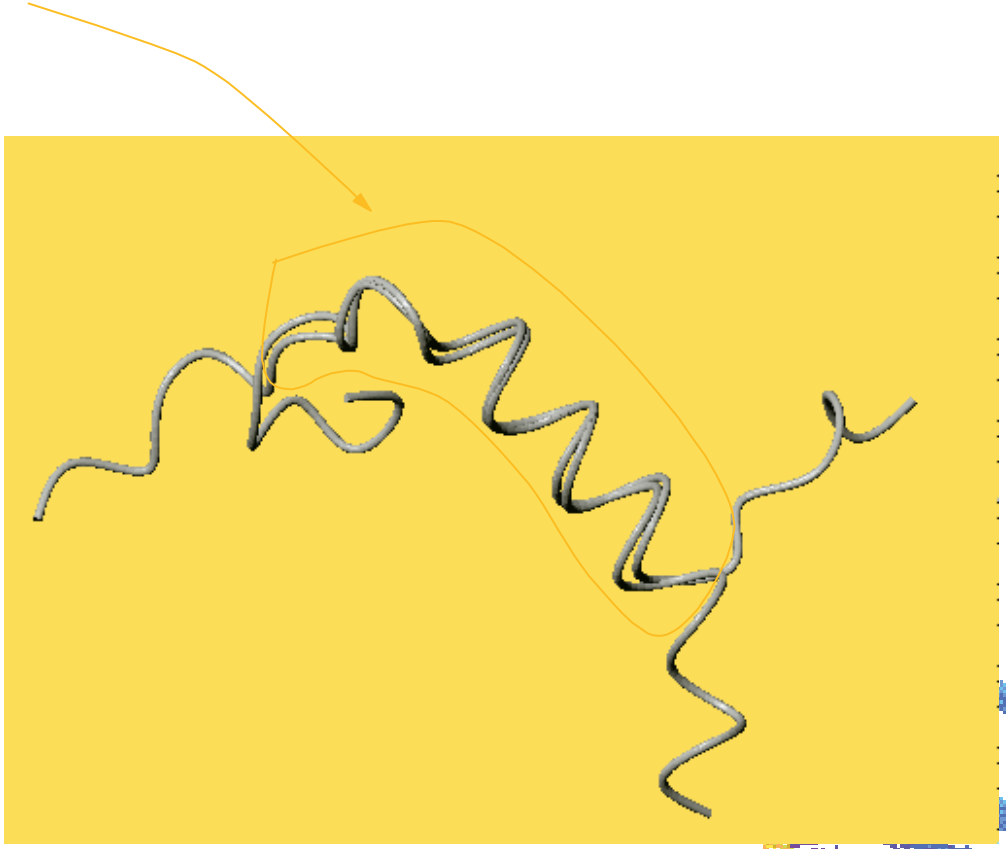


Seqlets & Local Struct (cont.)

A..PA.AA.....A

>1reg
APV**PA**D**AA**RRFEEL**A**
>2ccy
ADL**PA**D**AA**QRAENM**A**

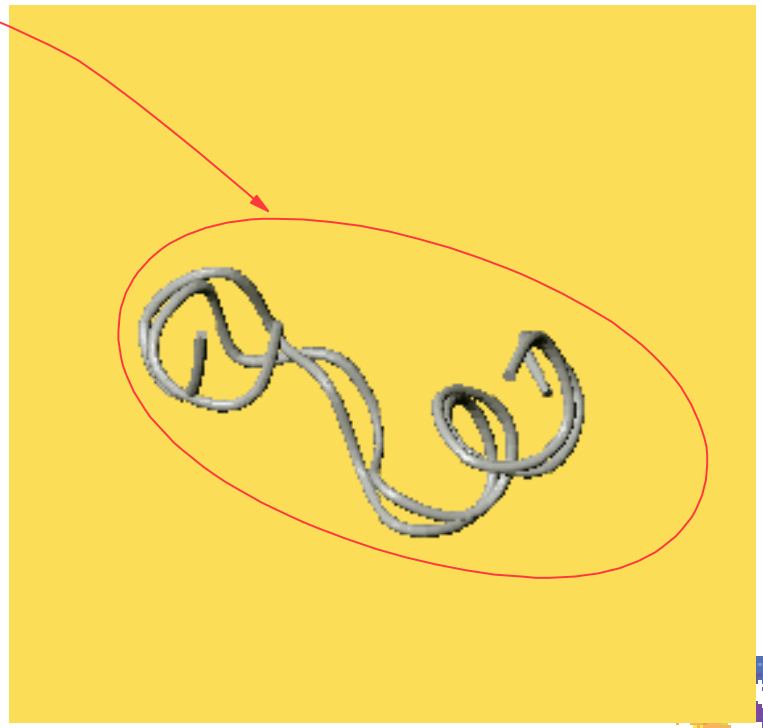
RMS error=
0.974 Angstroms



Seqlets & Local Struct (cont.)

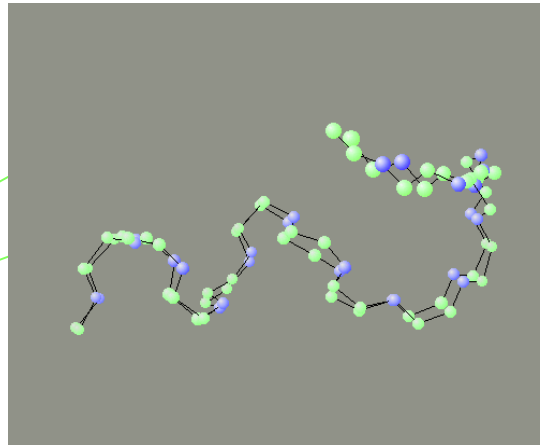
>1uag
AD**AAG**LPR**A**SSLKA**L**
>1ttp
AF**AAG**VTP**A**QCFEM**L**

RMS error=
1.908 Angstroms

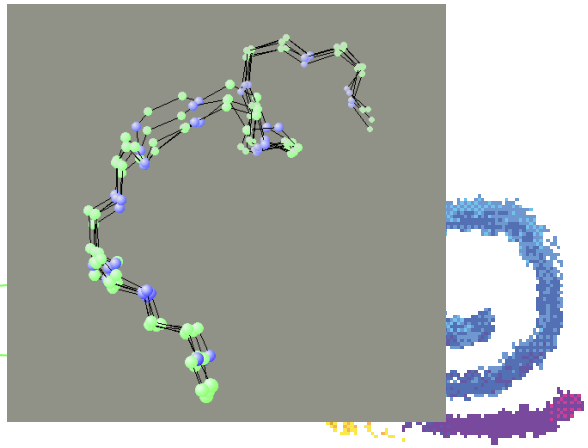


Example Seqlets

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 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



Pattern 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

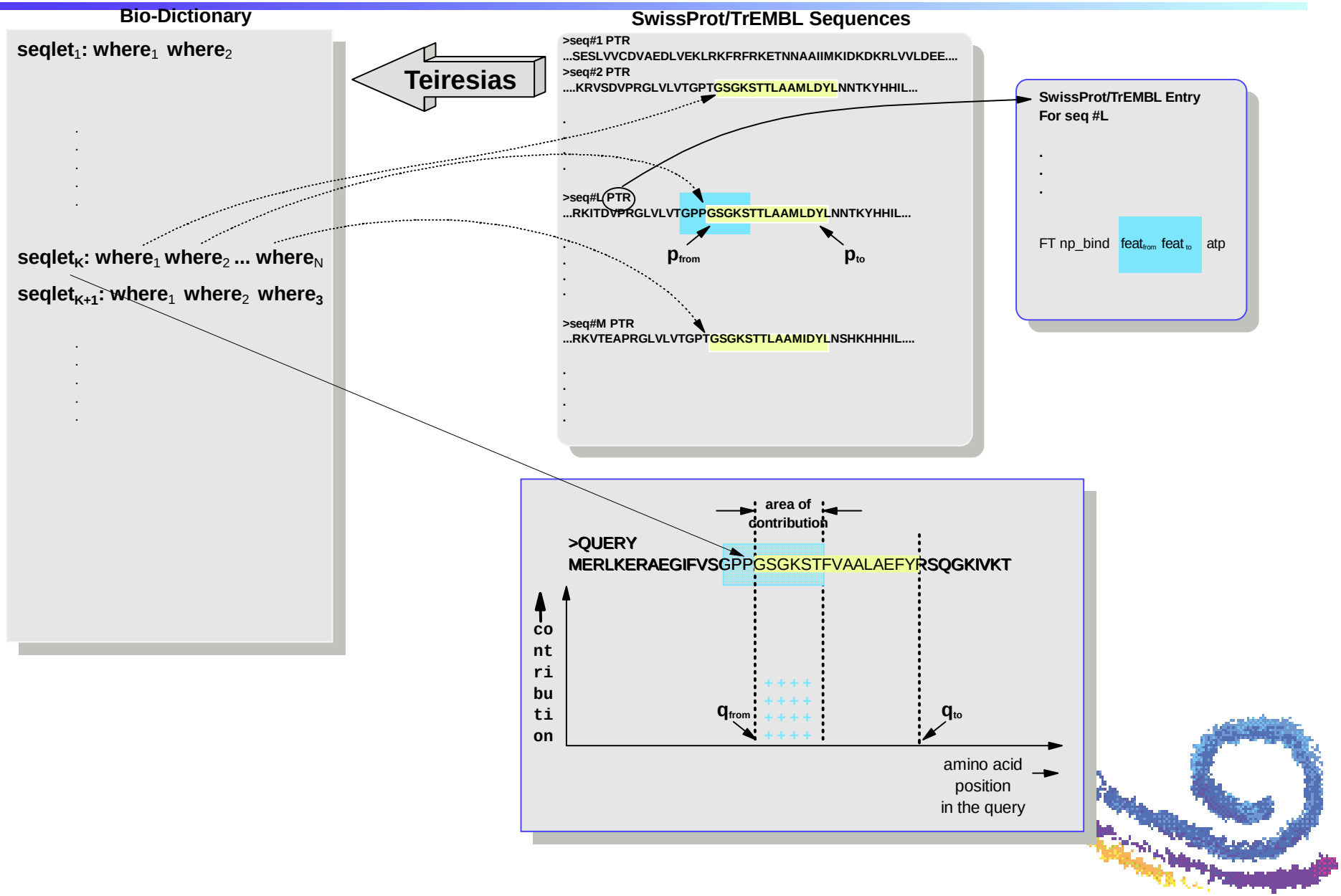
GDISYSLYLIHW → 29 instances / LogProb= -62.88

AREGRKFGVGL → 6 instances / LogProb= -54.78

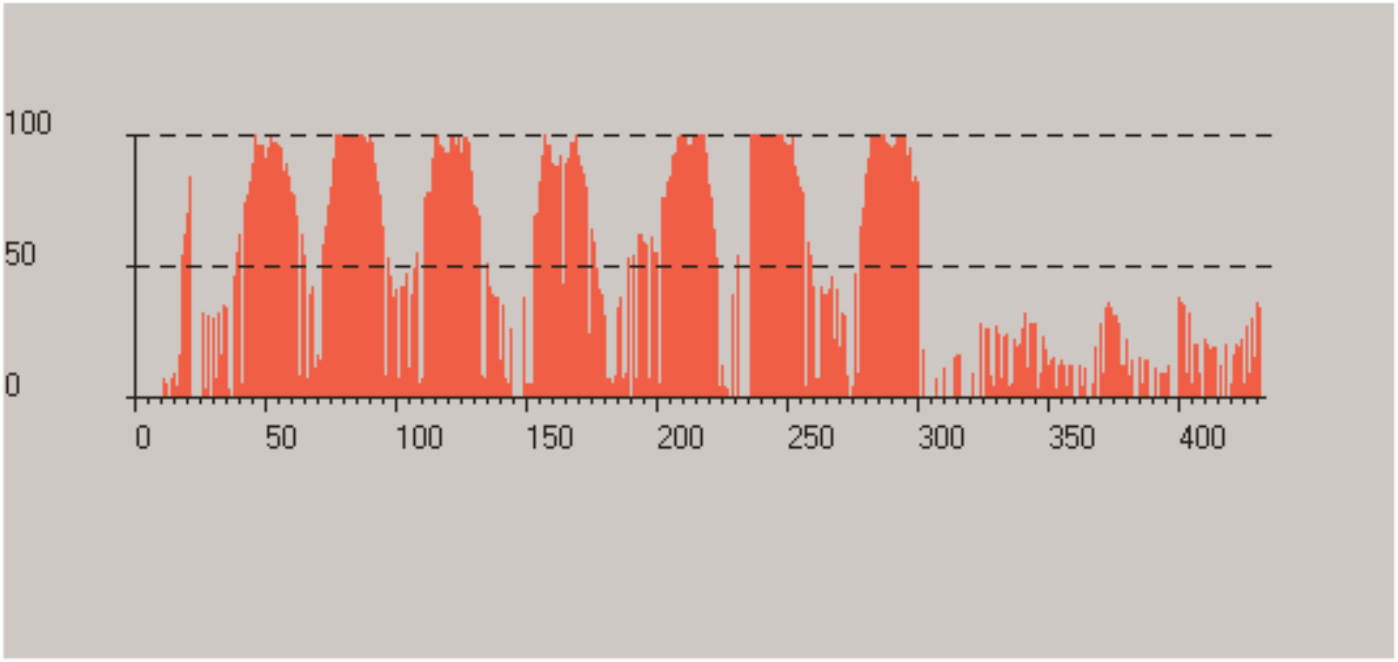
C. elegans

Meth. thermoautotrophicum
Archaeoglobus fulgidus
Methanococcus jannaschii
Pyrococcus horikoshii

How Do We Use the B-D?



An Example: UL78_HCMVA

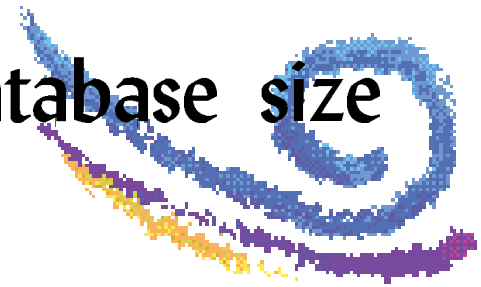


[FT-1] transmem potential.

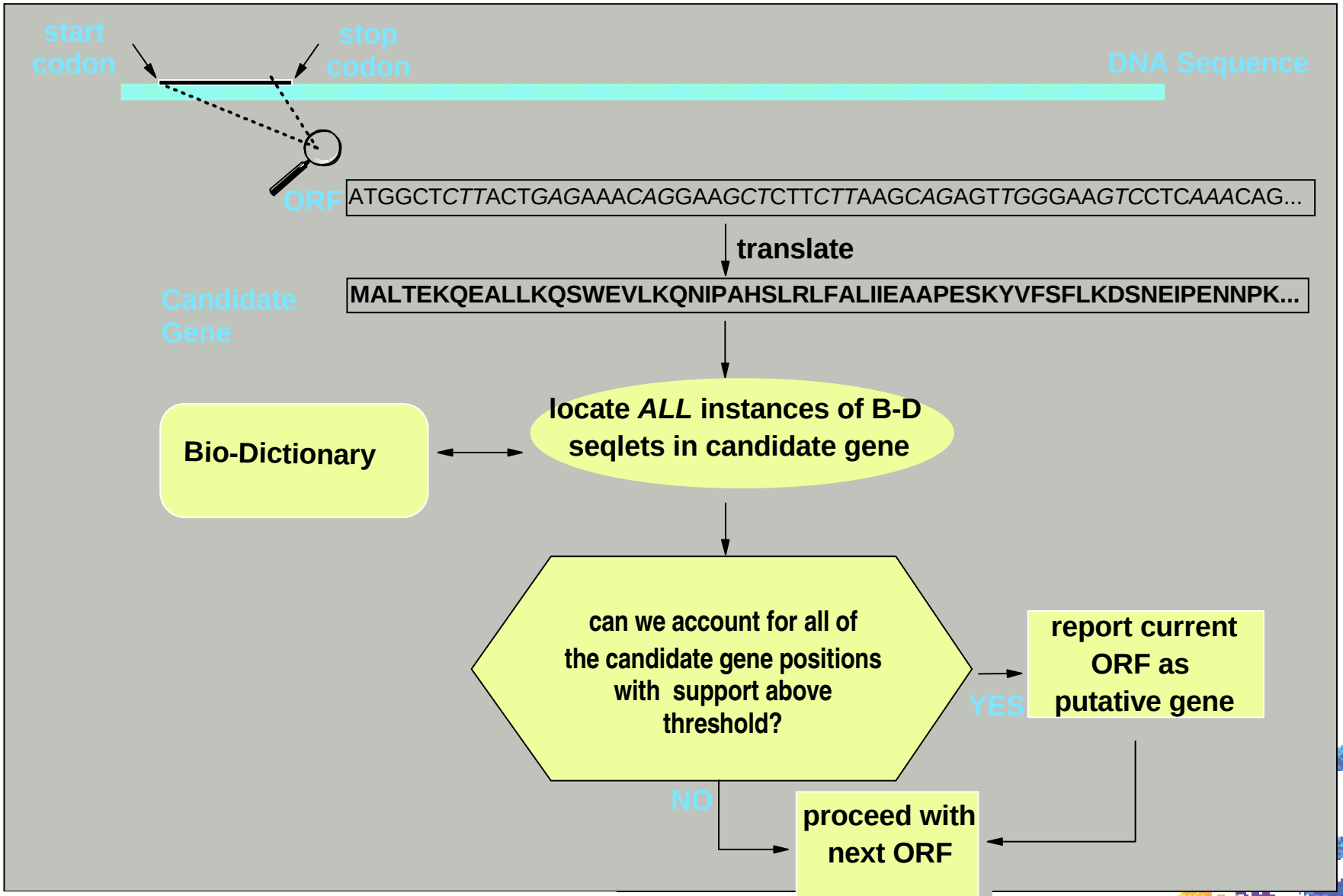
(Minimum, average, standard deviation) of this feature in related sequences = (9.00, 21.12, 1.75)

Benefits From This Approach

- can process very short inputs (6+ a.a.)
- no statistical filtering
- query components do not have to be in previously encountered order (solution to domain problem)
- can locate local/global similarities
- can locate 'narrow' & 'partial' attributes
- fast / exhaustive / objective
- result quality will further improve with database size



Prokaryotic Gene Discovery

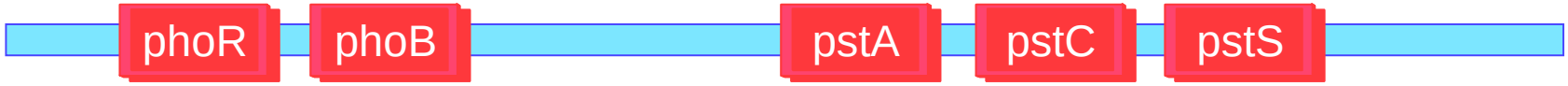


Shibuya, T. and I. Rigoutsos, (2002) Nucleic Acids Res. 30(12).

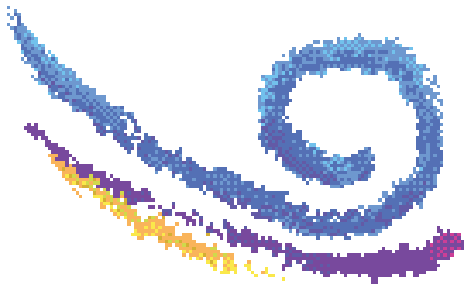
	Sensitivity & Specificity	Not Previously Reported Genes	Confirmed by FASTA	Confirmed by BLASTP	Confirmed by CD	Correct Start Site Prediction # of Genes	Correct Start Site Prediction Ratio of Genes
AF	95.3	113	52	54	34	1835	81.5
MJ	96.9	53	23	24	17	1476	89.9
MT	96.8	60	5	6	1	1499	83.9
PA	95.9	72	16	17	2	1306	77.7
AA	95.7	66	18	23	15	1279	88.3
BB	94.0	51	4	5	1	672	83.7
BS	96.4	147	36	42	10	2643	67.8
CJ	96.9	50	21	18	7	1375	87.2
CPc	97.0	32	23	23	6	838	83.0
CPa	93.9	68	15	15	6	888	85.2
CT	96.9	28	5	5	0	662	76.9
EC	97.0	128	45	45	16	3256	80.2
HI	96.8	54	45	44	27	1384	84.9
HP	96.0	63	35	35	18	1222	82.2
RP	96.9	26	9	11	6	688	85.7
SS	96.7	106	12	16	2	2423	80.5
TM	95.9	76	35	38	29	1300	73.9

Prokaryotic Gene Discovery (cont.)

H. influenzae



- Two operons involved in phosphate transport:
  - phoR + phoB form regulon
  - pstS/pstC/pstA/pstB form transport system

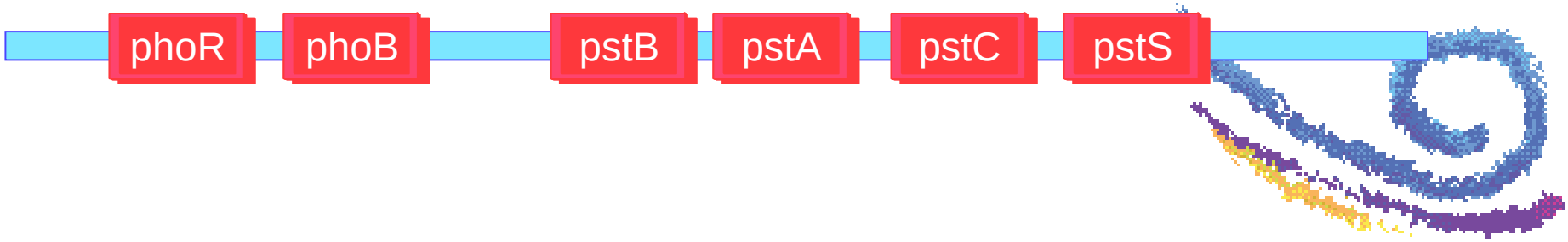


Prokaryotic Gene Discovery (cont.)

H. influenzae

our gene finder discovers pstB

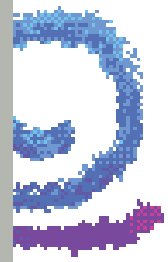
P. multocida



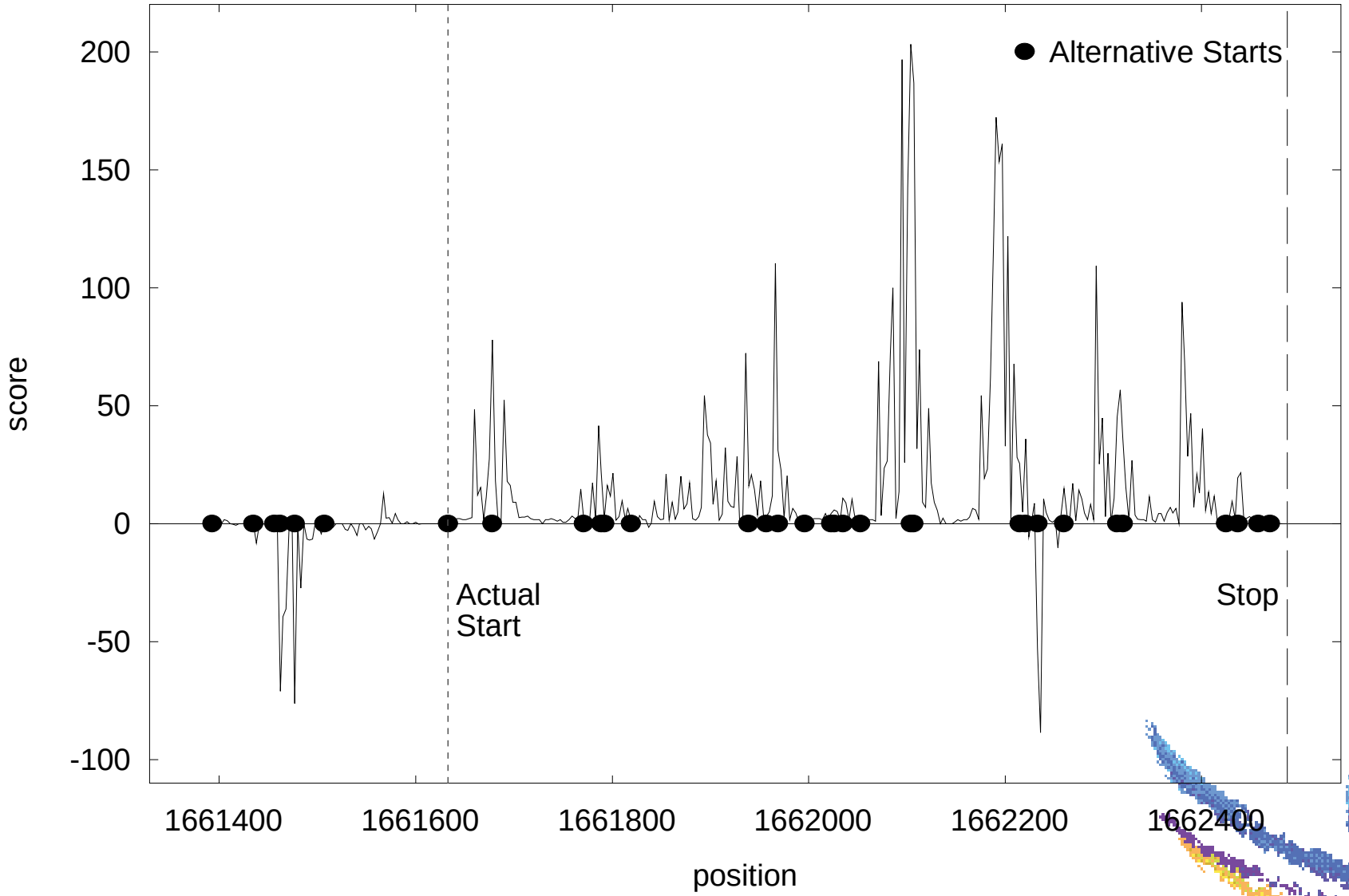
Q9PEZ3
from *X. fastidiosa*

Predicted Gene
from *A.fulgidus*

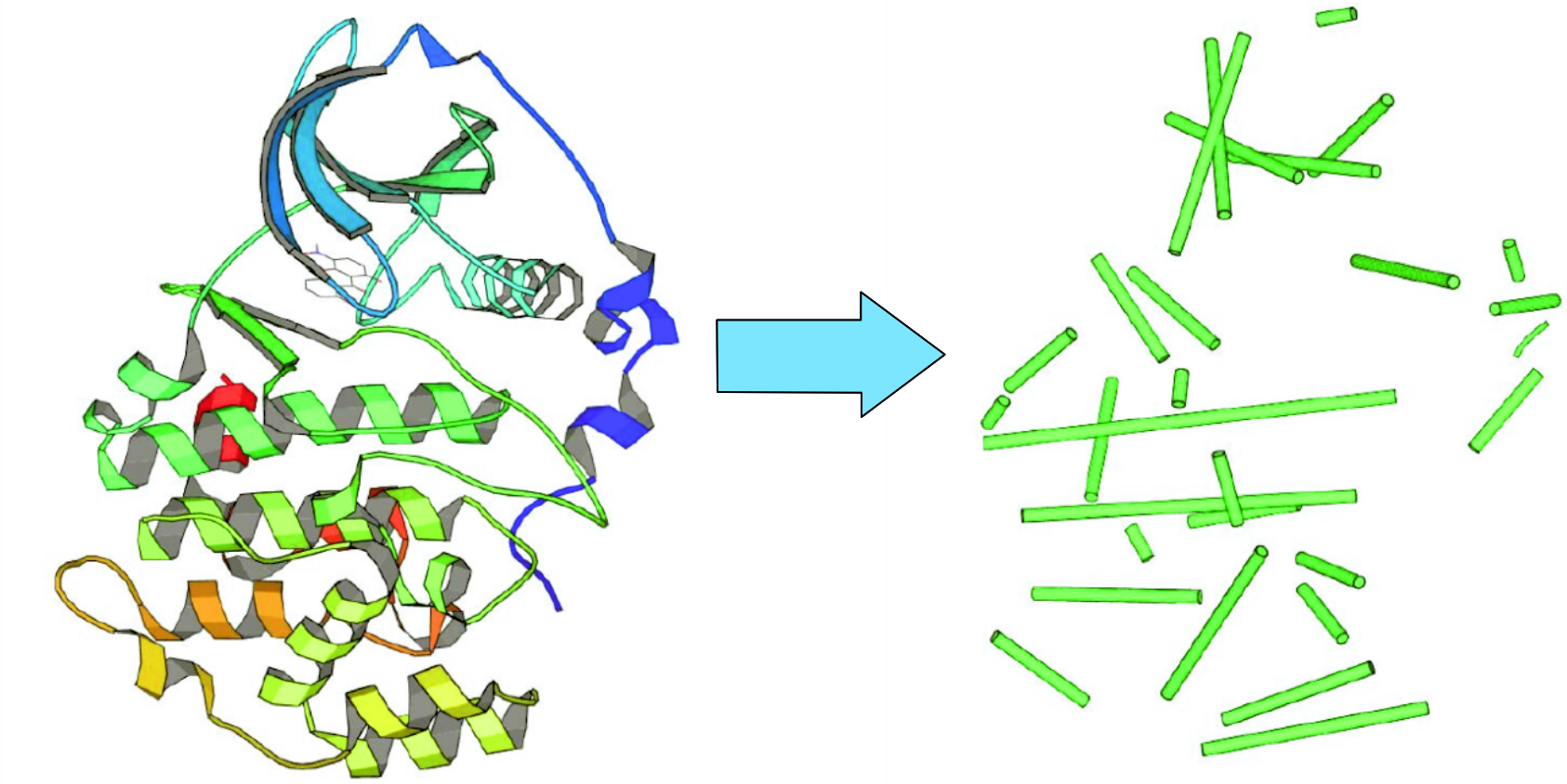
```
Q9PEZ3                          MSVDS-----VSRKH-DRIP-F-IRFLGVY-----LLFQWRRQCARLHRA
Pred_AFul_282847-284262_-       MDYCSGAALMVKKEIFEKIGGFDGRFKPAYYEDTDLCFSVRKMGYKVMYQ
                                *.  *     *.::  ::*  *  **  .*     * *. *:   ::

Q9PEZ3                          R----IALYSRRFG----QWIRLRAAVSADQSAPPSVAAVVSASATQTPT
Pred_AFul_282847-284262_-       PKSVIIHLEGATCGTDTSSGVKKFQEINRQKFYEKWKDTLLKHHYNPDPS
                                     * * .   *    . ::    :. ::      :::.   .  *:

Q9PEZ3                          N---TRC------ILLIDTVPPRPDRDSGSLRCHHLMHLMVCMGYKVVLH
Pred_AFul_282847-284262_-       NLFLARCRNGGKRILVIDHYVPTFDKDSGSYRMYNILKILIELGHCVTFI
                                *   :**     **:**   *  *:**** * :::::::: :*: *.:

Q9PEZ3                          CQERMPSAAEVMALRAIGVTTT--AVAGGFPSWLLTNPERYCAVVVCRYH
Pred_AFul_282847-284262_-       GDNLAKMEPYTSILQQIGIEVLYGPYTRSIEDYLKDHGKFFDIVILSRAP
                                 ::     . .  *: **: .   . : .: .:*  : : :  *::.*

Q9PEZ3                          LGLSWLPLLRAFVPDSLCILDTVDLHHLREQREAELRNLSGLRAAAAITR
Pred_AFul_282847-284262_-       IAEKHILAVRKYCSKAKIIFDTVDLHFLREMRRAELEKNDKVKELAEKLK
                                :. . :  :* : ..:  *:******.*** *.***.: . ::  *   :

Q9PEZ3                          RHELHAISCADLVWVVSPVERDYLARLLPQARVEVVPNMHDMVETIPPFV
Pred_AFul_282847-284262_-       NIELKLARLANLTLVVSPFEKELLLKEDPTLNVEVLSNIHEIKPPKNSFE
                                . **:    *:*. ****.*:: * :  *  .***:.*:*::  .  .*

Q9PEZ3                          SRHGFLFVGGSRHPPNVDAVRWLLSEIFPRIRERLPDAQLHLVGAGLAEA
Pred_AFul_282847-284262_-       NRKDIMFLGGFAHPPNIDAVKWFINDIFPKIKQQLPEVKFIIVGSNPPEE
                                .*:.::*:**  ****:***:*::.:***:*:::**:.:: :**:. .*

Q9PEZ3                          MQSQQMICPGVSFHGHVPEMAPLLHACRVSLAPLRFGAGVKGKISEALAY
Pred_AFul_282847-284262_-       IMS--LSSDDIIVTGYVRELEPYFESVRVFVSPLRYGAGVKGKIGEAMAH
                                : *  : . .: . *:* *: * :.: ** ::***:********.**:*:

Q9PEZ3                          GLPVVTTPEGAEGMYLRSGMDALISGDAEDLARQAVCAHEDLEVWQRLSD
Pred_AFul_282847-284262_-       GVPVVTTSIGGEGMGLIDGENALIADDPVEFAEKVVKLYTDKALWEKISM
                                *:*****. *.*** * .* :***:.*. ::*.:.*  : *  :*:::*

Q9PEZ3                          NGQQIIKQHFSLHSTRAALAVMLPH
Pred_AFul_282847-284262_-       RSIEHVKCNFSYNVAKNKIINII--
                                .. : :* :** : ::  :  ::
```

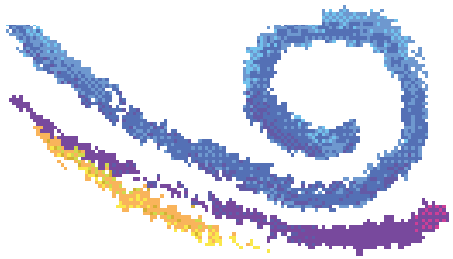


Prokaryotic Gene Discovery (cont.)

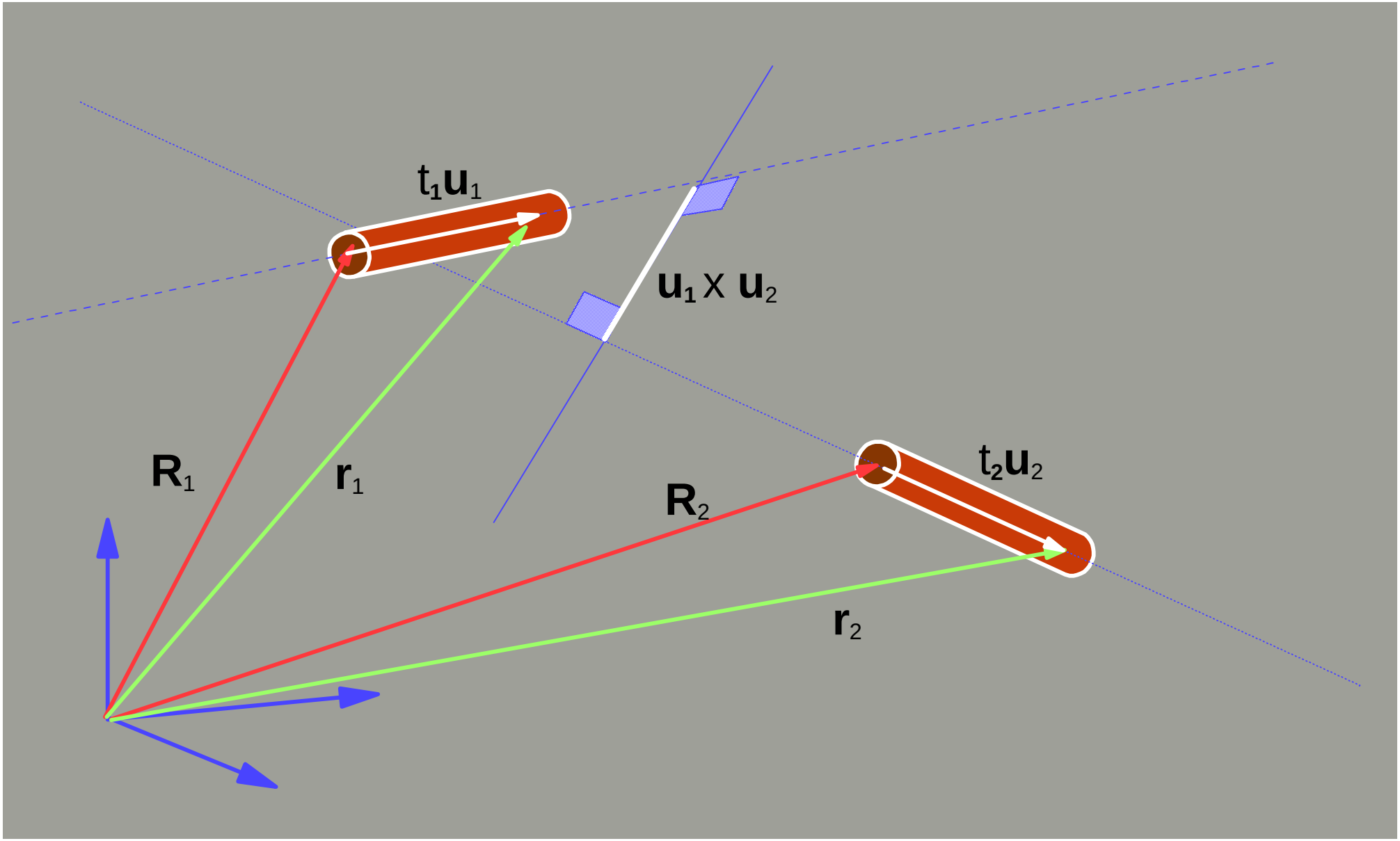


Biases in 3-dimensions

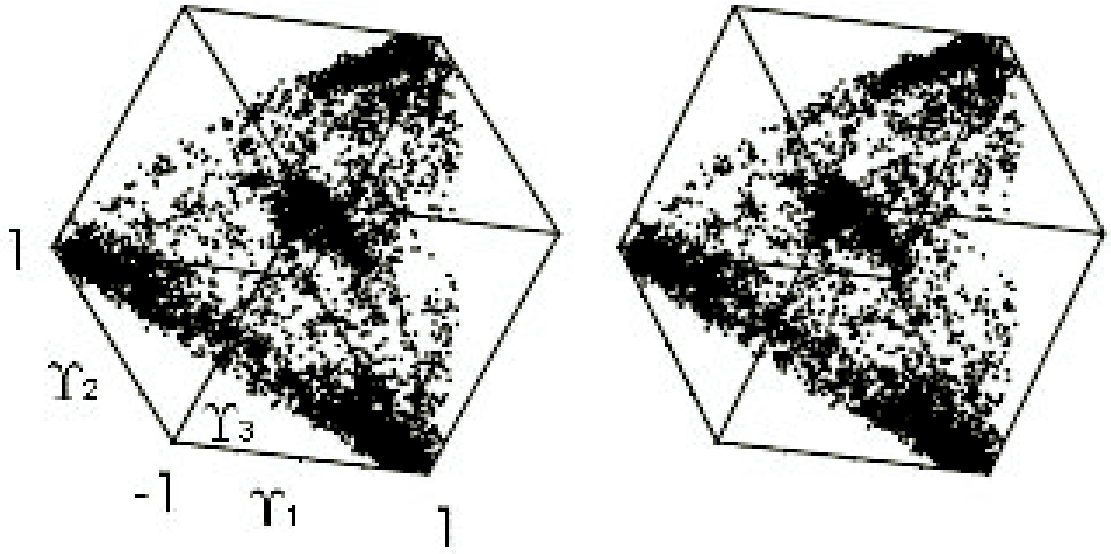




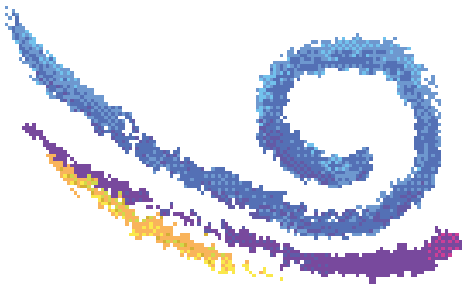
Identifying Interacting Secondary Structure



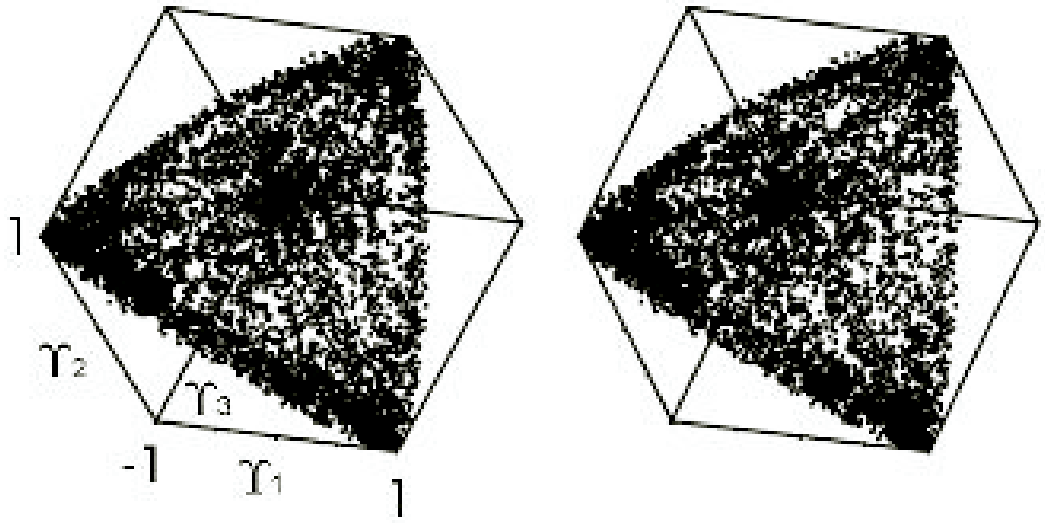
Biases in 3-dimensions (cont.)



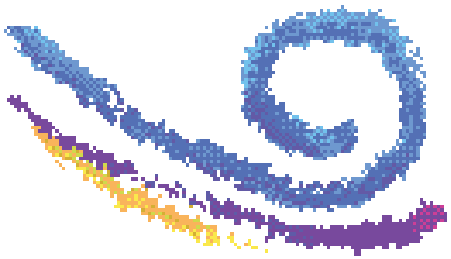
interactive



Biases in 3-dimensions (cont.)

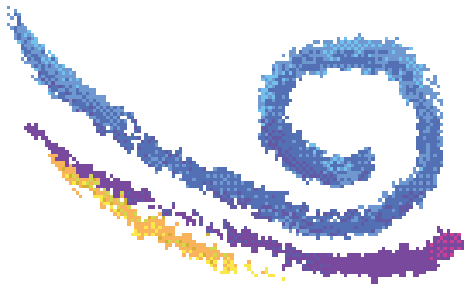


all



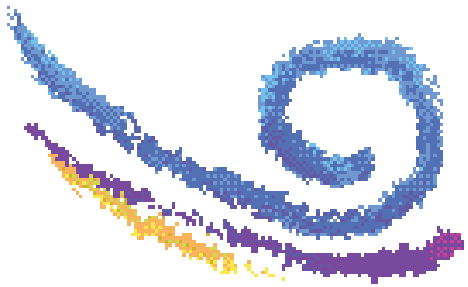
Biases in 3-dimensions (cont.)

- Previous studies:
  - focused on LOCAL interactions.
  - Folk Wisdom had been that non-interacting secondary structures are randomly oriented.

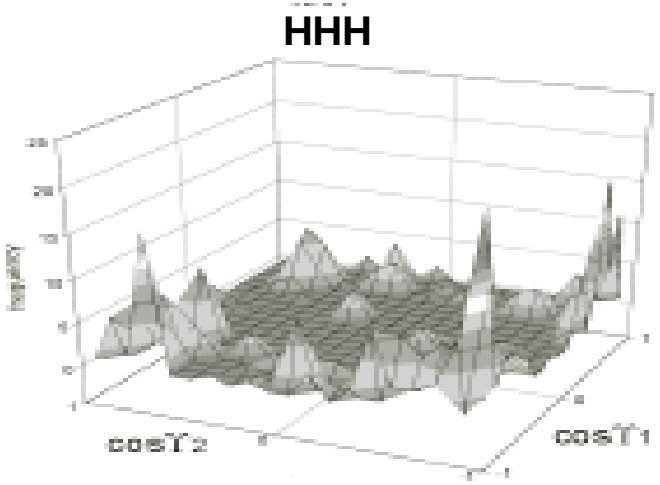


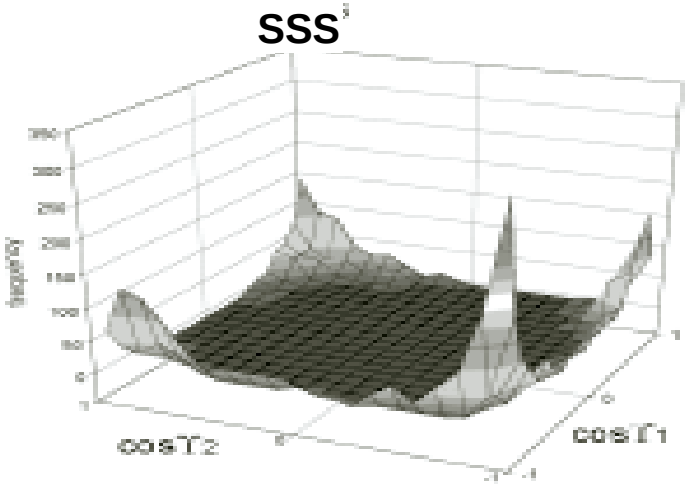
Biases in 3-dimensions (cont.)

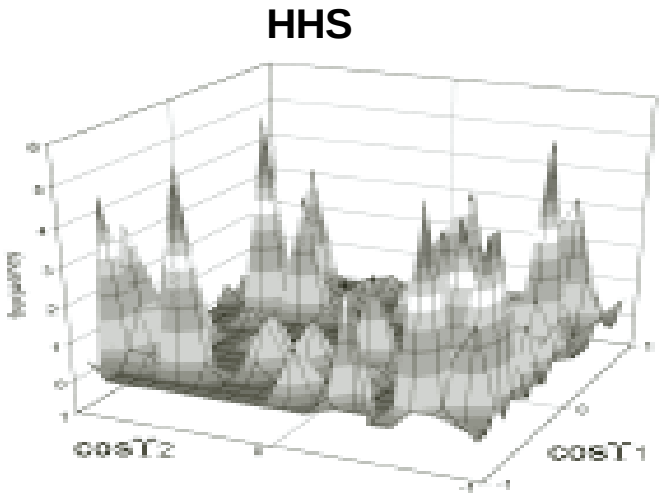
- Main Result:
  - we have shown that secondary structures exhibit a global packing bias (the secondary structures tend to be parallel or antiparallel) something that indicates "cooperativity"

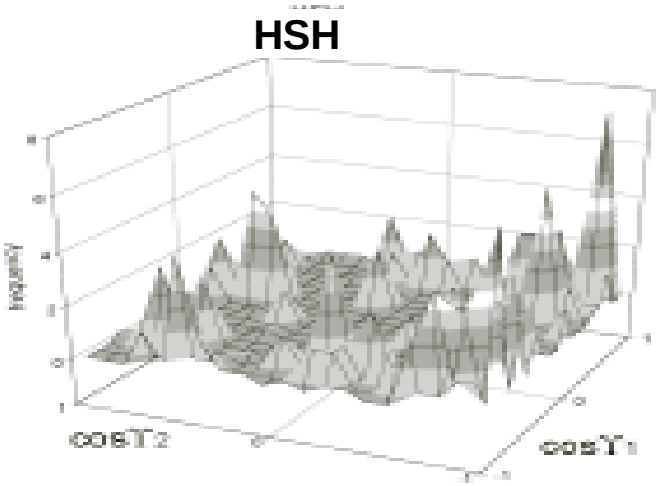


Marginal Distribution of Interacting Secondary Structure Axes









Marginal Distribution of Noninteracting Secondary Structure Axes