Analysis of Sequence Similarity

Plant Breeding 607

Cornell University

Spring ’00

Dave Schneider (instructor)
Mauricio la Rota (assistant)

November 13, 2000

Prerequisites and registration

- Prerequisites:
  - Basic biology
  - Basic genetics
  - Familiarity with computers.
- Permission of instructor required for registration.
- 1 credit, S-U only.

Staff

Instructor: Dave Schneider

Address:	632 Rhodes Hall
Email:	schneid@tc.cornell.edu
Telephone:	254-4510
Fax:	254-8888

Teaching assistant: Mauricio la Rota

Address:	622 Rhodes Hall
Email:	cml22@cornell.edu
Telephone:	255-0186

Schedule and venue

Lectures

Dates:	10/27/00 – 11/27/00
Times:	MWF 11:15-12:05
Location:	Morrison 342

Laboratory

Dates:	10/27/00 – 11/27/00
Times:	M 9:00-11:00, W 4:00-6:00, Th 4:00-6:00
Location:	Bradfield G-04

Assignments and grading

Weekly assignments will be distributed and collected on Mondays.

- Careful interpretation of computer laboratory exercises

- Emphasis on clear exposition of scientific ideas, not on “correctness.”

Letters and alphabets

1 Definition (Character) *A character is a symbol.*

Characters will typically be denoted by lower case Arabic letters, e.g. a, b, $\ldots$, z.

2 Definition (Alphabet) *An alphabet is a definite set of unique characters.*

Alphabets will typically be denoted by upper case Greek letters, e.g. $\mathbf{\Gamma}$, $\mathbf{\Theta}$, $\mathbf{\Sigma}$.

The size of an alphabet, $\mathbf{\Sigma}$, is denoted $|\mathbf{\Sigma}|$, and is assumed to be finite.

Examples of alphabets

Alphabets represent classes of physical objects, and characters represent particular instances of these objects.

- $\Sigma_{\mathsf{DNA}} = \{A, T, G, C\} \qquad \Sigma_{\mathsf{RNA}} = \{A, U, G, C\}$.
- $\Sigma_{\mathsf{protein}} = \left\{ \begin{array}{l} A, B, C, D, E, F, G, H, I, K, L, \\ M, N, P, Q, R, S, T, V, W, X, Y, Z \end{array} \right\}$.
- $\Sigma_{\mathsf{decimal}} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Strings and substrings

3 Definition (String) *A string over an alphabet Σ is a ordered set of characters selected from Σ.*

Strings will be denoted using lower case Greek characters, e.g. α, β, γ, …, ω. The length of α is just the number of characters in α and will be denoted $|\alpha|$. The empty string (i.e. the string with no characters) will always be denoted ϵ.

4 Definition (Substring) *A substring β of a string α is a subset of consecutive characters of α.*

By convention, ϵ is considered a substring of every string.

Sequences and subsequences

5 Definition (Sequence) *A sequence is an ordered list of characters selected from a definite alphabet.*

6 Definition (Subsequence) *A subsequence, τ, of a sequence σ is a subset of the characters of σ with order preserved.*

$$\sigma = (a_1, a_2, a_3, \ldots, a_N)$$

$$\tau = \left(a_i, a_j, a_k, \ldots, a_n\right).$$

where

$$1 \leq i < j < k < \ldots < n \leq N.$$

All substrings are subsequences, but subsequences with gaps are not substrings.

Syntax, grammar, and semantics

7 Definition (Syntatic content) *The abstract study of the arrangement of characters in strings or sequences over a definite alphabet.*

8 Definition (Grammar) *The study of the allowable arrangement of words in a language.*

9 Definition (Semantic content) *The meaning or interpretation of a string over a particular alphabet.*

Both of these concepts are essential in biology, and one must be careful to distinguish between them.

The Information Hierarchy

Knowledge This sequence codes for a cytochrome c that is expressed in brain tissue in the early embryo.

Information The sequence is . . .TATAACGTATTGC. . ..

Data The chromatogram generated by the sequencer is just a record of electrical signals generated by sensors.

A logical analysis of self-reproducing entities

In 1948, five years before the discovery of the structure of DNA by Watson and Crick, the famous mathematician John von Neumann discovered a very deep truth about the logical structure of any entity capable of self-reproduction. His central insight was that it is necessary to make an operational distinction between syntax and semantics, and these dual roles of sequences must be handled be separate infrastructures. This requirement arises from the need to avoid certain types of self-referential statements that are semantically ambiguous.

The following statement is true. The previous statement is false.

von Neumann’s recipe

A) Factory: A facility that collects raw materials and assembles them according to instruction streams supplied from B.

B) Duplicator: A facility that collects raw materials to duplicate instruction streams defined by D.

von Neumann's recipe, continued

C) Controller: A facility that coordinates the action of A and B. This controller must make sure that the D is properly translated by B into instructions suitable for A and also duplicated by B for the next "generation".

D) Instruction set: Complete instructions for ensuring that component C correctly coordinates the construction of a new copy of the entire system, A+B+C+D.

A logical view of biology

A) Ribosomes are universal translators that use instructions in the form of mRNA and consume aminoacyl tRNA to produce proteins. Note, both rRNA and protein components are encoded in DNA.

B) DNA polymerases to replicate DNA, and RNA polymerases to transcribe DNA into mRNA to serve as templates for protein synthesis used by ribosomes. All coded by DNA.

C) Gene regulation network and associated controlling molecules (repressors, promotors, etc.). Note, these molecules are encoded by DNA and produced by ribosomes.

D) The genetic materials, primarily DNA, secondarily RNA.

Sequence analysis in molecular biology

- Sequences analysis deals with the identity of objects, not energy, time, or other physical properties that are directly related to function. Therefore, sequence analysis is intrinsically syntactic and empirical in nature.

- Most of biology, including functional and structural genomics, is primarily semantic in nature.

- The primary intellectual hurdle is to properly interpret syntactical evidence as possible clues to semantic content.

The annotation problem

Annotation is the purported assignment of semantic content (i.e. function) to syntactic content (i.e. sequence).

- The lack of widely accepted controlled vocabularies, keywords, etc. make it difficult to find sequences with the desired annotation.
- Automated annotation methods are very widespread but frequently unreliable since they are essentially all based on syntactic analyses.
- Laboratory verification is an absolute necessity.

The stability problem

Grammatical correctness and semantic meaning can be greatly altered by small changes in syntax.

- Exchange of a proline for a glycine in a peptide chain.
- Mis-spelling in computer password.
- Errors in bank deposits or withdrawls.

Dave knows what he is talking about!

Dave doesn’t know what he is talking about!

Matching a 5S rRNA gene to dbEST

```
blastall -p tblastx -i gi6689418.seq -d blast-18-9-2000\est\est
TBLASTX 2.1.1 [Aug-8-2000]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= gi|6689418|emb|AJ245808.1|TNI245808 Tetraodon nigroviridis 5S rRNA gene
         (429 letters)

Database: blast-18-9-2000\est\est
           5,700,267 sequences; 2,262,334,554 total letters
```

Surprise!

```
                                                                   Score    E
Sequences producing significant alignments:                        (bits)  Value

gb|AI119137.1|AI119137 ue94d01.y1 Sugano mouse embryo mewa Mus m...    87  3e-016
gb|BE046547.1|BE046547 hn40c05.x1 NCI_CGAP_RDF2 Homo sapiens cDN...    84  3e-015
gb|AA472346.1|AA472346 vh05d01.r1 Soares_mammary_gland_NbMMG Mus...    82  8e-015
gb|AW491665.1|AW491665 UI-M-BH3-atq-h-07-0-UI.s1 NIH_BMAP_M_S4 M...    82  8e-015
gb|AW147957.1|AW147957 da01b05.x1 Xenopus laevis oocyte Xenopus ...    82  1e-014
gb|AW199201.1|AW199201 da17d08.x1 normalized Xenopus laevis gast...    80  4e-014
gb|AA534204.1|AA534204 nj21b07.s1 NCI_CGAP_AA1 Homo sapiens cDNA...    75  1e-012
gb|AW920107.1|AW920107 EST351515 Rat gene index, normalized rat,...    74  2e-012
gb|AA876181.1|AA876181 nx25c04.s1 NCI_CGAP_GC4 Homo sapiens cDNA...    74  2e-012
gb|AW839223.1|AW839223 CM0-LT0066-030300-264-h07 LT0066 Homo sap...    71  2e-011
gb|AI009130.1|AI009130 EST203581 Normalized rat embryo, Bento So...    62  8e-010
gb|AW058434.1|AW058434 wx20g11.x1 NCI_CGAP_Gas4 Homo sapiens cDN...    65  1e-009
gb|BE075398.1|BE075398 MR2-BT0589-230300-202-b12 BT0589 Homo sap...    63  5e-009
gb|H58310.1|H58310 yr25a04.r1 Soares fetal liver spleen 1NFLS Ho...    63  5e-009
dbj|AV599486.1|AV599486 AV599486 Bos taurus cartilage fetus Bos ...    61  2e-008
gb|AW200240.1|AW200240 da17d08.y1 normalized Xenopus laevis gast...    61  2e-008
gb|AA587509.1|AA587509 nn30a03.s1 NCI_CGAP_Gas1 Homo sapiens cDN...    58  1e-007
```

```
>gb|AI119137.1|AI119137 ue94d01.y1 Sugano mouse embryo mewa Mus musculus cDNA clone
           IMAGE:1498753 5’ similar to gb|K01594|RATRRA Rat 5S
           ribosomal RNA. gb|J01867|HUMRRA Human 5S (rRNA);
           gb:M13963 Mouse inhibitory G protein of adenylate
           cyclase, alpha chain (MOUSE);
          Length = 475

 Score = 82.6 bits (174), Expect = 6e-015
 Identities = 36/47 (76%), Positives = 39/47 (82%)
 Frame = +2 / -3

Query: 266 KTHSNGMKKLTAPGIPRRSPIQVLSRPDPA*LPRSDEIGRSQGGMAV 406
           +  S+   K  APGIPRRSPIQVL+RPDPA*LPRSDEIGR QGGMAV
Sbjct: 143 QVRSSERLKPAAPGIPRRSPIQVLTRPDPA*LPRSDEIGRVQGGMAV 3

 Score = 84.9 bits (179), Expect = 1e-015
 Identities = 35/40 (87%), Positives = 36/40 (89%)
 Frame = +1 / -2

Query: 289 KAYSTWYSQAVSHPSTKQARPCLASEIRRDRAFSGWYGRK 408
           KA ST YSQAVSHPST QARPCLASEIRRDRA SGWYGR+
Sbjct: 120 KACSTRYSQAVSHPSTNQARPCLASEIRRDRARSGWYGRR 1
```

```
>gb|BE046547.1|BE046547 hn40c05.x1 NCI_CGAP_RDF2 Homo sapiens cDNA clone
           IMAGE:3024584 3’ similar to gb|K01594|RATRRA Rat 5S ribosomal
           RNA. gb|J01867|HUMRRA Human 5S (rRNA);
          Length = 230

 Score = 82.2 bits (173), Expect = 8e-015
 Identities = 36/53 (67%), Positives = 39/53 (72%)
 Frame = +2 / +3

Query: 245 FTRQAGQKTHSNGMKKLTAPGIPRRSPIQVLSRPDPA*LPRSDEIGRSQGGMA 403
           F+    Q        + TAPGIPRRSPIQVL+RPDPA*LPRSDEIGR QGGMA
Sbjct: 66  FSHNPTQAERYGSAAEPTAPGIPRRSPIQVLTRPDPA*LPRSDEIGRVQGGMA 224

 Score = 82.6 bits (174), Expect = 6e-015
 Identities = 35/46 (76%), Positives = 37/46 (80%)
 Frame = +3 / +1

Query: 264 RKPTQMA*KSLQHLVFPGGLPSKY*AGPTLLSFRDQTRSGVLRVVW 401
           R+    A +SLQH VFPGGLPSKY* GPTLLSFRDQTRSG  RVVW
Sbjct: 85  RRNDTAAPRSLQHPVFPGGLPSKY*PGPTLLSFRDQTRSGAFRVVW 222
```

```
>gb|AW147957.1|AW147957 da01b05.x1 Xenopus laevis oocyte Xenopus laevis
          cDNA clone XENOPUS_SOURCE_ID:xlnoc001a10 3’ similar to
          gb|K02695|XELRRA X.laevis 5S ribosomal RNA.
          gb|M21176|XELRRAOLA (rRNA);
         Length = 489

 Score = 62.5 bits (130), Expect = 7e-009
 Identities = 28/39 (71%), Positives = 30/39 (76%)
 Frame = +2 / +2

Query: 290 KLTAPGIPRRSPIQVLSRPDPA*LPRSDEIGRSQGGMAV 406
           K T PGIPRRSPIQVL+RPD   L RSDEI   QGG+AV
Sbjct: 5   KPTTPGIPRRSPIQVLTRPDSVSLLRSDEIRHFQGGVAV 121

 Score = 77.6 bits (163), Expect = 2e-013
 Identities = 32/40 (80%), Positives = 34/40 (85%)
 Frame = +3 / +3

Query: 291 SLQHLVFPGGLPSKY*AGPTLLSFRDQTRSGVLRVVWP*A 410
           SL+HLVFPGGLPS+Y* GPTL  F DQTRSG  RVVWP*A
Sbjct: 6   SLRHLVFPGGLPSRY*PGPTLYRF*DQTRSGTFRVVWP*A 125
```

```
>gb|AA534204.1|AA534204 nj21b07.s1 NCI_CGAP_AA1 Homo sapiens cDNA
clone IMAGE:993109 3’ similar to contains Alu repetitive
        element; contains element PTR7 repetitive element ;
          Length = 607

 Score = 74.8 bits (157), Expect = 1e-012
 Identities = 29/43 (67%), Positives = 34/43 (78%)
 Frame = +1 / +3

Query: 274 LKWHEKAYSTWYSQAVSHPSTKQARPCLASEIRRDRAFSGWYG 402
           LK   K YSTW SQ +SHPST QAR CLAS+IR+D++ SGWYG
Sbjct: 303 LKNKFKTYSTWNSQPISHPSTNQARTCLASKIRKDQSHSGWYG 431

 Score = 55.1 bits (114), Expect = 1e-006
 Identities = 24/37 (64%), Positives = 29/37 (77%)
 Frame = +2 / +1

Query: 290 KLTAPGIPRRSPIQVLSRPDPA*LPRSDEIGRSQGGM 400
           K  APGIP +S IQVL+RP+PA* PRS++I   QGGM
Sbjct: 319 KPIAPGIPSQSLIQVLTRPEPA*PPRSEKISHIQGGM 429
```

```
>gb|AW920107.1|AW920107 EST351515 Rat gene index, normalized rat,
norvegicus, Bento Soares Rattus norvegicus cDNA clone
RGIGS51 5’ end
          Length = 601

 Score = 19.8 bits (37), Expect(2) = 6e-008
 Identities = 8/13 (61%), Positives = 9/13 (68%)
 Frame = +2 / -2

Query: 287 KKLTAPGIPRRSP 325
           +K TAPGIP   P
Sbjct: 132 QKPTAPGIPGGLP 94

 Score = 59.3 bits (123), Expect(2) = 6e-008
 Identities = 26/34 (76%), Positives = 27/34 (78%)
 Frame = +2 / -1

Query: 311 PRRSPIQVLSRPDPA*LPRSDEIGRSQGGMAVSA 412
           PR SPI VLS PDPA*LPRSDEIGR  G MAV +
Sbjct: 109 PRWSPIHVLSMPDPA*LPRSDEIGRVPGSMAVGS 8
```

```
>gb|AW058434.1|AW058434 wx20g11.x1 NCI_CGAP_Gas4 Homo sapiens
cDNA clone IMAGE:2544260 3' similar to contains element
A3R repetitive element ;
          Length = 402

 Score = 60.6 bits (126), Expect = 2e-008
 Identities = 27/45 (60%), Positives = 31/45 (68%)
 Frame = +2 / +2

Query: 272 HSNGMKKLTAPGIPRRSPIQVLSRPDPA*LPRSDEIGRSQGGMAV 406
           H   + K     I   SPIQV++RPDPA*LPRSDEI R QGGMA+
Sbjct: 41  HPGSISKKKKKEIVLSSPIQVVTRPDPA*LPRSDEIRRVQGGMAI 175

 Score = 64.8 bits (135), Expect = 1e-009
 Identities = 27/43 (62%), Positives = 30/43 (68%)
 Frame = +1 / +1

Query: 283 HEKAYSTWYSQAVSHPSTKQARPCLASEIRRDRAFSGWYGRKR 411
           H K      S   SHPS+ QARPCLASEIRRD+A SGWYG +R
Sbjct: 52  HLKKKKKGNSFKFSHPSSNQARPCLASEIRRDQARSGWYGHRR 180
```

```
>gb|AA343856.1|AA343856 EST49698 Gall bladder I Homo sapiens
cDNA 5’ end similar to serum amyloid A2, beta
          Length = 239

 Score = 49.2 bits (101), Expect = 7e-005
 Identities = 23/39 (58%), Positives = 25/39 (63%)
 Frame = +2 / -3

Query: 290 KLTAPGIPRRSPIQVLSRPDPA*LPRSDEIGRSQGGMAV 406
           K  APGIPR SPI   +RPDPA L R +EI     GMAV
Sbjct: 237 KPLAPGIPRCSPIPSTTRPDPAYLSRXEEIRHLXNGMAV 121

 Score = 56.5 bits (117), Expect = 4e-007
 Identities = 22/38 (57%), Positives = 24/38 (62%)
 Frame = +1 / -2

Query: 289 KAYSTWYSQAVSHPSTKQARPCLASEIRRDRAFSGWYG 402
           KA+STWYSQ  SHP   QARPCL    R D+    WYG
Sbjct: 238 KAFSTWYSQVFSHPKYYQARPCLPL*XRGDKTPXEWYG 125
```

```
  Database: blast-18-9-2000\est\est
    Posted date:  Sep 18, 2000  5:26 AM
  Number of letters in database: 2,262,334,554
  Number of sequences in database:  5,700,267

   Lambda     K      H
   0.318    0.135    0.401

Matrix: BLOSUM62
Number of Hits to DB: -1492598922
Number of Sequences: 5700267
Number of extensions: 40433766
Number of successful extensions: 882031
Number of sequences better than 10.0: 251
length of query: 143
length of database: 754,111,518
effective HSP length: 54
effective length of query: 88
effective length of database: 446,297,100
effective search space: 39274144800
effective search space used: 39274144800
frameshift window, decay const: 50,  0.1
T: 13
A: 40
```

Questions

- Is the query sequence really a 5S rRNA gene?
- Can rRNA genes contaminate EST experiments?
- Are there real proteins with subsequences that look like a translation of 5S rRNA?
- Is there a bug in BLAST?

Lessons

- An low E-value, even as low as 10^{-16}, does not guarantee biological significance.
- One can observe similarities, but cannot make causal connections.
- Assume all annotations are incorrect until proven otherwise by careful laboratory experimentation.

Matching 5S rRNA gene to random pseudo-ESTs

Generate random set of pseudo-ESTs

- Assume $P(A) = P(T) = P(G) = P(C) = 1/4$.
- Generate 2×10^5 strings with lengths sampled from a normal distribution with mean 350 and standard deviation of 50. Impose minimum length of 75.
- Run formatdb then blastn and see what happens . . .

BLASTN results

```
blastall -p blastn -d rnd -i ..\Exercises\5S_rna\gi6689418.seq -e 1.0
BLASTN 2.1.1 [Aug-8-2000]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= gi|6689418|emb|AJ245808.1|TNI245808 Tetraodon nigroviridis 5S
rRNA gene
         (429 letters)

Database: random
           200,000 sequences; 87,384,782 total letters
```

```
                                                                 Score    E
Sequences producing significant alignments:                      (bits)  Value

gnl|random|lrnd00151919                                             40  0.029
gnl|random|lrnd00115753                                             38  0.11
gnl|random|lrnd00089180                                             38  0.11
gnl|random|lrnd00046616                                             38  0.11
gnl|random|lrnd00178519                                             36  0.45
gnl|random|lrnd00156625                                             36  0.45
gnl|random|lrnd00137336                                             36  0.45
gnl|random|lrnd00109313                                             36  0.45
gnl|random|lrnd00003849                                             36  0.45

>gnl|random|lrnd00151919
          Length = 419

 Score = 40.1 bits (20), Expect = 0.029
 Identities = 20/20 (100%)
 Strand = Plus / Minus

Query: 118 tcatctcagcacatcattcc 137
           ||||||||||||||||||||
Sbjct: 394 tcatctcagcacatcattcc 375
```

More lessons

Garbage in, garbage out.

Anonymous

The purpose of computing is insight, not numbers.

R. Hamming

The purpose of sequence analysis is insight, not answers.

D. Schneider

The limits of intuition in modern biology

- You are now facing an flood of noisy data on an unprecedented scale.
- Your training has not equipped you (or anyone else) to interpret data of this type by intuition.
- Quantitative methods are an absolute requirement.

Hangman

_ _ _ _ _ _ _ _ _ _ _   _ _ _ _ _ _

Syntactic-statistical model of English text

- B. Hayes. A progress report on the fine art of turning literature into drivel, *Sci. Am.*, 249(5):16, 1983.
- R. W. Lucky. *Silicon Dreams: Information, Man and Machine*, St. Martin's Press, New York, 1989.

Statistical modeling of English text

- Compute of each letter separately, then doublets, triplets, quadruplets, quintuplets, sextuplets,. . ., in a given text.

- Convert frequencies to probabilities of individual letters, and conditional probabilities for substrings.

- Use a random number generator to generate a stream of characters from the conditional probability distributions.

First-Order Correlations

Tdory d neAeeeko,hs wieadad ittid eIa c i lodhgin un a a svmb i ee’ kwrdmn.

Clearly a monkey at a typewriter...

Second-Order Correlations

Le hoin. whan theoaromies out thengachilathedrid be we frergied ate k y wee ’ e the sle!

Perhaps it is a Welsh monkey...

Third-Order Correlations

’Weed. Thed to dre you and a dennie. A le men eark yous, the sle nown ithe haved saindy. If - it to to it dre to gre. I wall much. ’Give th pal yould the it going, youldn’t thave away, jostove mouble so goink steace, ’If take we’re do mennie.

- Capitalization is correct. Contractions are correct. Why?
- Quotation marks are not balanced. Why?

Fourth-Order Correlations

I can light, ’George tried in you and fire.’ ’Nothen it and I want yourse, George some other ther. There’s if his hand rolledad ther hisky, ’I little amonely we’re we’re with him the rain.

- Many short words are recognizable.
- Capitalization is still correct.
- Quotation marks are still not balanced.

Fifth-Order Correlations

’I...I’m not running.’ The ranch, work on the time. Do you because you get somethings spready told you just him by heat to coloured rabbits. That’s going grew it’s like a whisky place.

- Most words are correctly spelled.
- Quotation marks are still not balanced.
- Grammar is erratic.

Sixth-Order Correlations

Million mice because it two me we’ll sit by the future. We’ll steal it. ’Aren’t got it. ’About the fire slowly hand. ’I want, George,’ he asked nervoulsly: ’That’s fine. Say it too hard too forget other.

- It won’t win the Nobel Prize for Literature, but it’s not bad for a computer.
- Problems with balanced quotation marks and grammar remain. Why?

Strengths and Weaknesses of the Statistical Approach

Statistical methods can model syntax, but not grammar or semantics.

Quantitative analytical methods are available:

- Simple mathematics
- Elegant mechanism to incorporate “intuition”
- Widely useful in practice

A brief review of statistics

- Bounds

$$0 \leq P(x) \leq 1$$

- Sum rules

$$\sum_x P(x) = 1$$

$$P(x) + P(\overline{x}) = 1$$

$$\sum_y P(x|y)P(y) = P(x)$$

Bayes rule

$$P(x|y)P(y) = P(y|x)P(x)$$

Provides a mechanism to compare observed probabilities with *a priori* probabilities constructed from an particular model.

Information

10 Definition (Information content of events) *The information content in the occurence of an random event X is*

$$I_X = -\log_2\left(Pr\{x = X\}\right) = \log_2\left(1/Pr\{x = X\}\right).$$

One unit of information is called a “bit”.

- High probability implies low information content (redundant).
- Low probability implies high information content.

Properties

- Non-negative: $I_X \geq 0$ for all X.
- Monotonic: if $Pr(x = Y) > Pr(x = Z)$ then $I_Y < I_Z$.
- Probabilistic: numerical values dependent on the structure of the statistical model.

Reasonableness

Progress in science is the result of either:

- Observing and classifying events that have not been previously recorded.
- Providing predictive theories for previously unexplained or unpredictable phenomenon.

Both of these result in changes to the expected probabilities for the known set of possible experimental outcomes.

Examples

- If one is reading English text, then a “q” will certainly be followed by a “u”. Thus, one could omit the “u” with introducing ambiguity.

- If you are dialing a phone number, each correct digit incrementally increases the probability of dialing the desired number.

Entropy

11 Definition (Entropy of a probability distribution) *The entropy of a distribution of a random variable $X \sim p(x)$ is defined as*

$$H(p) = -\sum_x p(x) \log_2 [p(x)]$$

- Maximal when all events have equal probability.
- Related to redundancy or "compressibility".
- Does not depend on the nature of the events themselves, only on the distribution itself.

Information and entropy

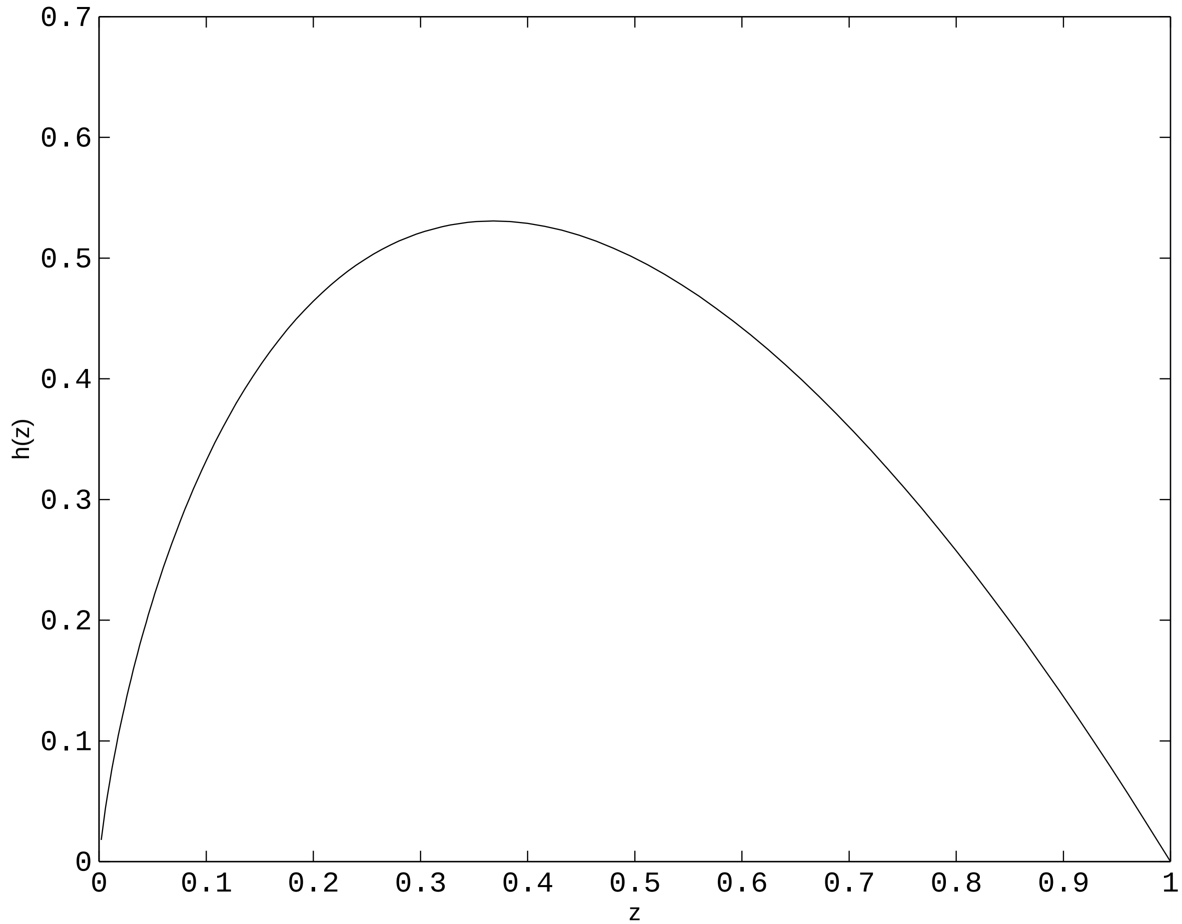
- Entropy is the expected value of the information,

$$H(X) = \sum_x p(x) I_x = \langle I_x \rangle$$

.

- Entropy is maximized when all events carry the same information. For discrete distributions, this means $p(x) = 1/N$ for all x and

$$H(X) < -\sum_{i=1}^{N} (1/N) \log_2(1/N) = \log_2(N) = H_{\mathrm{max}}.$$



Plot of $h(z) = -z \log_2(z)$ for $0 \leq z \leq 1$.

Joint entropy

12 Definition (Joint entropy) *The entropy of a joint distribution* $(X, Y) \sim p(x, y)$ *is defined as*

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log_2 [p(x, y)] .$$

As before, $H \geq 0$.

Joint entropy of statistically independent variables

If X and Y are statistically independent events then $(X, Y) \sim p(x, y) = q(x)r(y)$ so

$$H = -\sum_x \sum_y q(x)r(y) \log_2 [q(x)r(y)] .$$

This can be simplified using the properties of logarithms,

$$H = -\sum_x \sum_y q(x)r(y) \{\log_2 [q(x)] + \log_2 [r(y)]\}$$

Since the sums over x and y are independent, the order can be interchanged

$$H(X,Y) = -\left[\sum_y r(y)\right]\left[\sum_x q(x)\log_2(q(x))\right] - \left[\sum_x q(x)\right]\left[\sum_y r(y)\log_2(r(y))\right]$$

. Therefore, since $\sum_z p(z) = 1$,

$$H(X,Y) = -\sum_x q(x)\log_2[q(x)] - \sum_y r(y)\log_2[q(x)].$$

This is just the sum of the entropies of the two distributions, q and r,

$$H(X,Y) = H(X) + H(Y).$$

Conditional entropy

13 Definition (Conditional entropy) *If $(X, Y) \sim p(x, y)$, then conditional entropy is defined as*

$$H(X|Y) = -\sum_{x}\sum_{y} p(x, y) \log_2 [p(x|y)]$$

where $p(x|y)$ is the conditional probability distribution for event X given the data Y.

Note:

- Non-negativity: $H(X|Y) \geq 0$.
- Asymmetry: $H(X|Y) \neq H(Y|X)$.

Mutual information

14 Definition (Mutual information) *If $(X, Y) \sim p(x, y)$ with marginal distributions $X \sim q(x)$ and $Y \sim r(y)$, then the mutual information is defined as*

$$I(X;Y) = -\sum_x \sum_y p(x,y) \log_2 \left[\frac{p(x,y)}{q(x)r(y)} \right].$$

Conditioning reduced entropy, $H(X|Y) \leq H(X)$.

$$I(X;Y) = I(Y;X) = H(X) + H(Y) - H(X,Y)$$

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Degree of Bias

15 Definition (Position dependent bias) *Given a set of sequences over an alphabet $\boldsymbol{\Sigma}$ with position dependent probabilities $X \sim p_i(x)$, the bias at position i is defined as*

$$D_i = \log_2 |\boldsymbol{\Sigma}| - \sum_{X \in \Sigma} p_i(x) \log_2 [p_i(x)] \,.$$

Note, $\max D_i = \log_2 |\boldsymbol{\Sigma}|$ occurs if there is complete agreement in the "consensus" sequence.

Alignments, scoring, and substitution matrices

16 Definition (Substitution matrix) *A substitution matrix is a table of scores, $s_{x\,y}$, for the alignment of x and y in the alignment of two strings.*

For similarity searching,

- Close similarity $\leftrightarrow$ positive scores
- Indifference $\leftrightarrow$ zero scores
- Dissimilarity $\leftrightarrow$ negative scores

Biological relevance

- Large penalty for mismatches relative to rewards for matches leads to short, strong alignments

- Small mismatch penalties lead to long, weak alignments.

The mechanics of scoring

17 Definition (Nominal score) *The nominal score for the alignment of two sequences, α and β, is given by*

$$S(\alpha, \beta) = \sum_{x \in \alpha} s_{xy}$$

where (x, y) is the pairwise alignment of letters.

18 Definition (Normalized (bit) score) *The normalized score is defined by*

$$S' = \frac{\lambda S - \ln K}{\ln 2} = \lambda' S - \log_2 K$$

where λ and K are parameters selected by statistical simulations.

The parameter λ simply sets the overall scaling of scores.

Expected number of alignments

19 Definition (E value)

$$E = nm\, 2^{-S'}$$

is the expected number of alignments with bit score S' expected for a string of length n in a database of length m.

E values are additive for “statistically independent” databases.

A statistical model of alignments

Assume:

- *A priori* marginal distributions: Character i occurs with probability $P(i)$.
- Random strings should not produce useful alignments:

$$\sum_{i,j} P(i)P(j)s_{ij} < 0.$$

Target frequencies and entropies

20 Definition (Target frequencies)

$$q_{x\,y} = P(x)P(y)\,\text{exp}\left[\lambda_u s_{x\,y}\right]$$

where λ_u is selected such that

$$\sum_{x,y} q_{x\,y} = 1.$$

21 Definition (Relative entropy of substitution matrices)

$$H(s) = \sum_{x,y} q_{x\,y} s_{x\,y}$$

is the entropy of the substitution matrix s.

Is this really a relative entropy?

$$H(s) = \left(\frac{1}{\lambda_u}\right) \sum_{x,y} q_{x\,y} \log_2 \left[\frac{q_{x\,y}}{P(x)P(y)}\right].$$

This is the average information in each pairwise character alignment.

Places severe limits of scientific inference based on “distant” evolutionary relationships.

BLOSUM62

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
```

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

General characteristics of substitution matrices

Matrix type	Closely related	Distantly related
BLOSUM	Higher numbers	Lower numbers
PAM	Lower numbers	Higher numbers

BLOSUM numbers are related to percentage identity in the alignments from which substitution statistics were derived.

PAM numbers are related to a measure of divergence with a specific evolutionary model.

Characteristics of BLOSUM matrices

Matrix	Entropy	Expected score
BLOSUM30	0.1424	-0.1074
BLOSUM35	0.2111	-0.1550
BLOSUM40	0.2851	-0.2090
BLOSUM45	0.3795	-0.2789
BLOSUM50	0.4808	-0.3573
BLOSUM55	0.5637	-0.4179
BLOSUM60	0.6603	-0.4917
BLOSUM62	0.6979	-0.5209
BLOSUM65	0.7576	-0.5675
BLOSUM70	0.8391	-0.6313
BLOSUM75	0.9077	-0.6845
BLOSUM80	0.9868	-0.7442
BLOSUM85	1.0805	-0.8153
BLOSUM90	1.1806	-0.8887
BLOSUMN	1.5172	-1.1484

Characteristics of PAM matrices

Matrix	Entropy	Expected score
PAM10	3.43	-8.27
PAM20	2.95	-6.18
PAM30	2.57	-5.06
PAM40	2.26	-4.27
PAM50	2.00	-3.70
PAM60	1.79	-3.21
PAM70	1.60	-2.77
PAM80	1.44	-2.55
PAM90	1.30	-2.26
PAM100	1.18	-1.99
PAM120	0.979	-1.64
PAM140	0.820	-1.35
PAM160	0.694	-1.14
PAM180	0.591	-1.51
PAM200	0.507	-1.23
PAM250	0.354	-0.844
PAM300	0.254	-0.835
PAM350	0.186	-0.701

PAM-BLOSUM comparisons

- PAM matrices have lower expected scores for the BLOSUM matrices with the same entropy.
- BLOSUM matrices “generally perform better” than PAM matrices

What, if anything, does this mean in scientific terms???

Substitution matrices for protein-protein searching recommended by NCBI

Query length	Substitution matrix	Gap costs
< 35	PAM30	(9,1)
$35 - 50$	PAM70	(10,1)
$50 - 85$	BLOSUM80	(10,1)
> 85	BLOSUM62	(11,1)

Short sequences cannot have participate in long, weak alignments.

Gap costs must be tailored to the substitution matrix.

See `http://www.ncbi.nlm.nih.gov/BLAST/matrix_info.html`.

How should one proceed in practice?

- ESTs against ESTs, genomic sequence and proteins
- Full length cDNAs against genomic sequence and proteins
- Genomic sequence against proteins

Computers and algorithms

Computers are not intelligent in that their operation is completely limited by an externally supplied set of instruction.

These instructions must be supplied in a form of logical operations and must be be self-consistent and goal-directed.

22 Definition (Algorithm) *An algorithm is an finite set of instructions which can be executed by a computer to produce an output, possibly requiring additional input data.*

Classes of problems

- Exact and approximate matching of substrings
- Keyword searches (matches to member of a sets of strings)
- Regular languages and matching of regular expressions
- Exact and approximate matching of subsequences

Exact matching

23 Definition (Exact match of two strings) *Two strings*

$$\alpha = (a_1, a_2, \ldots, a_{|\alpha|})$$

and

$$\beta = (b_1, b_2, \ldots, b_{|\beta|})$$

match exactly if, and only if,

$$|\alpha| = |\beta|$$

and

$$a_i = b_i$$

for $1 \leq i \leq |\beta|$.

Exact matches will be denoted $\alpha = \beta$.

General scheme for finding exact matches of α in β

Preprocess strings to determine window shifts ;
Align the left end of the window with the left end of β
while (the right end of the window has not gone past the right end of β
    attempt match of α with the substring of β in the window
    if (found) report success ;
    shift window ;
endwhile

The naive approach

Query: she
Subject: ushers

u	s	h	e	r	s
s	h	e			
	s	h	e		
		s	h	e	
			s	h	e

Brute-force algorithm

```
m = |α| ;
n = |β| ;
i = 0 ;
while( i ≤ n − m )
    j = 1 ;
    while ( j ≤ m and a_j = b_{i+j})
        j = j + 1 ;
    endwhile
    if (j > m) output(i+1) ;
    i = i + 1 ;
endwhile
```

Characteristics of the brute-force algorithm

- Performs $n - m$ shifts
- May perform as many as m comparisons for each window.
- Total number of comparisons scales as nm in worst case.
- Average number of comparisons for random strings is a constant times n rather than nm.

  Typically, $m = 10^3$ and $n = 10^8$ so $nm = 10^{11}$.

Improvements to worst-case performance

- Preprocess strings to identify optimal shifting strategy.
  - String match
  - Character mismatches
- Scan for matches from right to left in window.
- Average performance depends on alphabet size.

Several strategies lead to guaranteed improvement in performance, but the details are unintelligible and unimportant for users.

24 Definition (Prefix) *A string* β *is a prefix of a string* α *if, and only if, there is another string* γ *such that* $\alpha = \beta\gamma$.

A prefix is constructed by deleting zero or more consecutive characters from the right end of a string.

25 Definition (Suffix) *A string* β *is a suffix of a string* α *if, and only if, there is another string* γ *such that* $\alpha = \gamma\beta$.

A suffix is constructed by deleting zero or more consecutive characters from the left end of a string.

References for exact matching algorithms

T. Lecroq. Experimental results on string matching algorithms. *Software-Practice and Experience* 25(7):727–765 (1995).

J. Tarhio and H. Peltola. String matching in the DNA alphabet. *Software-Practice and Experience* 27(7):851–861 (1997).

D. Gusfield. “Algorithms on Strings, Trees, and Sequences”. Cambridge Univ. Press (1997).

Approximate matching of strings and substrings

- Relax exact matching criteria to allow at most a fixed number of character mismatches.
- Insertions and deletions are **not** allowed.
- Algorithms can be viewed as generalizations of fast exact matching schemes.

These so-called k-mismatch problems are extremely important in practice because sequences have errors.

Example of approximate matching

Given $k = 2$ and

Query:	bend
Subject:	abentbananaend

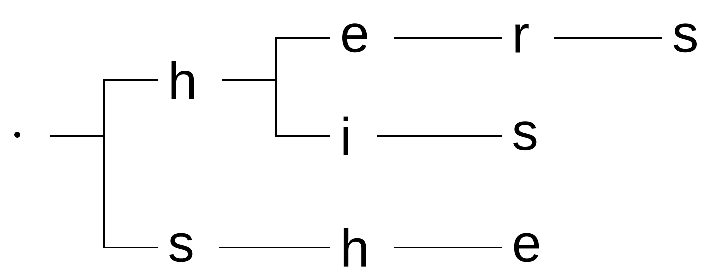
Approximate matches

Substring	**Mismatch count**
bent	1
bana	2
aend	1

Keyword searches

Find all occurences of members of fixed set of query strings in a subject string.

Query strings:	hers, his, she
Subject string:	ushers


·
h
e
r
s
i
s
s
h
e

Lexical and grammatical structure of strings

Strings with internal structure are of interest:

- Eukaryotic genes, prokaryotic operons
- Direct and inverted repeats in DNA sequences
- Tandem duplication of genes
- Secondary structure motifs in tRNA and protein

26 Definition (Language) *A language, $\mathcal{L}$, is a set of strings over a fixed alphabet Σ.*

Languages are extremely powerful tools that enable biologists to systematize their intuition and knowledge of structure, and convert it into practically useful analytical tools.

- Structural patterns
- Constraints

Example: Zinc fingers

The alphabet is Σ_{protein}.

Either E or D followed by
Either E or N followed by
L followed by
Either S or A or N followed by
Exactly two amino acids followed by
Either D or E followed by
Exactly one amino acid followed by
E followed by
L

How would you go about finding zinc fingers in a protein database?

Regular languages and regular expressions

Two languages, $\mathcal{M}$ and $\mathcal{N}$, over the same alphabet, Σ, can be combined by

Repetition $\mathcal{M}$, $\mathcal{M}\mathcal{M}$, $\mathcal{M}\mathcal{M}\mathcal{M}$, $\ldots$

Alternation $\mathcal{M}$ or $\mathcal{N}$

Concatenation $\mathcal{M}\mathcal{N}$, $\mathcal{N}\mathcal{M}$

These rules can be used to create complex languages from simpler languages.

A simple syntax for regular languages

Expression	Lexical match
c	only c
$[c_1 c_2 \ldots]$	any c_i
$[{}^\wedge c_1 c_2 \ldots]$	anything except one of c_i
$r \mid s$	either r or s, exclusively
rs	r followed by s
(r)	r
r^+	one or more copies of r
r^*	zero or more copies of r
$r?$	either zero or one copies of r

Example: Codons and “Triplet wobble”

Serine $AG(U|C)$

Phenyalanine $UU[AUGC]$

STOP $U(U(U|G)|GA)$.

The syntax of PROSITE patterns

The standard IUPAC one-letter codes for the amino acids are used.

The symbol x is used for a position where any amino acid is accepted.

Ambiguities are indicated by listing the acceptable amino acids for a given position, between square parentheses '[ ]'. For example: $[ALT]$ stands for exactly one of Ala or Leu or Thr.

Ambiguities are also indicated by listing between a pair of curly brackets '{ }' the amino acids that are not accepted at a given position. For example: $\{AM\}$ stands for any amino acid except Ala and Met.

Each element in a pattern is separated from its neighbor by a '$-$'.

Repetition of an element of the pattern can be indicated by following that element with a numerical value or a numerical range between parenthesis. Examples: $x(3)$ corresponds to $x - x - x$, and $x(2, 4)$ corresponds to $x - x$ or $x - x - x$ or $x - x - x - x$.

When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a '$<$' symbol or respectively ends with a '$>$' symbol.

Examples of PROSITE patterns

$$[AC] - x - V - x(4) - \{ED\}$$

This pattern is translated as: [Ala or Cys]-any-Val-any-any-any-any-any but Glu or Asp

$$< A - x - [ST](2) - x(0,1) - V$$

This pattern, which must be in the N-terminal of the sequence (‘¡’), is translated as: Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val.