Sequence Similarity and BLAST

Dave Schneider
USDA Agricultural Research Service

August 8, 2002

The Information Hierarchy

Knowledge This sequence codes for a hemoglobin that is expressed in human fetuses that binds O_2 more tightly than the adult human hemoglobin.

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

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

Data is often treated as disposable, information is usually worth keeping, and knowledge is precious.

The limits of intuition in modern biology

- Biology is facing an flood of information on an unprecedented scale.
- There is too much information to interpret type by intuition; quantitative methods are an absolute requirement.
- Intuition is still needed to go from information to knowledge.

Syntax, grammar, and semantics

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

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

3 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.

See the little phrases go,
Watch their funny antics.
The men who make them wiggle so
Are the teachers of Semantics.
The words go up, the words go round
And make a great commotion,
But all that lies behind the sound
hebutude Boeotian.

F. Windsor. *A Space Child's Mother Goose*. Simon and Schuster, 1958.

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 syntactic 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-----knowswhatheistalkingabout!
||||      |||| ||||||||||||||||||||
Davedoesntknow-whatheistalkingabout!
```

Similarity or homology?

- Similarity is a mathematical measure computed directly from syntactic information.

- Homology are related concepts are semantic in nature and not uniquely determined by syntax.

Similarity is what is measured by BLAST, inference of homology based on similarity requires the ability to model singular events in the past.

Computational Experiments

- Formulate testable hypothesis
- Design positive and negative controls
- Run experiment
- Assess significance of results

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!

Sequences producing significant alignments:	Score (bits)	E 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 contaminate EST experiments?
- Are there real proteins with subsequences that look like a translation of 5S rRNA?
- Everything is suspect because there is a bug in the code. . .

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

Hangman

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

What happened?

- What is the reasoning behind the order that guesses were made?
- How many guesses were required relative the the number of unknowns?
- What about prefixes, suffixes, and root words?

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

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.

This is the Cybernetics and Stuff
That covered the Chaotic Confusion and Bluff
That hung on the Turn of a Plausible Phrase
And thickened the Erudite Verbal Haze
Cloaking Constant K
That saved the Summary
Based on the Mummery
Hiding the flaw
That lay in the Theory that Jack built.

F. Windsor. *A Space Child's Mother Goose*. Simon and Schuster, 1958.

Alignments, scoring, and substitution matrices

4 Definition (Substitution matrix) *A substitution matrix is a table of scores for the alignment of two characters in the alignment of 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.

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.

5 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

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

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

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

Biological structures definable by regular expressions

- Protein motifs in nucleotide or protein sequences
- Restriction sites in genomic sequences
- TATA boxes in genomic sequences
- Codons in nucleotide sequences
- SSRs

Structures not definable by regular expressions

- Direct repeats of arbitrary length motifs
- Inverted repeats of arbitrary length motifs
- Secondary structure of tRNA

A qualitative change in perspective

Allowing insertions and deletions means:

- switching focus from substrings to subsequences
- using a new class of (expensive) algorithms

Selection of optimality criteria

- Minimize edit distance
- Maximize similarity measures
- Should one look for global or local optima?

Can one definition of optimality hold for both coding and non-coding regions?

Gapped alignments

Gaps are a mechanism to cope with extended regions dominated by mismatches and indels interspersed among regions of stronger alignment where matches predominate.

Local maximum similarity

The region of similarity is not required to extend the full length of the sequences.

Very similar to global maximum similarity algorithms except:

- All negative values are set to zero
- End of optimal alignment is determined by the location of the largest element.

The NCBI BLAST Suite

Original effort (circa 1990) reflected the confluence of three critical factors:

- Careful analysis of statistical significance of alignments
- Fast substring searching techniques
- Effective heuristics for identifying promising candidates

Fast substring searching (version 1)

Find all substrings of length w with scores above a fixed threshold, t:

- Construct all substrings of length w with scores greater than t
- Insert all such substrings into a keyword tree
- Use an Aho-Corasick style algorithm to quickly find all occurrences of "hits" in the database.

Segment pairs

6 Definition (Segment pair) *A segment pair is a pair of equal length substrings of the query and subject strings that are aligned without indels or gaps.*

7 Definition (Locally maximal segment pair) *A segment pair whose alignment score would decrease if the region of alignment is either shortened or lengthened.*

8 Definition (Maximal segment pair) *The segment pair with the largest score over all possible segment pairs in query and subject strings.*

Extending and scoring hits (version 1)

Restrict search for optimal alignment to a strip along the diagonal of the dynamic programming matrix.

Gaps are treated implicitly by combining scores for multiple segment pairs with high scores that give a statistically significant cumulative score.

Scoring parameters are computed directly from the scoring matrix.

Fast substring searching (version 2)

- Require two non-overlapping hits within a distance A along the same "diagonal" (default $A = 40$).

- Must decrease default value of t to maintain sensitivity. The default values are $w = 3$ and $t = 11$ for BLASTP.

Extending hits (version 2)

- Start from a single pair of aligned residues using a gapped dynamic programming algorithm to extend hits in both directions.
- Adaptivity eliminates the need to consider only a strip of the dynamic programming matrix.
- Halt extension when the score drops below a value of X lower than the best score obtained so far.
- Approximately 500 times slower than ungapped extension.

Statistical significance (version 2)

The values of statistical parameters are determined by computer simulation methods.

> *A drawback of this approach is that the program may not accept an arbitrary scoring system, for which no simulation has been performed, and still produce accurate estimates of the statistical significance.*

FASTA recomputes λ_{gapped} and K_{gapped} "on the fly".

Position-specific iterated BLAST

In protein motifs or profiles, some residues are more critical to structure and function than others (recall PROSITE patterns).

- Create an initial alignment with position-independent scoring matrix (20×20).
- Construct a position-specific matrix and associated target frequencies from initial results (weight sequences, fudge gaps, $L \times 21$).
- Repeatedly search the database with the position-specific scoring matrix, updating matrix and target frequencies after every iteration.

Thematic retrospection: Theory

- Information theory and statistics impose strict limits on interpretation.
- Linguistic approach to hierarchical complexity in biology.
- Algorithmic details are important but very refractory.

Thematic retrospection: Practice

- Clearly specify question to be addressed.
- Design experiments with proper controls.
- Purify reagents before use.
- Select correct tool(s).

Computer experiments produce hypotheses, not conclusions.